Where one term ends and the next begins

Measurement uncertainty, CMC, bias, TUR, decision rules, guardbands, and PFA/PFR

Henry Zumbrun, Morehouse Instrument Company, York, PA

Abstract

Conformity assessment in calibration runs on seven related terms: measurement uncertainty, bias, calibration and measurement capability (CMC), test uncertainty ratio (TUR), decision rules, guardbands, and the probabilities of false accept (PFA) and false reject (PFR). The governing standards define each term on its own, but no single document states where each one ends, how the terms connect, or when to use which. This paper presents the seven terms as one chain, in the order they are used. Measurement uncertainty is the input. Bias is corrected before the uncertainty is combined. The CMC is the floor under the reported uncertainty. The TUR screens whether the measurement is capable of supporting a decision. The decision rule states how uncertainty is accounted for. The guardband, $w = |TL - AL|$, carries out the rule's intent. PFA and PFR are the outputs. Each boundary is illustrated with force-calibration examples, with requirements drawn from ISO/IEC 17025:2017, JCGM 106, ILAC-G8, ILAC-P14, and UKAS LAB 48.

There is a messy mix of terms in conformity assessment, and no single place that says where each one stops, how they connect, and when to use which. This paper is the answer. It takes several articles and documents we have worked on, adds some new material, and puts everything together in one place. Section 1 is the map. Sections 2 through 7 each take one boundary that gets crossed most often and walk it with the same force-calibration examples used in the Decision Rule Guidance, the GUM in Plain English series from Morehouse, and the Guidance on Uncertainty Budgets for Force Measuring Devices. Each section stands alone, though the chain reads best in order.

Contents

Table of contents: right-click and choose Update Field to populate.

1. Seven terms, one chain

"If you choose not to decide, you still have made a choice." Neil Peart wrote that line for a song about free will, and we have quoted it in our decision rule guidance for years. It fits because every pass/fail statement is a decision, whether or not anyone admits to making one. And for those that say, what about non-binary decision rules, I will say, eventually someone needs to make a binary decision. Are you going to see x movie in the theatre? Your answers could be yes, no, maybe. Eventually, the movie leaves the theatres, and you saw it, or you did not. Okay, if it's Star Wars, the maybe could have lasted 50 years, though that is an exception to the rule, which, if we adjust the question to “Are you going to see the movie upon initial release?”, we solve the issue. These metrologists always try to get me with

something. So, I have to triple-check everything, and I can 100 % guarantee there's a mistake somewhere in all of this. Dilip, Craig, Jeff, Ryan, Andreas, Rob, Kevin, Klaus, Heather, Joseph, Eric, Phil, Paul, Cory, Ginger, Danae, John, Daniel, Chuck, Stephen, Seth, Howard, Shawn, Gary, Daniel, Mike, Tim, Paul, Han, Chewbacca, whatever your name is, or Greg C will find one for sure.

A customer/friend recently told me why the deciding is so hard. His words: there seems to exist a messy mix of measurement uncertainty, CMC, TUR, guardband, bias compensation, decision rule method, PFA and PFR, and there does not seem to exist a clear set of rules defining the boundary of each, their relation to each other, or when to use what.

He is right. The standards define each term well enough on their own. ISO/IEC 17025:2017 defines the decision rule. ILAC-G8 kind of defines the guardbanding rules. JCGM 106 defines the risks (likelihood and consequences). ILAC-P14 defines the rules for reporting the CMC. The GUM defines uncertainty, and the VIM defines most of the definitions. Nobody put them on one page in the order they actually get used. So here is that page.

The chain, in the order it runs

Think of a conformity decision as a short assembly line. Each term is a station. Each station takes something from the one before it and hands something to the one after it. Skip a station, or run them out of order, and the part that comes off the end is a pass/fail statement nobody can defend. Below is

the map, and it's not perfect. It’s here to give a generalized sense of it all.

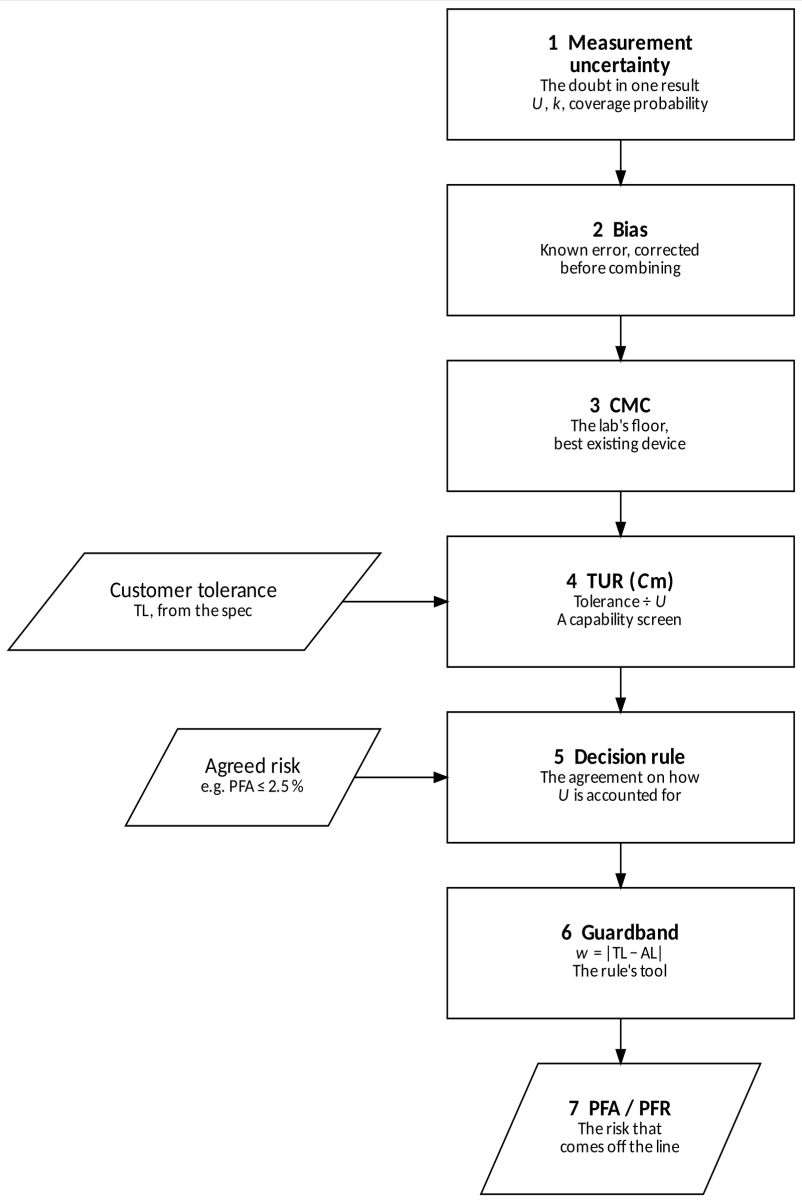


Figure 1. The seven stations in the order they run. Two things enter from outside the lab: the customer's tolerance and the risk the customer agrees to carry.

Station 1 is **measurement uncertainty**. This is the doubt that remains after you have done everything right: the reference standard's uncertainty and drift, repeatability, resolution, temperature, alignment, and the rest, combined by the GUM method into one number with a coverage factor. It belongs to one measurement result. It is the raw material for everything downstream. **No uncertainty, no decision, and**

per ISO/IEC 17025:2017 Clause 7.6 and the VIM, no traceable result either, and that uncertainty should be backed up by statistical process control, trend analysis, and demonstrating via proficiency testing that your claims are what they are. ISO/IEC 17025:2017 Clause 7.7 turns that from good practice into a requirement: monitor the validity of results, record the data so trends are detectable, and compare your performance with other laboratories through proficiency testing or interlaboratory comparisons. The budget is the claim. An E_n ratio below one is the evidence, and Section 2 covers what that takes.

Example: Your GPS says you will arrive at 6:12, and you read it as 6:12 give or take a few minutes without thinking about it. An uncertainty budget writes down that give-or-take: what it is made of, how the pieces combine, and how sure you are of the total.

Station 2 is **bias**, and it is the one station that is not about doubt at all. Bias is the known, systematic difference between what your instrument reads and what the reference says. The GUM is direct on this (Clause 3.2.3): identify systematic effects, correct for them, then carry the uncertainty of the correction. Bias compensation means applying that correction to the measured value. It happens before the uncertainty is combined, and it changes the value, not the doubt. Section 3 covers why people get this backward.

Example: A marksman with bias hits the same wrong spot every time. That is the easiest shooter at the range to fix: a few clicks of windage and the group moves onto the bull. The shooter who sprays shots all over the target is the one who will have a very large uncertainty budget.

Station 3 is the **CMC**, the Calibration and Measurement Capability on a laboratory's scope of accreditation. The CMC uncertainty parameter (one part of several on the scope) is the smallest uncertainty the lab can offer a customer under normal conditions, for the best existing device it is likely to see. It is a floor, not a promise. ILAC-P14 requires that the uncertainty on your certificate include the contributions of your own device, so the number on a certificate is normally larger than the CMC, never smaller. Section 2 draws that line.

Example: A sprinter's personal best is a time set on a perfect track, in perfect weather, with nothing on his back. The CMC is the personal best: ideal conditions, best existing device. The number on your certificate is what he runs carrying your instrument, and it is always slower.

Station 4 is the **TUR**, the Test Uncertainty Ratio, or in JCGM 106 language, the measurement capability index (C_m is the preferred term as TUR has been botched and has several definitions). It is the tolerance of the item under test divided by the expanded uncertainty of the calibration process. It takes the uncertainty from Station 1 and the tolerance from the customer and tells you whether the measurement is capable enough to make the decision at all. It is a screening ratio. It is not a decision rule, and Section 4 explains why a 4:1 on a purchase order does not settle anything by itself.

Example: Nobody checks a 1 mm gap with a ruler marked in centimeters. TUR is that instinct written as a ratio: how fine the measuring process is compared with the tolerance it is judging. And owning a

micrometer still does not tell you where to call the gap good or bad, which is why a healthy ratio is not a decision.

Station 5 is the **decision rule** itself. ISO/IEC 17025:2017 Clause 3.7 defines it as a rule that describes how measurement uncertainty is accounted for when stating conformity with a specified requirement. It is an agreement between the lab and the customer (Clause 7.1.3), documented with its risk (Clause 7.8.6.1), and reported on the certificate (Clause 7.8.6.2). The standard requires that a rule exist and be stated. It does not require any particular rule. Section 5 is the selection guide.

Example: In tennis, a ball that touches the line is in. Agreed before the first serve, printed in the rulebook, the same call for every player. That is a decision rule. Deciding how to treat the close ones after you have seen the shot is what a lab does when the certificate carries no rule.

Station 6 is the **guardband**. A guardband is the distance between the tolerance limit and the acceptance limit, $w = |TL - AL|$. It is the tool most decision rules use to carry out their intent. You don't pick its width on a whim. It comes from the uncertainty and the risk you agreed to accept. Section 6 gives the one formula that produces it.

Example: The speed limit is 65; you do not fully trust your speedometer, so you hold 62. That margin is a guardband: the acceptance limit pulled in from the tolerance limit by your doubt, so a reading that passes leaves you comfortably legal. How far to pull it in is not taste. It falls out of the speedometer's uncertainty and how much ticket risk you will accept.

Station 7 is **PFA and PFR**, the probability of false accept and the probability of false reject. These are the outputs. A specific-risk rule calculates them for the one result in front of you. A global-risk rule is designed so that they stay below a target across a whole population. Either way, they are what comes off the end of the line. Section 7 explains why a 2 % global PFA and a 50 % specific PFA can describe the same instrument at the same time.

Example: An airport metal detector has one dial and two ways to fail. Set it hot and the line backs up with belt buckles: false reject. Set it cool and the real thing walks through: false accept. PFA and PFR are that dial read off a calibration bench, and turning it to shrink one number grows the other.

Traceability is a property of the result, not the instrument

One more term rides the whole chain and gets defined wrong more often than any station on it. The VIM (JCGM 200:2012, 2.41) defines metrological traceability as the property of a measurement result whereby the result can be related to a reference through a documented unbroken chain of calibrations, each contributing to the measurement uncertainty**. Every word in that definition is doing work**. Property of a result, not of an instrument, a laboratory, or a sticker. A documented unbroken chain of calibrations, not a single certificate. Each link contributes to the measurement uncertainty, which is why Station 1 is where traceability lives or dies.

Most certificates get this backward. A typical one devotes a section titled Traceability to an equipment list: a force standard with a NIST number and a calibration due date, a standard indicator with a NIST number of its own, a test indicator with another, and a sentence certifying that the force measurements are traceable to primary standards at NIST. The format treats traceability as something the hardware carries, renewed each time the hardware is recalibrated.

NIST disagrees with the certificates. Its position is that only measurement results are traceable, and that merely having an instrument calibrated, even by NIST, is not enough to make the result obtained with that instrument traceable. The NIST numbers prove even less: NIST GMP 13 states that test report numbers are issued solely for administrative purposes, and holding one provides no assurance or evidence that a measurement is traceable. A calibration result is strictly valid at the time and place it was performed. The moment the standard goes back on the shelf, drift starts working on it. What carries traceability forward is the process wrapped around the instrument: the uncertainty budget at Station 1, the monitoring and control charts Clause 7.7 requires, and a documented chain in which every link states its uncertainty.

So, when a certificate's traceability case is a list of NIST numbers, ask the questions that matter instead. Does it identify the reference standards used and their uncertainties? Does it report measured values with an expanded uncertainty, a coverage factor, and a coverage probability? Can the provider produce the unbroken chain from your result back to the SI? Yes to all three is traceability. A NIST number with a due date is an inventory record. You can use our NIST number, 8675309, for force or torque (NIST does not do torque), and tell 'em Tommy T gave it to you.

If you want proof instead of a number, it lives in the BIPM Key Comparison Database (kcdb.bipm.org). The KCDB lists every calibration and measurement capability an NMI has had internationally reviewed and approved under the CIPM MRA, searchable by metrology area, quantity, range, and country. Run a CMC search for the parameter, force, torque, or anything else, and you can see whether the institute actually supports it, over what range, and at what uncertainty. If the parameter is not there, the institute's name on a certificate is not evidence of traceability for that measurement.

Morehouse Instrument Company
1742 Sixth Ave.
York, PA 17403-2675
(717) 843-0081

One table to possibly keep

Term	What it is	Who sets it	What it feeds	What it is not
Measurement uncertainty	The quantified doubt in one result, with *k* and a coverage probability	The lab, from its budget	TUR, guardband, PFA/PFR	Error. Accuracy. A spec-sheet number
Bias	A known systematic error, corrected before the uncertainty is combined	Found by comparison to a reference	The corrected measured value	A component you pad into uncertainty
CMC	The lab's best-case uncertainty on its scope, for the best existing device	The lab, approved by its accreditation body	The floor under every certificate uncertainty	The number on your certificate
TUR (C_m)	Tolerance divided by calibration process uncertainty	Both: the customer's tolerance, the lab's uncertainty	A precondition for simple acceptance; an input to Method 6	A decision rule
Decision rule	The agreed statement of how uncertainty is accounted for in pass/fail	Customer and lab, at contract review	The guardband and the certificate statement	Optional once a conformity statement is made
Guardband	The offset from the tolerance limit to the acceptance limit	Calculated from U and the agreed risk	The acceptance limits	A number you choose independently of uncertainty
PFA / PFR	The probability of a wrong accept or wrong reject. Also answers to consumer's risk and producer's risk, Type II and Type I error, FAR and FRR	Produced by the rule, or targeted by it	The customer's decision on disposition	A property of the instrument alone

Note: the proper definition of Test Uncertainty Ratio (TUR) is equivalent to JCGM 106's definition of Measurement Capability Index.

