Evaluation of Out-of-Tolerance Risk in Measuring and Test Equipment

Speaker/author: Rick Walker
Fluke Corporation, Hart Scientific Division
799 E Utah Valley Dr.
American Fork, Utah, U.S.A.
phone: 801-763-1600
e-mail: rick.walker@hartscientific.com

Abstract

In the interest of providing reliable and accurate measuring and test equipment to customers, a manufacturer or calibration service provider should understand and control out-of-tolerance risk—the probability an instrument is actually out of tolerance after passing calibration. While the equipment is expected to be in tolerance when it leaves the calibration laboratory, it is just as important that it remain in tolerance during use. A manufacturer also does not want to see too many products fail calibration or be found out of tolerance later at recalibration. Pass yields must be high. Risks and yields are dependent on product specifications, guard bands, calibration uncertainties, and product characteristics, such as long-term stability, ambient temperature sensitivity, and linearity. This paper provides formulae for calculating various relevant risks and yields. Examples are given. Calculations are verified by computer simulation.

1. Introduction

A manufacturer of measuring and test equipment must ensure that its products are reliable and accurate, especially if they are to be used as calibration standards. Instruments must be well designed, properly calibrated, and given realistic specifications. There must be low risk that an instrument is operating outside specifications when it leaves the factory. Likewise, there should be high confidence in the product as it experiences a variety of conditions in the field and that it will be found in tolerance when it is retested after use.

At the same time, it is in the best interest of the manufacturer, and perhaps the customer too, that out-of-tolerance risks not be unnecessarily low. Otherwise the product might cost more than it should. It might be over-designed, have understated specifications, or be calibrated with an unduly expensive system. It might also mean that production yields could be improved with more liberal acceptance limits.

Managing out-of-tolerance risks requires an understanding of the various product and calibration errors present during the calibration cycle. Then, with proper analysis, risks can be evaluated. Calibration equipment and procedures, including guard band limits, can be selected to provide adequate confidence without unnecessary costs. The manufacturer might realize that test uncertainty ratios can be less strict than otherwise assumed, perhaps even 2:1. Or it might be found that better product specifications are feasible. A thorough out-of-tolerance risk (OTR) analysis can lead to improved reliability for the consumer while offering cost savings and marketing advantage for the producer.

2. Out-of-tolerance conditions

Having an instrument not meeting specifications is bad for the owner or operator of an instrument and bad for the manufacturer. It can be costly. The owner may need to take special action to mitigate the problem, perhaps even recall work done. The manufacturer may lose customers' trust. Clearly, it is a situation to be avoided. Without due consideration, out-of-tolerance conditions may be much too likely to occur.

Consider the following scenario. A supplier delivers a new measuring instrument to a customer. The customer will be using it to test other equipment and is depending on it for quality. The product has been tested and calibrated, and the calibration report shows that its error is just within specifications. After some time, the owner wishes to have the instrument tested to make sure it is still in tolerance. The instrument is sent to a national metrology institute (NMI) that offers very low test uncertainties. The NMI reports back to the owner that the instrument has failed its tests.

Reasons the instrument may fall out of tolerance are varied:

- Calibration error during the original calibration happened to cause the product to appear more accurate than it really was. The product was actually out of tolerance when it was delivered to the customer.
- The product has experienced drift since it was originally calibrated. Though it may have been in tolerance at first, now it is not.
- The product is sensitive to ambient temperature, and the NMI tested it at the edge of its specified temperature range, where it failed. The original calibration was performed near the center of the range.
- The product has nonlinearity. The NMI has chosen measurement points that differ from those at which the product was calibrated by the manufacturer. The product may be in tolerance at the manufacturer's calibration points but out at others.

In a similar scenario, equipment is returned to the manufacturer for routine calibration. The manufacturer reports that it appears to be out of tolerance. This might occur for some of the same reasons as before, but also because the calibration error for the as-found test is enough to cause the product to appear out of tolerance when it really is not.

The manufacturer must take measures to prevent such situations. While there might never be absolute certainty equipment will always perform within specifications, the risk of an out-of-tolerance condition when it is operated as intended should be reasonably low.

Out-of-tolerance risk can be reduced by using stricter acceptance criteria during calibration. The trade-off is that more products are likely to fail calibration. Yields, for both original calibration and retesting, must be weighed.

In summary, the following risks should be managed:

- Maximum immediate OTR must be low: Of products that pass calibration, those that show the greatest error must still have acceptably low probability of actually being out of tolerance.
- First-pass yield must be high: A high percentage of products should pass calibration.
- Maximum field OTR must be low: Of products that pass calibration, those that show the greatest error must still have acceptably low probability of being out of tolerance over time and under various operating conditions in the field.
- Maximum retest OTR must be low: Of products passing retest, those that show the greatest error must still have acceptably low probability of being out of tolerance.
- Retest yield must be high: A high percentage of products should pass retest.

3. Maximum individual risk

An important point is that the risk to be managed is not population risk but maximum individual out-of-tolerance risk. For some types of products there is nothing to distinguish one unit from the next, and individual risk cannot be differentiated from population risk. Such is not the case with measuring and test equipment. Each instrument is individually tested; its error is measured and reported. Those with the largest observed error have the highest risk of being out of tolerance.

