Handbook 44 Tests: A Risk Assessment
Ross Andersen, New York (retired)
Comments to S&T Blocks B-1 and B-2
December 11, 2018

> *Human beings are notoriously bad at guessing probabilities, and a bad guess in many conflict contexts can be exceedingly costly.*
> *Norman Schultz*
>
> *It often costs resources to avoid risks. Unreasonably cautious policies promise to be exceedingly expensive and/or extremely low on the cost/benefit scale.*
> Norman Schultz

The NCWM is debating the important subject of suitability of transfer standards. The NIST Working Group on Alternative Test Methods is working in a similar area. What do we demand of our testing standards and our test methods? This paper attempts to provide some needed information on the risks involved. As the quotes above indicate, knowing the risks can be very helpful in avoiding overly costly conclusions one way or the other. The first quote deals with taking too much risk and the second with overly avoiding risk.

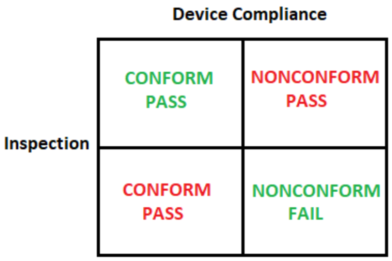
The place to begin is with an understanding of what Handbook 44 tests are. To place this in context, consider the last three links in the traceability chain leading to the retail gas pump. In the lab, they use mass standards and a gravimetric procedure to test a slicker plate standard. Next they use the slicker plate and a volume transfer procedure to test the test measure. In the field, they use the test measure and a volume transfer procedure to test the gas pump. I want to compare and contrast these three steps in terms of "calibration" and "verification." Here I am using the terms as defined in the UWML in Handbook 130.

I think of "calibration" as an attempt to define something precisely and "verification" as an attempt to assess if something is suitable for a purpose. Essentially, calibration tells what something is and verification decides if it is good enough. The signs I look for in the Report of Test are either four crucial items (reference conditions, measured value at the reference conditions, uncertainty, and confidence interval) for calibration, or a statement of conformity for verification.

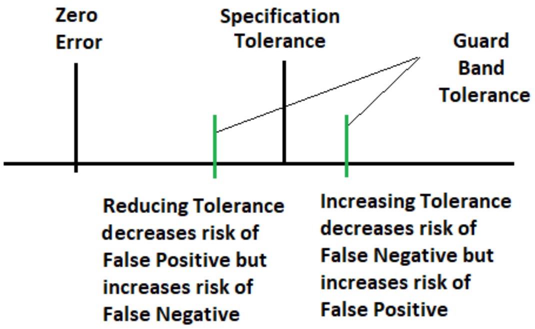
The test of the slicker plate is a probably calibration. The ROT for the slicker will show that the delivery from the slicker at 60 °F was measured as 5.0008 gal with an uncertainty of 0.0005 gal at 95% confidence. You may see no judgment of conformity and the uncertainty follows the GUM (ISO Guide to the Expression of Uncertainty in Measurement). A typical GUM formula would be
$U = k * \sqrt{}(u_s^2 + s_b^2 + u_o^2 + u_b^2)$. The k term is a coverage factor that defines the included probabilities, normally 2 for a little over 95% confidence. The u_s term is uncertainty from the standard and comes from the uncertainty reported in the calibration of the reference standard. The s_b term is uncertainty from the test procedure and comes from an analysis of control charts of representative check standards. The u_o term is uncertainty from other sources and comes from a systems analysis. The u_b term is uncertainty from biases and comes from analysis of proficiency tests.

The test of the 5 gallon test measure is probably verification. The ROT will show the item conforms to NIST Handbook 105-3. The ROT may also provide the four crucial items for a calibration at the reference mark at5 gal. We may learn that, in addition to conformance to Handbook 105-3, the delivery of the test measure at 60 °F was measured as 5 gal (+0.1 cu in) with an uncertainty of 0.3 cu in at 95% confidence. If both the conformity statement and the calibration data are provided, it is a hybrid, perhaps better stated as a calibration/verification. However, it is important to note that with instruments with a range of indication it is common to calibrate at more than one point to establish uniformity of the measuring scale. In our case, the conformity with 105-3 includes a test of the measuring scale(s).