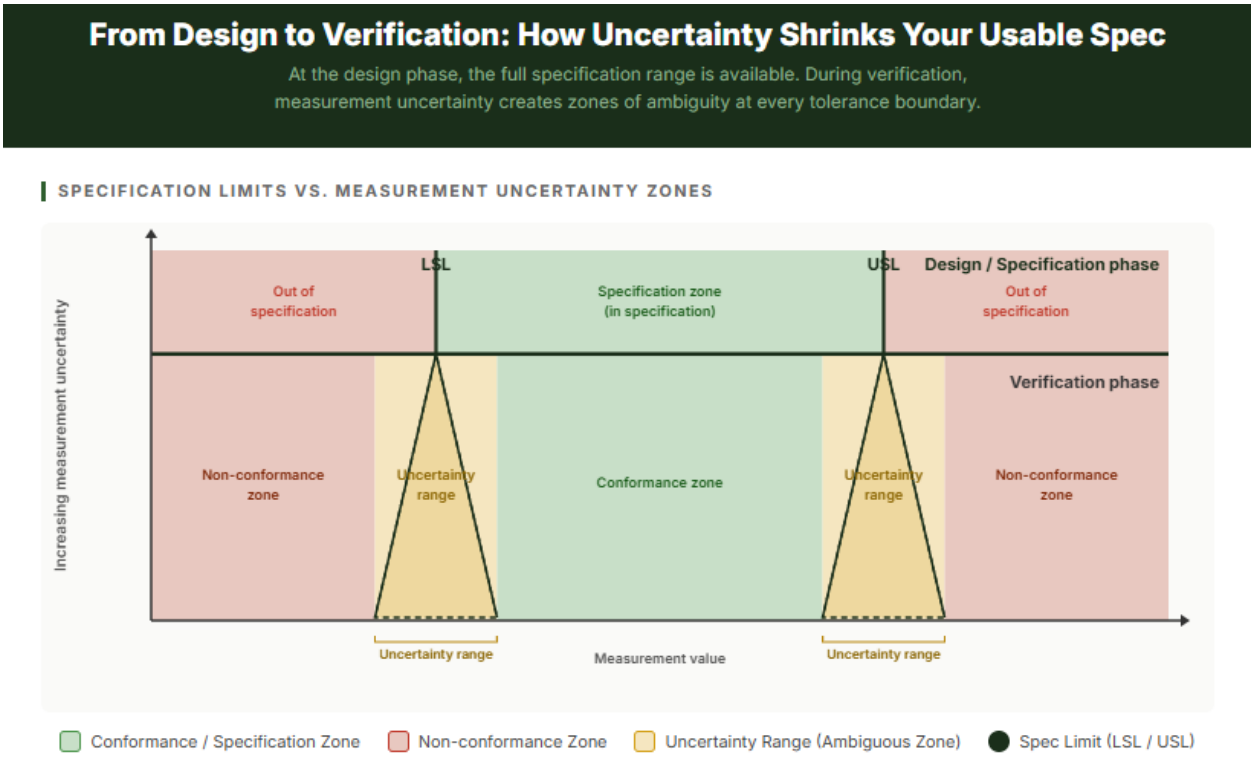


Figure 2. What you specify on paper and what you can verify in practice are not the same thing. Uncertainty turns each sharp limit into a band of ambiguity, and the wider the band, the less of the specification is usable.

When to use what, in one paragraph

If the customer does not ask for a pass/fail, report the measured value and its uncertainty and stop. No rule is required. If they do ask, the rule has to exist before the measurement is made. **When the TUR is high, and the consequence of a wrong accept is small, simple acceptance with a stated TUR or uncertainty precondition can be defended. When you have little or no history of how that instrument population behaves, you should likely use a specific-risk guardband, because it is conservative and needs only the uncertainty and the result.** When a program invokes ANSI/NCSL Z540.3 and its 2 % limit, and you lack reliability data, you can use the managed guardband (Method 6, which, if the TUR is 4.6:1, will limit your risk to 2 %). When you have real end-of-period reliability data, global-risk methods will tell you where the acceptance limits should sit to protect yield. When a false reject is the expensive mistake, guarded rejection exists for exactly that. Section 5 turns this paragraph into a tree.



Walk it backwards before you walk it forwards

The chain above runs left to right, which is how a lab experiences it. A customer should read it right to left. Our Start With the End in Mind paper puts the whole method in two questions: how good does the measurement need to be, and how good can it be made? The first is a business question about what a wrong decision costs. The second is a technical question about what your equipment, your process, and your provider can deliver. Answer the first, and the chain tells you what to demand at every station upstream.

Here is the arithmetic in its simplest form. If you want to run simple acceptance against a global PFA of 2 % or less, you probably need a TUR of about 4.6:1, which means the calibration process uncertainty has to be at or below the tolerance divided by 4.6. A 0.10 % tolerance needs $U \leq 0.0217$ %. A 0.05 % tolerance needs $U \leq 0.0109$ %. A 0.01 % tolerance needs $U \leq 0.0022$ %. If the lab you are using cannot get there, you have three choices, and the chain shows all of them: add a guardband at Station 6, improve the capability at Station 1 or Station 3, or bring reliability data to Station 7 and justify the risk at the population level. What you cannot do is skip the question and hope.

A practical check

Pull three recent certificates that say PASS from the drawer or digital file system where they were filed, to never be looked at again 😊. For each one, find the decision rule, the guardband, and the uncertainty with its coverage factor. If any of the three is missing, the chain was broken somewhere between Station 1 and Station 7, and the pass is an opinion with a decimal point. The remaining sections show where to look, and if the k's are not *k*'s, there's no MU, or nothing on the certificate says the result is traceable, maybe question your calibration provider a little.

2. Uncertainty, CMC, and the number on your certificate

Pull up any accredited lab's scope of accreditation, and you will find a line for force: perhaps 0.002 % of applied, or 0.025 % of applied, or 0.05 % of applied. Many people read that line and assume it is the uncertainty they will get. Then the certificate arrives with a larger number, and the phone rings.

Three different numbers get called "uncertainty" in that conversation. They are related; however, they are not the same.

Measurement uncertainty is a property of one result

The GUM builds uncertainty one measurement at a time. Write the model. Size every input by data (Type A) or by what you already know (Type B). Combine by root-sum-of-squares. Carry the degrees of freedom. Expand by a coverage factor you can defend. For the force-measuring instrument we walked through in the GUM in Plain English series, the result at 5000 lbf was U = 1.71 lbf, k = 2.57, approximately 95 % confidence.

That number belongs to that instrument, on that machine, at that point. It includes the instrument's own repeatability (0.58 lbf, three-quarters of the variance), the reference standard's stability (0.29 lbf), and the reference standard's calibration uncertainty (0.05 lbf), which barely registered once squared. Change the instrument and the number changes.

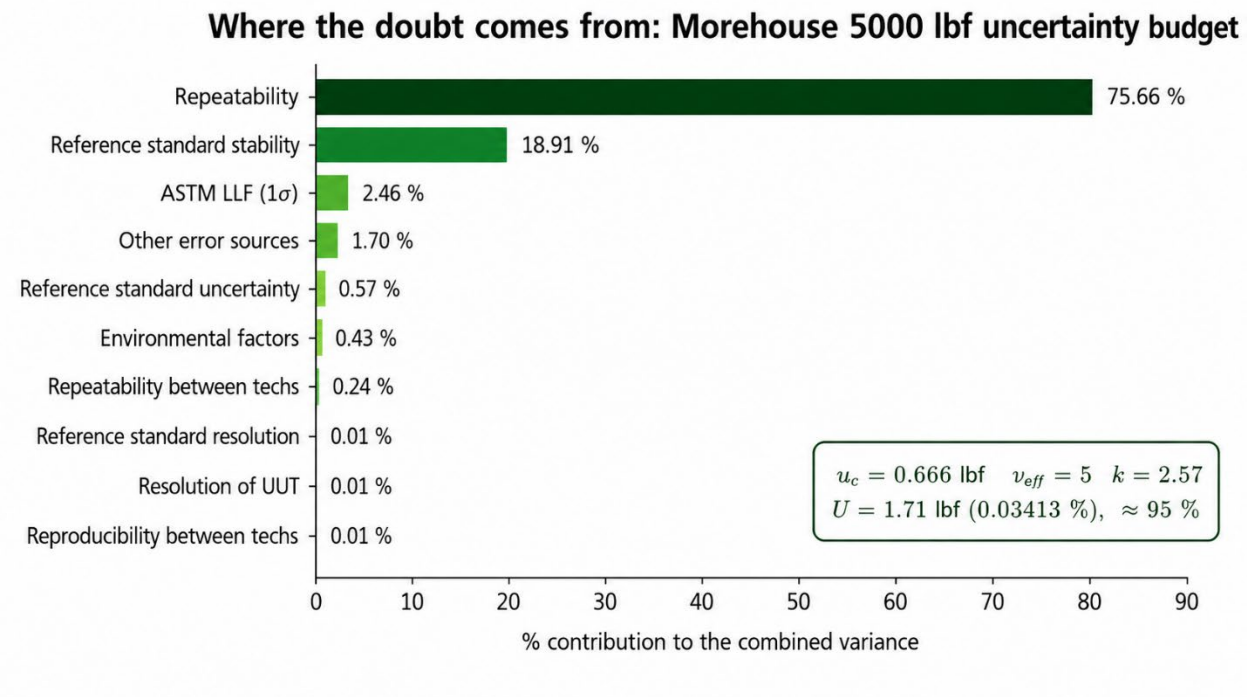


Figure 3. Where the doubt comes from in the 5000 lbf budget. Repeatability sets three-quarters of the variance; the reference standard's own calibration uncertainty barely registers.

In this situation, the issue is likely an operator problem or poor repeatability in what is being tested. The solution is to buy a better load cell for comparison, or have the operator learn to use the equipment properly. What is interesting is that the MU budgets quickly tell you where to focus your efforts to improve your measurement capability.

CMC is a property of the laboratory

The Calibration and Measurement Capability is defined in ILAC-P14 as the capability available to customers under normal conditions, as described in the lab's scope of accreditation. It is evaluated for the best existing device: the most stable, highest-resolution force transducer and indicator the lab is realistically going to see. It is the lab's best case.

Two rules follow from that definition. ILAC-P14 states the first in its CMC policy (Section 4) and the second in its policy on uncertainty in calibration (Section 5).

First, the CMC must include the contributions of the reference standards used. A lab cannot claim a CMC equal to or lower than the uncertainty of the lab that calibrated its references, because that reference uncertainty enters the budget as one positive term among several. The reference you choose sets the floor. ISO/IEC 17025:2017 Clause 7.6 says the same thing from the other side: the lab must evaluate uncertainty for all calibrations, and math does not run uphill.

Second, the uncertainty stated on a customer's certificate must cover the same contributions as the CMC, with the best-existing-device components replaced by the customer's device. The reported uncertainty is therefore normally larger than the CMC. If a certificate shows an uncertainty smaller than the lab's CMC, something is wrong with the certificate or with the scope.

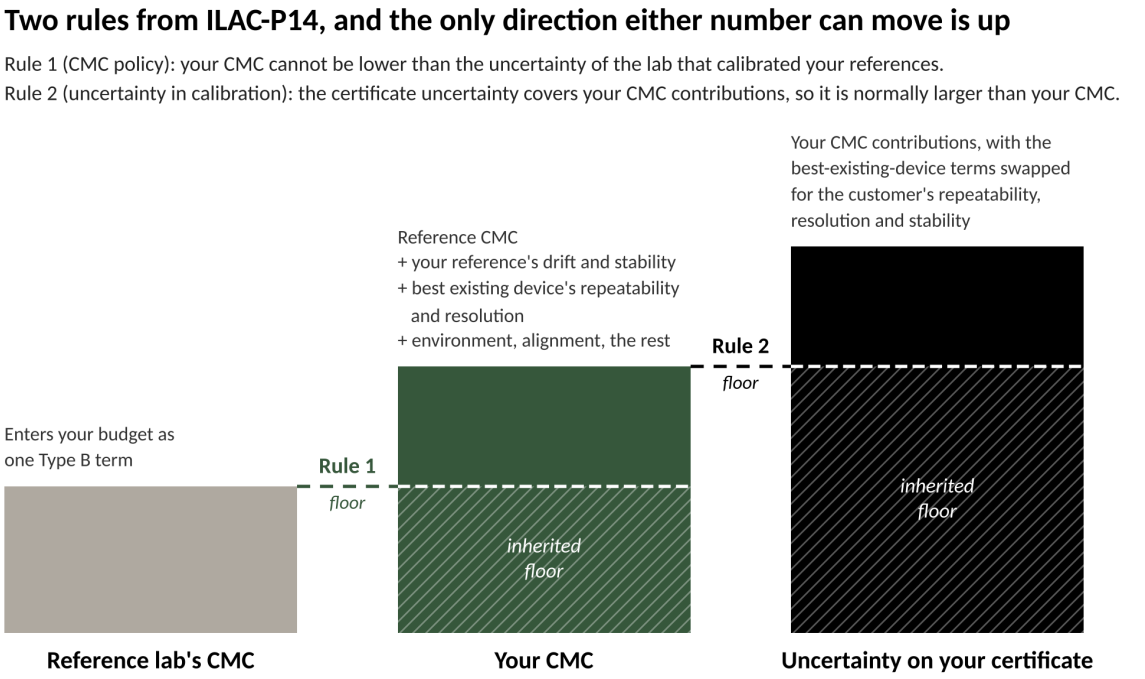


Figure 4. The two ILAC-P14 rules drawn as a staircase. The reference laboratory's CMC is the floor under your CMC, and your CMC is the floor under every certificate you issue. Each step adds contributions to the one it inherits, and nothing on the way down the chain can make the number smaller.

The certificate number is where the two meet

So the chain reads: the reference lab's CMC becomes one Type B term in your budget. Your budget, evaluated with the best existing device, becomes your CMC. Your CMC becomes the floor under every certificate you issue. Each certificate carries your CMC contributions plus the customer's instrument's repeatability, resolution, and the rest.

Watch what that does at the top of the chain. Calibrate the same instrument against deadweight primary standards known to 0.0016 % and the expanded uncertainty at 100 000 lbf comes out near 3.78 lbf. Calibrate it against a secondary machine known to 0.025 %, and the same point jumps to about 25.23 lbf. Same instrument, same care, roughly 6.7 times the uncertainty, purely from the standard above you.

