Uncertainty Analysis: What is it we're uncertain of?[1]

Authors
Dr. Howard Castrup
President, Integrated Sciences Group
hcastrup@isgmax.com

Suzanne Castrup
VP, Engineering, Integrated Sciences Group
scastrup@isgmax.com

Abstract

Recently, there has emerged a growing confusion regarding what to include in an uncertainty analysis of a measurement result and what to leave out. This paper provides guidelines for identifying error sources which are relevant to uncertainty analysis and those which are not. The guiding question in such identifications is "what is it we're uncertain of?" This question is explored within the context of various activities with differing immediate objectives. Such activities include conformance testing, measurement decision risk analysis, capability statement development, hypothesis testing and equipment parameter tolerancing.

1 Introduction

1.1 Purpose

At ISG, we frequently receive "calls for help" from customers and other colleagues who have established valid uncertainty analysis procedures, only to be challenged by accreditation assessors less knowledgeable than themselves. This frequently leads to attempts to update an assessor who steadfastly clings to a flawed understanding, often acquired through rudimentary training or opinions from "experts" who have lost sight of the guiding question of uncertainty analysis, namely, what is it in a measurement result that we're uncertain of?

This paper is motivated in part by the not inconsiderable and unnecessary administrative problems caused for calibration labs which refuse to substitute invalid uncertainty analysis procedures for valid ones. In attempting to provide some explicit guidelines for valid procedures, we find there are two main points to be elaborated.

1. The term "measurement error," along with its companion term "true value," are essential to addressing the guiding question.[2]
2. The errors and uncertainties to include (and exclude) in an uncertainty analysis depends on the objective of the measurement being made.

1.2 Background

In this paper, we focus on the identification of error sources and uncertainties which are relevant to alternative measurement activities and objectives. Accordingly, we begin by defining

---

[1] Presented at the 2012 NCSLI Workshop & Symposium, Sacramento, July 30.

[2] The value of these terms has been argued by Ehrlich and Dybkaer [1].

measurement error and uncertainty. For the latter, we need to again ask the guiding question. The obvious answer is "how close is our measured value to the actual or "true" value of an artifact or parameter?"

1.3 Definitions

1.3.1 True Value

For most test or calibration purposes, the following definition is adequate:

A physical property we seek to estimate through measurement.

In appendix B.2.3 of the GUM [2], true value is defined as

A value consistent with the definition of a given particular quantity that would be obtained by a perfect measurement.

In a discussion concerning the impact of measurement conditions on true value changes during a measurement session, Kirkup and Frenkel [3] offer the following:

The value we would obtain for a completely specified measurand if we could use an ideal instrument in a completely specified environment.

In the references cited above and in many other publications on measurement uncertainty analysis, it is generally understood that the true value of an artifact or parameter is unknowable and can, at best, only be estimated.[3]

1.3.2 Measurement Error

In the above definitions of true value, reference is made to an "ideal instrument" or "perfect measurement." In practice, neither exists, since all instruments produce error and all measurements exhibit error. To get a handle on this, we define measurement error as the difference between a value obtained by measurement and the corresponding true value. This definition derives from the simple equation

$$x_{meas} = x_{true} + \varepsilon_{meas} , \tag{1}$$

where

x_{meas}	=	value obtained by measured
x_{true}	=	true value at the instant of measurement
ε_{meas}	=	measurement error.

With this relation and the guiding question, we can readily see that the uncertainty in a measurement result can be thought of as the lack of knowledge of the sign and magnitude of measurement error.[4] This lack of knowledge will be given a mathematical definition presently.

---

[3] In practice, the true value may sometimes be equated with the value that would be obtained by an NMI, such as NIST, under conditions identical to those under which the measured value was obtained. A somewhat related definition is offered by Fornasini [4] as applied to published fundamental constants for many didactic applications [5]: *A value that is accepted, sometimes by convention, as having an uncertainty suitable for a given application.*

In this paper, such values are regarded as "conventional values" rather than "true values."

2 True Value and Measurement Error

As mentioned above, use of the term “true value” has recently encountered some resistance.[5] The ostensible reason is that the term cannot be defined independently of other terms, e.g., “measured value” and “measurement error.” Use of the term “measurement error,” has also been discouraged since it cannot be rigorously defined independently of “measured value” and “true value.” Since “true value” and “measurement error” are both useful and understood intuitively, it would seem counterproductive to avoid their use. There are at least four reasons why.

First, it is important to realize, that there are many quantities in the physical sciences, such as time, length and mass, that also defy definition independently of other quantities, yet are well understood and necessary for the pursuit and practice of science.

Second, like time, length and mass, true value and measurement error are also well understood until we attempt to define them in some stand-alone fashion. As Thomas Aquinas once remarked “I know what time is, but if someone asks me what time is, I don’t know what it is.” Also, like time, length and mass, the concepts of true value and measurement error are necessary for cogent measurement uncertainty analysis. This is evident when we focus on what it is we’re uncertain of in a measurement result. In the discussion following Eq. (1), we defined measurement uncertainty as the lack of knowledge of the sign and magnitude of measurement error. This qualitative definition can be quantified by further defining measurement uncertainty as the standard deviation of the measurement error probability distribution. Without getting into the rationale for this definition, which is a subject in itself, suffice it to say that it provides a guidepost for uncertainty estimation and combination.[6]

Third, before the introduction of the GUM, the subject of uncertainty analysis was referred to as “error analysis” in which many useful concepts and methods were developed over the years. For example, the uncertainty model of Eq. (16) of Section 5.2.2 of the GUM [2] is easily constructed by simply taking the variance of an error model which is arrived at by applying small error theory to a multivariate or “indirect” measurement [7, 8, 9].[7] Additionally, by applying the methods of error analysis, it was possible to develop a rigorous method for computing the degrees of freedom for Type B uncertainty estimates [10] and a variant of the Welch-Satterthwaite relation for correlated errors [11].

Fourth, the concept of measurement error facilitates the development of error models from which uncertainty models may be constructed. This is discussed later under “Computing Uncertainty.”

---

[4] It should be noted that Eq. (1) applies to a single measurement made at a given instant of time. All three quantities in the equation may differ from measurement to measurement. This is discussed further in Section 4.

[5] The GUM discourages the concept of true value [2, D.3.5]. However, to avoid confusion, the concept is encouraged in the VIM [6].

[6] In section E.5.4 of the GUM a statement is made to the effect that avoiding the use of the concepts of true value and error eliminates the confusion between error and uncertainty. Defining uncertainty as is done here *clearly* eliminates confusion and has the advantage of promoting cogent and consistent uncertainty analyses.

[7] As will be shown later, working from a rigorous error model leads to the straightforward development of a rigorous expression of uncertainty referred to as an “uncertainty model.”

3 Common Measurement Error Sources

Measurement error sources are present in making a direct measurement of an artifact or parameter or of a component of a multivariate measurement (see Section 4.3). Included in the list are[8]

- the bias of a reference parameter
- random error or "repeatability" in the measurement process
- resolution error of the measurement reference and/or the unit under test (UUT)
- operator bias in using the measurement reference, the UUT or ancillary equipment
- influence of environmental factors
- computation error
- shipping and handling error
- digital processing error
- other.

As will be seen presently, the uncertainty in *the bias of the UUT at the time of measurement is not included in the list*. This is because, for the measurement activities considered in this paper, the point of the activity is to "measure" the values of biases of UUT parameters, estimate the uncertainties in the measurements and act on the analysis results [12, 13].

Several of the errors in the list may include contributions from both the measurement reference and the unit under test (UUT). Exceptions are reference parameter bias, which is solely a property of the measurement reference, and shipping and handling error, which results from the UUT's response to stress. If the objective of a measurement is the evaluation of the reference measurement system, it may be advisable to somehow compensate for the UUT contributions. In conformance testing or any other calibration or test in which a decision is made regarding the status of the UUT, such compensation is to be avoided.

4 Computing Uncertainty

4.1 Variance And Uncertainty

As stated previously, the uncertainty in a measurement is the uncertainty in the measurement error, expressed as the standard deviation of the error probability distribution. This standard deviation is just the square root of the "mean square error" or *variance* of the distribution.

Obtaining the variance of an error distribution can be accomplished by applying a "variance operator," denoted "var," to the error ε_{meas} of Eq. (1):

$$\text{variance in } \varepsilon_{meas} = \text{var}(\varepsilon_{meas}) \,. \tag{2}$$

The variance operator is a mathematical tool the emerges naturally and simply from basic probability considerations. The details of its development and use are routinely given in upper division college statistics text. For our purposes, we state that, for a given measurement,

[8] For detailed descriptions of these error sources, see Ref [14].

$$\begin{aligned} \text{uncertainty in } x_{meas} &= \text{ uncertainty in } \varepsilon_{meas} \\ &= \sqrt{\text{var}(\varepsilon_{meas})} = u_{meas} \, . \end{aligned} \tag{3}$$

4.2 Direct Measurements

4.2.1 Combined Error

The error ε_{meas} is comprised of one or more error sources, such as are listed above under Common Measurement Error Sources. Then, for a direct measurement, we can write

$$\varepsilon_{meas} = \varepsilon_{ref} + \varepsilon_{ran} + \varepsilon_{res} + \varepsilon_{op} + \varepsilon_{env} + \cdots, \tag{4}$$

where

$\varepsilon_{ref} =$ measurement reference bias

$\varepsilon_{ran} =$ random error or "repeatability" of the measurement process $= \varepsilon_{ran,ref} + \varepsilon_{ran,UUT}$

$\varepsilon_{res} =$ resolution error $= \varepsilon_{res,ref} + \varepsilon_{res,UUT}$

$\varepsilon_{op} =$ operator bias $= \varepsilon_{op,ref} + \varepsilon_{op,UUT}$

$\varepsilon_{emv} =$ error due to environmental or ancillary factors

The “ref” and “UUT” labels in the subscripts flag error contributions from the measurement reference and the UUT, respectively. Note that, while ε_{ran} is separated into measurement reference and UUT components, these components are rarely individually distinguishable in practical measurement situations.