To illustrate the difference between population OTR and maximum individual OTR, suppose that ten customers receive equipment from a supplier. Nine receive instruments that have only 1% probability of being out of tolerance. The tenth receives an instrument that has 10% OTR. The population risk is only 1.9%. But that is of little consolation to the one customer who realizes the instrument he purchased has a high risk. *No customer should receive an instrument that is known to have high risk of being out of tolerance*. A consequence of this policy may be that most customers receive products with OTR much less than the maximum acceptable risk. A preferred choice for owners would be to know with absolute certainty their equipment is in tolerance, but that is generally not possible. A supplier can only ensure that every product has low probability of being out of tolerance.

Mathematically, assuming symmetric limits, population risk can be written as $P(|E| > L \mid |T| \leq A)$, the probability that the magnitude of the true error (E) is greater than the specification limit (L) given that the magnitude of the observed error (T) is *at or below* an acceptance limit (A). Instead, what is important is maximum individual risk (R):

$$R = P\left(|E| > L \mid |T| = A\right) \tag{1}$$

This is the conditional probability that the magnitude of the true error is greater than the specification limit when the magnitude of the observed error is *at* the acceptance limit.

4. Guard band

Out-of-tolerance risk is dependent on several factors. The tighter the specifications, the more challenging it is to ensure products meet those specifications. Calibration uncertainty and drift and other product related errors also lead to increased OTR. These factors may be difficult or costly to change.

One factor that can easily be changed is the acceptance limit, A. By making the acceptance limit tighter than the product specification, the risk that a passing product is actually out of tolerance is reduced. In practice, the specification is multiplied by a guard band factor (g) that is less than 1 ($A = g \cdot L$). The product does not pass if the observed error is outside the guard band limit, even though it may still be within published specifications. This provides some margin to allow for unknown calibration error and possible future variation of performance.

The guard band should be set just low enough to control risk, perhaps with a little extra margin, but not so low that yields are poor. This requires a careful mathematical analysis, based on estimated uncertainties for product and calibration errors.

5. Risk analysis for independent test process

To see how guard band affects OTR, consider the following. Suppose when a product is tested its observed error is T, which includes some test error, τ, which has a probability density function (PDF) that is $p_\tau(x)$. The true error of the product, E, is unknown but has an estimated PDF before testing that is $p_E(x)$. The observed error is the sum of the true error and test error. So what is the probability that the magnitude of E exceeds specification limit L after testing?

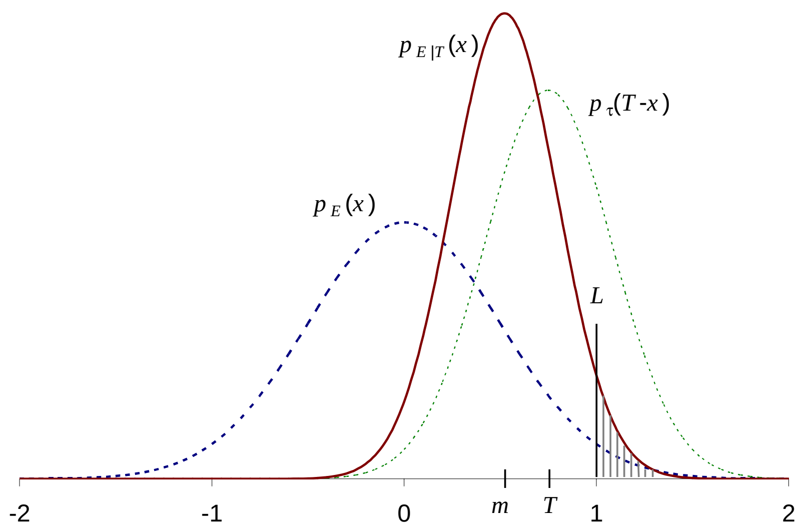

Figure 1. Product, test, and conditional PDFs.

Figure 1 illustrates the situation. The a priori PDF of the true error, before testing, is shown as $p_E(x)$ (with mean 0). When the error is observed as T, considering only possible test error, the PDF for the true error would be $p_\tau(T - x)$. The PDF for the true error *E given observed error T* combines the two. It is, in essence, the product of the two distributions (scaled to make the cumulative probability 1). The resulting distribution has a mean $m(T)$ between 0 and T and a smaller standard deviation than either of the component PDFs. The probability that the true error is outside specification $\pm L$ is the area under the curve that is to the right of L and left of $-L$.

It should be clear that this analysis is based on the assumption that the product error and the test error are independent and uncorrelated. The errors also have no bias, meaning their means, or expected values, are 0. And finally, the variables are assumed to follow normal distributions, as characterized by the following:

$$p_n(x,m,\sigma)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-m)^2}{2\sigma^2}} \tag{2}$$