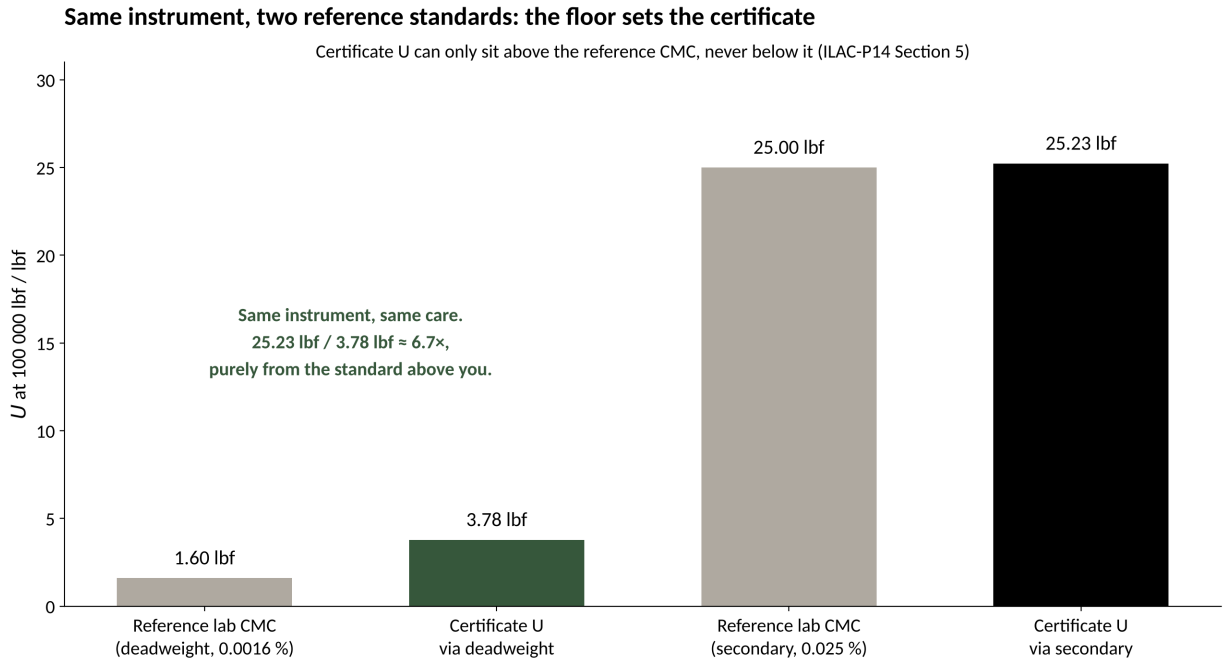


Figure 5. The reference laboratory's CMC is the floor. The certificate uncertainty sits above it in both cases; the only difference is how high the floor was to begin with.

The boundary

Measurement uncertainty is the doubt in one result. CMC is the lab's best-case doubt for the best device it is likely to see. The certificate uncertainty is the lab's CMC contributions plus your instrument's own. Three numbers, one chain, and the only direction any of them can move as you go down the chain is up. Figure 4 is the whole part in one picture; if you remember nothing else, remember that the staircase only goes up.

Two mistakes cross this boundary. The first is quoting the CMC as if it were the certificate uncertainty, which overstates capability and, as the Decision Rule Guidance notes, understates every PFA and guardband derived from it. The second is treating the certificate uncertainty as if it were a spec-sheet accuracy. UKAS LAB 48 spends a full page on the word "accuracy" and concludes that the VIM treats it as a qualitative concept with no established methodology for assigning a number. Uncertainty has one. Use it.

The number is not done until another laboratory agrees

Everything above is arithmetic the lab performs on itself. ISO/IEC 17025:2017 does not let it stop there. Clause 7.7.1 requires a procedure for monitoring the validity of results, with the data recorded in such a way that trends are detectable and, where practicable, analyzed with statistical techniques. The clause's own list of monitoring tools includes check or working standards with control charts, intermediate

checks on measuring equipment, replicate calibrations, and intralaboratory comparisons. A control chart that starts walking is the earliest sign that the budget no longer describes the process.

Clause 7.7.2 pushes the check outside the building: the laboratory shall monitor its performance by comparison with the results of other laboratories, through participation in proficiency testing, in interlaboratory comparisons other than proficiency testing, or both. ISO/IEC 17043 supplies the definitions the standard leans on. An interlaboratory comparison is the organization, performance, and evaluation of measurements on the same or similar items by two or more laboratories under predetermined conditions. Proficiency testing is the evaluation of participant performance against pre-established criteria by means of those comparisons.

The pass criterion for a calibration comparison is the E_n ratio: the difference between your result and the reference laboratory's value, divided by the root-sum-of-squares of the two expanded uncertainties. $|E_n| \leq 1$ means the two results agree within the uncertainties both parties claimed. $|E_n| > 1$ means at least one of the claimed uncertainties is wrong, and Clause 7.7.3 says what happens next: the monitoring data shall be analyzed, and when results fall outside pre-defined criteria, action shall be taken to prevent incorrect results from being reported.

This is the part the budget alone cannot do. A measurement uncertainty claim is not finished when the root-sum-of-squares is computed and the coverage factor is applied. It is finished when the lab demonstrates an E_n ratio below one against an independent reference, through an ILC or PT, and keeps demonstrating it period after period with trend data showing the process is stable. Until then, the CMC on the scope and the uncertainty on the certificate are claims waiting on evidence.

If you want to see how a CMC is built for force, the Guidance on Uncertainty Budgets for Force Measuring Devices lists every contributor A2LA R205 expects, and the free Morehouse CMC workbook does the arithmetic, degrees of freedom and coverage factor included.

3. Bias is not uncertainty, and compensation is not padding

Picture two bathroom scales. The first reads 2 lb heavy every single time, dead consistent. The second is perfectly centered on average but jumps around by a pound either way. The first scale has a large error and small scatter. The second has near-zero error and large scatter. They are not the same problem, and you do not fix them the same way.

That picture is the whole of this part. Everything else is the standards agreeing with it.

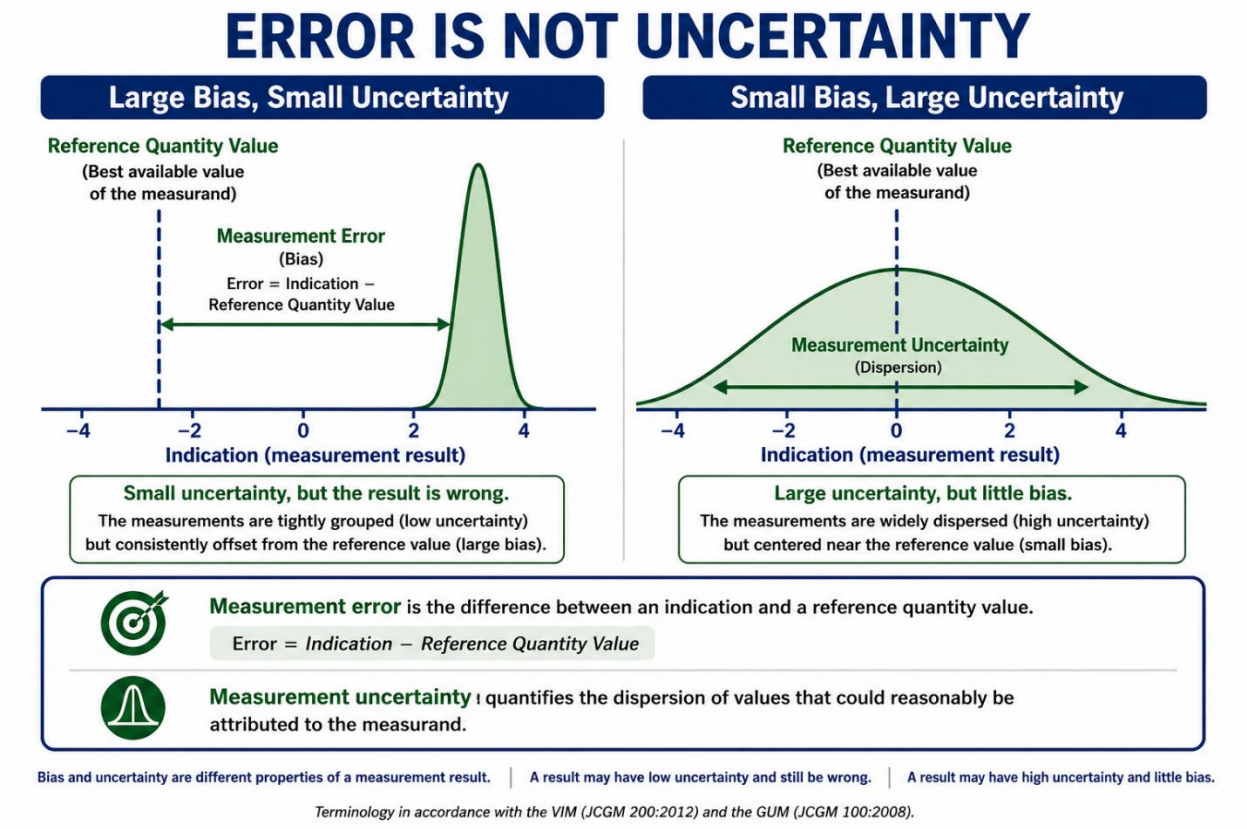


Figure 6. Large error with small scatter on the left; small error with large scatter on the right. Different problems, different fixes.

Error is a known number. Uncertainty is a quantified doubt.

The VIM (2.16) defines measurement error as the measured value minus a reference value. Measured against a calibrated standard, that difference is a bias you can find and correct. Uncertainty, by contrast, characterizes the dispersion of values that could reasonably be attributed to the measurand. One is a direction and a size. The other is a width.

The GUM makes the point in a line worth taping to the wall (Clause 3.3.1, note): a corrected result can sit very close to the true value, and so carry a negligible error, and still have a large uncertainty. Small error, big uncertainty. The two move independently.

What bias compensation actually means

The GUM's instruction (Clause 3.2.3) has three steps. Identify the systematic effect. Correct for it. Carry the uncertainty of the correction in the budget.

In force work, the correction usually arrives as the calibration itself. An ASTM E74 calibration fits a curve to the instrument's response; using that curve instead of the nominal sensitivity is bias compensation. The doubt that remains in that fit, the lower limit factor, a statistical estimate of the error in forces computed from the calibration equation, goes into the budget as a Type A term once it is divided by its coverage factor of 2.4. That is the sequence: fix the value, then account for the doubt in the fix.

What the GUM forbids is the shortcut (Clause 6.3.1, note): skipping the correction and padding the uncertainty to cover it. If you leave a known bias in place, the honest alternative is the one the Decision Rule Guidance states plainly: add it to the expanded uncertainty, U + bias. Linear addition is the price of not correcting.

Where bias sits relative to TUR

The ANSI/NCSL Z540.3 Handbook settles a question people argue about. The TUR denominator is the 95 % expanded uncertainty of the calibration process, including all the influences of the item being calibrated, except for the bias of the item being calibrated. The unit's bias is what the calibration is there to find. It is the thing being measured, not a source of doubt about the measurement. So, it stays out of the denominator and shows up instead as the measured deviation that the decision rule will judge.

That is the boundary. Bias lives in the value column. Uncertainty lives in the doubt column. The decision rule compares the value column against the tolerance, using the doubt column to decide how much margin to keep.

What this costs when it goes wrong

Load a 50 000 lbf tension link with the wrong pin, a 1.75 in pin where the design calls for a 50 mm (1.97 in) pin, and it reads 49 140 lbf. That is an 860 lbf miss, 1.72 % of applied, and no amount of uncertainty accounting repairs it. It is a bias you remove by using the right adapter. A lab that instead widens its uncertainty to cover a known adapter effect does two things wrong at once: it leaves the error in the value and inflates the doubt, so every downstream guardband grows and every acceptance zone shrinks, with no gain in truth.

The reverse mistake also happens. A lab corrects a bias and then forgets to carry the uncertainty of the correction. Its budget is too small, its TUR too flattering, and its PFA understated.

The two questions to ask

Is this effect something I can determine and remove? Then it is a bias: correct it, and budget the uncertainty of the correction. Is it something that could reasonably push the result either way, by an amount I can only bound? Then it is uncertainty: size it, assign a distribution, combine it.

Ask those two questions of every line in your budget. The ones that answer yes to the first do not belong there.

4. TUR is a screen, not a rule

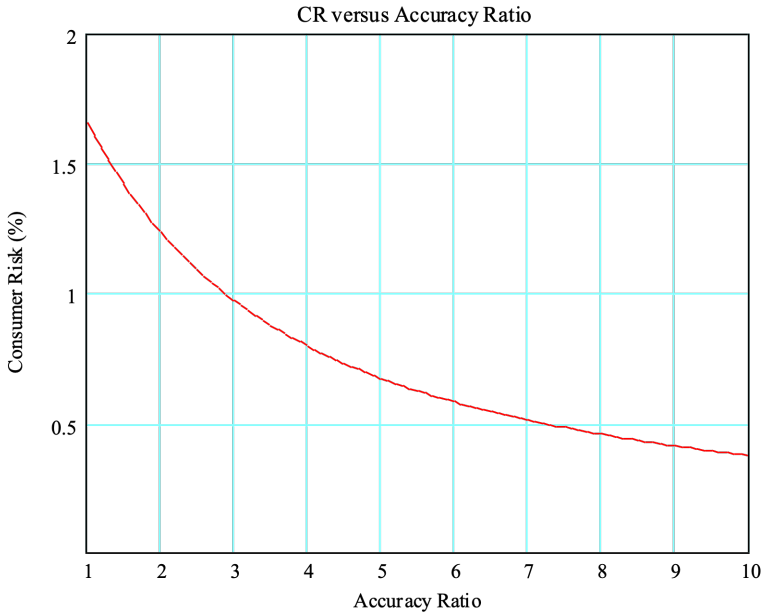


Figure 7. Consumer risk against accuracy ratio. The curve passes through roughly 1 % consumer risk near a 3:1 ratio, the relationship Hayes worked on.

Jerry Hayes of the United States Navy had a problem in the mid-1950s. Measurement decision risk had been worked out on paper by Eagle, Grubbs, and Coon, and almost nobody could use it. There was no computing power to speak of. So, Hayes accepted a consumer's risk of about 1 % and worked out that roughly a 3:1 accuracy ratio would deliver it. Assuming a 95 % confidence level for each, Hayes used specifications for the test ratio: the numerator was the unit under test's specifications, and the denominator the calibration standards' specifications. The intent was to use procedure controls to mitigate other measurement errors during testing. Not fully trusting that all specifications were equal, Hayes and Crandon increased the ratio to 4:1. Three became four. The Navy adopted 4:1, and much of the metrology community followed. **The 4:1 rule therefore provides a specific level of risk only under explicit assumptions and conditions.**

That rule was always meant as a placeholder until the math got easier. The math got easier decades ago. Yet, the placeholder is still on purchase orders, and some argue to keep the 4:1 TAR rule. TAR, Test Accuracy Ratio, is not a typo for TUR: the original rule compared accuracy specifications, not uncertainties, because specifications were what was simple to use at the time. In some fairness, many manufacturers of metrology equipment lived on reputation in those days, so their specifications were taken at face value.

What TUR is

The Test Uncertainty Ratio is the tolerance of the item under test divided by the expanded uncertainty of the calibration process. ILAC-G8 writes it as TUR = TL/U. ANSI/NCSL Z540.3 writes the same thing with the span of a symmetric tolerance over twice the expanded uncertainty. JCGM 106 and UKAS LAB 48 call it the measurement capability index, C_m, and LAB 48 defines C_{95} = (TU − TL)/(2·U_{95}). For a symmetric tolerance, every one of those gives the same number.

The denominator is where labs go wrong. It is not the reference standard's accuracy; that was the old Test Accuracy Ratio, and it ignored everything else. It is the whole calibration process uncertainty: the reference's CMC, the unit under test's resolution and repeatability, the environment, and the other contributors, combined in quadrature, excluding only the unit's own bias. ILAC-P14 asks for the same contents on the certificate. If your TUR denominator is smaller than your certificate uncertainty, you have calculated a TAR and called it a TUR.

What TUR does

TUR answers one question: is this measurement process capable enough to make the decision at all? It is a screening ratio, and it screens at the population level. Dobbert's managed guardband work, and the Reese and Harben paper that followed it, showed that at a TUR of 4.6:1 or better the worst-case global false accept risk stays below 2 % with no guardband at all. That is useful. It is what UKAS LAB 48 means when it says the role of uncertainty under simple acceptance is to act as a precondition: PASS when the value is within tolerance AND $C_{95} \geq 5$, FAIL otherwise.