Another crucial issue with conformity is the decisions rules. In the 2017 revision to ISO 17025 I understand that testing labs will now have to describe the decision rules in the ROT. These rules define how the lab addressed measurement uncertainty in making tolerance compliance decisions. With uncertainty we always face the possibility of erroneous decisions. If we pass a non-conforming device we have a false positive, in the upper right of the figure. If we fail a conforming device we have a false negative, in the lower left. While we can't eliminate erroneous decisions, we can take steps to reduce one of the two risks. This is called "guard banding."


Device Compliance

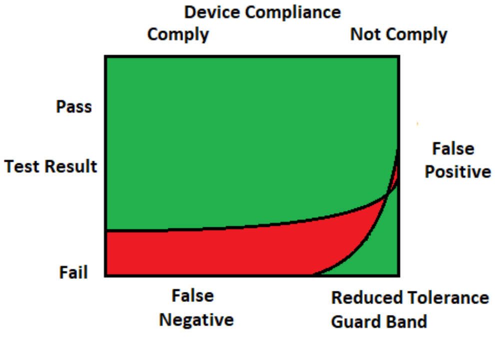
Inspection		
	CONFORM PASS	NONCONFORM PASS
	CONFORM PASS	NONCONFORM FAIL


Zero Error
Specification Tolerance
Guard Band Tolerance
Reducing Tolerance decreases risk of False Positive but increases risk of False Negative
Increasing Tolerance decreases risk of False Negative but increases risk of False Positive

In the state labs they guard band when applying tolerances to field standards by reducing the tolerance by the measurement uncertainty. This form reduces risk of false positives but increases risk of false negatives. To help visualize the guard banding at the state labs, consider the depiction below. The baseline is formed by the line coming down from the upper right with about 5% non-complying devices to the right of the dividing line. Yet the lab has reduced the tolerance for enforcement by about 20%.

The false positives are significantly reduced to near 0%, but the false negatives are probably 15% of the population. This is not a tragedy since it only means the technician has to adjust and retest the measure to get it back in compliance.

In contrast, enforcement of highway weight limits work in the other direction. The official recognizes that the portable scales used have a 2% tolerance and thus they increase the overweight tolerance by this 2% variability. This form of guard band decreases risk of a false negative (an unfair citation) but increases the risk of having trucks over the legal limit by small amounts. Handbook 44 sometimes uses this approach and sometimes doesn't.



We can now move to the field and see that the test of the gas pump is verification, hence our terminology initial and subsequent verification. It is not calibration in that we may not have uncertainty for the field standard (if it was only verified) and we presently have no control charts or proficiency tests on which to estimate an uncertainty of the field test. The crucial question though is what is the inspector verifying in a Handbook 44 test? The universal response I get to that question is "We're verifying conformance with NIST Handbook 44."

However, I respond, "I don't believe that's what our field inspectors do." I usually get strange looks after saying this. I'll offer some evidence to support my position. Exhibit 1, look carefully at a field inspection report for devices that pass and the approval seal that is applied to the device. Where on these official documents do you find a statement "conforms to NIST Handbook 44?" I'm betting you won't find one, and rightly so.

Exhibit 2, look carefully at the official documents for a device that fails. I'm betting that your stop-use order clearly states "This device fails to conform to NIST Handbook 44 section XX."

Exhibit 3, Handbook 44 Introduction, paragraph B.

> **B. Purpose.** The purpose of these technical requirements is to eliminate from use, weights and measures and weighing and measuring devices that give readings that are false, that are of such construction that they are faulty (that is, that are not reasonably permanent in their adjustment or will not repeat their indications correctly), or that facilitate the perpetration of fraud, without prejudice to apparatus that conforms as closely as practicable to the official standards.