4.2.2 Combined Uncertainty — The Variance Addition Rule

Applying the variance operator to Eq. (4) gives

$$\begin{aligned} \text{var}(\varepsilon_{meas}) = \text{var}(\varepsilon_{ref}) + \text{var}(\varepsilon_{ran}) + \text{var}(\varepsilon_{res}) + \text{var}(\varepsilon_{op}) + \text{var}(\varepsilon_{env}) + \cdots \\ +2\,\text{cov}(\varepsilon_{ref}, \varepsilon_{ran}) + 2\,\text{cov}(\varepsilon_{ref}, \varepsilon_{res}) + 2\,\text{cov}(\varepsilon_{ref}, \varepsilon_{op}) + 2\,\text{cov}(\varepsilon_{ref}, \varepsilon_{env}) \\ + \cdots + 2\,\text{cov}(\varepsilon_{ran}, \varepsilon_{res}) + \cdots \end{aligned}$$

the cov(ε_i, ε_j) covariance terms represent contributions to var(ε_{meas}) from interactions or *correlations* between the ith and jth error sources. For practical purposes, it is usually convenient to express covariances in terms of variances and correlation coefficients $\rho_{i,j}$, i.e.,

$$\text{cov}(\varepsilon_i, \varepsilon_j) = \rho_{i,j}\sqrt{\text{var}(\varepsilon_i)\,\text{var}(\varepsilon_j)} = \text{cov}(\varepsilon_j, \varepsilon_i) \, .$$

Using this relation, together with Eq. (3) yields

$$\begin{aligned} u_{meas}^2 &= u_{ref}^2 + u_{ran}^2 + u_{res}^2 + u_{op}^2 + u_{env}^2 + \cdots \\ &\quad + 2\rho_{ref,ran} u_{ref}^2 u_{ran}^2 + 2\rho_{ref,res} u_{ref}^2 u_{res}^2 + 2\rho_{ref,op} u_{ref}^2 u_{op}^2 + \cdots \end{aligned} \tag{5}$$

Generally, with direct measurements, the correlations between error sources are zero, and Eq. (5) becomes

$$u_{meas}^2 = u_{ref}^2 + u_{ran}^2 + u_{res}^2 + u_{res}^2 + u_{op}^2 + u_{env}^2 + \cdots \tag{6}$$

4.3 Multivariate Measurements

4.3.1 Combined Error

For a multivariate or "indirect" measurement, the quantity of interest is a function of more than one directly measured quantity or "component." For example, letting the variable y represent the value of an N-component quantity and the variables $x_1, x_2, \cdots, x_N$ represent the directly measured variables, we express y as a function of these variables in a "system equation"

$$y = f(x_1, x_1, \cdots, x_N). \tag{7}$$

If the errors $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_N$ of the direct measurements are small such that the product of any two errors is negligible, then Eq. (7) can be used to construct an error model as a Maclaurin series expansion of y to first order according to

$$\begin{aligned}\varepsilon_y &= \left(\frac{\partial f}{\partial x_1}\right)\varepsilon_1 + \left(\frac{\partial f}{\partial x_2}\right)\varepsilon_2 + \cdots + \left(\frac{\partial f}{\partial x_N}\right)\varepsilon_N \\ &= c_1\varepsilon_1 + c_2\varepsilon_2 + \cdots + c_N\varepsilon_N ,\end{aligned} \tag{8}$$

where $c_1, c_2, \cdots, c_N$ are called *sensitivity coefficients* and the errors ε_i are the errors in the direct measurements of x_i, $i = 1,2, \cdots, N$ constructed as in Eq. (4).

4.3.2 Combined Uncertainty

Using Eq. (8), together with the variance addition rule gives

$$\begin{aligned}u_y^2 = c_1^2u_1^2 + c_2^2u_2^2 + \cdots c_N^2u_N^2 + 2\rho_{1,2}c_1c_2u_1u_2 \\ + 2\rho_{1,3}c_1c_3u_1u_3 + \cdots + 2\rho_{N-1,N}c_{N-1}c_Nu_{N-1}u_N.\end{aligned} \tag{9}$$

In Eq. (9), the uncertainties u_i, $i = 1,2, \cdots , N$ are the uncertainties, expressed as in Eq. (6), of each of the N direct measurements.

Note that, in many multivariate measurements, while correlations between the errors of a direct measurement are zero, the correlations between error components are not always zero. For example, consider obtaining a measurement of the area A of a rectangular plate by measuring its length L and width W using a tape measure. Suppose, to simplify the discussion, the only error source in each measurement is the tape measure bias. Since it is reasonable to assume that the bias is essentially the same for both measurements, we set $\varepsilon_{L,bias} = \varepsilon_{W,bias} \equiv \varepsilon_b$ for short. In this simple example, the system equation is

$$A = LW , \tag{10}$$

where L and W are the directly measured components. By Eq. (8), we have the error model

$$\varepsilon_A = c_L\varepsilon_L + c_W\varepsilon_W$$

where

$$c_L = \left(\frac{\partial A}{\partial L}\right) = W ,$$

and

$$c_W = \left(\frac{\partial A}{\partial W}\right) = L \, .$$

Since we simplified the example by setting $\varepsilon_{L,bias} = \varepsilon_{W,bias} = \varepsilon_b$, the error model becomes

$$\begin{aligned}\varepsilon_A &= W\varepsilon_L + L\varepsilon_W \\ &= (L+W)\varepsilon_b \, ,\end{aligned}$$

and, from Eq. (9), we get

$$u_A^2 = W^2 u_L^2 + L^2 u_W^2 + 2\rho_{L,W} LW u_L u_W \, .$$

In this case, $u_L = u_W = \text{var}(\varepsilon_b) \equiv u_b$ and this expression becomes

$$u_A^2 = \left(W^2 + L^2 + 2\rho_{L,W} LW\right) u_b^2 \, . \tag{11}$$

Since the same tape measure is used to measure both L and W, the two measurements are not independent of one another and, accordingly, the correlation coefficient between ε_L and ε_W is nonzero. Such a correlation between component errors is called a *cross-correlation* [9]. In this case, since the same tape measure is used to measure both length and width, we can safely set $\rho_{L,W} = 1$, and get

$$u_A^2 = \left(L^2 + W^2 + 2LW\right) u_b^2 = \left(L+W\right)^2 u_b^2 \, ,$$

so that

$$u_A = \left(L+W\right) u_b \, . \tag{12}$$

To a good approximation, we can set $L = L_{meas}$ and $W = W_{meas}$ and write

$$u_A \cong \left(L_{meas} + W_{meas}\right) u_b \, .$$

To see the importance of cross-correlations, it is interesting to consider an example in which the length and width measurements are made using different tape measures drawn randomly from an inventory. In this case, the biases may be considered to be independent of one another and we set $\rho_{L,W} = 0$. Then Eq. (11) becomes

$$u_A^2 = \left(W^2 + L^2\right) u_b^2$$

and

$$u_A = \sqrt{L^2 + W^2}\, u_b \, . \tag{13}$$

Comparison of Eqs. (12) and (13) shows that, if a nonzero $\rho_{L,W}$ is ignored, the estimated uncertainty may be considerably different from what is appropriate. When the cross-correlation coefficient is positive, the estimate will be too small. Conversely, if the coefficient is negative, the estimate will be too large.[9]

5 Relevant Error Sources

At stated at the outset, whether an error is to be included in an uncertainty analysis, depends on the objective of the measurement activity, i.e., the intended use of the uncertainty estimate. In

[9] The proof of this is left as an exercise for the reader.

this paper, we consider five basic activities: (1) conformance testing, (2) measurement decision risk analysis, (3) hypothesis testing, (4) capability statement development, and (5) equipment parameter tolerancing.

5.1 Conformance Testing

A conformance test is one in which a decision is made as to whether the value of an attribute is within its tolerance limits or not. The attribute value is estimated by a measurement or set of measurements, the result of which is taken to approximate the attribute's true value under the conditions of the measurement. Inevitably, the estimate is made with some unknown measurement error. As discussed earlier in Section 4, we quantify the uncertainty in the measurement as the standard deviation of the error probability distribution.[10]

In conformance testing, the relevant error sources for direct measurements are those indicated in Eq. (4). For a given measurement, some error sources explicitly included in Eq. (4) may not be relevant. In addition, the list for a given measurement may include several error sources not explicitly included. Assembling the list is a case-by-case exercise. To reiterate from earlier, the one error source never to be included in conformance testing is the bias of a UUT artifact or parameter, since this is the quantity we estimate by measurement. In other words, the measurement uncertainty is the uncertainty in the estimate of the bias obtained by measurement and does not include any pre-measurement estimate of the UUT bias uncertainty.[11]

5.1.1 What are we Testing?

It is important to focus on the fact that what we are testing is the conformance with specifications of the attributes or parameters of a UUT. The information we seek in this context is *not* the quality of the measurement system, useful though that information may be in a different context.[12]

Since the measurement result includes measurement error, we often cannot simply make an in- or out-of-tolerance proclamation from the measurement result alone. Instead, what we may be justified in doing is estimating a confidence level that the attribute is in conformance with specifications. The confidence level is typically computed using the estimated standard deviation of the measurement error distribution and the degrees of freedom of the estimate.[13] If the confidence level is sufficiently high, the UUT attribute or parameter may be considered to be in compliance with specifications.

---

[10] In this paper, the term standard deviation is preferred over the term "standard uncertainty" to emphasize that this quantity is a statistic of an error distribution and not just a term used as a heuristic reference devoid of statistical content. See Ref [10]

[11] Clearly, its inclusion would be both frivolous and constitute "double dipping."

[12] If the uncertainty in the bias of the measurement reference is estimated prior to measurement, the measurement result of a conformance test can be used to "update" this uncertainty using Bayesian analysis [7, 15, 16] (see Section 5.2.5.3).

[13] Other uses of the uncertainty in the measurement result are possible. For example, this uncertainty is key to estimating measurement decision risk and to developing attribute tolerance limits. The former are important measurement quality metrics and the latter can be established to ensure some expectation of passing inspection before shipping.

This computation is analogous to computing confidence limits from a standard uncertainty estimate, the degrees of freedom for the estimate and a stated confidence level, except in this case, the confidence limits are the attribute's tolerance limits and the confidence level is what is computed [17].

It is important to realize that in estimating this confidence level, all relevant sources of error that contribute to the total measurement error must be included.[14], [15]

5.1.2 Contributions to the Uncertainty Estimate

The guiding light in deciding which error sources to include is the list of factors that affect our confidence that the UUT attribute is in-tolerance. In the case of random error or repeatability, it is not relevant whether the random error of the measurement result is due to fluctuations in the measurement system or fluctuations in the value of a UUT attribute. What is relevant is that such fluctuations impact our knowledge of the value of the attribute and the confidence that the attribute is in conformance with specifications.

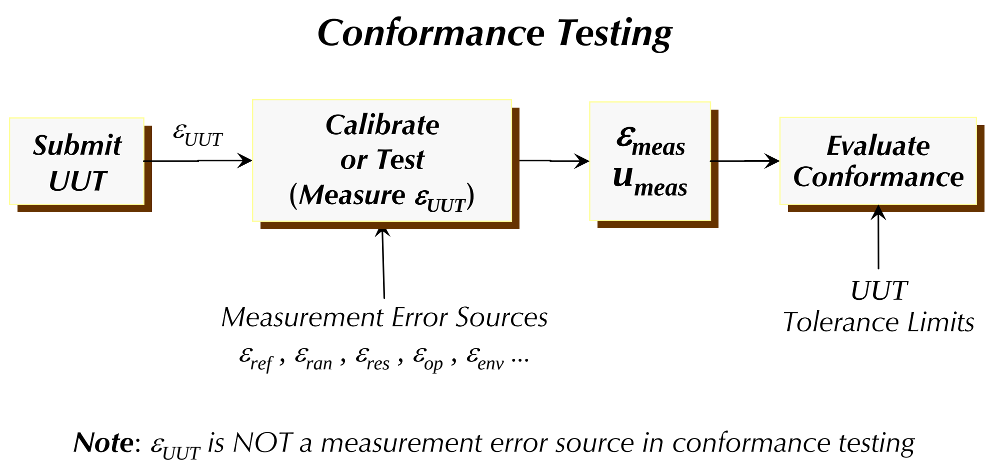

Figure 1. Stages Involved in Conformance Testing. Shown is a provisional listing of relevant error sources accompanying testing or calibration. Note that the bias ε_{UUT} is not included as a measurement error source.