What TUR does not do

It does not change the risk of the one decision in front of you. Picture the measurement result as a bell curve centered on the observed value, with its width set by the uncertainty. Move that curve so its center sits exactly on the tolerance limit. Half the area is outside. The specific probability that this item is actually out of tolerance is about 50 %, and it is 50 % at a TUR of 2:1, at 4:1, and at 10:1. A higher TUR narrows the curve, so fewer results land close enough to the limit for this to matter, and the guardband needed to control the risk gets smaller. But at the limit, the coin flip is the coin flip.

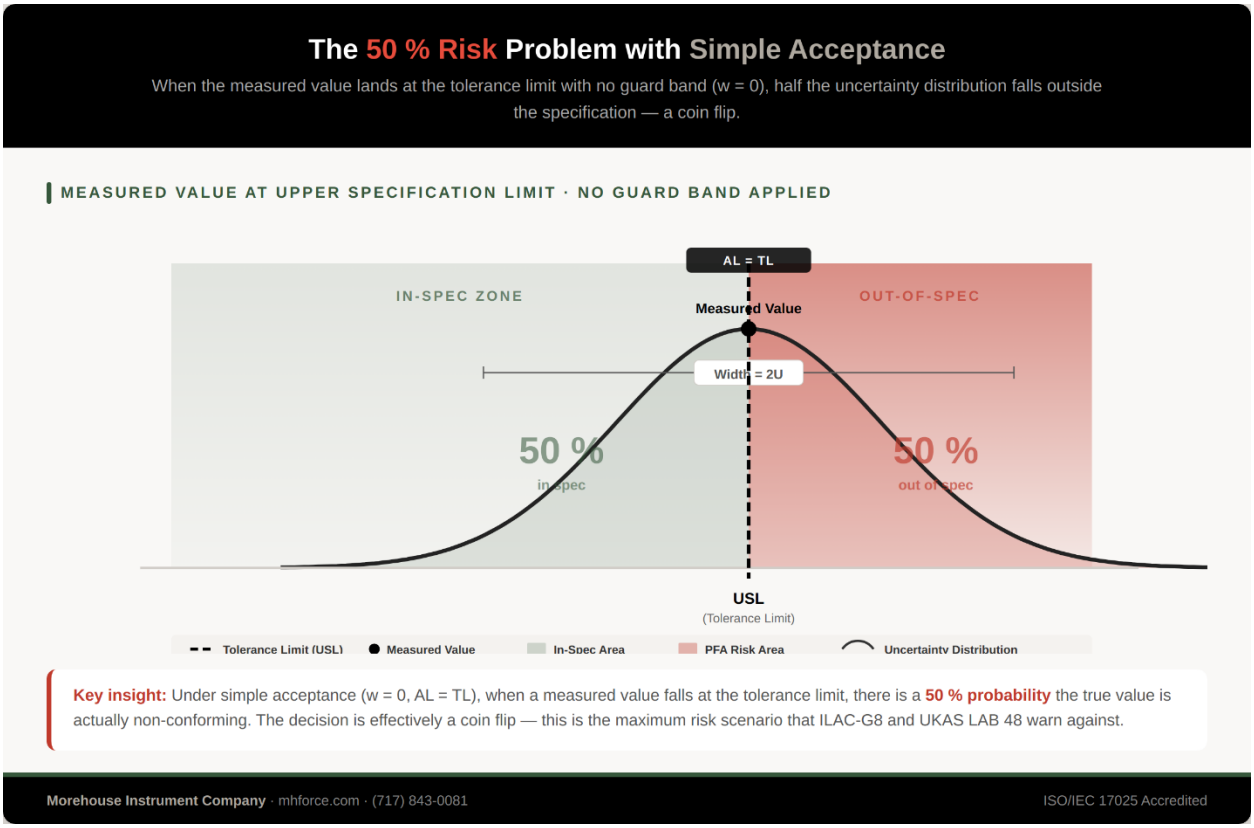
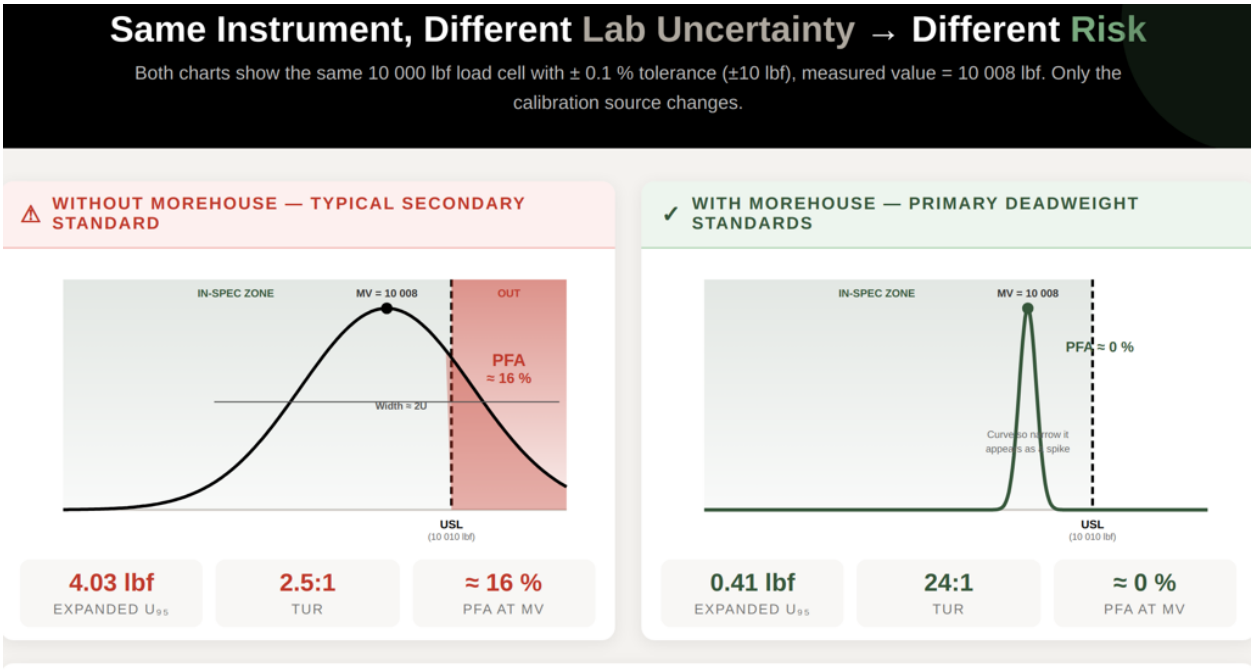


Figure 8. The 50 % problem. Under simple acceptance, a measured value exactly at the tolerance limit has half its distribution outside the specification, whatever the TUR.

That coin flip has a name: simple acceptance, often called shared risk. The measured value is reported, no guardband is applied (w = 0), and anything at or inside the tolerance passes. ISO/IEC 17025 permits it only when the rule is explicitly defined and agreed with the customer, and the customer accepts a probability of false acceptance that can reach 50 % at the tolerance limit. The standard does not prescribe an acceptable risk level; it requires that the risk be understood and documented.



PARAMETER	TYPICAL LAB	MOREHOUSE	DIFFERENCE
Expanded U_{95} at 10 000 lbf	≈ 4.03 lbf	≈ 0.41 lbf	**≈ 10× lower**
CMC (lab reference)	4.00 lbf	0.16 lbf	**25× lower**
Repeatability	0.379 lbf	0.057 lbf	**6.6× lower**
TUR (±10 lbf tolerance)	2.5:1	24:1	**10× higher**
PFA at MV = 10 008 lbf	**≈ 16 %**	**≈ 0 %**	Far lower risk
Guard band (2.5 % specific risk)	3.95 lbf	0.40 lbf	**10× smaller**
Usable acceptance zone	±6.05 lbf (60 %)	±9.60 lbf (96 %)	**+36 % recovered**

Figure 9. Same instrument, same tolerance, same measured value of 10 008 lbf. Only the calibration source changes, and with it the uncertainty, the TUR, the guardband, and the usable acceptance zone.

Here is the same load cell, the same ±0.1 % tolerance on 10 000 lbf, the same measured value of 10 008 lbf, under two calibration sources. Against a secondary standard, $U \approx 4.03$ lbf, $k = 2$, and the TUR is about 2.5:1. The curve spills past the upper limit, and the specific PFA at 10 008 lbf is about 16 %. To pull that below 2.5 % you need a guardband of about 3.95 lbf, which eats nearly 40 % of the tolerance and leaves an acceptance zone of ±6.05 lbf. Against deadweight primary standards, $U \approx 0.41$ lbf, $k = 2$, the TUR is about 24:1, the guardband shrinks to about 0.40 lbf, and the acceptance zone is ±9.60 lbf, almost the full specification. The TUR did not make the decision. It set how much of the tolerance you get to keep.

TUR is one of two doors, not the only one

There is a second route to the same 2 % global target, and it has nothing to do with TUR. NASA HDBK 8739.19-4 Annex 4 and the paper Using Reliability to Meet Z540.3's 2 % Rule show that if the end-of-period reliability for a family of like instruments is 89 % or better, the unconditional false accept risk across that population stays at or below 2 % whatever the measurement uncertainty. Our Start With the End in Mind paper calls these the dual dominance conditions: bench-level dominance at TUR ≥ 4.6:1, global dominance at EOPR ≥ 89 %. Either one alone is sufficient. That is why a lab with a modest TUR and good reliability records can meet a Z540.3 program requirement, and why a lab with a superb TUR and no records still has to compute the guardband. The two doors open onto the same room, and both are population statements. Neither changes the specific risk of the result at 10 008 lbf.

The boundary

TUR is an input to the decision rule, and sometimes a precondition inside it. It is never the rule. Specifying TUR ≥ 4:1 on a purchase order with no decision rule tells the lab how capable to be and tells it nothing about how to treat a result at 10 008 lbf. The Decision Rule Guidance puts it this way: a stated TUR is a global risk control. Specific risk governs the correctness of an individual pass/fail decision. Both are legitimate. They answer different questions.

TUR ≥ 4:1 by itself is a wish. UKAS LAB 48 writes simple acceptance rules the same way: measurement uncertainty enters as a pre-condition that must be met before the acceptance decision, and its example rule reads PASS when the measured value is within tolerance and the capability constraint is met, FAIL otherwise.

5. Which decision rule, and when

"Would you tell me, please, which way I ought to go from here?" Alice asks the Cheshire Cat. "That depends a good deal on where you want to get to," the Cat says. Decision rule selection is the same conversation. The standards do not pick a rule for you because the right rule depends on where you want to get to: how much risk you can carry, what data you have, and what a wrong call costs. If your rule is currently simple acceptance by default, adopted because nobody ever chose anything, then almost any deliberate choice is better than that.

ISO/IEC 17025:2017 is deliberately silent on method. It requires that a rule be defined, agreed with the customer when it is not inherent in the specification (Clause 7.1.3), documented with its risk (Clause 7.8.6.1), and reported (Clause 7.8.6.2). The Decision Rule Guidance summarizes the consequence: guardbands, probability calculations, and TUR are tools, not requirements. That leaves the choosing to you. Here is how to choose.

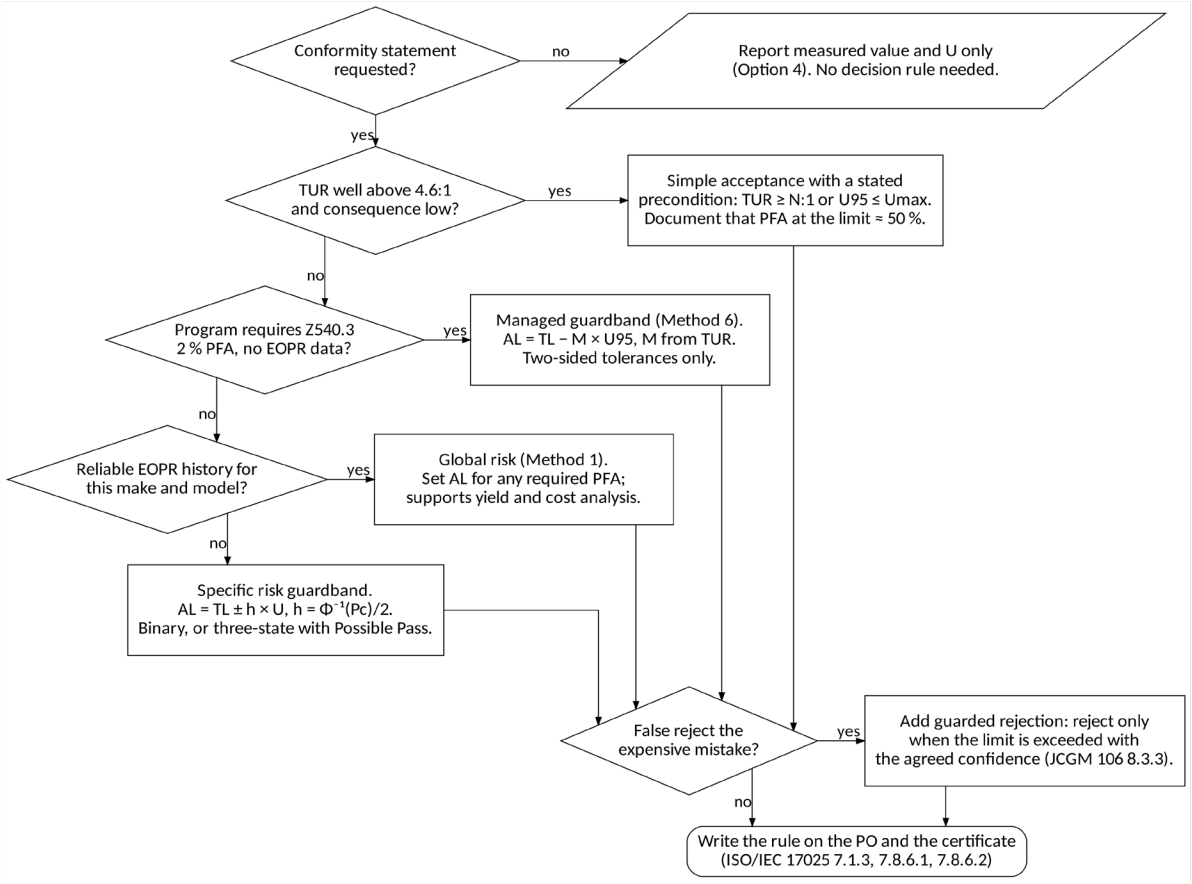


Figure 10. Decision rule selection, as the questions a lab actually asks. Adapted from the ILAC-G8 selection guide with the reliability-data question made explicit.

Question 0: Is a conformity statement needed at all?

UKAS LAB 48 makes a point most labs forget. Conformity decisions are not a default requirement of ISO/IEC 17025. For most calibrations there is no prior requirement to report pass or fail. If the customer does not ask, report the measured values and uncertainties, state the specification as a fact if they want it, and let them decide at their convenience. No decision rule is needed, and none should be implied. This is Option 4 in the supplier language later in this part, and it is the correct default for a lab calibrating other labs' reference standards.

Question 1: Is the TUR high and the consequence low?

If yes, simple acceptance can be defended, provided it carries a precondition. ILAC-G8 defines simple acceptance as AL = TL, and the guidance indicates that without a constraint on uncertainty it may not satisfy Clause 3.7 at all. So write the constraint in: PASS when within tolerance AND TUR ≥ 4:1, or AND $U_{95} \leq 2$ N. Document that specific PFA near the limit can approach 50 %, and have the customer agree to carry it. Use this when the TUR is well above 4.6:1, when the customer explicitly wants it and understands the risk, or when a wrong accept costs little.