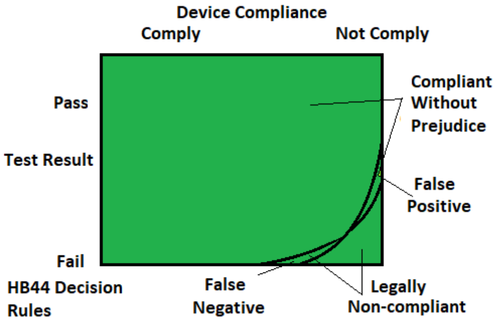
Notice that the Handbook is clearly focused on non-conformance. It talks about rejecting rather than approving devices. This may be subtle but it is crucial. The last part of the paragraph, beginning with "without prejudice," is of particular import. We reject devices for non-conformance with both technical requirements and performance requirements, but after the "without prejudice" it only mentions performance. You can't use the word "closely" with regard to technical requirements. I am not a lawyer so please verify what follows with your legal counsel. The term "without prejudice" is a legal term and I have found it means that the conformity that comes after the term without prejudice has not been fully established. Literally the inspector has not made a definitive decision on compliance. Another key feature of without prejudice is the inspector cannot be held liable for his compliance decision later on if the device is found to be non-compliant. That's a good thing for them and the W&M departments. To understand why HB44 hedges its bets this way, consider the following.

Exhibit 4. Handbook 44 General Code paragraph G-S.5.4.

> **G-S.5.4. Repeatability of Indications.** – A device shall be capable of repeating, within prescribed tolerances, its indications and recorded representations. This requirement shall be met irrespective of repeated manipulation of any element of the device in a manner approximating normal usage (including displacement of the indicating elements to the full extent allowed by the construction of the device and repeated operation of a locking or relieving mechanism) and of the repeated performance of steps or operations that are embraced in the testing procedure.

This section starts with a blanket statement requiring compliance and then includes additional clarifying information. The first part of the clarification deals with use of the device in a "manner approximating normal usage." Contrast that with the HB 105-3 that specifies accuracy only at 60 °F. If you use a 5 gallon stainless test measure at temperatures other than 60 °F, you make errors since it is not at reference conditions. Just a 10 °F difference results changes in the measure exceeding the lab uncertainty. The gas pump, however, has no reference condition and must perform within tolerance at all conditions of normal usage. This means it must be within tolerance on the hottest and coldest days of the year, the days when it measures the most dense summer gas and the least dense winter gas, etc. The inspector gets to do a few tests on some random day. The people who wrote these sections of Handbook 44 understood that a few tests can't tell us enough to make statements of positive compliance with such a broad specification. Minimal tests however, are fully sufficient for decisions of non-compliance. If you think about it, that is what law enforcement is all about because we rarely have the resources to verify conformance. Law enforcement is not to confirm compliance, but rather to detect and properly react to non-compliance. As a result there is an unavoidable risk of stuff getting past us, both false positives and false negatives.

There is still a little more to G-S.5.4. The last clause deals with the repeated performance of steps and operations embraced in the testing procedure. These are Handbook 44's decision rules. The Handbook is requiring the device to perform within tolerance with the inclusion of the uncertainty in the testing procedure. Note that this includes uncertainties in both the field standard and the testing method. The evidence is clear that HB44 increases the tolerance when guard banding as most law enforcement does. In several codes (e.g. Codes 3.34., 3.38., and 3.39) there is even more expansion of the tolerances when using "transfer standards." However, there is no evidence of any guard band within the basic HB44 tolerance structure.



Now consider how the combination of paragraphs B and G-S.5.4. impacts the compliance decision. The figure to right is similar to the one for the 5 gallon test measure. However, you'll notice that there are no incorrect decisions! The entire figure is green with no red areas. Pursuant to Introduction paragraph B, things that pass the tests (including false positives) are considered compliant without prejudice. We pass them and allow them to continue in service since we have no conclusive evidence of non-compliance on which to take action. Pursuant to G-S.5.4., those devices found out of tolerance in our tests (regardless of the test method uncertainty) are legally non-compliant and we have grounds to remove them from service.

Also notice that the proportion of false positives and false negatives depicted are roughly the same. This is often referred to as balanced risk and an important feature of HB 44 and OIML (See OIML D-20). What we will want to know going forward is how these proportions of false positives to false negatives change as we change test standards and test methods. This is a branch of statistics, called Bayesian Statistics, seeks to embrace uncertainty but at the same time use new evidence (our inspections) to provide confidence in our decisions. Handbook 44 addresses uncertainty in the field standard but does not address the uncertainty of the test method per G-S.5.4. It is this void that prompts many efforts to avoid risk. Before evaluating that risk, let's look more closely at the Handbook's 1/3 rule.