Knowing $p_E(x)$ and $p_\tau(x)$ the conditional PDF for E given T can be evaluated. This is the probability of both E and corresponding τ that results in T [$\tau = T - E$], divided by the probability of T in any case:

$$p_{E|T}(E)=\frac{p_E(E)p_\tau(T-E)}{p_T(T)} \tag{3}$$

Since $T = E + \tau$, the PDF of T, the denominator of the equation, is the convolution of the PDFs of E and τ:

$$p_T(T)=\int_{-\infty}^{\infty}p_E(x)p_\tau(T-x)\,dx \tag{4}$$

Random variables E and τ have standard deviations σ_E and σ_τ respectively. When their PDFs are multiplied, as is done in both the numerator and denominator, the exponents of the two normal distribution functions are added. Through algebraic manipulation the exponent becomes

$$-\frac{\left(x-T\left(\frac{\sigma_\tau{}^2}{\sigma_E{}^2}+1\right)^{-1}\right)^2}{2\left(\sigma_E{}^{-2}+\sigma_\tau{}^{-2}\right)^{-1}}-\frac{T^2}{2\left(\sigma_E{}^2+\sigma_\tau{}^2\right)}$$

The second term in the exponent can be moved outside the exponential and integral as a factor that is constant relative to variable x or E. It divides out and the integral in the denominator becomes 1. The reduced equation is a normal distribution function with mean and variance as follows:

$$m(T)=T\left(\frac{\sigma_\tau{}^2}{\sigma_E{}^2}+1\right)^{-1}$$
$$\frac{1}{\sigma^2}=\frac{1}{\sigma_E{}^2}+\frac{1}{\sigma_\tau{}^2} \tag{5}$$

Interestingly, to one familiar with electrical resistance networks these equations appear much like those that describe the equivalent circuit of a voltage divider, with error variances replacing resistances, and means (or the observation) replacing potentials (Figure 2). The electrical analogue can be helpful for deriving the equations characterizing conditional probabilities, especially when numerous variables are involved.

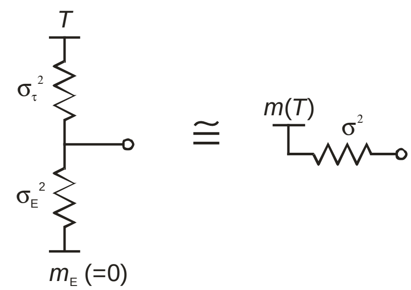

Figure 2. Electrical analogue of conditional PDF.

Finally, the risk $R(T)$ that E is between symmetric limits $-L$ and L when the observed error is T is

$$R(T)=\left(2-\Phi\left(\frac{L-m(T)}{\sigma}\right)-\Phi\left(\frac{L+m(T)}{\sigma}\right)\right)\cdot 100\% \tag{6}$$

Function Φ is the cumulative normal distribution function. (In Microsoft® Excel, the function NORMDIST( ) with cumulative option TRUE can be used.)

As an example, consider the case where the reported error is at the specification limit ($T = L = 1$, k=2) and test uncertainty ratio (TUR) is 4:1 ($\sigma_E = 0.5$, $\sigma_\tau = 0.125$). Out-of-tolerance risk is then calculated to be 31.4%. If instead T is 75% of L, the risk becomes 0.8%. The smaller guard band greatly reduces OTR.

Table 1 below shows various guard bands ($g = A / L \cdot 100\%$) required for several OTR and TUR values, assuming a specification that is $2\sigma_E$ and calibration uncertainty coverage factor k=2.

Table 1. Guard band vs. TUR and target risk.

TUR	**1% OTR**	**2%**	**5%**	**10%**
5:1	80.3%	83.1%	87.2%	90.9%
4:1	76.3%	79.8%	85.1%	89.7%
3:1	70.2%	75.0%	82.2%	88.6%
2:1	60.0%	67.6%	79.0%	89.2%

These results may be surprising. With a 2:1 TUR and an 80% guard band (the resulting margin being less than half the calibration uncertainty) risk of the product being out of tolerance might be supposed to be quite high. But maximum OTR is only about 5%. There are two reasons for this. First, only one of the two tails of the probability curve weighs heavily, since the peak is closer to one of the specification limits. More importantly, knowing that the actual error tends toward 0, a large observed error is likely partly due to test error, so the actual error is probably smaller than observed (a conclusion reached previously by Deaver [1]). Experience also supports this. Products that barely fail calibration the first time tend to pass when tested again.

Given a particular maximum acceptable risk, it may be prudent to target a somewhat lower risk and set a smaller guard band than the calculation suggests. This provides some extra margin to allow for possible abnormal product or calibration errors, such as when a standard used to calibrate the equipment is found to be out of tolerance itself. This can help to avoid unpleasant corrective actions.

The foregoing analysis, again, assumes the product and calibration errors are independent and uncorrelated—the product is manufactured and adjusted, if necessary, using a different process than the final test. However, such is generally not the case with measuring and test equipment. Typically, an instrument is adjusted based on measurements taken by the calibration system, and then the same system is used to test and report as-left errors. As a consequence, product and calibration errors are somewhat correlated. Furthermore, the analysis did not yet consider errors manifest subsequent to calibration, such as product drift. A more thorough treatment is needed.

6. Product and calibration errors

A typical calibration cycle will involve a variety of inaccuracies, of both the product and the calibration system. For this study, total true product errors are denoted E, product error components are ε, test system error components are τ, and observed errors are T. Each error has an associated mean and standard deviation, or standard uncertainty. For most product and calibration system error components the means are 0 because any significant known bias will have been corrected. Uncertainties associated with individual or combinations of errors can be estimated using familiar methods of uncertainty analysis [2]. Different errors affect the various risks and yields in different ways, so they must be distinguished.