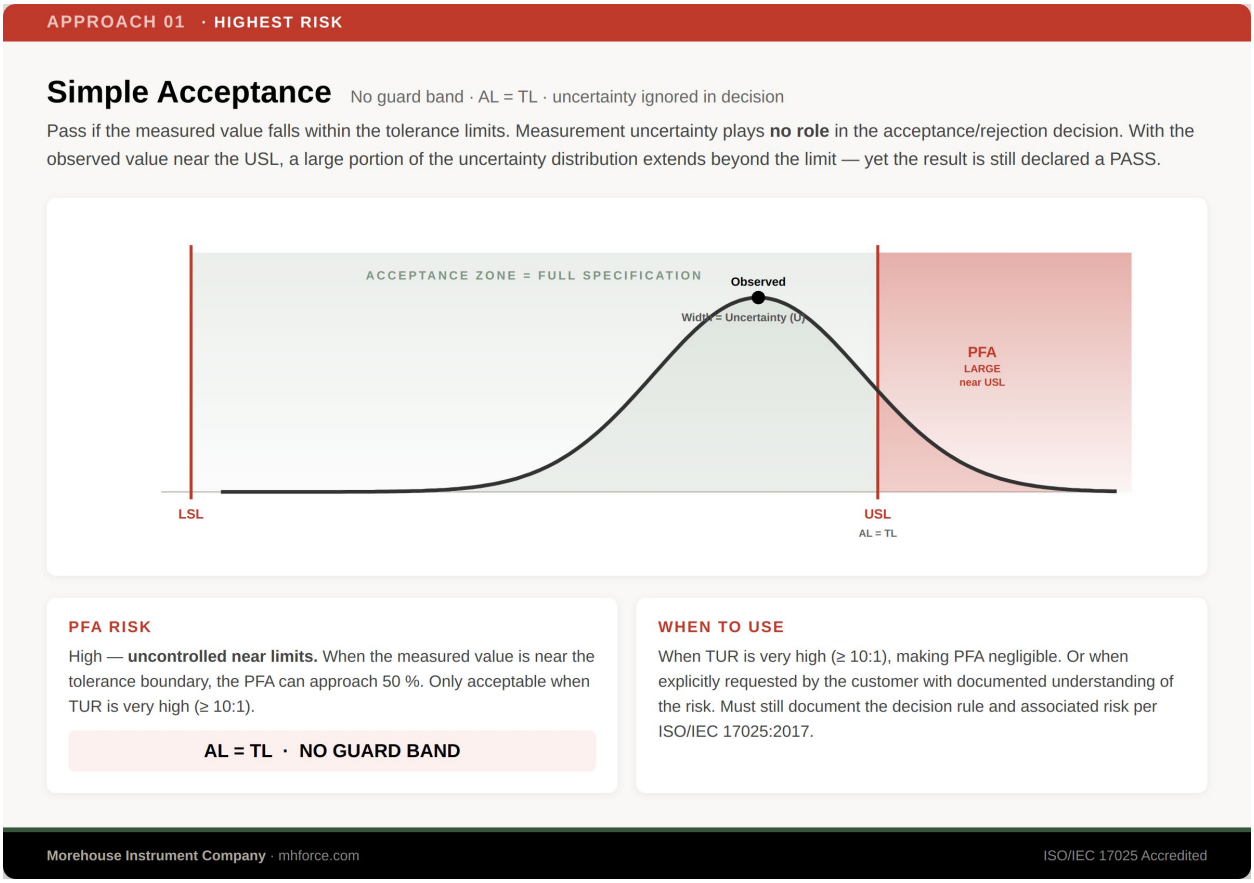


Figure 11. Simple acceptance: acceptance limit equals tolerance limit, and uncertainty plays no role in the decision itself.

Question 2: Do you have reliable history on this instrument population?

If not, use specific risk with a guardband. JCGM 106 defines specific risk in Bayesian terms, and JCGM GUM-5:2026 Example 10 confirms that when prior information is meagre, the single-distribution calculation comes out very close to the full Bayesian answer. That is most calibration work: one instrument, one result, no defensible population data. The rule reads: PASS when the specific PFA ≤ 2.5 % per side, otherwise FAIL. Section 6 gives the arithmetic. This method is conservative in the right direction, needs only the uncertainty and the result, and works for any risk level the customer names.

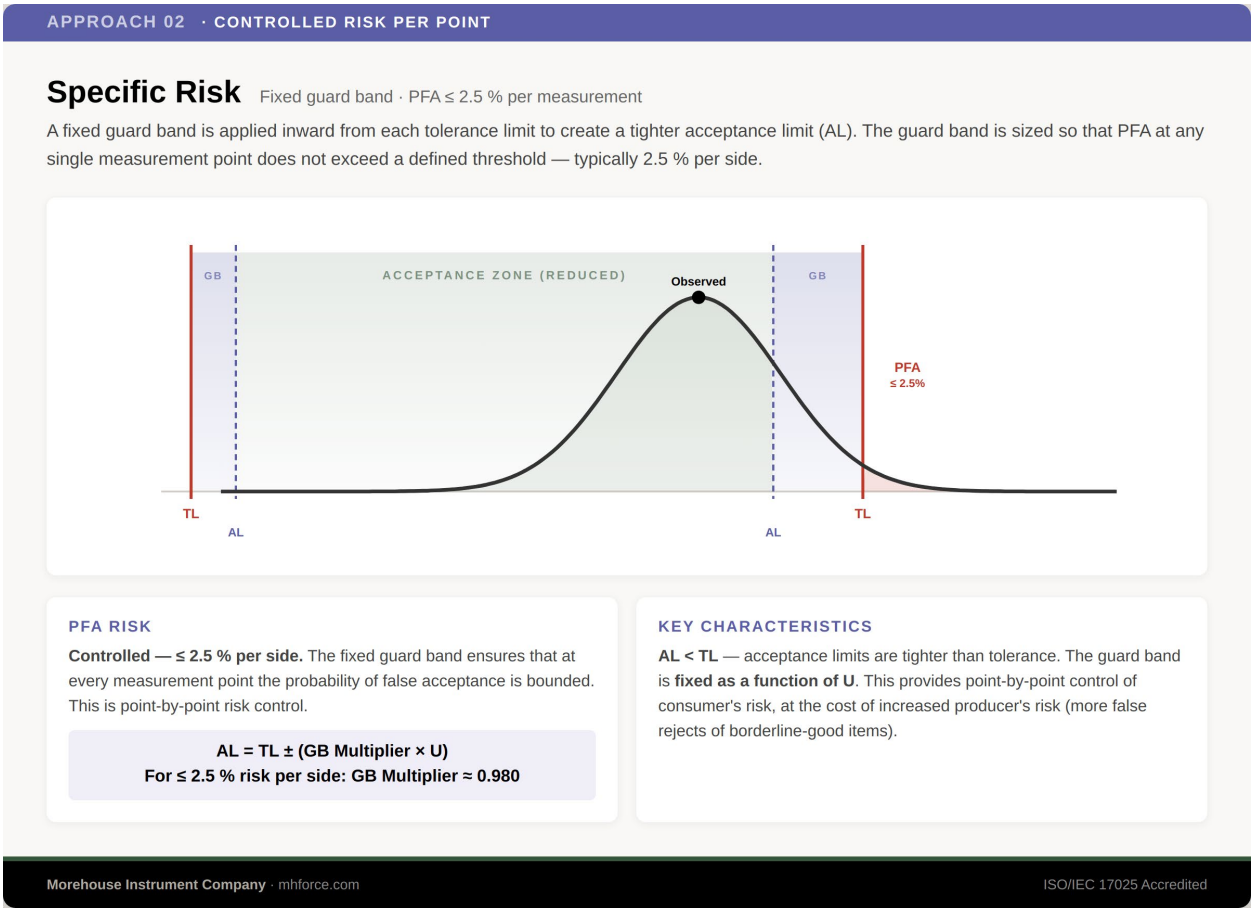


Figure 12. Specific risk: a fixed guardband sized so the PFA at any single measurement point does not exceed the agreed level.

If yes, you have real end-of-period reliability data for that make and model at that interval; global risk methods open up. Method 1 estimates unconditional PFA at the test-point level and lets you set acceptance limits for any required risk. It is the most flexible and the hardest to implement. It is also the route to the yield and cost analysis in Section 7.

Question 3: Does the program invoke the withdrawn ANSI/NCSL Z540.3's 2 % limit, with no reliability data?

Then use the managed guardband, Method 6, developed by Michael Dobbert. It needs no EOPR. It computes a multiplier M from the TUR, applies a guardband of $M \times U_{95}$, and holds the worst-case global PFA under 2 % across all combinations of TUR and out-of-tolerance rate. At TUR ≥ 4.6:1 the multiplier goes to zero and no guardband is required. Note that it is hard-coded to about 2 % and does not apply to one-sided tolerances.

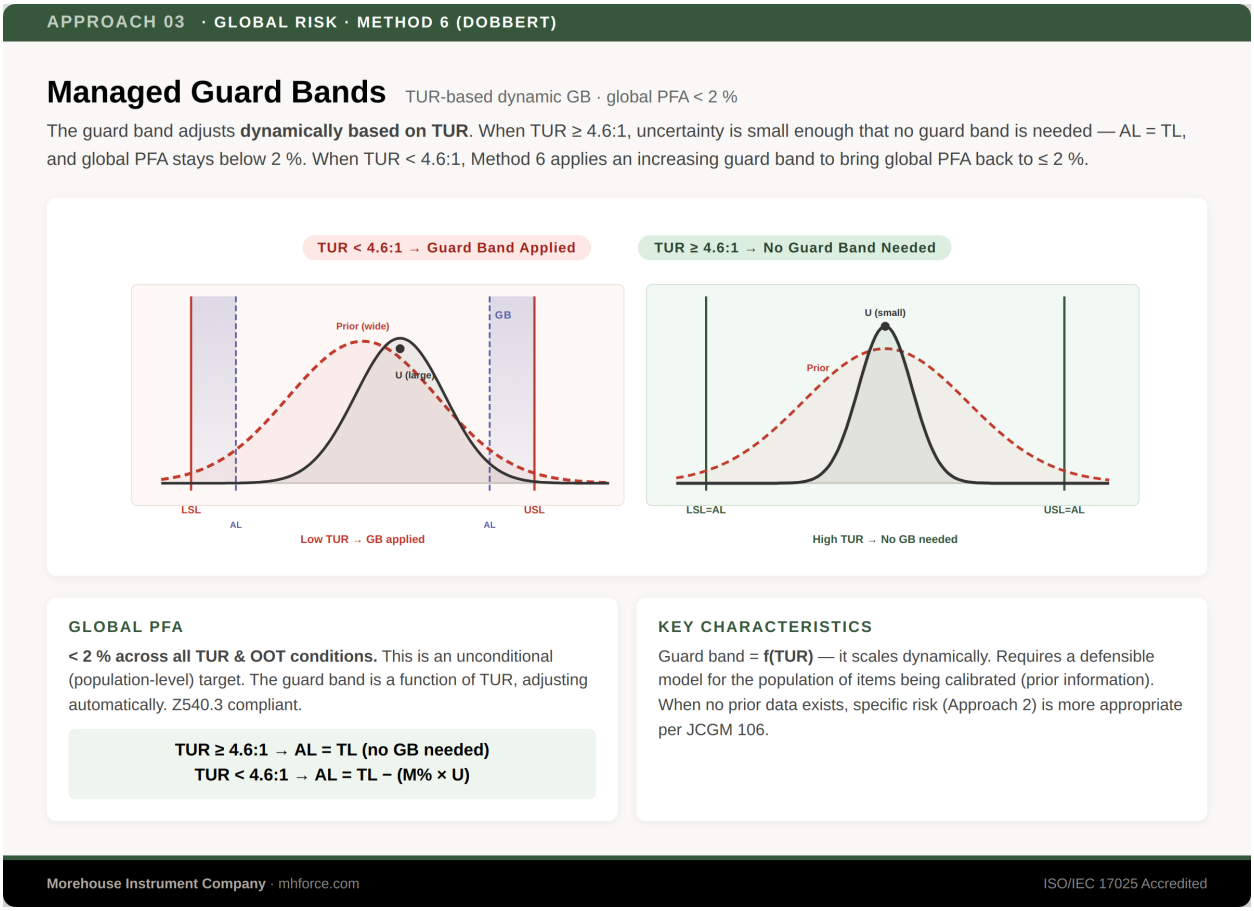


Figure 13. Managed guardband (Method 6): the guardband adjusts with TUR to hold global PFA below 2 % and vanishes at 4.6:1.

Question 4: Is a false reject the expensive mistake?

Then consider guarded rejection. JCGM 106 Section 8.3.3.1 describes setting the rejection limit outside the tolerance so that a rejected item is almost certainly non-conforming. It is the speed-camera rule: the officer wants 99.9 % confidence you were speeding before writing the ticket. In the JCGM example, a 100 km/h limit measured with 2 % relative standard uncertainty becomes a rejection threshold of about 107 km/h. Use it when you need clear evidence a limit has been exceeded before taking a negative action.

Question 5: Binary, or something in between?

A binary rule forces every result into pass or fail. ILAC-G8 and LAB 48 both allow multi-state rules, and a three-outcome version is often the most honest: PASS when specific PFA is at or below the agreed level, FAIL when PFA is 50 % or more, and Possible Pass in between, with the result, the uncertainty, and the

calculated PFA reported so the customer can decide disposition. Use this when a maximum PFA has not been specified, and the region near the limit is genuinely ambiguous.

Put it on the purchase order and the certificate

The best rule is the one the end user writes down. Three examples from the Decision Rule Guidance, each a complete instruction:

Simple acceptance with a precondition: Issue a PASS when the measured value is within ±10 N of the applied force and $U_{95} \leq 2$ N; otherwise FAIL. (Notice the 5:1 TUR requirement?)

Specific risk, binary: Issue a PASS when the PFA using specific risk is ≤ 2 %; otherwise FAIL.

Specific risk, multi-state: Issue a PASS when the PFA using specific risk is ≤ 2 %, a FAIL when PFA > 50 %, otherwise a Possible Pass.

A supplier can offer the same choices as a menu. Option 1, simple acceptance with a TUR or capability precondition. Option 2, a conservative guardband of one full expanded uncertainty, the older ILAC-G8 rule. Option 3, Method 6 with PFA under 2 %. Option 4, measured result and uncertainty only. State the default. No option is inherently superior; the suitability depends on the customer's needs, the contract, and the risk they are willing to carry.

The boundary

The decision rule is the agreement. It is not the TUR (that is a precondition inside some rules), not the guardband (that is the tool the rule uses), and not the PFA (that is what the rule controls). Pick it before the measurement is made, write it where the customer and the assessor can both find it, and apply it the same way every time.

6. Guardbands, and the one formula that sets them

Ask a field goal kicker about the wind. The goal posts never move. But a kicker facing a strong crosswind aims inside the upright to compensate for the drift he knows is coming. That aim point is the acceptance limit. The distance between the aim point and the upright is the guardband. The heavier the wind, the farther inside the post he must aim to keep the same chance of a good kick.

Measurement uncertainty is the wind. That one picture tells you everything a guardband is and is not. It is not a number you choose on feel. It is the distance you must aim inside the limit, given how hard the wind is blowing and how sure you need to be.

Definition first

ILAC-G8 defines the guardband as the interval between a tolerance limit and the corresponding acceptance limit, $w = |TL - AL|$. The tolerance limit is the customer's specification. The acceptance limit is the tighter line the lab actually uses to say PASS. The gap between them is the guardband.

A guardband deliberately trades some false rejects for fewer false accepts. You will reject more borderline-good instruments. You will accept fewer borderline-bad ones. How much of each depends entirely on how wide it is, and how wide it is depends on two things: the uncertainty and the risk you agreed to accept.