As an example, consider the calibration of a UUT using a CMM. In this, a probe is applied and a value is obtained. Suppose that the value lies within the UUT's tolerance limits. In this case, we might be inclined to declare the UUT in-tolerance. We now repeat the measurement and obtain a value that is out-of-tolerance. We scratch our heads and ask, "which is it?" In-tolerance or out? To help answer this question, we could take a sample of measurements from which we would obtain a mean value and a sample standard deviation, the latter of which would be included in the combined uncertainty of the measurement.

[14] Other considerations, such as attribute value uncertainty growth during use and customer receiving inspection false reject risk may also be factored in.

[15] The handling of error sources to develop a total error in a measurement is described in detail for various measurement scenarios [12].

It might be said that this is an extreme example. Perhaps, but not an improbable one. Regardless, repeatability must be included in the uncertainty estimate. Why? Because measurement errors, random or otherwise, can add up, and it's the uncertainty in the total error that must be estimated.

We focus specifically on repeatability in this discussion, since some have advocated excluding it from consideration on the grounds that the observed random error is mostly due to the UUT rather than the measurement system. But, we repeat, it is the conformance of the UUT that is being tested, not the quality of the measurement system. Accordingly, excluding repeatability uncertainty may constitute a glaring omission, especially for cases where random error comprises the dominant uncertainty component.[16]

Recently, an ISO Technical Standard, ISO/TS 23165:2006, has been released promoting the exclusion of repeatability uncertainty in making conformance testing decisions. This practice is just plain wrong. Possible negative impacts include flawed estimates of in-tolerance probability, overly optimistic estimates of measurement decision risk, unrealistic capability statements or unsupportable equipment tolerances.

We realize that some would like to sweep random error under the rug, for whatever reasons. Nevertheless, in real-world conformance testing, apart from certain exceptions, it must be included as an error source.

5.1.2.1 Exceptions

Now and then, we encounter examples where the random error in a measurement is hidden by the granularity of displayed values, as is portrayed in Figure 2 (a). In cases represented by Figure 2 (b), variations in the measurement result are perceptible as repeatability, but repeatability need not be included as an error source, since these variations are due to resolution error, which is separately accounted for.

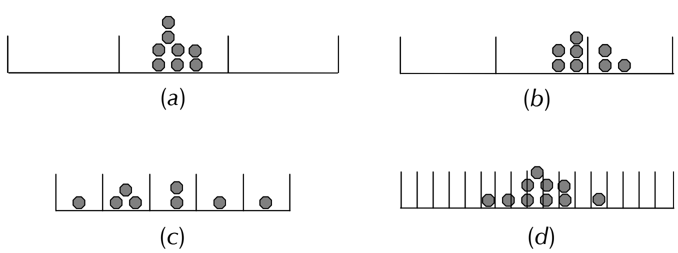

Figure 2. Repeatability and resolution error. Shown are four idealized examples of perceived variations being due to random error or resolution error or both..

In the example shown in Figure 2 (c), perceived variations are mostly due to random error. However, resolution error must be also be included, since its effect on displayed values is not trivial. In the example of Figure 2 (d), resolution error may be excluded as virtually negligible.

Whether to interpret sample variations as due to repeatability or resolution is a case-by-case issue. The point being, it is clearly not justified to exclude one or the other as a blanket policy.

[16] In our customer support efforts, we find this to often be the case.

5.1.2.2 Single Measurements vs. Measurement Samples

If a mean value $\overline{x}_{meas}$ and sample standard deviation s_n are calculated from a sample of measurements of size n, then, Eq. (1) yields

$$\overline{x}_{meas} = \overline{x}_{true} + \overline{\varepsilon}_{meas} , \tag{14}$$

where

$$\overline{x}_{meas} = \frac{1}{n}\sum_{i=1}^{n} x_i . \tag{15}$$

In Eq. (15), x_i is the ith sampled measured value, $i = 1,2, \cdots, n$. The sample standard deviation s_n is given by

$$s_n = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \overline{x}_{meas}\right)^2} . \tag{16}$$

The sample standard deviation in Eq. (16) is an estimate of the uncertainty due to random error for a single measurement. This quantity is useful for characterizing the random error of single measurements of a UUT attribute or parameter made with a specific measurement process under specific conditions.

If an action is to be taken, based on the mean value in Eq. (15), then the sample standard deviation s_n is not the correct uncertainty estimate for random error. Instead, we must use the uncertainty in the mean value, given by

$$s_{\overline{x}} = \frac{s_n}{\sqrt{n}} . \tag{17}$$

This is referred to in statistics texts as the standard deviation of the *sampling distribution* [3].

5.1.3 Conformance Testing Uncertainty Analysis Reports

Reports of the analysis of the uncertainty in a conformance test should include a breakdown of the relevant error sources. To be thorough, the following details should accompanying each direct measurement uncertainty estimate:[17]

- Error source name
- Uncertainty estimate
- Degrees of freedom
- Error probability distribution
- Uncertainty estimate type (A, B or A,B)
- Comments, if applicable

It is important to note that, if the mean value of a sample of measurements is reported, then the random error uncertainty estimate must be the uncertainty in the mean, as given in Eq. (17).

[17] For an in-depth description of the elements of uncertainty analysis reports, see Ref [18].

5.2 Measurement Decision Risk Analysis

In this section, methods for computing measurement decision risk are presented within the framework of both process-level and bench-level analyses.[18]

In the following sections, the calibration error for a given measurement scenario is denoted ε_{cal} and the uncertainty in this error is u_{cal}.[19] The uncertainty u_{cal} is obtained by taking the variance of ε_{cal}, i.e.,

$$u_{cal} = \sqrt{\mathrm{var}(\varepsilon_{cal})} \,. \tag{18}$$

5.2.1 Process Level Analysis

Process-level analysis employs what is referred to as "the Classical Method" [17]. With this alternative, risks are evaluated for a UUT attribute test point prior to testing or calibration by applying an expected UUT attribute in-tolerance probability and assumed calibration or test measurement process uncertainties. With process-level risk control, test limits called "guardband limits" are developed in advance if needed and may be incorporated in calibration or test procedures. Measured values observed outside guardband limits may trigger some corrective action, such as adjustment or repair, reduction in status or disposal.

5.2.2 Bench Level Analysis

Bench-level analysis includes methods referred to as the "Bayesian Method," the "Confidence Level Method" and the "TUR method" [17]. Bench-level methods control risks in response to measured equipment attribute values obtained during test or calibration.[20] With bench-level methods, guardband limits are superfluous, since corrective actions are triggered by the on-the-spot computation of risk or other measurement quality metric.

5.2.3 Risk Analysis Variables

The basic probability functions and definitions of both process level and bench level analysis are developed in Sections A.1.1 – A.1.3 of the Appendix.

5.2.4 The Classical Method

The Classical Method provides process level decision risk control in that risk estimation can assist in making equipment adjustment or repair decisions using nominal criteria. Estimates obtained using the classical method are also useful in making equipment procurement decisions, adjusting calibration intervals, and setting end-of-period measurement reliability targets.

5.2.4.1 Error Sources

The error sources of the Calibrate or Test activity shown in Figure 3 are the same as those discussed in conformance testing. In addition to these error sources, risk analysis using the

---

[18] A freeware application called *RiskGuard* [19] is available from www.isgmax.com for developing some experience with all three methods.

[19] Specific combinations of measurement process errors comprising ε_{cal} have been described [13].

[20] For detailed descriptions of measurement decision risk analyses, performed within the context of four calibration scenarios, see Ref [13].

Classical Method also requires an *a priori* estimate of the UUT bias uncertainty u_{UUT}. Note that, although the *a priori* u_{UUT} estimate is required, it is used merely as a risk analysis parameter and is not a contributor to the measurement process uncertainty.

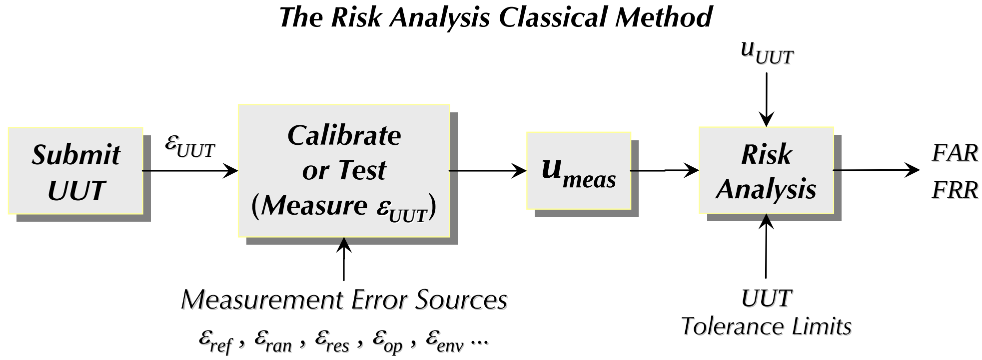

Figure 3. The Classical Method of risk analysis. Shown are the basic steps involved in measurement decision risk analysis using the Classical Method. The terms *FAR* and *FRR* represent false accept risk and false reject risk, respectively. The term u_{UUT} represents the *a priori* bias uncertainty of the UUT . It is not a contributor to measurement process uncertainty, but serves as a risk analysis parameter.

5.2.4.2 Estimating Risk

The estimation of measurement decision risk using the Classical Method is presented in Section A.1.4 of the Appendix.

Figure 4 shows an example of the Classical Method for a case where the maximum allowable *FAR* is 1%. Since the computed *FAR* is greater than the 1% maximum, the use of ± 9.6627 mV guardbands may be called for.[21]

5.2.5 The Bayesian Method

The Bayesian risk analysis methodology was developed by Castrup [23] and Jackson [24] in the '80s and later published with the label SMPC (Statistical Measurement Process Control) [15].

With the Bayesian Method, a risk analysis is performed for accepting or rejecting a specific UUT attribute based on *a priori* knowledge and on a “post-test” measured or sample mean value for the attribute taken during testing or calibration.