Errors can result from imperfect calibration, of course. They can also arise from changes within the instrument, such as **drift** of electronic components. An instrument is usually less accurate near the end of its calibration interval than at the beginning. These effects must not be underestimated. During transit and routine handling, equipment may experience bumps and vibration, temperature changes, and extreme humidity, all of which affect stability. Estimates for long-term stability might be obtained by studying critical components, from experience, or from testing. As further experience is gained, estimates can be revised. Drift depends on time, so if there are specifications for different time intervals, such as 90-day and 1-year, drift should be estimated for each. Drift may tend to be in one direction, having a mean that is not 0. Here drift error is ε_d, with mean m_d and uncertainty u_d. To ensure worse-case probabilities are calculated, the absolute value $|m_d|$ is applied in the equations, which also take the positive symmetric guard band limit.

Product errors can also be caused by nonlinearity, sensitivity to ambient temperature or humidity, or other effects. These errors can be much larger in the field than in the calibration lab, where conditions are more controlled. Again, estimates might be obtained by analysis, experience, or testing. These combined **field** errors are denoted ε_f, with uncertainty u_f.

Calibration errors will be separated into three groups. The various calibration errors are manifest at different times during the calibration cycle and affect the risks and yields in different ways. For lack of better terms, *random*, *systematic*, and *alignment* distinguish the types of calibration errors. These labels should not be confused with the Type A and Type B designations used in uncertainty budgets. In this study the method used to obtain an uncertainty component is of little relevance.

Random calibration errors vary during the calibration process and are different for every measurement. The standard and the equipment under test may both have significant measurement noise. Uncertainty can be reduced by taking an average of a series of readings. An estimate of the combined random uncertainty should be the result of a detailed uncertainty analysis of the calibration process, focusing on those errors that vary during tests. Here random calibration error is τ_r. with uncertainty u_r.

Unknown **systematic** calibration errors remain constant during a calibration. An example of this is the difference between the assigned value of a standard resistor and its unknown true resistance. An estimate of the combined systematic uncertainty is the result of a detailed uncertainty analysis of the calibration process, focusing on those errors that are constant during tests. Systematic error is τ_s, with uncertainty u_s.

Over a long period of time, such as the calibration interval, the systematic error may vary. A standard may have drifted, been recalibrated, or exchanged. The following analysis takes this into consideration using a **systematic error variability** factor v_s. A value of 0 means the systematic error is expected to be the same every time the product is tested. A value of 1 means systematic errors are entirely uncorrelated. A value of 0.5 means systematic error consists of equal parts constant and random variances.

When measuring and test equipment is calibrated, adjustments are often made to correct for observed error. In many cases perfect **alignment**, even if there were no random or systematic calibration errors, is not possible. Perhaps there are fewer adjustment variables than there are calibration points, with residual nonlinearity errors remaining after a best-fit adjustment. In some equipment there is only an indirect relationship between adjustments and measurement error. Alignment could be an imperfect, iterative process, where at some time it is considered good enough. Some instruments are adjusted with potentiometers, which by their mechanical nature have limited precision. Alignment might even be skipped if observed error is deemed sufficiently small. In this study alignment error is τ_a, with uncertainty u_a.

Realize that uncertainties may be different at different points in the instrument's measurement range. This can lead to different risks, yields, and guard bands at different points.

7. Calibration cycle

The steps in a typical calibration cycle, with associated errors, are outlined as follows:

1. The measuring and test equipment begins with some error, E_0.

2. The equipment is tested to measure the error so it can be corrected. The test is affected by systematic and random errors. So the observed error is

$$T_{\mathrm{A}} = E_0 + \tau_{\mathrm{s1}} + \tau_{\mathrm{r1}} \tag{7}$$

3. The equipment is adjusted in an attempt to correct the error. The immediate true error after adjustment is

$$\begin{aligned} E_{\mathrm{I}} &= E_0 - T_{\mathrm{A}} + \tau_{\mathrm{a}} \\ &= -\tau_{\mathrm{s1}} - \tau_{\mathrm{r1}} + \tau_{\mathrm{a}} \end{aligned} \tag{8}$$

4. The equipment is tested again to report its as-left error. The test process has the same systematic error as before, but a different random error. The observed immediate error is

$$\begin{aligned} T_{\mathrm{I}} &= E_{\mathrm{I}} + \tau_{\mathrm{s1}} + \tau_{\mathrm{r2}} \\ &= -\tau_{\mathrm{r1}} + \tau_{\mathrm{a}} + \tau_{\mathrm{r2}} \end{aligned} \tag{9}$$

5. While in the field the equipment is subject to additional errors due to drift and other product imperfections, such as ambient temperature sensitivity and nonlinearity. The equipment's true error in the field is

$$E_{\mathrm{F}} = -\tau_{\mathrm{s1}} - \tau_{\mathrm{r1}} + \tau_{\mathrm{a}} + \varepsilon_{\mathrm{d}} + \varepsilon_{\mathrm{f}} \tag{10}$$

6. At the end of the calibration interval the equipment is returned for routine calibration. With drift having occurred but other conditions controlled, the total error is now

$$E_{\mathrm{R}} = -\tau_{\mathrm{s1}} - \tau_{\mathrm{r1}} + \tau_{\mathrm{a}} + \varepsilon_{\mathrm{d}} \tag{11}$$

7. At retest, the test system has different systematic error and random error from the original calibration. The observed error at retest is

$$T_{\mathrm{R}} = -\tau_{\mathrm{s1}} - \tau_{\mathrm{r1}} + \tau_{\mathrm{a}} + \varepsilon_{\mathrm{d}} + \tau_{\mathrm{s3}} + \tau_{\mathrm{r3}} \tag{12}$$

Having estimates of the uncertainties associated with the various errors, probability distributions and probabilities for the various risks and yields of interest can be calculated.