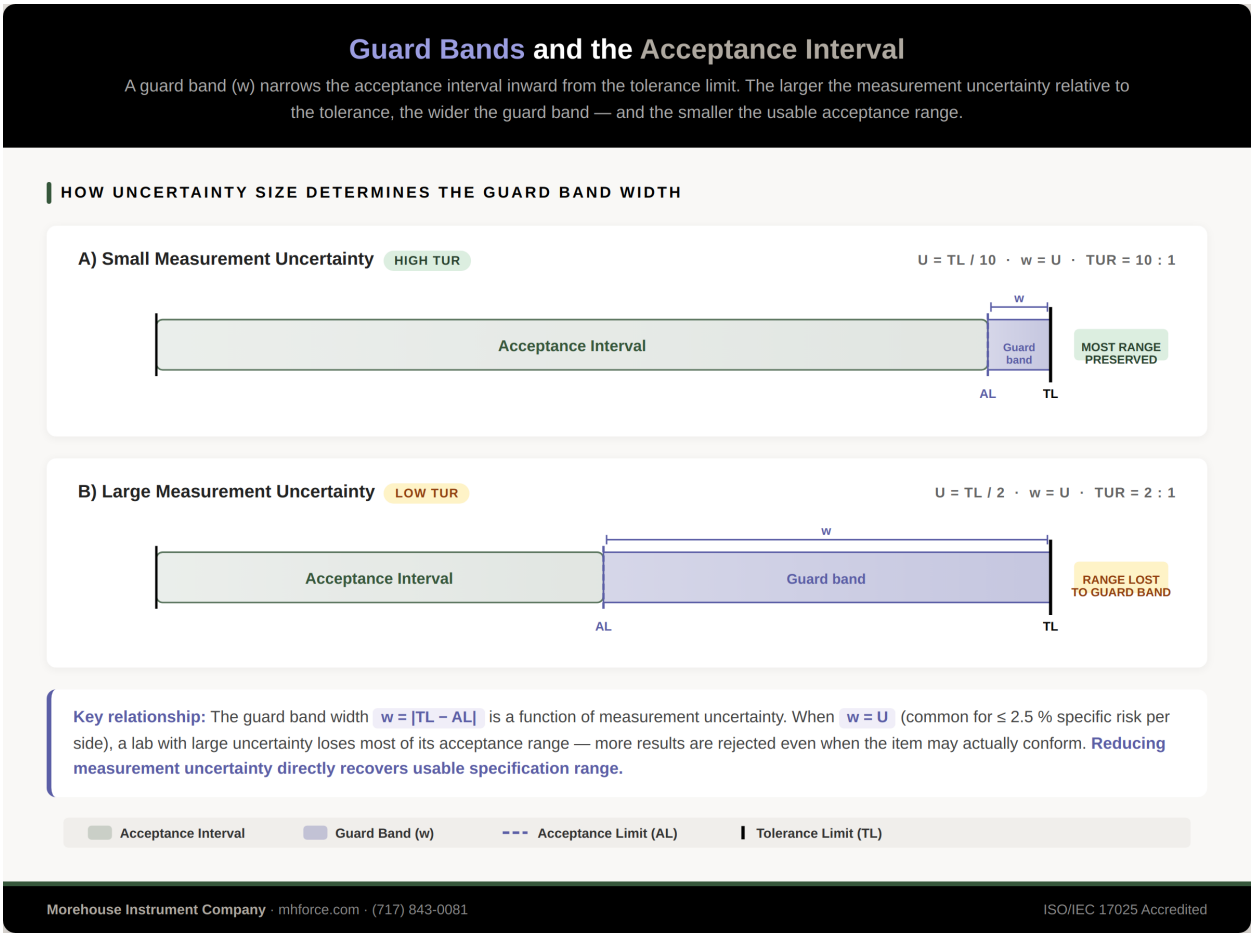


Figure 14. The guardband narrows the acceptance interval inward from the tolerance limit. The larger the uncertainty relative to the tolerance, the wider the guardband and the smaller the usable acceptance range.

The one formula

ASME B89.7.4.1 gives the method, and the Decision Rule Guidance calls it the de facto method because it works for any risk level a customer names. The guardband is a multiple of the expanded uncertainty, $w = h \times U$, and the multiplier is

$$h = \Phi^{-1}(Pc) / 2$$

where Φ^{-1} is the inverse standard normal function, and Pc is the probability of conformance you want at the acceptance limit. In Excel, that is NORM.S.INV(Pc)/2. Then

Acceptance limit = Tolerance limit ± $h \times U$

With the sign always chosen so the acceptance limit moves inward from the tolerance limit.

Worked example, two-sided.

- Tolerance: ±1 unit.
- Uncertainty: standard uncertainty 0.125 units, so U = 0.250 units at k = 2.
- Risk: no more than 2.5 % per side, so Pc = 0.975 and h = NORM.S.INV(0.975)/2 = 0.980.
- Acceptance limits: upper = 1 − 0.980 × 0.250 = 0.755 units, lower = −0.755 units. Any value between them passes.

A value of 0.82 is within tolerance and outside the acceptance limit; under a binary rule it fails, and under a three-state rule it is a Possible Pass with a calculated PFA of about 7.5 %.

Worked example, one-sided at 5 %.

- Tolerance and uncertainty: the same as the two-sided example.
- Risk: the customer accepts 5 %, so h = NORM.S.INV(0.95)/2 = 0.822.
- Acceptance limit = 1 − 0.822 × 0.250 = 0.794 units.

The multiplier changed because the risk changed, not because the tolerance is one-sided.

The shortcuts, and what they assume

UKAS LAB 48 points to a simple starting rule: make the guardband equal to the expanded uncertainty, so w = U95 and the multiplier is 1. In plain terms, move the acceptance limit inside the tolerance limit by one full uncertainty. Do that and the false accept risk at the limit is about 2.5 %. That holds for a one-sided tolerance, and for a two-sided tolerance when C_{95} is roughly 1.5 or better. This is the older ILAC-G8 rule, and at k = 2 it is slightly safer than the h = 0.980 version above. The acceptance limit lands two standard uncertainties inside the tolerance instead of 1.96, so the risk works out to 2.275 % instead of 2.5 %.

Note: the trade-off is more false rejects. A guardband pulls the acceptance limit inward, so more good instruments fail at the limit and come back for unnecessary rework.

Dobbert's managed guardband takes a different route. Its multiplier depends on the TUR: $M = 1.04 - \exp(0.38 \ln(Cm) - 0.54)$. At a TUR of 2:1, M = 0.282 and the acceptance limits land at ±0.859 for a ±1 tolerance with U = 0.5. At 4.6:1, M reaches zero. In any practical case, this guardband is narrower than the specific-risk one at the same TUR, because it is designed to cap the global PFA at 2 % over a population, not the specific PFA of one result.

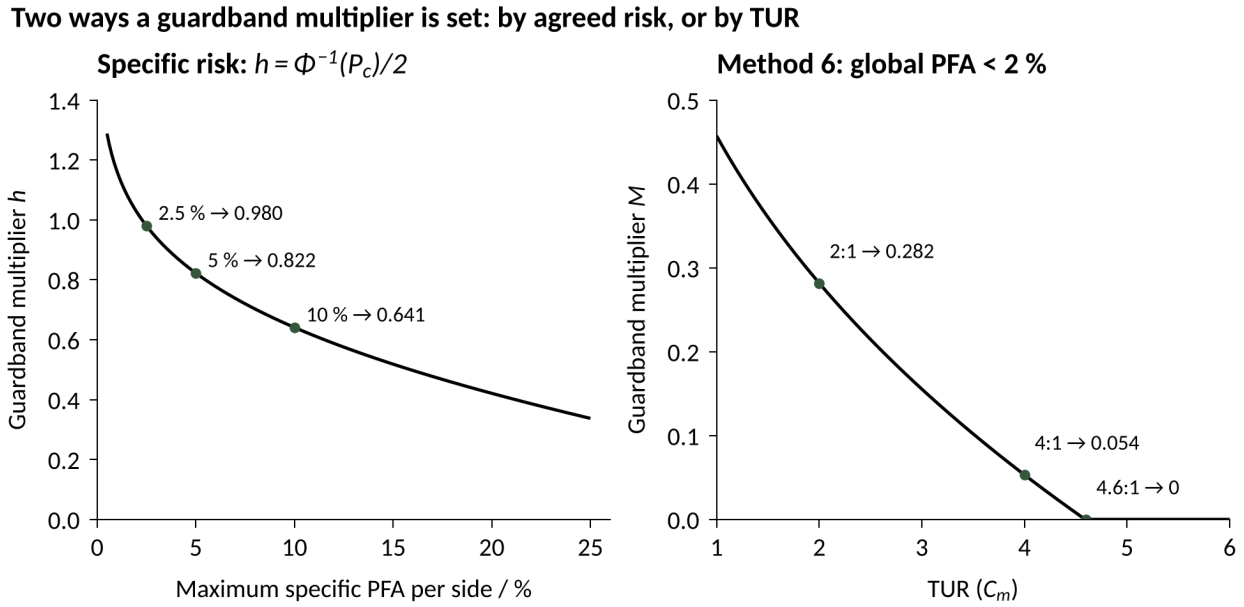


Figure 15. Two ways a guardband multiplier is set. Left: the specific-risk multiplier falls as the customer accepts more risk per side. Right: the Method 6 multiplier falls as TUR rises and reaches zero at 4.6:1.

The guardband tax, by TUR

The width of the guardband is a tax on the tolerance, and the rate depends on which rule you choose and how capable the measurement is. The table below comes from the Guard Band Analysis sheet in the Morehouse True Cost Calculator, for a ±10 lbf tolerance on a 10 000 lbf point.

TUR	*U* / lbf	Specific-risk guardband (2.5 % per side)	Tolerance consumed	Method 6 guardband	Tolerance consumed
2:1	5.00	4.90 lbf	49 %	1.41 lbf	14 %
3:1	3.33	3.27 lbf	33 %	0.52 lbf	5 %
4:1	2.50	2.45 lbf	25 %	0.13 lbf	1 %
4.6:1	2.17	2.13 lbf	21 %	0	0 %
10:1	1.00	0.98 lbf	10 %	0	0 %
24:1	0.42	0.41 lbf	4 %	0	0 %

Read the 2:1 row and the choice becomes concrete. A specific-risk rule hands half the tolerance back to the uncertainty; Method 6 takes a seventh of it. Both are legitimate. One protects every individual decision, the other protects the population average, and the difference between them is exactly the width of the guardband. Read the 24:1 row and the choice stops mattering: at a high enough TUR, every

rule leaves nearly the whole tolerance usable, which is the quiet argument for buying lower uncertainty instead of arguing about rules.

If you are thinking about who your calibration provider is, this is where calibration by Morehouse primary standard deadweight machines matters, provided your equipment is repeatable and has sufficient resolution. Our new model 6100 indicator, paired with an Ultra Precision load cell and calibrated by deadweight primary standards, yields some of the lowest measurement uncertainties available for a transfer standard. Only a few national metrology institutes (NMIs) and very few manufacturers offer anything better, and those that do often cost at least five times more for roughly 0.002 % lower total measurement uncertainty.

The payoff is simple: lower uncertainty raises your TUR, and a higher TUR means the guardband takes a smaller bite out of your tolerance. More of the tolerance stays usable, and fewer measurements land in the gray zone near the limits where wrong calls happen. That cuts both kinds of mistakes. A false accept passes a bad instrument, and the error travels into every measurement that instrument makes afterward. A false reject fails a good instrument and sends it back for rework that was never needed. Both are costly, and both can tarnish your reputation, though there is a strong argument that the false accept is worse, because the problem leaves the building with your name on it. **Only Elvis should leave the building.**

Three traps

Trap one: a guardband with no uncertainty behind it. A fixed fraction of the tolerance, applied without reference to U, is a tighter tolerance, not a guardband, because it has no relationship to the measurement's doubt. Trap two: a guardband of zero presented as a decision. The Decision Rule Guidance is blunt: stating that a guardband of $w = 0$ has been applied means no guardband has been applied. Say simple acceptance, and say what precondition makes it acceptable. Trap three: a guardband wider than the tolerance. When U approaches the tolerance, $h \times U$ consumes the whole acceptance zone. That is not a formula problem. It is the measurement telling you it cannot make this decision. Improve the capability or report the value and uncertainty without a conformity statement.

The boundary

A guardband is the output of two inputs: the uncertainty you evaluated and the risk you agreed to. Change either one and the guardband changes. It is not a decision rule; it is the tool a decision rule uses. It is not a TUR; the TUR tells you how wide the wind band is relative to the posts, and the guardband is where you aim. Keep the formula, keep the inputs, and the number takes care of itself.

The calculations and the curves are interactive at decisionrules.mhforce.com.

7. PFA and PFR, the two numbers at the end of the line

A kicker makes 97 % of his field goals from the 35-yard line. This kick is from the 35, and it sails wide. Based on the population data alone, you would have said it was likely to go in. Based on the one kick in front of you, it did not. Both statements are true. They answer different questions.

That is the entire difference between global risk and specific risk, and it is why a customer can be told their instrument passed at 2 % PFA and still have been handed a coin flip.

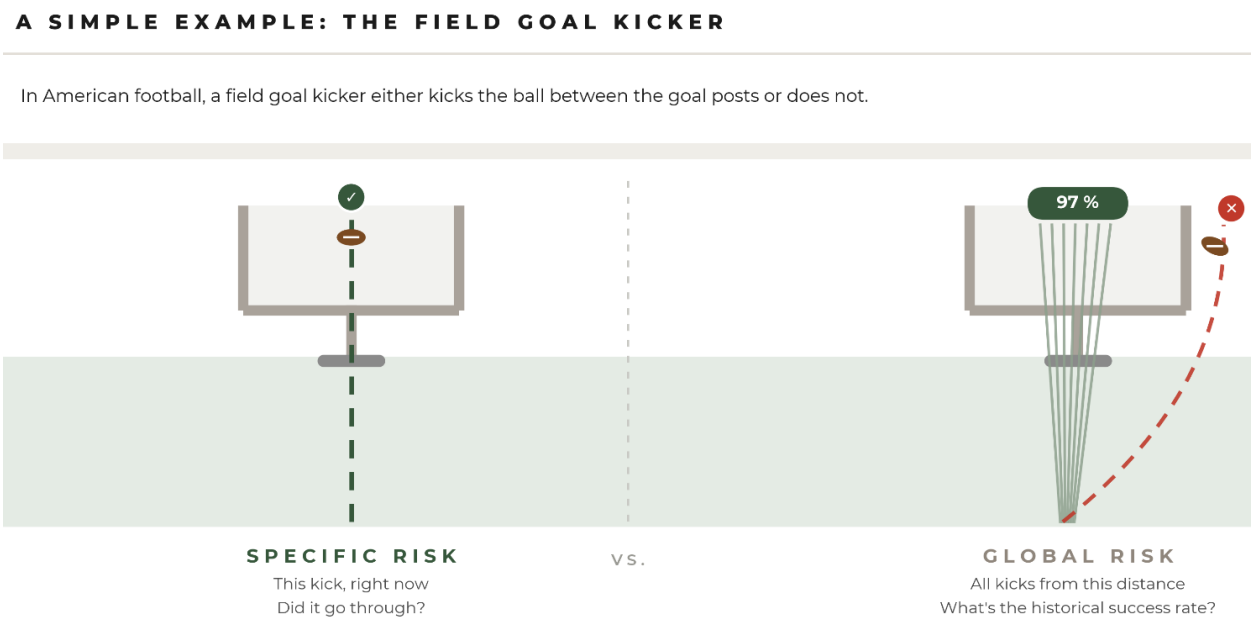


Figure 16. Specific risk is the one kick in front of you. Global risk is the kicker's record from that distance.