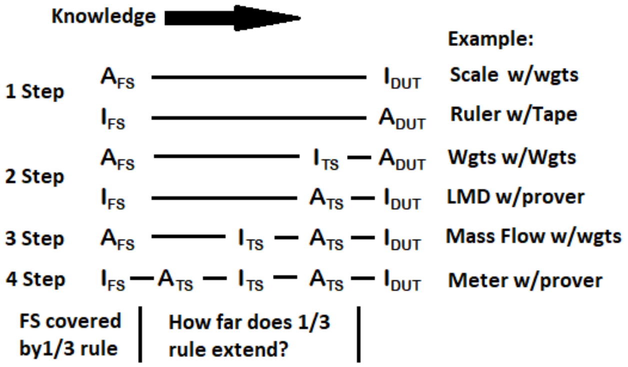
In Fundamental Considerations paragraph 3.2. we find the requirement that test apparatus must meet the 1/3 rule. When used without correction, as most field standards are, this means the error of the standard and its uncertainty combined must be less than 1/3 of the applied tolerance. We don't want to have two sets of standards in the field so they are specified to the tightest tolerance, i.e. acceptance tolerance. Thus, the rule becomes a 1/6 rule relative to most maintenance tolerances. We can only view the entire population in reference to maintenance tolerance. However, the NIST 105 series specifications may go beyond the 1/3 requirement. For example, the 5 gallon test measure and most Class F test weights are actually 1/5 of AT and 1/10 of MT. That's considerably tighter.

I propose to look at the tests in HB44 in broad overview. The simplest test involves one step, and a good example is the use of a field standard weight to test a scale. Note that the device under test (DUT) scale is an instrument and the field standard (FS) weight is an artifact. Actually it is the scale that measures the weight, as it is always the instrument measuring the artifact, i.e. $A_{FS} – I_{DUT}$. We can also have a one step comparison using a measuring tape instrument to measure an artifact yardstick, i.e. $I_{DUT} – A_{FS}$.

It would be great if we could verify all commercial devices using this type of one-step direct comparison, but that's not realistic. Many tests require multiple steps. For example, in two steps we can compare an artifact (FS) weight to an artifact (DUT) commercial weight, i.e. $A_{FS} – I_{TS} – A_{DUT}$. The middle step involves a transfer standard (TS) instrument as a comparator. In the inverse procedure, we can use two steps to compare an instrument (FS) test measure to an instrument (DUT) gas pump, i.e. $I_{FS} – A_{TS} – I_{DUT}$.

There are also three step and four step procedures in the mix. We test mass flow meters and belt scales by a three step procedure. In the middle we find two transfer standards, one an instrument and one an artifact, i.e. $A_{FS} – I_{TS} – A_{TS} – I_{DUT}$. For the mass flow test the first is a reference scale and the second is the liquid of the volume transfer. We test some meters using a four step procedure. In this case there are three transfer standards, i.e. I_{FS} – A_{TS} – I_{TS} – A_{TS} – I_{DUT}. For the master meter test, the instrument in the middle is the master meter and the artifacts on either side are the liquids of the volume transfers.

In the figure at right we can now evaluate the meaning of the 1/3 rule in practice. Clearly we understand that the reference standard, the field standard must meet the requirement, but how far beyond that does the rule extend?


Knowledge
Example:
1 Step
A_{FS} I_{DUT} Scale w/wgts
I_{FS} A_{DUT} Ruler w/Tape
2 Step
A_{FS} I_{TS} — A_{DUT} Wgts w/Wgts
I_{FS} A_{TS} — I_{DUT} LMD w/prover
3 Step
A_{FS} I_{TS} — A_{TS} — I_{DUT} Mass Flow w/wgts
4 Step
I_{FS} — A_{TS} — I_{TS} — A_{TS} — I_{DUT} Meter w/prover
FS covered by1/3 rule
How far does 1/3 rule extend?