8. Immediate risk

Immediate risk is the probability a product is actually out of tolerance when testing reports that the error is at the guard band limit. This is the maximum *false accept* risk. Deriving it is a three-step process. First the conditional probability for $(-\tau_{\mathrm{r1}} + \tau_{\mathrm{a}})$ given $(-\tau_{\mathrm{r1}} + \tau_{\mathrm{a}}) + \tau_{\mathrm{r2}} = g \cdot L$ is obtained

(g the calibration guard band, L the specification limit). Then the PDF for E_{I}, which is $(-\tau_{r1} + \tau_a) - \tau_{s1}$, is characterized. Finally, the probability that $|E_{\mathrm{I}}|$ is greater than L is calculated.

The mean and variance for $(-\tau_{r1} + \tau_a)$ are obtained by applying Equation 5 (Section 5) replacing σ_τ with u_r, σ_E^2 with $(u_r^2 + u_a^2)$, and T with $g{\cdot}L$. Then the variance of systematic error is added. The result is a mean and standard deviation as follows:

$$m_{\mathrm{I}} = g \cdot L\left(\frac{u_{\mathrm{r}}^2}{u_{\mathrm{r}}^2 + u_{\mathrm{a}}^2} + 1\right)^{-1}$$
$$\sigma_{\mathrm{I}}^2 = \left(\left(u_{\mathrm{r}}^2 + u_{\mathrm{a}}^2\right)^{-1} + u_{\mathrm{r}}^{-2}\right)^{-1} + u_{\mathrm{s}}^2 \tag{13}$$

The immediate risk is then

$$R_{\mathrm{I}} = \left(2 - \Phi\left(\frac{L - m_{\mathrm{I}}}{\sigma_{\mathrm{I}}}\right) - \Phi\left(\frac{L + m_{\mathrm{I}}}{\sigma_{\mathrm{I}}}\right)\right) \cdot 100\% \tag{14}$$

The maximum acceptable immediate OTR must be decided. As written in ANSI/NCSL Z540.3-2006 [3], false-accept probability must not exceed 2%. A target OTR of 1% will allow some additional margin for unexpected errors. In practice, guard bands may end up being even lower to maintain acceptable field risk and retest yields.

9. First-pass yield

First-pass yield (FPY) is the expected percentage of products that will pass calibration. The PDF of observed error T is considered a normal distribution with mean 0 and variance that is the sum of the component variances, which are u_r^2, u_a^2, and u_r^2 again:

$$\sigma_T^2 = 2u_{\mathrm{r}}^2 + u_{\mathrm{a}}^2 \tag{15}$$

A product passes if $|T|$ is $g{\cdot}L$ or less. So first-pass yield is

$$Y_{\mathrm{T}} = \left(2 \cdot \Phi\left(\frac{g \cdot L}{\sigma_T}\right) - 1\right) \cdot 100\% \tag{16}$$

Of those products that exceed the guard band limit and fail calibration, few will actually be out of tolerance. A higher false-reject rate is the trade-off for lower OTR. But the manufacturer must deal with the failures. An instrument might be re-adjusted or repaired. If the first-pass yield is high, such costs may be minimal as few products will fail.

An acceptable yield limit must be decided. This may depend on the type of product, production volume, and cost. With high-profit margin, low-volume products, FPY of 98% or even as low as 95% might be acceptable. Low-margin, high-volume products might require FPY much closer to 100%.

As manufacturing proceeds, actual yield can be measured and compared to the original estimate as part of a quality-control process. If it is seen to differ substantially, assumptions and estimated uncertainties should be reviewed.

10. Field risk

An important metric too often overlooked is **field risk**. This is the probability a product becomes out of tolerance in the field, when errors due to drift (ε_d) and other variables (ε_f) are manifest. These errors affect the mean and variance as follows:

$$\begin{aligned} m_F &= m_I + |m_d| \\ \sigma_F{}^2 &= \sigma_I{}^2 + u_d{}^2 + u_f{}^2 \end{aligned} \tag{17}$$

Again, this is with the calibration observed error at the guard band. The field risk is then

$$R_F = \left(2 - \Phi\left(\frac{L - m_F}{\sigma_F}\right) - \Phi\left(\frac{L + m_F}{\sigma_F}\right)\right) \cdot 100\% \tag{18}$$

The acceptable limit for field risk should generally be no higher than 5%, which corresponds to a k=2 coverage factor commonly used in uncertainty budgets. For applications where reliability is particularly critical, lower risk might be called for. The manufacturer might also want to target a lower value considering that the quality of parts and materials could vary in the future.