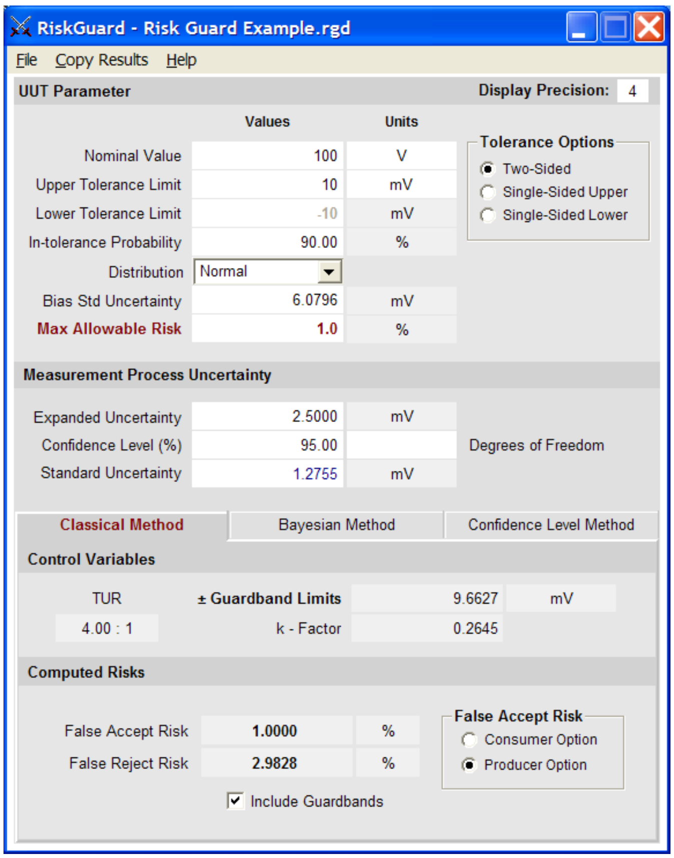

Figure 4. A risk analysis example. Shown are the results of a RiskGuard 2.0 analysis using the Classical Method [19]. In the case shown, the maximum allowable risk is 1%, necessitating the use of guardband limits of ±9.6627 mV.

[21] The applicability of guardband limits is discussed in Ref [17].

The post-test or *a posteriori* knowledge, when combined with the *a priori* knowledge, allows the computation of the quantities of interest, such as UUT and reference attribute biases, bias uncertainties and pre-test in-tolerance probabilities [15, 22, 23, 24].

5.2.5.2 Estimating *FAR*, *FRR*, ε_{UUT} and u_{UUT}

In addition to estimating false accept risk and false reject risk, the process shown in Figure 5 also produces post-test estimates of ε_{UUT} and u_{UUT}, based on a measured value or sample mean value for ε_{UUT}. The measurement error sources for the scenario shown in Figure 5 are the same as in conformance testing. As with the Classical Method, an *a priori* estimate of u_{UUT} is also required. Estimating *FAR*, *FRR* and the post-test estimates of ε_{UUT} and u_{UUT} is described in Section A.1.5.2 of the Appendix.

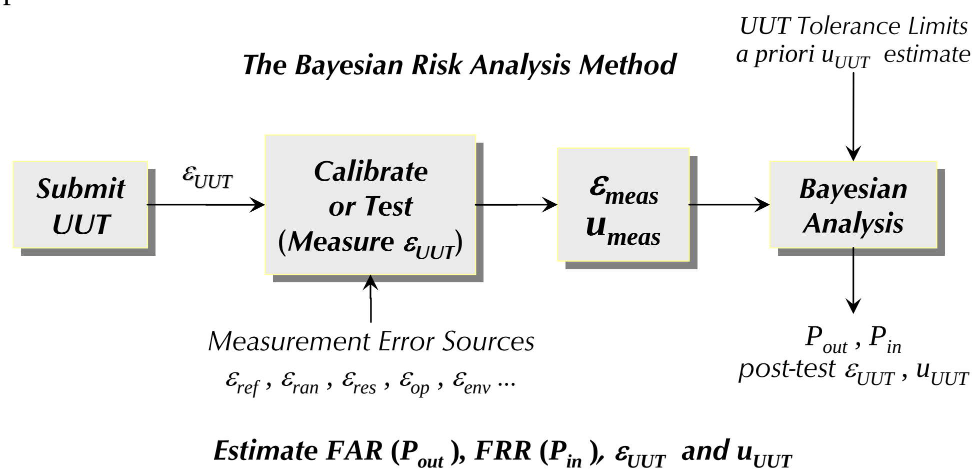

Figure 5. Bayesian Method of Risk Analysis. Shown are the basic steps involved in using the Bayesian Method to obtain *FAR*, *FRR*, and post-test estimates of the variables ε_{UUT} and u_{UUT}. The *a priori* u_{UUT} input is the *a priori* estimate of the bias uncertainty of the UUT. It is not a contributor to measurement process uncertainty, but serves as a risk analysis parameter. The measurement error sources are the same as in conformance testing.

An example of the Bayesian Method is shown in Figure 6. The analysis results are obtained using the UUT tolerance limits, measurement process uncertainty and *a priori* measurement data of Figure 4, in which a 1% maximum allowable *FAR* is enforced. In the Figure 6 example, the measured bias of the UUT attribute is considered acceptable, since the computed in-tolerance probability corresponds to a *FAR* less than 1%.

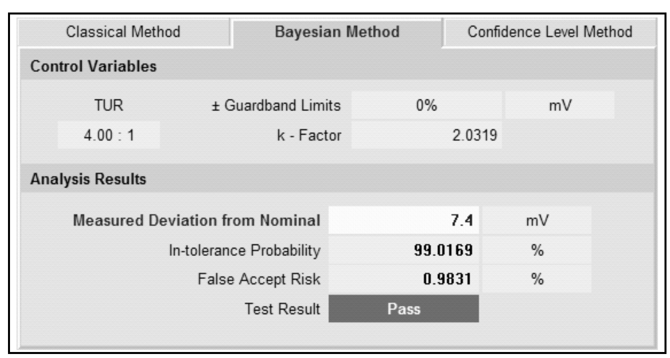

Figure 6. Example results with the Bayesian Method. Shown are RiskGuard 2.0 [19] Bayesian analysis results for the *a priori* measurement data of Figure 4. In the example, since the calculated in-tolerance probability corresponds to a False Accept Risk less than the 1% allowable maximum the UUT attribute is acceptable.

5.2.5.3 Estimating ε_{ref} and u_{ref}

The process shown Figure 7 produces post-test estimates of ε_{ref} and u_{ref}, based on a measured value of or sample mean value for ε_{UUT}. The measurement error sources for the scenario shown in Figure 7 are the same as in conformance testing with ε_{ref} replaced by ε_{UUT}. An *a priori* estimate of u_{ref} is also required. Estimating *FAR*, *FRR* and the post-test estimates of ε_{ref} and u_{ref} is described in Section A.1.5.3 of the Appendix.

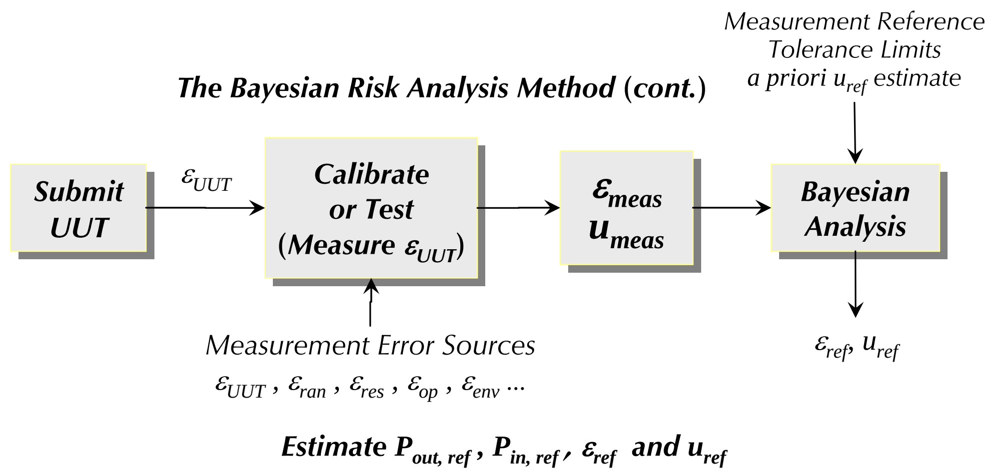

Figure 7. An alternative application of the Bayesian method. Shown are the basic steps involved in using the Bayesian method to obtain post-test estimates of ε_{ref} and u_{ref}. The *a priori* u_{ref} input is the *a priori* estimate of the bias uncertainty of the measurement reference. It is not a contributor to measurement process uncertainty, but serves as a risk analysis parameter.

5.2.6 The Confidence Level Method

The Confidence Level Method, represented in Figure 8, is used to estimate the confidence that a UUT attribute bias is in-tolerance, based on a UUT measured or sample mean value, taken during testing or calibration, together with an estimated uncertainty in the measurement process. Like the Bayesian Method, the Confidence Level Method is also a bench-level method. The relevant error sources are the same as for Conformance Testing.

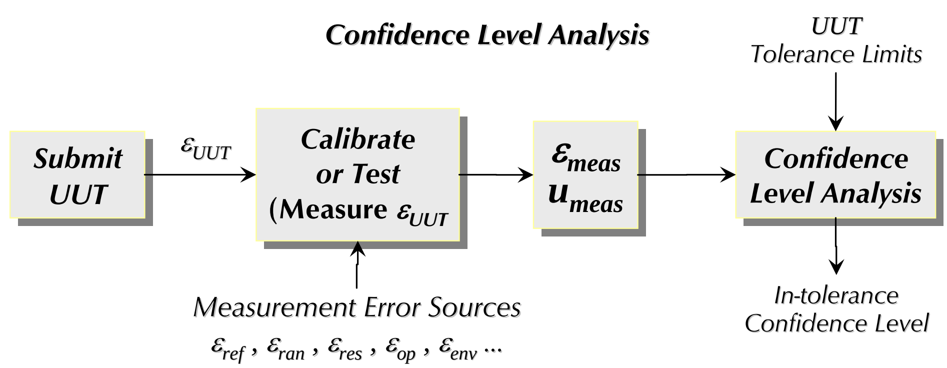

Figure 8. The Confidence Level Method. Shown are the basic steps involved in applying the Confidence Level Method to obtain an estimate of the confidence that a UUT attribute or parameter is in-tolerance. The error sources are the same as in conformance testing.

The confidence level method is distinguished from the classical and Bayesian methods in that the result of the analysis is an in-tolerance confidence level, rather than an in-tolerance probability. The method is applied when an *a priori* u_{UUT} estimate is not available or feasible. As such, it is not a true "risk control" method, but rather an application of the results of measurement uncertainty analysis.[22]

5.2.6.1 Confidence Level Estimation

Estimating confidence level risk control metrics is discussed in Section A.1.6 of the Appendix.