Some proponents seem to think the rule applies to all the transfer standards, at least the instruments. This position is difficult to support when you see that Handbook 44 increases tolerances in several codes when using transfer standards. If the transfer standard had to meet the 1/3 requirement, why would you increase the tolerance when using it? That notion that transfer standards have to meet the 1/3 requirement also seems like revisionist history. Handbook 44 has never been interpreted that way in the almost 70 years it's been in use. I think this reaction is just uninformed risk aversion. Paragraph G-S.5.4. always included the measurement uncertainty in the tolerance.

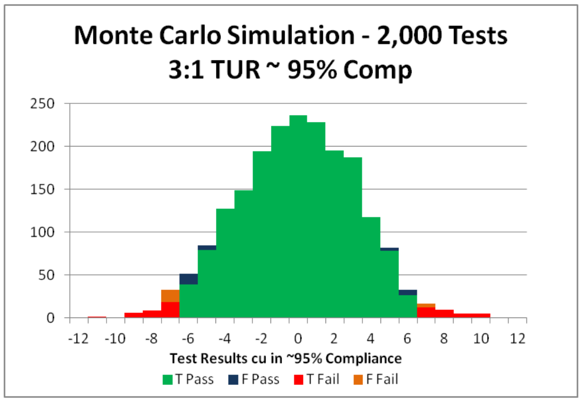
Next, to apply the 1/3 rule to any of the transfer standards in this diagram, you have to have uncertainty estimates. Where are the control charts and proficiency tests to produce these estimates? Other comments on these items also address this question. Finally, try to imagine what kind of check standard you would use for each control chart that cannot be a DUT. The variability of the DUT has to be excluded if you are to allow the DUT the full use of its tolerance. Finally think about that in terms of what kind of workload would be required for each field inspector to collect and maintain that information for every single transfer standard they use! Just think in terms of reference scales. Each one is unique and would require a separate control chart. Wouldn't those pushing to apply the 1/3 rule to transfer standards have to justify the benefits vs the costs?

I'll offer a simpler take on the issue that I believe follows Handbook 44, the way it was written and the way it should be interpreted. I contend that all of the transfer standards in the diagram above are part of the test procedure and Handbook 44 sets no restriction on the accuracy of the test method. Go back to the GUM formula discussed earlier and consider the s_b term. This is the standard uncertainty of the test procedure. In normal situations this includes elements of variability of the standard, the test method, and the DUT. For example, the standard is being used in a different environment than in the lab and the DUT also contributes random variation and smaller effects of influences from the test environment. This is why it is virtually impossible to separate out the test procedure uncertainty from the DUT. I think Handbook 44 never intended for us to have to do that.

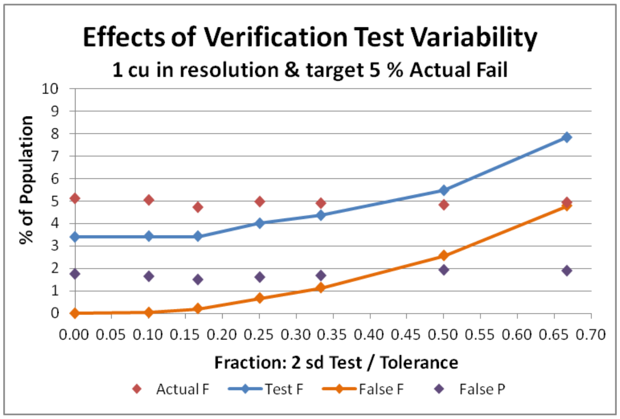
What we have to relearn is the wisdom that the creators of Handbook 44 built into our specification. I believe they not only talked to statisticians, but that they listened. They understood that the risks that people are trying to avoid are not that severe. I'd like to bring that wisdom to light, and I want to do it with what is called Monte Carlo simulation. If we look at compliance data of the population of devices we like to see a nice bell shaped, or normal curve, centered at zero error and with a rejection rate of about 5% of the population. That curve includes both the false positives and the false negatives. Our problem is that testing does not help us identify these false outcomes, and as I've suggested, we want to know more about them.