11. Retest risk and guard band

When equipment is returned to the manufacturer or calibration lab for recalibration, the owner will want to know whether or not the instrument is in tolerance. If the error is observed to be near the specification limit, the condition may be difficult to determine because of calibration uncertainty. Only if the error is well within the specification limit can it be declared in tolerance with high confidence. For this reason a retest guard band, g_R, (which may be different from the calibration guard band) is used. If the error is outside the retest guard band but within the specification limit, the condition is called *indeterminate pass* or *marginal*.

The retest guard band should be selected to provide adequately low **retest risk**, the probability that the true error at the time of retest exceeds the specification limit when the observed error is at the retest guard band limit. Evaluation of this risk requires solving the conditional probability as before, but with additional components. A precise calculation would include the condition that the error reported by the original calibration was within the calibration guard band. This would, however, make the mathematics much more complicated. If the first-pass yield is high there is little consequence of disregarding this condition. The simpler calculation will produce a result that errs slightly high, meaning the actual risk will be slightly lower than estimated.

The calibration uncertainty now includes the variable systematic uncertainty. The product error uncertainty includes variable systematic uncertainty, random uncertainty, alignment uncertainty, and drift uncertainty. Field error uncertainty is not included since it is assumed that conditions are controlled. The results are as follows:

$$m_{\mathrm{R}} = g_{\mathrm{R}} \cdot L\left(\frac{v_{\mathrm{s}}u_{\mathrm{s}}^2 + u_{\mathrm{r}}^2}{v_{\mathrm{s}}u_{\mathrm{s}}^2 + u_{\mathrm{r}}^2 + u_{\mathrm{a}}^2 + u_{\mathrm{d}}^2} + 1\right)^{-1} + |m_{\mathrm{d}}|\left(\frac{v_{\mathrm{s}}u_{\mathrm{s}}^2 + u_{\mathrm{r}}^2 + u_{\mathrm{a}}^2 + u_{\mathrm{d}}^2}{v_{\mathrm{s}}u_{\mathrm{s}}^2 + u_{\mathrm{r}}^2} + 1\right)^{-1}$$

$$\sigma_{\mathrm{R}}^2 = \left(\left(v_{\mathrm{s}}u_{\mathrm{s}}^2 + u_{\mathrm{r}}^2 + u_{\mathrm{a}}^2 + u_{\mathrm{d}}^2\right)^{-1} + \left(v_{\mathrm{s}}u_{\mathrm{s}}^2 + u_{\mathrm{r}}^2\right)^{-1}\right)^{-1} + (1 - v_{\mathrm{s}})u_{\mathrm{s}}^2 \tag{19}$$

The retest risk is then

$$R_{\mathrm{R}} \approx \left(2 - \Phi\left(\frac{L - m_{\mathrm{R}}}{\sigma_{\mathrm{R}}}\right) - \Phi\left(\frac{L + m_{\mathrm{R}}}{\sigma_{\mathrm{R}}}\right)\right) \cdot 100\% \tag{20}$$

A maximum acceptable retest risk of 5% would be consistent with the typical coverage factor, and a somewhat lower value would allow some extra margin.

12. Retest yield

Retest yield is the expected percentage of products just meeting the calibration guard band that will also pass retest. There are two yield values of interest. Retest pass yield is the percentage that are expected to be within the retest guard band. Retest marginal yield is the percentage within the specification limit. These yields should be considered at the time specifications, calibration uncertainties, and guard bands are decided.

Retest pass yield is the conditional probability that $|T_{\mathrm{R}}| < g_{\mathrm{R}} \cdot L$ given $T_{\mathrm{I}} = g \cdot L$. The constant component of the systematic error is hidden, and the variable component is manifest twice. Mean and standard deviation are as follows:

$$m_{\mathrm{RY}} = m_{\mathrm{I}} + |m_{\mathrm{d}}|$$

$$\sigma_{\mathrm{RY}}^2 = \left(\left(u_{\mathrm{r}}^2 + u_{\mathrm{a}}^2\right)^{-1} + u_{\mathrm{r}}^{-2}\right)^{-1} + 2v_{\mathrm{s}}u_{\mathrm{s}}^2 + u_{\mathrm{d}}^2 + u_{\mathrm{r}}^2 \tag{21}$$

The retest pass yield is then

$$Y_{\mathrm{RP}} = \left(\Phi\left(\frac{g_{\mathrm{R}} \cdot L - m_{\mathrm{RY}}}{\sigma_{\mathrm{RY}}}\right) + \Phi\left(\frac{g_{\mathrm{R}} \cdot L + m_{\mathrm{RY}}}{\sigma_{\mathrm{RY}}}\right) - 1\right) \cdot 100\% \tag{22}$$

and the retest marginal yield is

$$Y_{\mathrm{RM}} = \left(\Phi\left(\frac{L - m_{\mathrm{RY}}}{\sigma_{\mathrm{RY}}} \right) + \Phi\left(\frac{L + m_{\mathrm{RY}}}{\sigma_{\mathrm{RY}}} \right) - 1 \right) \cdot 100\% \tag{23}$$

The manufacturer may find it necessary to reduce the calibration guard band (making m_{I} smaller) to achieve desired retest yields. Acceptable yield limits must be decided. Limits of 95% for marginal yield and 90% for pass yield may be sufficient. But consider, often some corrective action is required by the owner or operator of the instrument if it is reported out of tolerance, and this can be quite costly. If the application calls for a highly reliable standard, and the impact of a failure is great, targeting a retest yield much closer to 100% might be prudent. Also, if a product has multiple independent measurement functions or ranges, like in a digital voltmeter, required yields for each function should be higher so that the overall yield is reasonable.