5.2.6.2 Applying Confidence Level Estimates

As with the Bayesian method, corrective action may be called for if a computed confidence level P_{in} is less than a predetermined specified limit. Let the minimum allowable in-tolerance confidence level be denoted C_{min}. Then corrective action is called for if the computed in-tolerance confidence level is less than C_{min}.[23]

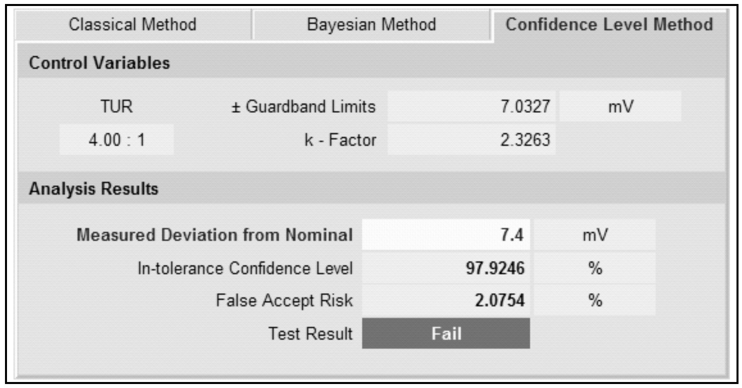

Figure 9. Example results with the Confidence Level Method. Shown are RiskGuard 2.0 [19] Confidence Level analysis results for the *a priori* input data of Figure 4. In the example, the UUT attribute is rejected, since the computed confidence level corresponds to a False Accept Risk greater than the 1% allowable maximum.

Figure 9 shows the results of the Confidence Level Method for the *a priori* input data of Figure 4. It can be seen by comparing Figure 9 with Figure 6, that applying the Confidence Level Method may yield a different decision than the Bayesian Method for the identical UUT tolerance limits, measurement uncertainty and measurement data.[24]

5.2.7 The TUR Method

Over the past five decades, the control of measurement decision risk has been attempted by specifying a nominal lower limit for the ratio of the tolerance limits of the UUT to the measurement uncertainty of the test or calibration process [25, 26]. These requirements provided some loose control of measurement decision risk but were not unambiguously defined or standardized. This state of affairs changed with the definition of an explicit relative accuracy requirement, published in ANSI/NCSL Z540.3-2006 [27]. This standard requires that, where it is not practical to compute false accept risk, the measurement’s "test uncertainty ratio" or TUR, shall be greater than or equal to 4:1.

The Z540.3 TUR definition is given in Section A.1.7 of the Appendix.

---

[22] The Confidence Level Method is not included in the current edition of the Z540.3 Handbook [32]. Hopefully, this will be rectified in future editions.

[23] If the Z540.3 nominal false accept risk requirement is adhered to, C_{min} = 0.98.

[24] This point is discussed in detail in [13].

5.2.7.1 Relevant Error Sources

The error sources needed to estimate a TUR, shown in Figure 10, are the same as those needed for Conformance Testing. Figures 4, 6 and 9 show examples with a computed TUR of 4.0.

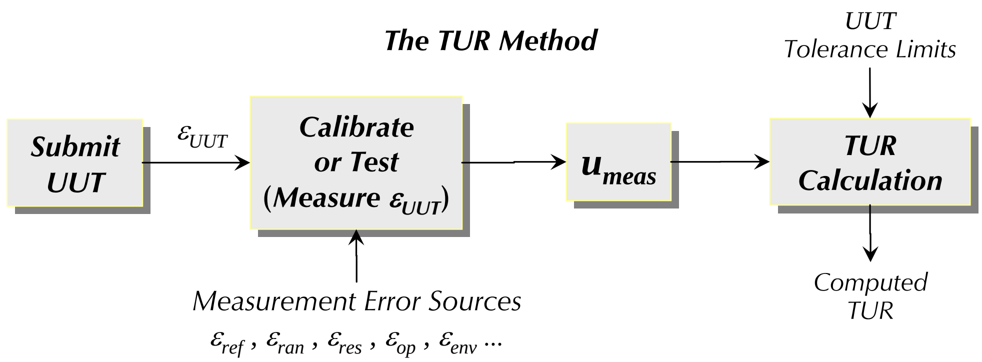

Figure 10. The TUR Method. The measurement error sources are the same as those for Conformance Testing. The computed TUR is the ratio of the sum of the span of the UUT attribute tolerance limits to twice the 95% expanded uncertainty of the measurement process. The measurement error sources are the same as in conformance testing.

5.2.7.2 A Note on Computing TUR

While the computation of the Z540.3 TUR appears on the face of it to be simple and straightforward, some difficulty has emerged since its publication. This has mainly arisen in cases where the UUT tolerance limits are specified in such a way that one or more of the specifications used in computing the measurement process uncertainty are also folded into the UUT tolerance limits. In short, some uncertainty contributions in the TUR numerator are also contained in the denominator.

This has been seen by some to present a conundrum which prevents achieving a TUR of 4:1 or better. This may be, but the definition is clear. If a tolerance limit includes information also pertinent to the estimation of u_{cal}, this is unfortunate but unavoidable, since the published UUT tolerance limits constitute a contractual quality guarantee to the UUT user, and the measurement process uncertainty is what it is.

5.2.8 Hypothesis Testing

Uncertainty estimates can be used to perform statistical tests concerning the populations they characterize. In such testing, the error sources are generally the same as those for conformance testing. In hypothesis testing, however, uncertainty estimates for two or more measurement processes may be employed. An application familiar to practitioners of metrology is the round-robin comparison of laboratory results. The applicable hypothesis testing methodology is given in Section A.2 of the Appendix.

5.2.9 Developing Capability Statements

The results of the calibrations performed by a commercial cal lab are used to determine equipment attribute in- or out-of-tolerance conditions. Statements of measurement capability made by these labs include uncertainty proclamations. While the tacit objective of capability statements is an expression of measurement quality, in practice the confidence level for in- or out-of-tolerance decisions issued by the lab are affected by the total uncertainty in the measurement result, which may include resolution, repeatability or other contributions from the UUT. In such cases, a rigorous capability statement pertaining only to the measurement quality of the lab is often not possible. In this way, UUT properties can emerge as an inconvenience in making a viable uncertainty estimate for a capability statement. However, since these properties may influence the confidence in reporting conformance, not including them can lead to misrepresentations.

For this reason, some of our customers have taken to issuing UUT-dependent values. For the sake of discussion, We may call these “worst case,” “typical case” and “best case,” depending on the class of UUT under consideration. It is acknowledged that, in a commercially competitive market, being honest in this way may constitute a hindrance in terms of loss of business. This is a dilemma. Forthright statements can be penalized while abbreviated statements are rewarded. The same can be said for instrument manufacturers who wish to issue realistic specifications but are in competition with manufacturers that do otherwise.

We feel that ignoring errors in measurement results due to UUT influences is not the solution. If the aim is to produce a meaningful standard for conformance testing, these influences must be included. Such inclusion should be mandatory for labs offering calibration services.

5.2.10 Equipment Parameter Tolerancing

Figure 11 depicts a set of activities that may be relevant to the development, distribution and use of a UUT parameter. These activities carry with them the potential of obtaining information that can be employed in arriving at publishable tolerance limits for the parameter in question.

In Figure 11, the variables ε_{UUT} and u_{UUT} (1) are the bias and bias uncertainty of an equipment parameter emerging from a manufacturing activity. The variables ε_{UUT} and u_{UUT} (2) are the same quantities emerging from first article testing. The variables ε_{UUT} and u_{UUT} (3) are the UUT parameter’s bias and bias uncertainty as input to the using activity’s receiving inspection.

The sequence in Figure 11 is discussed in the following sections.

5.2.10.1 Produce UUT

This activity include analyses accompanying the production of the UUT [29]. These analyses may include the following

- Engineering Analysis
- Component testing
- Board level testing
- Module level testing
- Preliminary development of tolerance limits

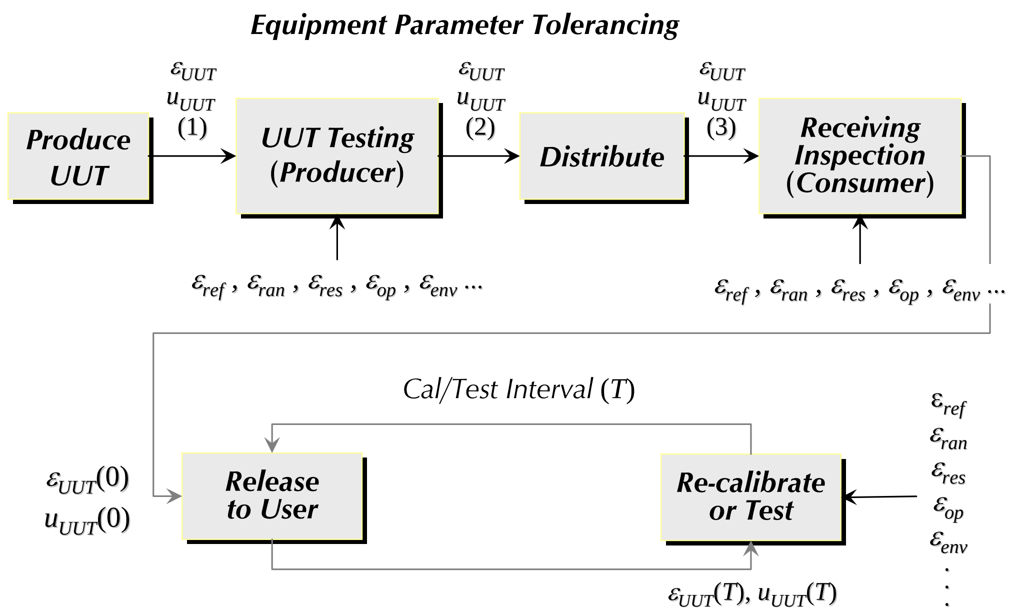

Figure 11. Relevant errors for Parameter Tolerancing. These include the measurement process errors which accompany conformance testing as well as error changes and uncertainty growth over cal/test intervals [8]. In the sequence, the qualifiers (1), (2) and (3) indicate measurement error sources for the producer's first article testing and the consumer's receiving inspection testing, respectively. The variable T represents the time elapsed since a previous test or calibration, often synonymous with the calibration interval.

5.2.10.2 UUT Testing (Producer)

This activity consists of first article testing [29]. The variables ε_{UUT} and u_{UUT} (1) are the input parameter bias and bias uncertainty, and the variables ε_{UUT} and u_{UUT} (2) are the output values.

The variables ε_{ref} , ε_{ran} , ε_{res} , ε_{op} , ε_{env}, $\cdots$ are possible error sources accompanying first article testing. The sources listed and implied are applicable to a direct measurement of the value of the UUT parameter in question. If this value is obtained as a multivariate measurement, as described in this paper, ε_{UUT} would be composed of error components, each of which would be composed of a combination of direct measurement errors.

5.2.10.3 Receiving Inspection (Consumer)

This activity involves the testing of the UUT parameters as received by the recipient using organization. The error sources involved in such testing should be considered to be equivalent to those of the producer's Calibrate and Test activity. However, the sign and magnitude of the testing errors and the uncertainty of the result reflect the interaction of the UUT parameter with the measurement process of the receiving inspection activity. These variables, labeled $\varepsilon(0)$ and $u(0)$ are taken to be applicable to the beginning of the usage of the UUT.

An estimate of the parameter bias uncertainty $u(0)$ is important to parameter tolerance development in that its magnitude is related to the risk of falsely rejecting a conforming parameter. Since this risk is a strong function of the parameter tolerance limits, it is prudent for a producer to estimate $u(0)$, compute the applicable false reject risk and then adjust tolerance limits to hold this risk to an acceptable level.