The definitions and the direction of each

PFA, the probability of false accept or acceptence, is the chance that an item declared conforming is actually out of tolerance. JCGM 106 calls it consumer's risk; statisticians call it a Type II error; ANSI/NCSL Z540.3 and the audit world shorten it to FAR, the false accept risk. PFR, the probability of false reject, is the chance that an item declared non-conforming is actually in tolerance. That is producer's risk, a Type I error, or FRR, the false reject risk. A guardband moves one up as it moves the other down. Choosing where to set the acceptance limit is, as JCGM 106 says, a business or policy decision that depends on the consequences.

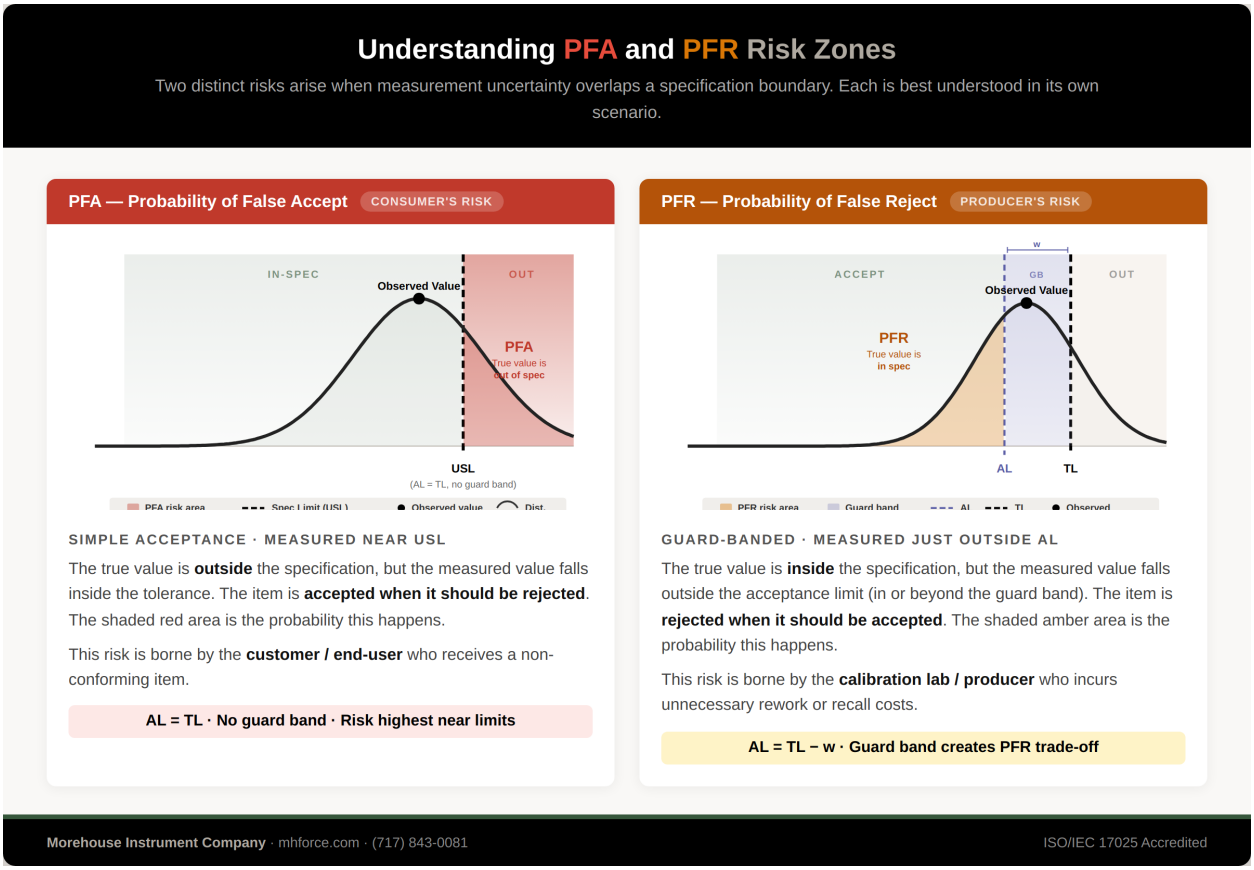


Figure 17. PFA is borne by the customer who receives a non-conforming item; PFR is borne by the lab or producer who rejects a conforming one. A guardband trades one for the other.

Specific risk: this result

Picture the bell curve centered on your measured value, its width set by the standard uncertainty. The area of that curve lying outside the tolerance limits is the specific PFA for this result. Excel does it in one line:

```
Risk upper = NORM.DIST(measured value, upper tolerance limit, standard uncertainty, TRUE)

Risk lower = 1 – NORM.DIST(measured value, lower tolerance limit, standard uncertainty, TRUE)
```

A 10 000 N load cell with a tolerance of ±0.1 % of full scale reads 10 000 N. The expanded uncertainty is 10 N, k = 2, so the standard uncertainty is 5 N.

- Risk upper = NORM.DIST(10 000, 10 010, 5, TRUE) = 2.275 %.
- Risk lower is the same, so the total specific PFA = 4.55 %.

Run the same formula with the reading at 10 008 N and the upper tail alone climbs to about 34 %.

UKAS LAB 48 makes the same point with a worked example, and our simplified decision rules guidance carries the risk arithmetic for it in a table that should be on every quality manager's wall. Tolerance ±1 unit, simple acceptance.

- A reading of 0.5 with U = ±0.1 passes with a PFA near zero.
- The same 0.5 with U = ±0.8 passes with a PFA near 11 %.
- A reading of 0.9 with U = ±0.8 passes with a PFA near 40 %.
- A reading of 0.99 passes at about 49 %.

Every one of them says PASS on the certificate.

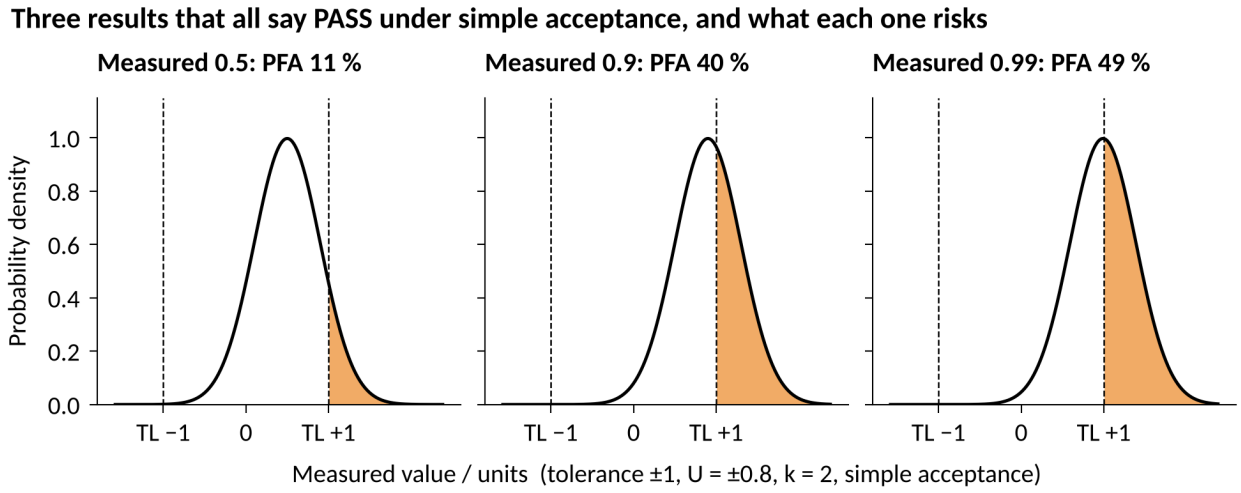


Figure 18. Three results, one tolerance, one uncertainty. Simple acceptance passes all three; the shaded tail is the specific PFA each one carries.

Specific risk needs nothing but the uncertainty and the result. It is the right framework when you have meagre prior information, which JCGM 106 and GUM-5 Example 10 both confirm is the usual case, and when a cautious approach to false accepts is required.

Global risk: this process

Global PFA asks a different question: across everything we calibrate of this kind, what fraction of the truly bad ones do we accept? It is an average over the whole population. To compute it, you need the measurement uncertainty and a second distribution, the a priori spread of the items themselves, which comes from history: end-of-period reliability for that make and model at that interval. JCGM 106 Section 9.5.2.2 gives the double integral. SunCal, Mathcad, and the functions in the Decision Rule Guidance appendix solve it.

The catch is what the average hides. ILAC-G8 and the Decision Rule Guidance both make the point that an instrument passing a 2 % global PFA criterion may carry a specific false accept risk approaching 50 % at the tolerance limit. The population looks fine. The one result at the edge is still a coin flip. GUM-5 Example 10 shows the reverse as well: a shared-risk rule on a real air-quality population came out at 0.5 % global consumer's risk and 0.9 % global producer's risk, because most of the population sat well inside the limit. Both numbers are honest. Neither one tells you what happened to the item in your hand.

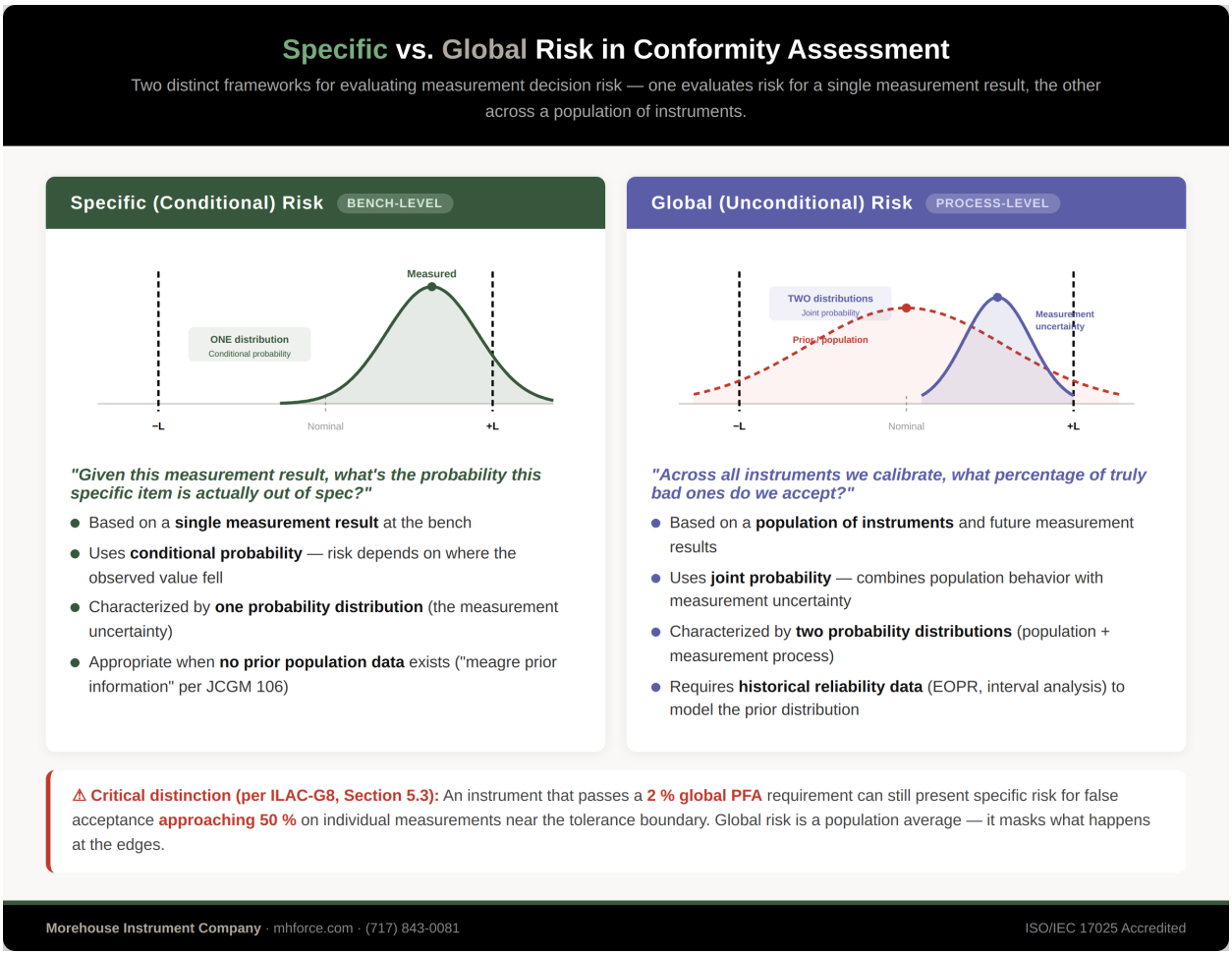
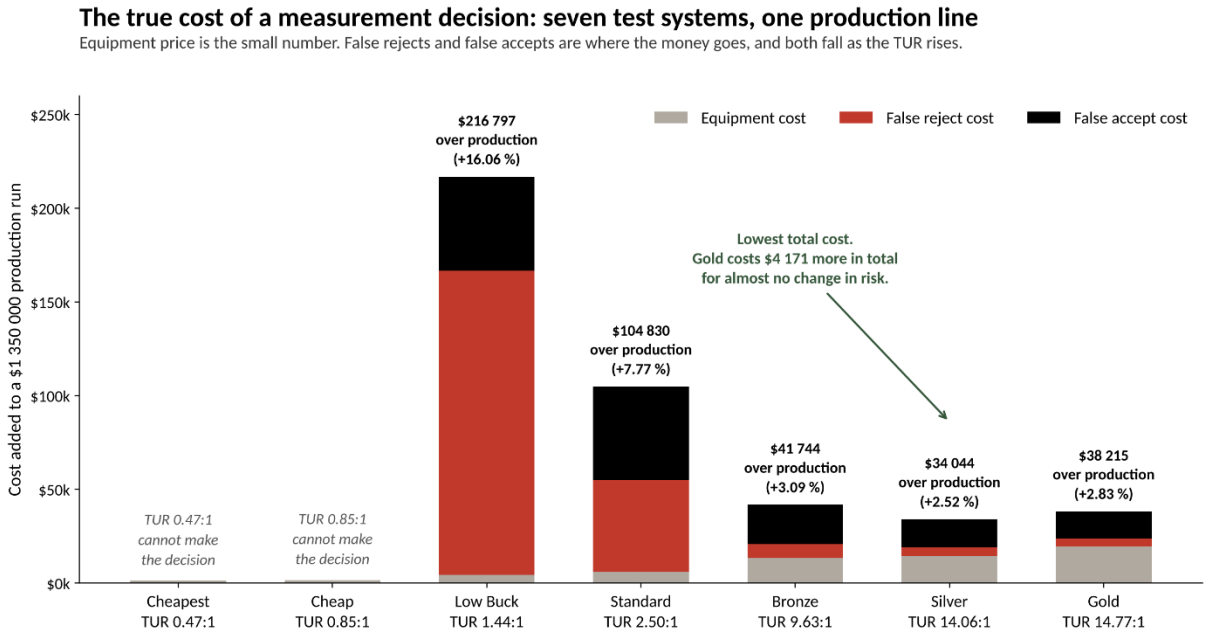


Figure 19. Specific (conditional) risk evaluates one result against one probability distribution. Global (unconditional) risk combines the measurement distribution with the population's prior distribution.