Again, not all products will have this high probability of failing retest, only those that happen to be near the guard band limit during the original calibration. It might be of interest to know the population retest yield—of all products that pass calibration, the expected percentage that will be found in tolerance at retest. Unfortunately, this calculation involves the integral of cumulative normal distribution functions, which can be difficult to evaluate using spreadsheet software. Math software might be used instead. The equation for population retest yield is as follows:

$$Y_{\mathrm{RMP}} = \frac{\displaystyle\int_{-g\cdot L}^{g\cdot L} p_n\left(T, 0, \sigma_{\mathrm{FP}}\right) \left(\Phi\left(\frac{L - m_{\mathrm{I}}(T)}{\sigma_{\mathrm{RY}}} \right) + \Phi\left(\frac{L + m_{\mathrm{I}}(T)}{\sigma_{\mathrm{RY}}} \right) - 1 \right) dT}{2 \cdot \Phi\left(\dfrac{g \cdot L}{\sigma_{\mathrm{T}}} \right) - 1} \cdot 100\% \tag{24}$$

$$m_{\mathrm{I}}(T) = T \left(\frac{u_{\mathrm{r}}^{2}}{u_{\mathrm{r}}^{2} + u_{\mathrm{a}}^{2}} + 1 \right)^{-1}$$

13. Example 1

As an example risk analysis, consider a measuring instrument that has a specified accuracy of 25 ppm. Calibration provides a TUR of approximately 4:1 (k=2 coverage), with systematic errors mostly uncorrelated over the calibration interval. Because of nonlinearity, alignment is an imperfect best fit. Drift and ambient temperature sensitivity are expected to occur in the field. Standard uncertainties of the various calibration and product errors are as follows:

Specification (L):	25 ppm
Test, random (u_{r}):	1.2
Test, systematic (u_{s}):	2.8
Systematic error variability (v_{s}):	0.7
Alignment (u_{a}):	6.0
Product drift, mean (m_{d}):	1.6

Product drift, std dev (u_d):	2.6
Field errors (u_f):	1.4

Assume that maximum acceptable risks are 2% for immediate risk, 5% for field risk, and 5% for retest risk. Desirable yields are at least 97% for first-pass, 90% for retest pass, and 95% for retest marginal yield. The calibration guard band is 75% (a typical value). The retest guard band is 90%. Applying the foregoing equations, probabilities are calculated to be as follows:

Calibration guard band (g):	75%
Retest guard band (g_R):	90%
Immediate risk:	1.1%
First-pass yield:	99.7%
Field risk:	*10.4%*
Retest risk:	4.1%
Retest pass yield:	*73.5%*
Retest marginal yield:	*88.1%*
Population retest yield:	99.95%

This shows that risk of the least accurate products being initially out of tolerance is low, but then too many of them will drift out of tolerance during use or fail retest. In this case a 75% guard band is insufficient to account for drift and ambient temperature effects. Changing the calibration guard band to 55% will bring the worse-case field risk to well under 5% and greatly improve retest yield. The trade-off is that first-pass yield will reduce to just over 97%.

14. Example 2

In another example, consider a temperature calibrator that has a specified accuracy of 0.4°C. The limiting factor of the specification is the calibration system uncertainty, and a TUR of 2:1 (k=2 coverage) is being considered. Standard uncertainties of the errors are as follows:

Specification (L):	0.4°C
Test, random (u_r):	0.028
Test, systematic (u_s):	0.094
Systematic error variability (v_s):	0.5
Alignment (u_a):	0.02
Product drift, mean (m_d):	0.038
Product drift, std dev (u_d):	0.052
Field errors (u_f):	0.032

Maximum acceptable risks are 2% for immediate risk, 5% for field risk, and 5% for retest risk. Yields should be at least 96% for first-pass yield, 90% for retest pass yield, and 95% for marginal yield. Based on the risk analysis, guard bands are set at 50% and 90% for calibration and retest respectively. The calculated probabilities are as follows:

Calibration guard band (g):	50%
Retest guard band (g_R):	90%

Immediate risk:	0.19%
First-pass yield:	99.999%
Field risk:	1.7%
Retest risk:	3.0%
Retest pass yield:	96.3%
Retest marginal yield:	98.4%
Population retest yield:	99.94%

The results show that in this case a 2:1 TUR is sufficient.

15. Computer simulation

Another way to estimate the risks and yields is by Monte Carlo computer simulation. Random variables are created to simulate the various errors. A large population of samples is collected. Error samples are added to produce product errors and observed errors. Numbers of samples exceeding limits or meeting particular criteria are counted. Finally, percentages are calculated. This approach can verify the results of direct calculation. It can also be helpful when errors follow unusual distributions and direct calculation would be more difficult.