5.2.10.4 Re-calibrate or Test

As with the impact of $u(0)$ on acceptance testing, the parameter bias uncertainty $u(T)$ experienced at the end of a calibration interval T has an impact on the false reject risk resulting from re-calibration or test. Again, it is prudent for a producer to estimate $u(T)$ for reasonable or extreme values of T, compute the applicable false reject risk and adjust tolerance limits to hold this risk to an acceptable level. In estimating $u(T)$, account must be taken of both the value of the calibration interval T and the parameter bias uncertainty growth rate [8] during anticipated conditions of use.

Once the toleranced parameter has been put into use, opportunities for feedback from users may lead to refinements in uncertainty estimates for use in modifying tolerance limits, if needed [29].

6 Summary

The two main points of Section 1.1 have been discussed at length. With regard to the first point, it has been argued that the terms “error” and “true value” are essential to and useful for

- developing error models for direct and multivariate measurements
- identifying and applying correlations between measurements
- developing uncertainty models by operating on error models using the variance operator
- evaluating UUTs for conformance with specifications
- estimating measurement decision risk
- estimating risk metrics, such as TUR,
- testing hypotheses
- developing capability statements
- developing equipment parameter tolerances.

With regard to the second point, it has been shown that the list of specific errors and uncertainties to include in an analysis depends on the objective of the analysis. The objectives considered were those associated with the activities of conformance testing, measurement decision risk analysis, hypothesis testing, developing capability statements and developing equipment parameter tolerances.

To assist assessors and test or calibration laboratories, the mistake of treating *a priori* estimates of UUT bias uncertainty as a measurement process error source in conformance testing was pointed out. It was also argued that the uncertainty due to random error should not be categorically excluded in an estimate of combined measurement error. In this, it is important to note that expediency is neither a substitute for nor a guarantee of validity.

References

[1] Ehrlich, C. and Dybkaer, R., “Uncertainty of Error: The Error Dilemma,” *Proc. NCSLI Workshop & Symposium*, Washington DC, 2011.

[2] BIPM JCGM 100:2008, *Evaluation of Measurement Data — Guide to the expression of uncertainty in measurement*, JCGM, 2008.

[3] Kirkup, L. and Frenkel, R., *An Introduction to Uncertainty in Measurement: Using the GUM*, Cambridge University Press, Cambridge, June 2006.

[4] Fornasini, P., *The Uncertainty in Physical Measurements,* Springer Science+Business Media, New York, 2008.

[5] Mohr, P., et al., "CODATA Recommended Values of the Fundamental Physical Constants: 2006," *Rev. Mod. Phys.*, **80**, June 2008.

[6] BIPM JCGM 200:2008, *International vocabulary of metrology — Basic and general concepts and associated terms* (VIM), JCGM, 2008.

[7] NASA Metrology and Calibration Working Group, *Metrology – Calibration and Measurement Processes Guidelines*, NASA Reference Publication 1342, Sec. 6.4 and Appendix D, Jet Propulsion Laboratory, Pasadena, 1994.

[8] NASA-HNBK-8739.19-3, *Measurement Uncertainty Analysis Principles and Methods,* NASA Measurement Quality Assurance Handbook **Annex 3**, Spring 2010.

[9] Castrup, H., "Uncertainty Analysis and Parameter Tolerancing," Presented at the NCSL Workshop & Symposium, Dallas, July 1995.

[10] Castrup, H., "Estimating Category B Degrees of Freedom," *Proc. Meas. Sci. Conf.*, Anaheim, January 2000.

[11] Castrup, H., "A Welch-Satterthwaite Relation for Correlated Errors," *Proc. Meas. Sci. Conf.*, Anaheim, March 2010.

[12] Castrup, H. and Castrup, S., "Uncertainty Analysis for Alternative Calibration Scenarios," *Proc. NCSLI Workshop & Symposium*, Orlando, August 2008.

[13] Castrup, H., "Decision Risk Analysis for Alternative Calibration Scenarios," *Proc. NCSLI Workshop & Symposium*, Orlando, August 2008.

[14] Castrup, S., "Comparison of Methods for Establishing Confidence Limits and Expanded Uncertainties," Proc. Meas. Sci. Conf., Pasadena, March 2010.

[15] Castrup, H., "Analytical Metrology SPC Methods for ATE Implementation," *Proc. NCSL Workshop and Symposium*, Albuquerque, July 1991.

[16] NASA-HNBK-8739.19, *Measurement Quality Assurance Handbook*, Spring 2010.

[17] Castrup, H., "Risk Analysis Methods for Complying with Z540.3," *Proc. NCSLI Workshop & Symposium*, St. Paul, August 2007.

[18] Castrup, S., "Important Elements of an Uncertainty Analysis Report," *Proc. NCSLI Workshop & Symposium*, Providence, July 2010.

[19] ISG, *RiskGuard 2.0*, www.isgmax.com/risk_freeware.htm, © 2007, Integrated Sciences Group.

[20] Castrup, H., *Evaluation of Customer and Manufacturer Risk vs. Acceptance Test In-Tolerance Level*, TRW Technical Report No. 99900-7871-RU-00, April 1978.

[21] Castrup, H., “An Examination of Measurement Decision Risk and Other Measurement Quality Metrics,” *Proc. NCSLI Workshop & Symposium*, San Antonio, July 2009.

[22] NASA-HNBK-8739.19-4, *Estimation and Evaluation of Measurement Decision Risk*, NASA Measurement Quality Assurance Handbook **Annex 4**, Spring 2010.

[23] Castrup, H., "Intercomparison of Standards: General Case," SAI Comsystems, D.O. 4M03, Dept. of the Navy Contract N00123-83-D-0015, 16 March 1984.

[24] Jackson, D., *Instrument Intercomparison: A General Methodology*, Analytical Metrology Note AMN 86-1, U.S. Navy Metrology Engineering Center, NWS Seal Beach, January 1, 1986; and "Instrument Intercomparison and Calibration," Proc. Meas. Sci. Conf., Irvine, January 1987.

[25] MIL-STD 45662A, *Calibration Systems Requirements*, U.S. Dept. of Defense, 1 August 1988.

[26] ANSI/NCSL Z540-1:1994, *Calibration Laboratories and Measuring and Test Equipment – General Requirements*, July 1994.

[27] ANSI/NCSL Z540.3-2006, *Requirements for the Calibration of Measuring and Test Equipment*, August 2006.

[28] *Uncertainty / SPC Analysis Training*, © 1995-2012, Integrated Sciences Group, All Rights Reserved.

[29] NASA-HNBK-8739.19-2, *Measuring and Test Equipment Specifications*, NASA Measurement Quality Assurance Handbook, **Annex 2**, Spring 2010.

[29] Bayes, T., “An Essay towards solving a Problem in the Doctrine of Chances,” *Phil. Trans.* **53**, 1763, 376-98.

[30] Eagle, A., "A Method for Handling Errors in Testing and Measuring," *Industrial Quality Control*, March, 1954.

[31] Grubbs, F. and Coon, H., "On Setting Test Limits Relative to Specification Limits," *Industrial Quality Control*, March, 1954.

[32] NCSLI, *Handbook for the Application of ANSI/NCSL Z540.3-2006*, NCSL International, February 2009.

[33] ISO 17043, *Conformity Assessment – General Requirements for Proficiency Testing*, Geneva, Switzerland.

Related Reading

ANSI/NCSL Z540-2, *U.S. Guide to the Expression of Uncertainty in Measurement*, 1st Ed., 1997, Boulder.

ASME B89.7.4.1-2005, *Measurement Uncertainty and Conformance Testing: Risk Analysis*, ASME Technical Report, February 3, 2006.

Cousins, R., “Why Isn’t Every Physicist a Bayesian?” *Am. J. Phys.*, **63**, No. 5, May 1995.

NASA-HNBK-8739.19-5, *Establishment and Adjustment of Calibration Intervals*, NASA Measurement Quality Assurance Handbook, **Annex 5**, Spring 2010.

Sommer, K., et al., " A Consideration of Correlations in Modeling and Uncertainty Evaluation of Measurements," *Proc. NSCLI Workshop & Symposium*, Salt Lake City, July 2004.

Toman, B., "A Bayesian Approach to Estimation of a Key Comparison Reference Value," *Proc. NSCLI Workshop & Symposium*, Salt Lake City, July 2

Appendix

A.1 Measurement Decision Risk Analysis

A.1.1 Risk Analysis Variables

The basic set of variables that are used in estimating measurement decision risk is shown in Table A-1. Lower and upper UUT tolerance limits are labeled L_1 and L_2, and lower and upper UUT acceptance limits are denoted A_1 and A_2. The symbols $\mathcal{L}$ and $\mathcal{A}$ are defined as $\mathcal{L} = [-L_1, L_2]$ and $\mathcal{A} = [-A_1, A_2]$.

Table A-1. Risk Variables Nomenclature

Variable	**Definition**
ε_{UUT}	the bias of the UUT attribute value at the time of calibration
u_{UUT}	the uncertainty in ε_{UUT}, i.e., the standard deviation of the probability distribution of the population of ε_{UUT} values.
$-L_1$ and L_2	the tolerance limits for ε_{UUT}
$-A_1$ and A_2	the "acceptance" limits (test limits) for ε_{UUT}
$\mathcal{L}$	the range of values of ε_{UUT} from $-L_1$ to L_2 (the UUT tolerance limits)
$\mathcal{A}$	the range of values of ε_{UUT} from $-A_1$ to A_2 (the UUT acceptance limits)
δ	a measurement (estimate) of ε_{UUT}
ε_{cal}	total error in the measurement of δ [25]
u_{cal}	the uncertainty in ε_{cal} (same as the uncertainty in ε_{meas})

A.1.2 Probability Relations

Measurement decision risk analysis consists of computing various probability functions. The basic probability functions are given in Table A-2. In constructing these functions, we make use of a notation in which the $\in$ operator reads "belongs to" or "is "included in." Likewise, the $\notin$ operator reads "does not belong to" or "is excluded from." In addition, we denote the occurrence of an event by the symbol E and the non-occurrence of an event by the symbol $\bar{E}$. We also express probabilities in standard probability notation, in which the function $P(E_1,E_2)$ denotes the probability that events E_1 and E_2 will *both* occur, and the function $P(E_2|E_1)$ denotes the probability that event E_2 will occur, *given* that event E_1 has occurred.

[25] For many, but not all, measurement scenarios, $\varepsilon_{cal} = \varepsilon_{meas}$ [12].