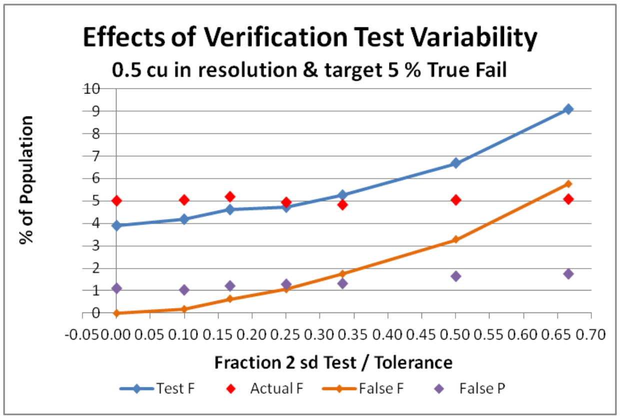
What Monte Carlo does is use simulations to give us real data to analyze. We use our random simulator to pick a device from its population, pick a test from its population, and round them to the resolution of the test. For example, the device we pick is +4.792348cu in and the test is 1.5296402 cu in. Combined and rounded we get a +6 cu in result although the result is almost +6.5 cu in.

In the simulation to right, I used a 5% rejection rate for the device and a variability in the test of 1/3 the maintenance tolerance. For a gas pump that means a 2 sigma variability of the test of +/-2 cu in. I used some knowledge of the system to estimate this situation in testing gas pumps. At the tolerance boundary we see a small but almost equal number of false positives and false negatives. Remember that the orange false negatives are legally non-compliant and the blue false positives are without prejudice.



The next phase was to look at how changes to the test variability impacted the outcome. In these simulations I increased the tests to 20,000 and plotted the rates for device failures, test failures, false positives and false negatives over a range of test variability. I also ran a simulation with the same device population but a 0.5 cu in resolution for the test measure. We can separate the false positives and false negatives since we have the actual device result from the random section of the device. I started with test variability of 0 cu in and went to 2/3 the tolerance or 2 sigma variability of 4 cu in. The plots appear below.

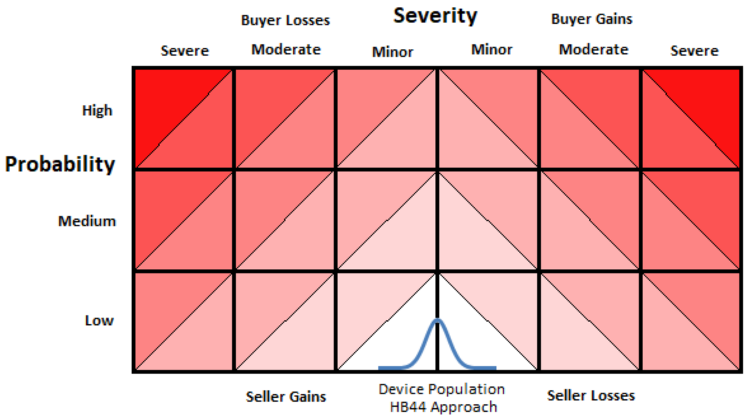




The original simulation at the top of the page is essentially at 0.33 on the X axis in these plots. The other test points you see reflect variability of 0, 1/10, 1/6, 1/4, 1/3, 1/2, and 2/3 of the maintenance tolerance. The first thing to notice is the device failure rate before the test hovers right at 5%, as predicted. Next, take note of the false positives. They are almost flat, reflecting only very small changes as the test becomes more variable. Most important is the false negatives. As the test variability increases this is where the impact is felt. The blue line showing the test failures tends to track in parallel with the false negatives. These are penalizing device owners for the fault of the test. This is why Handbook 44 guard bands by increasing tolerances when some transfer standards are used.

The critical area to look at is the range along the X axis between about 0.2 and 0.4 where there is less than 1% change in failure rate for a doubling of the test variability. What I believe the authors of Handbook 44 understood is that this is where most of our test methods fall and thus have little impact on the failure rate. Remember that the variability associated with the test measure verification is typically less than 1/10 the maintenance tolerance and you must combine that with the test variability as the root-sum-square. For example, if the total variability is 1/3 (0.33) and the field standard variability is 1/10 (0.1), then the variability of the test method is 0.31. Thus the test method contributes to the total uncertainty by a 9:1 ratio, or saying it another way, 90% of the total variability is due to the test method vs only 10% from the standard. Even with that large contribution, the balanced risk (false positives vs false negatives) keeps the failure rate consistent with the devices exclusive of the test.