A concern with simulation is that counts must be large enough to gather reasonable statistics. If the first-pass yield is very close to 100%, there may be too few samples near the guard band. Also, simulated observed errors will never be exactly at the guard band limit. Instead, samples must be counted if they are reasonably close to the limit, within some window. The window must be large enough to produce sufficient counts, but not so large that results are skewed.

In producing the random variables, accuracy is critical. Samples must not be correlated or biased and must have the correct standard deviation. Computer random number generators produce uniformly distributed samples that have finite resolution and recurrence periods. They often have a slight bias, with 0 being a possible value while 1 is excluded. A good random number function will minimize the effects of these limitations. Uniformly distributed random variables can be converted into normally distributed variables using the Box-Muller transform.

When calculating risks when an error component has a nonzero mean (m_F), it is important that worse-case probabilities are obtained. The absolute value of the mean is used, and statistics are calculated only for observed errors near the positive guard band limit.

In summary, the simulation program takes the following steps:

1. In a loop that repeats a large number of times, do the following:
   a. Generate new normally distributed random values for each error, with the appropriate standard deviations and mean $|m_F|$ applied (see Section 6).
   b. Sum errors to simulate the various true product errors and test results (see Section 7).
   c. Maintain counts of the various cases:
      i. Total number of cases
      ii. Number of cases where the immediate observed error is near the positive guard band limit

iii. Number of cases where the immediate observed error is near the positive guard band limit and the immediate true error is outside the specification limits (immediate risk)
iv. Number of cases where the immediate observed error is between the guard band limits (first-pass yield)
v. Number of cases where the immediate observed error is near the positive guard band limit and the true error in the field is outside the specification limits (field risk)
vi. Number of cases where the immediate observed error is between the calibration guard band limits and the retest observed error is near the positive retest guard band limit
vii. Number of cases where the immediate observed error is between the calibration guard band limits, the retest observed error is near the positive retest guard band limit, and the true error at retest is outside the specification limits (retest risk)
viii. Number of cases where the immediate observed error is near the positive guard band limit and the retest observed error is between the retest guard band limits (retest pass yield)
ix. Number of cases where the immediate observed error is near the positive guard band limit and the retest observed error is between the specification limits (retest marginal yield)
x. Number of cases where the immediate observed error is between the calibration guard band limits and the retest observed error is between the specification limits (population retest yield)

2. Calculate the various risk and yield percentages using the applicable counts.

16. Simulation of Example 1

A simulation was run to confirm the results of Example 1 (Section 13). A population of $2 \cdot 10^8$ samples was collected. The guard band window was ±1.0%. Results are shown below in Table 2.

Table 2. Example 1 simulation results.

	Calculation	Simulation	Counts
Immediate risk:	1.1%	1.1%	742 / 69801
First-pass yield:	99.7%	99.7%	199472305 / 200000000
Field risk:	10.4%	10.4%	7289 / 69801
Retest risk:	4.1%	3.1%	3175 / 101601
Retest pass yield:	73.5%	73.3%	51166 / 69801
Retest marginal yield:	88.1%	88.0%	61456 / 69801
Population retest yield:	99.95%	99.93%	199323653 / 199472305

The simulation results closely match those of direct calculation. The largest discrepancy is with retest risk. This is because the calculation disregards the condition that products pass calibration in the first place, whereas the simulation applies the condition. The calculation slightly overestimates the retest risk.

17. Conclusion

Out-of-tolerance risk analysis requires estimating uncertainties using techniques familiar to metrologists. Once relevant uncertainties are quantified, calculations of out-of-tolerance risks and calibration yields are straight-forward. This can provide confidence that equipment will perform according to specifications until due for recalibration, and will be found in tolerance when retested. It offers a method for intelligently choosing guard bands for acceptance limits.

Motivation for researching and writing this paper arose from a case in which achieving a 4:1 test uncertainty ratio was not practical. The question was, could a lower TUR, even approaching 2:1, still provide an adequate calibration, so that the risk to the customer of the equipment being out of tolerance was still reasonably low? As presented here, analysis showed that a 2:1 TUR, given proper guard bands, could indeed provide an acceptable calibration. It also showed that using the tighter guard bands would not significantly compromise calibration and retest pass yields.

Acknowledgements

Much is owed to Tom Wiandt, Director of Metrology at Fluke–Hart Scientific, for many patient discussions on the subjects of calibration, uncertainties, guard bands, and quality control. His contributions to this paper, both direct and indirect, are numerous. There are many others in this field who deserve proper recognition for their pioneering work. Among them is David Deaver of Fluke, who has made valuable contributions.

References

1. D. Deaver, Guardbanding and the World of ISO Guide 25: Is There Only One Way?, NCSL, 1998
2. ANSI/NCSL Z540-2-1997, U.S. Guide to the Expression of Uncertainty in Measurement
3. ANSI/NCSL Z540.3-2006, Requirements for the Calibration of Measuring and Test Equipment
4. S. Dwyer and G. Cenker, K=2 and the Normality Assumption of Test Equipment: Empirical Evidence, NCSLI, 2007
5. H. Castrup, Risk Analysis Methods for Complying with Z540.3, NCSLI, 2007
6. S. Mimbs, Measurement Decision Risk—The Importance of Definitions, NCSLI, 2007
7. M. Dobbert, Understanding Measurement Risk, NCSLI, 2007