Where the dollars are



Equipment	Equipment cost	False reject cost	False accept cost	Total cost	Over production	Payback vs Low Buck	TUR
Acme Cheapest	$1 225	N/A	N/A	N/A	N/A	N/A	0.475:1
Acme Cheap	$1 650	N/A	N/A	N/A	N/A	N/A	0.852:1
Acme Low Buck	$4 405	$162 392	$50 000	$1 566 797	+16.06 %	Baseline	1.441:1
Acme Standard	$6 000	$48 830	$50 000	$1 454 830	+7.77 %	5.2 days	2.500:1
Acme Bronze	$13 481	$7 263	$21 000	$1 391 744	+3.09 %	18.9 days	9.634:1
Acme Silver	**$14 315**	**$4 779**	**$14 950**	**$1 384 044**	**+2.52 %**	**19.8 days**	**14.065:1**
Acme Gold	$19 429	$4 536	$14 250	$1 388 215	+2.83 %	30.7 days	14.771:1

Production cost $1 350 000 for 100 000 pieces at 1500 Ω ± 0.2 Ω, $13.50 to build and test, $35 to retest a reject. Grey rows: TUR below 1:1, the uncertainty exceeds the tolerance, so no conformity decision is possible and no risk cost can be assigned.
Payback is the time for the lower false reject and false accept cost to recover the extra equipment price over Acme Low Buck. Source: Morehouse True Cost Calculator, Decision Rule Guidance resistor example.

Figure 20. The true cost of a measurement decision. Seven test systems on the same 1500 Ω ± 0.2 Ω production run. Equipment price is the smallest slice of every bar; false rejects and false accepts are where the money goes, and both fall as TUR rises.

Global risk is where the money is, and Figure 20 shows it. Each bar is the same production run measured with a different test system: the gray slice is what the equipment costs, the red slice is false rejects, and the black slice is false accepts. The cheapest systems that can make the decision at all are the most expensive to own. The Decision Rule Guidance works the numbers behind it with a resistor example: 100 000 pieces a month at 1500 Ω ± 0.2 Ω, $13.50 to build and test each piece, $35 to retest a reject, and a test system with a capability index of 2.5. With that setup, 6 838 good pieces a month failed when they should have passed. Between the pieces wrongly rejected and the $35 retest on each one, the bill came

to $746 706 against a $1 350 000 production run, more than half the cost of making the parts. And when the rejects were retested, 44.41 % of them passed: nearly half the failures were never failures at all. A better reference instrument raised the capability index toward 9.6 and pushed the expected yield to 99.91 %. That is the drop you see moving right across Figure 20: better measurement pays for itself long before the equipment price matters.

GUM-5 Example 10 puts the trade in standards language. Tightening the acceptance limit enough to drive global consumers' risk from 0.5 % down to 0.002 % raised producers' risk from 0.9 % to 5.8 %. Relaxing it to cap producer's risk at 0.1 % pushed consumer's risk from 0.5 % to 1.7 %. There is no free direction.

What a false reject does on its way downstream

A false reject does not stay in the lab. Here is the chain of events from our PFR webinar. A customer's load cell has a tolerance of ±5 lbf at 10 000 lbf, 0.05 % of applied. They send it to a provider whose CMC is 0.05 %, so the TUR is at best 1:1. The instrument reads 10 009 lbf. The provider calls it out of tolerance and adjusts it to read 10 000. Now suppose a second lab, with a TUR of 10:1, calibrates the same instrument when it is next due and finds it reading 9 991 lbf. Out of tolerance again, and adjusted again. Had the first provider left it alone, the instrument would most likely have been sitting near 10 000 all along. The first adjustment was a false reject acted upon, and it planted a 9 lbf bias in a good instrument.

Then it propagates. The customer uses that instrument as a reference to calibrate 0.1 % devices. They load it until it reads 10 000, which means only 9 991 lbf is actually applied, and every device under it inherits the error. Some of those devices fail. Some get adjusted. The lab that made the first bad call never sees any of it, and unless someone is tracking out-of-tolerance rates and asking why they rose, nobody does. The cost ladder of a false reject runs retest, then return, then recall, and each rung is an order of magnitude more expensive than the one before.

The rule you pick sets your rework bill

The Morehouse True Cost Calculator makes the trade concrete for a force lab. Take a ±0.1 % tolerance at 10 000 lbf, an EOPR of 89 %, 200 calibrations a year, and $87.50 per false reject event (35 minutes at $150 an hour). Under Method 6, a machine at 2.2:1 carries a PFR near 10.4 %, which is about $1 820 a year in rework. A machine at 3.3:1 drops to 3.7 % and $643. At 5:1 it is 2.5 % and $440. Deadweight at 50:1 is 0.5 % and $83. A machine at 1.25:1 is 54 % and $9 453, and at 1:1 the PFR reaches 74 %: three of every four good instruments rejected. Same tolerance, same reliability, same rule. Only the uncertainty changed.

Now hold the machine fixed and change the rule. For the 2.2:1 machine, annual rework runs about $913 under a global rule set for 2 % PFA, $1 820 under Method 6, and $5 947 under a specific-risk guardband at 2.5 % per side. For deadweight, the same three rules give $83, $83, and $159. The ordering is always global, then Method 6, then specific, because each rule is more protective of the individual decision

than the last and pays for it in false rejects. That ordering is not a flaw in any of the rules. It is the trade you are choosing when you pick one, and the calculator exists so you can see the price before you sign the purchase order.

Note: The cost calculator allows you to vary all of your labor costs, calibration costs, and measurement uncertainty to calculate as close as possible to your actual true costs.

The boundary

PFA and PFR are outputs. A specific-risk rule calculates them for one result. A global-risk rule is designed to hold them under a target for a population, and needs reliability data to do it. They are not properties of the instrument, the lab, or the TUR on their own; they are what you get when uncertainty, bias correction, capability, a decision rule, and a guardband have all done their jobs in order.

Which brings the paper back to where it started. Seven terms, one chain. Uncertainty is the input. Bias is corrected before it. CMC is the floor under it. TUR screens it. The decision rule is the agreement about it. The guardband is the tool. PFA and PFR are the result. Keep them in that order and the messy mix becomes a line you can walk, and defend, from the first reading to the last word on the certificate.

If you want the full treatment with every formula and example, the Decision Rule Guidance is free at mhforce.com, and the calculations run interactively at decisionrules.mhforce.com. Read it, pick your rule, and write it down before the next instrument hits the machine.

Appendix A: Frequently asked questions

Seven parts generate questions. Here are the ones that show up most, with short answers and a pointer to the part that carries the long version.

It's also the credits section, hence part 8 of 7, as not many will stay for the credits, and this one's ending is all the references used, courtesy of Claude, as there were a lot, and what used to take me a day is now done in minutes, talking about generating the FAQ's and references from the previous content.

Is TUR ≥ 4:1 a decision rule?

No. It is a screen, and sometimes a precondition inside a rule, but a rule has to say what happens to the result. PASS when the value is within tolerance and TUR ≥ 4:1, otherwise FAIL, is a decision rule. The ratio by itself is a wish. Section 4 has the history and the UKAS LAB 48 support.

What is simple acceptance?

In tolerance means pass, with no guardband (w = 0). ISO/IEC 17025 permits it when the rule is explicitly defined and agreed with the customer, and the customer accepts a false accept risk that can reach 50 % for a value at the tolerance limit. Figure 8 shows the coin flip.

What is the difference between PFA and PFR?

PFA is the chance a bad instrument passes. PFR is the chance a good one fails. The false accept leaves the building and spreads its error into everything the instrument measures next. The false reject stays home and runs up a rework bill. Section 7 prices both.

What is the difference between specific and global risk?

Specific risk is about the one result in front of you: given this reading and this uncertainty, what is the chance this call is wrong. Global risk is about the population: across everything calibrated under the rule, what fraction of calls are wrong. Global risk needs reliability data, such as end-of-period reliability, to calculate.

How big should a guardband be?

Start from the risk you can carry, not from a habit. The multiplier is h = NORM.S.INV(Pc)/2: a 2.5 % risk per side gives h = 0.980, and 5 % gives 0.822. If you want a simple default, w = U95 is the old ILAC-G8 rule and holds the false accept risk near 2.5 %.

Does a guardband make the risk go away?

No. It moves risk from one side of the deal to the other. Tighten the acceptance limits and false accepts fall while false rejects rise. GUM-5 Example 10 and the guardband tax table in Section 6 both show it. There is no free direction.

Is bias part of the uncertainty?

No. A known bias gets corrected, and the uncertainty of that correction goes into the budget. Inflating the uncertainty to cover an error you know about is padding, not caution. Section 3 draws the line.

My provider has a great CMC. Does that settle it?

CMC is the floor, not your number. The uncertainty on your certificate includes your instrument's behavior: resolution, repeatability, and how it was used. A great CMC with a noisy instrument still yields a large uncertainty. Section 2 walks through it.

Who picks the decision rule?

You and your calibration provider, together, in writing, before the work is done. ISO/IEC 17025 requires the rule to be agreed with the customer and identified on the certificate. Doing whatever the lab did last time is also a rule; it is just one that nobody chose.

Where do I start?

Write down the tolerance you actually need. Get the real uncertainty, not the brochure number. Pick the rule by the risk you can carry, run it through the calculator at decisionrules.mhforce.com, and put it in the purchase order. Then hold everyone to it, including us.

Seven parts, twenty figures, one ask: pick your rule and write it down. The Cat cannot tell you which way to go. Get in the fast lane, Grandma, the bingo game is ready to roll, and Elvis has just left the building.

References

Sources cited in this paper, listed alphabetically. JCGM documents follow the citation form the BIPM recommends, with the member-organization author line and the JCGM DOI. Clause numbers given in the text refer to these editions.

A2LA. *R205 — Specific Requirements: Calibration Laboratory Accreditation Program.* American Association for Laboratory Accreditation, Frederick, MD.

ANSI/NCSL Z540.3-2006 (R2013). *Requirements for the Calibration of Measuring and Test Equipment.* NCSL International, Boulder, CO, 2006.

ASME B89.7.4.1-2005. *Measurement Uncertainty Conformance Testing: Risk Analysis.* The American Society of Mechanical Engineers, New York, 2005.

ASTM E74-18. *Standard Practices for Calibration and Verification for Force-Measuring Instruments.* ASTM International, West Conshohocken, PA, 2018. https://doi.org/10.1520/E0074-18

BIPM. *The BIPM key comparison database (KCDB).* Bureau International des Poids et Mesures, Sèvres. https://www.bipm.org/kcdb/

BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP and OIML. *Evaluation of measurement data — Guide to the expression of uncertainty in measurement.* Joint Committee for Guides in Metrology, JCGM 100:2008. https://doi.org/10.59161/JCGM100-2008E

BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP and OIML. *Evaluation of measurement data — The role of measurement uncertainty in conformity assessment.* Joint Committee for Guides in Metrology, JCGM 106:2012. https://doi.org/10.59161/JCGM106-2012

BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP and OIML. *International vocabulary of metrology — Basic and general concepts and associated terms (VIM), 3rd edition.* Joint Committee for Guides in Metrology, JCGM 200:2012. https://doi.org/10.59161/JCGM200-2012

BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP and OIML. *Guide to the expression of uncertainty in measurement — Part 5: Examples.* Joint Committee for Guides in Metrology, JCGM GUM-5:2026. https://doi.org/10.59161/JCGMGUM-5-2026

Carroll, L. *Alice's Adventures in Wonderland.* Macmillan, London, 1865. https://www.gutenberg.org/ebooks/11

Cenker, G., Zumbrun, H. and Shah, D. *Decision Rule Guidance.* Morehouse Instrument Company, York, PA. https://mhforce.com/documentation-tools/

Dobbert, M. "A Guard-Band Strategy for Managing False-Accept Risk." *Proceedings of the NCSL International Workshop & Symposium,* Orlando, FL, 2008.

Eagle, A. R. "A Method for Handling Errors in Testing and Measuring." *Industrial Quality Control,* Vol. 10, pp. 10–15, March 1954.

Grubbs, F. E. and Coon, H. J. "On Setting Test Limits Relative to Specification Limits." *Industrial Quality Control,* Vol. 10, pp. 15–20, March 1954.

Harris, G. L. "GMP 13 — Good Measurement Practice for Ensuring Traceability." In: *Selected Laboratory and Measurement Practices and Procedures to Support Basic Mass Calibrations,* NIST IR 6969, National Institute of Standards and Technology, Gaithersburg, MD, 2019. https://doi.org/10.6028/NIST.IR.6969-2019

Hayes, J. L. *Factors Affecting Measuring Reliability.* U.S. Naval Ordnance Laboratory, Technical Memorandum No. 63-106, Corona, CA, 1955.

ILAC-G8:09/2019. *Guidelines on Decision Rules and Statements of Conformity.* International Laboratory Accreditation Cooperation, 2019. https://ilac.org/publications-and-resources/ilac-guidance-series/

ILAC-P14:09/2020. *ILAC Policy for Measurement Uncertainty in Calibration.* International Laboratory Accreditation Cooperation, 2020. https://ilac.org/publications-and-resources/ilac-policy-series/

ISO/IEC 17025:2017. *General requirements for the competence of testing and calibration laboratories.* International Organization for Standardization, Geneva, 2017. https://www.iso.org/standard/66912.html

ISO/IEC 17043:2023. *Conformity assessment — General requirements for the competence of proficiency testing providers.* International Organization for Standardization, Geneva, 2023. https://www.iso.org/standard/80864.html

Mimbs, S. M. "Using Reliability to Meet Z540.3's 2 % Rule." *NCSL International Workshop & Symposium,* National Harbor, MD, 2011. NASA KSC-2011-151.

Morehouse Instrument Company. *Decision rules interactive calculator.* https://decisionrules.mhforce.com

Morehouse Instrument Company. *Guidance on Uncertainty Budgets for Force Measuring Devices.* York, PA. https://mhforce.com/documentation-tools/

Morehouse Instrument Company. *GUM in Plain English* (series). York, PA. https://mhforce.com/documentation-tools/

Morehouse Instrument Company. *Morehouse CMC Workbook.* York, PA. https://mhforce.com/documentation-tools/

Morehouse Instrument Company. *Morehouse True Cost Calculator* and *The True Cost of Measurement Decisions.* York, PA. https://mhforce.com/documentation-tools/

Morehouse Instrument Company. *Start With the End in Mind.* York, PA, 2026. https://mhforce.com/wp-content/uploads/2026/05/Start_With_The_End_in_Mind.pdf

Morehouse Instrument Company. *Probability of False Reject (PFR)* (webinar). York, PA. https://mhforce.com/documentation-tools/

NASA-HDBK-8739.19-4. *Estimation and Evaluation of Measurement Decision Risk.* NASA Measurement Quality Assurance Handbook, Annex 4. National Aeronautics and Space Administration, Washington, DC, 2010. https://standards.nasa.gov/standard/NASA/NASA-HDBK-873919-4

NCSL International. *Handbook for the Application of ANSI/NCSL Z540.3-2006 — Requirements for the Calibration of Measuring and Test Equipment.* NCSL International, Boulder, CO, 2009.

Peart, N. "Freewill." Performed by Rush on *Permanent Waves.* Anthem Records, 1980.

Reese, P. and Harben, J. "Risk Mitigation Strategies for Compliance Testing." *NCSLI Measure,* Vol. 7, No. 1, 2012.

UKAS. *LAB 48: Decision Rules and Statements of Conformity,* Edition 5. United Kingdom Accreditation Service, Staines-upon-Thames, 2024. https://www.ukas.com/wp-content/uploads/2023/05/LAB-48-Decision-rules-and-statements-of-conformity.pdf

About the author

Henry Zumbrun is a non-feathered bi-pod who stands upright and talks. He likes Metallica and the Grateful Dead; he tries to play guitar, and might be a goober. Oh, and he won some awards in the metrology field; he's written some books, and he's too old to have received participation trophies, though he should have had a room full of them.