Table A-2. Risk Computation Nomenclature

Risk Analysis Function	**Definition**
$P(\varepsilon_{UUT} \in \mathcal{L})$	the *a priori* probability that $-L_1 \leq \varepsilon_{UUT} \leq L_2$. This is the probability that the UUT attribute is in-tolerance at the time of measurement.
$P(\varepsilon_{meas} \in \mathcal{A})$	the probability that a measured UUT bias δ satisfies the condition $-A_1 \leq \delta \leq A_2$. This is the probability that a measured value of ε_{UUT} will be observed to be in-tolerance.
$P(\varepsilon_{UUT} \in \mathcal{L}, \varepsilon_{meas} \in \mathcal{A})$	the probability that $-L_1 \leq \varepsilon_{UUT} \leq L_2$ and $-A_1 \leq \delta \leq A_2$. This is the joint probability that a UUT attribute will be in-tolerance *and* will be observed to be in-tolerance.
$P(\varepsilon_{meas} \in \mathcal{A} \mid \varepsilon_{UUT} \in \mathcal{L})$	the probability that, if $-L_1 \leq \varepsilon_{UUT} \leq L_2$, then $-A_1 \leq \delta \leq A_2$. This is the conditional probability that an in-tolerance attribute will be observed to be in-tolerance.
$P(\varepsilon_{UUT} \notin \mathcal{L}, \varepsilon_{meas} \in \mathcal{A})$	the probability that ε_{UUT} lies outside $\mathcal{L}$ and that $-A_1 \leq \delta \leq A_2$. This is the joint probability that a UUT attribute will be out-of-tolerance *and* will be observed to be in-tolerance.
$P(\varepsilon_{UUT} \in \mathcal{L}, \varepsilon_{meas} \notin \mathcal{A})$	the probability that $-L_1 \leq \varepsilon_{UUT} \leq L_2$ and δ lies outside $\mathcal{A}$. This is the joint probability that a UUT attribute will be in-tolerance *and* will be observed to be out-of-tolerance.
$P(\varepsilon_{UUT} \notin \mathcal{L} \mid \varepsilon_{meas} \in \mathcal{A})$	the probability that ε_{UUT} lies outside $\mathcal{L}$ *given* that $-A_1 \leq \delta \leq A_2$. This is the conditional probability that a UUT attribute observed to be in-tolerance will be out-of-tolerance.

Table A-3 shows the equivalence of the probability functions in Table A-2 with generic risk probability functions. In Table A-3, the variable E_L represents the event that the UUT attribute is in-tolerance and E_A represents the event that the attribute is *observed* to be in-tolerance.

Table A-3. Risk Analysis Probability Representations

Description	**Risk Analysis Function**	**Basic Probability Representation**
Probability that a UUT attribute is in-tolerance	$P(\varepsilon_{UUT} \in \mathcal{L})$	$P(E_L)$
Probability that the measurement result (measured bias) is accepted as being in-tolerance	$P(\delta \in \mathcal{A})$	$P(E_A)$
Probability that the UUT attribute is in-tolerance and accepted as being in-tolerance	$P(\varepsilon_{UUT} \in \mathcal{L},\ \delta \in \mathcal{A})$	$P(E_L, E_A)$
Probability that an in-tolerance UUT attribute will be accepted as in-tolerance	$P(\delta \in \mathcal{A} \mid \varepsilon_{UUT} \in \mathcal{L})$	$P(E_A \mid E_L)$
Probability that the UUT attribute is not in-tolerance and is accepted as being in-tolerance,	$P(\varepsilon_{UUT} \notin \mathcal{L}, \delta \in \mathcal{A})$	$P(\bar{E}_L, E_A)$
Probability that the UUT attribute is in-tolerance and rejected as being out-of-tolerance, i.e., *FRR*	$P(\varepsilon_{UUT} \in \mathcal{L}, \delta \notin \mathcal{A})$	$P(E_L, \bar{E}_A)$
Probability that an accepted UUT attribute is out-of-tolerance	$P(\varepsilon_{UUT} \notin \mathcal{L} \mid \delta \in \mathcal{A})$	$P(\bar{E}_L \mid E_A)$

A.1.3 Applicable Probability Density Functions

The probability functions of Tables A-2 and A-3 are mathematically represented by the probability density functions (pdfs), shown in Table A-4. These pdfs relate random variables of interest to their probability of occurrence.

Table A-4

Risk Analysis Probability Density Functions[26]

pdf	**Description**
$f(\varepsilon_{UUT})$	pdf for the UUT bias at the time of calibration
$f(\delta)$	pdf for the measurement result
$f(\delta, \varepsilon_{UUT})$	pdf for the joint distribution of δ and ε_{UUT}
$f(\delta \mid \varepsilon_{UUT})$	pdf for the conditional distribution of δ given ε_{UUT}
$f(\varepsilon_{UUT} \mid \delta)$	pdf for the conditional distribution of $\varepsilon_{UUT,}$ given a measured value δ

A.1.4 The Classical Method

Using the definitions in Tables A-1 through A-4, the probability definitions used in the Classical Method can be written [7, 13, 17, 20, 21, 22]

$$P(E_L) = \int_{-L_1}^{L_2} f(\varepsilon_{UUT})\, d\varepsilon_{UUT}\ , \tag{A-1}$$

$$\begin{aligned} P(E_L, E_A) &= \int_{-L_1}^{L_2} \int_{-A_1}^{A_2} f(\delta, \varepsilon_{UUT})\, d\delta\, d\varepsilon_{UUT} \\ &= \int_{-L_1}^{L_2} \int_{-A_1}^{A_2} f(\delta \mid \varepsilon_{UUT}) f(\varepsilon_{UUT})\, d\delta\, d\varepsilon_{UUT}\ , \end{aligned} \tag{A-2}$$

and

$$\begin{aligned} P(E_A) &= \int_{-\infty}^{\infty} \int_{-A_1}^{A_2} f(\delta, \varepsilon_{UUT})\, d\delta\, d\varepsilon_{UUT} \\ &= \int_{-\infty}^{\infty} \int_{-A_1}^{A_2} f(\delta \mid \varepsilon_{UUT}) f(\varepsilon_{UUT})\, d\delta\, d\varepsilon_{UUT}\ . \end{aligned} \tag{A-3}$$

With the use of the probability notation of Section A.1.2, these relations yield convenient expressions for false accept risk *FAR*, defined as the probability that a UUT attribute is out-of-tolerance and accepted, and the term *FRR*, defined as the probability that a UUT attribute is in-tolerance and rejected:

$$FAR = P(\bar{E}_L, E_A) = P(E_A) - (E_L, E_A) \tag{A-4}$$

and

$$FRR = P(E_L, \bar{E}_A) = P(E_L) - (E_L, E_A)\, . \tag{A-5}$$

---

[26] To be more rigorous with respect to notation, each pdf would have its own letter designator or subscript to distinguish its functional form from other pdfs. Such rigor is laudable but leads to a more tedious notation than we already have. It is hoped that the distinct character of each pdf will be apparent from its context of usage.

From these expressions, practitioners of statistics will recognize *FAR* and *FRR* as *consumers' risk* and *producer's risk*, respectively [30, 31].

A.1.4.1 Classical Method Risk Computation

With the Classical Method, it is ordinarily assumed that the measurement result δ is normally distributed with a mean value of ε_{UUT} and a standard deviation u_{cal}. Under this assumption, Eqs. (A-1), (A-2) and (A-3) become

$$\begin{aligned} P(E_L) &= \frac{1}{\sqrt{2\pi}u_{UUT}} \int_{-L_1}^{L_2} e^{-\varepsilon_{UUT}^2/2u_{UUT}^2} d\varepsilon_{UUT} \\ &= \Phi\left(\frac{L_1}{u_{UUT}}\right) + \Phi\left(\frac{L_2}{u_{UUT}}\right) - 1. \end{aligned} \tag{A-6}$$

$$\begin{aligned} P(E_L, E_A) &= \frac{1}{\sqrt{2\pi}u_{cal}} \int_{-L_1}^{L_2} \int_{-A_1}^{A_2} e^{-(\delta-\varepsilon_{UUT})^2/2u_{cal}^2} f(\varepsilon_{UUT})\, d\delta\, d\varepsilon_{UUT} \\ &= \int_{-L_1}^{L_2} \left[\Phi\left(\frac{A_1+\varepsilon_{UUT}}{u_{cal}}\right) + \Phi\left(\frac{A_2-\varepsilon_{UUT}}{u_{cal}}\right) - 1 \right] f(\varepsilon_{UUT})\, d\varepsilon_{UUT}\,, \end{aligned} \tag{A-7}$$

and

$$\begin{aligned} P(E_A) &= \frac{1}{\sqrt{2\pi}u_{cal}} \int_{-\infty}^{\infty} \int_{-A_1}^{A_2} f(\varepsilon_{UUT}) e^{-(\delta-\varepsilon_{UUT})^2/2u_{cal}^2}\, d\delta\, d\varepsilon_{UUT} \\ &= \int_{-\infty}^{\infty} \left[\Phi\left(\frac{A_1+\varepsilon_{UUT}}{u_{cal}}\right) + \Phi\left(\frac{A_2-\varepsilon_{UUT}}{u_{cal}}\right) - 1 \right] f(\varepsilon_{UUT})\, d\varepsilon_{UUT}\,, \end{aligned} \tag{A-8}$$

where Φ is the normal distribution function available in most spreadsheet applications.

The variable ε_{UUT} may follow any number of plausible probability distributions. In all cases, ε_{UUT} is assumed to have a zero mean value[27] and a standard deviation of u_{UUT}. Like the variable δ, ε_{UUT} is often assumed to be normally distributed. For such cases, Eqs. (A-6), (A-7) and (A-8) become

$$\begin{aligned} P(E_L) &= \frac{1}{\sqrt{2\pi}u_{UUT}} \int_{-L_1}^{L_2} e^{-\varepsilon_{UUT}^2/2u_{UUT}^2} d\varepsilon_{UUT,b} \\ &= \Phi\left(\frac{L_1}{u_{UUT}}\right) + \Phi\left(\frac{L_2}{u_{UUT}}\right) - 1, \end{aligned} \tag{A-9}$$

$$P(E_L, E_A) = \frac{1}{\sqrt{2\pi}u_{UUT}} \int_{-L_1}^{L_2} \left[\Phi\left(\frac{A_1+\varepsilon_{UUT}}{u_{cal}}\right) + \Phi\left(\frac{A_2-\varepsilon_{UUT}}{u_{cal}}\right) - 1 \right] e^{-\varepsilon_{UUT}^2/2u_{UUT,b}^2}\, d\varepsilon_{UUT}\,, \tag{A-10}$$

and

[27] A somewhat tacit assumption implied in this statement is that the expectation value of a measurement result is equal to the true value.

$$P(E_A) = \frac{1}{2\pi u_{UUT} u_{cal}} \int_{-\infty}^{\infty} \int_{-A_1}^{A_2} e^{-(\delta-\varepsilon_{UUT})^2/2u_{cal}^2} e^{-\varepsilon_{UUT}^2/2u_{UUT}^2} d\delta \, d\varepsilon_{UUT}$$
$$= \Phi\left(\frac{A_1}{u_A}\right) + \Phi\left(\frac{A_2}{u_A}\right) - 1, \quad \text{(A-11)}$$

where

$$u_A = \sqrt{u_{UUT}^2 + u_{cal}^2} \, . \quad \text{(A-12)}$$

A.1.5 The Bayesian Method

A.1.5.1 Bayes' Theorem

The Bayesian Method was derived from Bayes' theorem [23, 29]. In its simplest form, using the definitions of Tables 2 and 4, this theorem is given by

$$f(\varepsilon_{UUT} \mid \delta) = \frac{f(\delta \mid \varepsilon_{UUT}) f(\varepsilon_{UUT})}{f(\delta)} . \quad \text{(A-13)}$$