There are a few additional things to consider. Risk is not just a matter of probabilities. Risk is the juxtaposition of probability and severity. I created a 3 x 3 risk matrix a little different from the norm. Normally you assign risks in the nine boxes created. Highest risks occur where you have high probability and high severity. I chose to draw the risk boundaries on the diagonal. In addition, we protect both buyer and seller. One's risk of loss is the others risk of gain, and vice versa. The risk matrix appears below.



For reference consider that the probability scale (Y axis) is linear with 0% at the bottom and 100% at the top. The X axis from the center line changes exponentially. The ranges from 0% to ~2% minor, ~2% to ~200% moderate, and ~200% to ~5,000% severe. For example, consider a $30 gasoline purchase. It's minor if the measurement is off $0.30 (1%). It's moderate if you get water in your gas your loss could be $800 (26%) for repairs. It's severe if you get severely burned from a device fault and incur medical and lost wages approaching $100,000 (3,333%). It's the severe risks in the darker reds that we need to sweat, not the minor ones in the white area.

My challenge was to place the device population appropriately in the risk matrix. I can't see any way the measurement risks get outside the lowest risk area at the bottom center. Think about that in terms Handbook 44 uses in the Fundamental Considerations 2.2. Tolerances are chosen to prevent serious harm to either buyer or seller and to keep costs of equipment (and testing) reasonable. Serious harm does not begin immediately outside the tolerance limit. Handbook 44 was taking a more global approach with an understanding that 100% compliance was unattainable. Also if you think about the risks increasing diagonally from center bottom to top right and left, you see that the balanced risks are very fairly balanced among buyers and sellers. The risk of getting a measurement with an error of twice the tolerance is counterbalanced by the ever decreasing risk of encountering that measurement.

This risk matrix lets us see the real issue. It is NOT variability. The critical issue is test method bias. That's what prompted this entire enterprise. One new test method produced a result where the mean of the bell did not match that of the official's test. The company offering the service marketed that bias to prospective clients as an edge on competitors. For me this is a blatant violation of the charge in the UWML §11(e) to "assure equity among buyers and sellers." Just as in package checking, reasonable variations are to be recognize (and they need not all be the same). Most important, you have to come to the same mean value.

Conclusions:

1. The authors of Handbook 44 were well aware that test variability (when within reasonable limits) was not the crucial factor in enforcement testing. Handbook 44 kept the focus on the mean value.
2. Transfer standards are NOT field standards and never have been. The 1/3 rule applies solely to the reference standards (field standards) used to initiate the testing procedure. Handbook 44 placed no specific requirements on transfer standards as they are part of the test procedure.
3. Any attempt to extend the 1/3 rule to include transfer standards will have significant impact on productivity of our field staff as they now have to maintain an extensive library of control charts and participate in proficiency tests for every test method using transfer standards in field enforcement.
4. The Monte Carlo simulations show that there is virtually no return for investment in making improvements in test methods that are already effective. This even brings the lab practice of guard banding field standards into question as it probably has an insignificant impact on the commercial device population.
5. The Work Group on Alternative Standards should begin to develop evaluation protocols to compare test methods with a focus on ensuring alternatives give equivalent results based on the mean value, with cursory evaluation to prevent excessive variability in the test procedure. Remember that excess variability results in increased false negatives rather than false positives.
6. The NCWM should be considering the practice of other standards setting organizations like ASTM and AOAC, where an official test method is specified and alternatives are permitted. Alternatives must demonstrate corresponding results (i.e. to the mean) relative to the official method. Here it is important to note that the choice of an official method is, at least in part, arbitrary, as each test method may involve some bias. The work group on alternative test methods should consider establishing the means to evaluate the correspondence between test methods so that alternatives can also be listed with the official.

Additional sources of similar information:

K.-D. Sommer, M. Kochsiek, W. Shultz, Error Limits and Measurement Uncertainty in Legal Metrology, https://pdfs.semanticscholar.org/f5ec/95bc8d9faff4bc77d7ec7294f9013927ad4a.pdf

W. Schultz, K.-D. Sommer, Uncertainty of Measurement and Error Limits in Legal Metrology, OIML Bulletin XL (4) (Oct. 1999) p. 5-15

OIML D-20, Initial and Subsequent Verification of Measuring Instruments and Processes.