A.1.5.2 Estimating *FAR*, *FRR*, ε_{UUT} and u_{UUT}

Assuming normally distributed variables in Eq. (A-13), the post-test pdf for ε_{UUT} is given by

$$f(\varepsilon_{UUT} \mid \delta) = \frac{1}{\sqrt{2\pi} u_\beta} \exp\left\{-\left[(\delta - \varepsilon_{UUT})^2 / 2u_{cal}^2 + \varepsilon_{UUT}^2 / 2u_{UUT}^2 - \delta^2 / 2u_A^2\right]\right\}$$
$$= \frac{1}{\sqrt{2\pi} u_\beta} e^{-(\varepsilon_{UUT}-\beta)^2/2u_\beta^2} , \quad \text{(A-14)}$$

where the variables are as defined in Table A-3. The parameter β in Eq. (A-14) is a post-test estimate of the *a priori* value of ε_{UUT}, and u_β is a post-test estimate of the *a priori* value of u_{UUT}, given by

$$\beta = \frac{u_{UUT}^2}{u_A^2} \delta \quad \text{(A-15)}$$

and

$$u_\beta = \frac{u_{UUT} u_{cal}}{u_A} . \quad \text{(A-16)}$$

An estimate of the UUT attribute in-tolerance probability $P_{UUT,in}$ is obtained by integrating $f(\varepsilon_{UUT} \mid \delta)$ from $-L_1$ to L_2

$$P_{UUT,in} = \frac{1}{\sqrt{2\pi} u_\beta} \int_{-L_1}^{L_2} e^{-(e_{UUT,b}-\beta)^2/2u_\beta^2} de_{UUT,b}$$
$$= \Phi\left(\frac{L_1 + \beta}{u_\beta}\right) + \Phi\left(\frac{L_2 - \beta}{u_\beta}\right) - 1. \quad \text{(A-17)}$$

The probability $P_{UUT,in}$ is the false reject risk *FRR* if the UUT attribute is rejected. If the attribute is accepted without correction, the false accept risk is just[28]

$$FAR = 1 - P_{UUT,in} \,. \tag{A-18}$$

A.1.5.3 Estimating ε_{ref} and u_{ref}

The first step in obtaining a post-test estimate of the uncertainty in ε_{ref} is to define a new uncertainty term

$$u_{process} = \sqrt{u_{cal}^2 - u_{ref}^2} \,. \tag{A-19}$$

Next, a calibration uncertainty is defined that would apply if the UUT were calibrating the reference

$$u'_{cal} = \sqrt{u_{UUT}^2 + u_{process}^2} \,. \tag{A-20}$$

With the Bayesian Method, the post-test pdf for the *a priori* value of ε_{ref} is obtained by modifying Eq. (A-15) by replacing δ with $-\delta$, u_{cal} with u'_{cal}, and the *a priori* estimate of u_{UUT} with the *a priori* estimate of u_{ref} to get

$$f(\varepsilon_{ref} \mid \delta) = \frac{1}{\sqrt{2\pi}u_\alpha} e^{-(\varepsilon_{ref} - \alpha)^2 / 2u_\alpha^2} \,, \tag{A-21}$$

$$\alpha = -\frac{u_{MTE,b}^2}{u_A^2}\delta \,. \tag{A-22}$$

and

$$u_\alpha = \frac{u_{ref} u'_{cal}}{u_A} \,. \tag{A-23}$$

From Eq. (A-21), we see that the estimates of the bias of the calibration reference attribute and the uncertainty in this estimate are α and u_α, respectively.

Since we have the necessary expressions at hand, we can also estimate the in-tolerance probability $P_{ref,in}$ of the reference. Letting $-l_1$ and l_2 represent lower and upper tolerance limits for the reference attribute, this probability is obtained by integrating the pdf $f(e_{ref} \mid \delta)$ in Eq. (A-23) from $-l_1$ to l_2 to get

$$\begin{aligned} P_{ref,in} &= \frac{1}{\sqrt{2\pi}u_{ref}} \int_{-l_1}^{l_2} e^{-(e_{ref} - \alpha)^2 / 2u_\alpha^2} de_{MTE,b} \\ &= \Phi\left(\frac{l_1 + \alpha}{u_\alpha}\right) + \Phi\left(\frac{l_2 - \alpha}{u_\alpha}\right) - 1. \end{aligned}$$

[28] Strictly speaking, the Bayesian false accept and false reject risks of Eqs. (A-17) and (A-18) are not the *FAR* and *FRR* defined in Eqs. (A-4) and (A-5), in which both risks are joint probabilities. The risks of Eqs. (A-17) and (A-18) are *conditional* probabilities, since both are probabilities conditional on the measurement result δ.

A.1.6 The Confidence Level Method

With the confidence level method, we represent values of the population from which δ was obtained with a random variable ζ, and assume that ζ is normally distributed with mean δ and standard deviation u_{cal}. We then obtain an in-tolerance confidence from[29]

$$\begin{aligned} P_{UUT,in} &= \frac{1}{\sqrt{2\pi}u_{cal}} \int_{-L_1}^{L_2} e^{-(\zeta-\delta)^2/2u_{cal}^2} d\zeta \\ &= \Phi\left(\frac{L_1+\delta}{u_{cal}}\right) + \Phi\left(\frac{L_2-\delta}{u_{cal}}\right) - 1. \end{aligned} \quad \text{(A-24)}$$

A.1.7 The TUR Method

A.1.7.1 The Z540.3 Definition

Z540.3 defines TUR as the span of the UUT tolerance limits divided by twice the "95%" expanded uncertainty of the measurement process. A caveat is provided in the form of a note stating that this requirement applies only to two-sided tolerances. Mathematically, the TUR so defined is stated as

$$\text{TUR} = \frac{L_1 + L_2}{2U_{95}}, \quad \text{(A-25)}$$

where U_{95} is the expanded uncertainty of the measurement process multiplied by a coverage factor k_{95} that presumably corresponds to a 95% confidence level

$$U_{95} = k_{95} u_{cal}. \quad \text{(A-26)}$$

In Z540.3, $k_{95} = 2$.

In addition to restricting the applicability of Eq. (A-25) to the calibration of UUT attributes with two-sided tolerance limits, Z540.3 also advises that Eq. (A-25) is strictly valid only in cases where the tolerance limits are symmetric, i.e., where $L_1 = L_2$. In such cases, the UUT attribute tolerance limits could be expressed in the form $\pm L$, and we would have[30]

$$\text{TUR} = \frac{L}{U_{95}}. \quad \text{(A-27)}$$

A.2 Hypothesis Testing

A.2.1 Stating the Hypothesis

A classic example of a hypothesis test is one in which two sample means are compared to see if they are significantly different [28]. The hypothesis being tested is a statement that both samples belong to the same population. If this hypothesis is rejected with some level of statistical significance, α, the samples are said to be incompatible with one another with $(1 - \alpha)\times 100\%$ confidence.

---

[29] Cases where the UUT attribute has a single-sided upper or a single-sided lower tolerance limit can be accommodated by setting $L_1 = \infty$ or $L_2 = \infty$, respectively.

[30] For other limitations of the TUR Method, see Ref [17].

For example, suppose we want to test whether a given lab is in agreement with a higher-level reference lab. The hypothesis H_0 states that the two sample means come from the same population and is written

$$H_0 : \quad \bar{x}_{lab} = \bar{x}_{ref} \,,$$

where

$\bar{x}_{lab}$ - mean value obtained by the test lab from a sample of measurements of an artifact

$\bar{x}_{ref}$ - mean value obtained by the reference lab from a sample of measurements of the same artifact

If H_0 is rejected, the test lab sample mean is pronounce defective or “significantly different” from the reference lab sample mean.

A.2.3 Constructing the Test Statistic

The recommended statistic to employ to test H_0 is one in which we assume that a test variable t_c is t-distributed, where t_c is given by[31]

$$t_c = \left| \frac{\bar{x}_{lab} - \bar{x}_{ref}}{\sqrt{u_{lab}^2 + u_{ref}^2}} \right| ,$$

with

$$u_{lab}^2 = \frac{s_{lab}^2}{n_{lab}} + u_{lab,other}^2 \,,$$

and

$$u_{ref}^2 = \frac{s_{ref}^2}{n_{ref}} + u_{ref,other}^2 \,,$$

where

s_{lab} - standard deviation of the sample of measurements taken by the test lab
s_{ref} - standard deviation of the sample of measurements taken by reference lab
n_{lab} - sample size of the sample of measurements taken by the test lab
n_{ref} - sample size of the sample of measurements taken by the reference lab
$u_{lab,other}$ - combined measurement uncertainty of other test lab measurement errors
$u_{ref,other}$ - combined measurement uncertainty of other reference lab measurement errors

It is important that both labs take into account errors from all sources. For instance, suppose that repeatability error is mistakenly excluded as an error source by the reference lab or the test lab. If so, then the denominator of t_c will be smaller that it should be relative to the estimated

[31] Historically, what has been applied instead of the statistic t_c is a variable E_n, given by [33]

$$E_n = \frac{|\,\bar{x}_{lab} - \bar{x}_{ref}\,|}{\sqrt{U_{lab}^2 + U_{ref}^2}} \,,$$

where the uncertainties in the denominator are expanded uncertainties equal to two times the standard uncertainties. The lab fails is $E_n > 1$. This test may be thought of as a crude version of the test recommended in this paper.

difference in bias between the two labs. This will make the value of t_c overly large and the hypothesis could be erroneously rejected, i.e., the lab could mistakenly fail the test. Another possibility is that the test lab will correctly divide its sample standard deviation by the square root of its sample size and the reference lab will not.[32] This could again resulting in an erroneous failure. Of course, the reverse situation is also possible, leading to a smaller than appropriate value of t_c, possibly resulting in an erroneous acceptance.

A.2.4 Choosing the Critical Statistic

To test the value of t_c, we need to choose a critical statistic that corresponds to a desired confidence level C for testing the hypothesis and the degrees of freedom of the combined uncertainty. The degrees of freedom is given by

$$\nu = \frac{(u_{lab}^2 + u_{ref}^2)^2}{\dfrac{u_{lab}^4}{\nu_{lab}} + \dfrac{u_{ref}^4}{\nu_{ref}}},$$

and the critical statistic is just the t-statistic $t_{\alpha,\nu}$, where $\alpha = 1 - C$.

A.2.5 Performing the Test

If $t_c > t_{\alpha,\nu}$ then we say that $\bar{x}_{lab} \neq \bar{x}_{ref}$ with confidence C. Otherwise, we accept the hypothesis H_0 that measurements made by the two labs belong to the same population.[33]

[32] This is needed because what is being compared is the difference between mean values. The rationale is covered in statistics texts in discussions of the "sampling distribution."

[33] This test is appropriate for interlaboratory comparisons or "round robins," in which the reference lab is referred to as the "pivot" lab and there are two or more test labs.