Let's Talk about Bias: Measurement Bias

What can happen when we use an accuracy specification and assume all the measurements are centered in relation to the specification limits? It is a typical problem in the metrology community, where many papers assume a centered process or Measurement.

When the Measurement deviates from the true value, it is said to have bias. More specifically, measurement bias refers to systematic errors in a measurement or measurement process that consistently cause the measured values to deviate from the true value of the quantity being measured.

Measurement bias can be caused by various factors, such as the design or calibration of the measurement equipment, the skill of the operator, or the conditions under which the Measurement is made. Measurement bias can lead to inaccurate or unreliable calibration and test results, affecting the quality and integrity of the data and leading to incorrect conformity assessments.

Making a conformity assessment might mean the measured value could be anywhere within the specification. In cases of simple acceptance, the measured value could even be at the tolerance limit.

The reason this matters is that when a known bias is ignored, meaning not corrected or not included in the Statement of Measurement Uncertainty on the Calibration Certificate, measurement traceability may not be fully achieved, and all subsequent measurements are suspect.

In this paper, we will discuss the importance of correcting for any bias in relation to the location of the Measurement to ensure metrological traceable measurements and adherence to ISO/IEC 17025:2017 requirements.

The location of the Measurement and Bias

Why do we care about the location of the Measurement if the device is within tolerance? If a device has a specification of 0.1 % of full scale and the calibrating laboratory reports a value within 0.1 %, the device is "Within Tolerance," In reality, it depends on all parties being in agreement per contractual requirements (contract review) on how measurement uncertainty is being taken into account via an acceptable and agreed-upon decision rule.

It also depends on the uncertainty of the Measurement and whether the lab performing the calibration followed the proper calculations in evaluating the Uncertainty of Measurement (UOM) when making a statement of conformity.

Nominal Value	10000.0
Lower specification Limit	9990.0
Upper Specification Limit	10010.0
Measured Value	10009.0
Measurement Error	9.0
Std. Uncert. (k=1)	0.085
Total Risk	0.00%
Upper Limit Risk	0.000%
Lower Limit Risk	0.000%
TUR =	58.78943644
Cpk=	5.999032319
TAR=	62.5
Simple Guard Band (Subtract Uncertainty)	
Guard Band LSL	9990.170
Guard Band USL	10009.8299
Percent of Spec	98.30%
Guard Band Limits for Risk of	2.500%
Guard Band LSL	9990.167
Guard Band USL	10009.833
Percent of Spec	98.33%

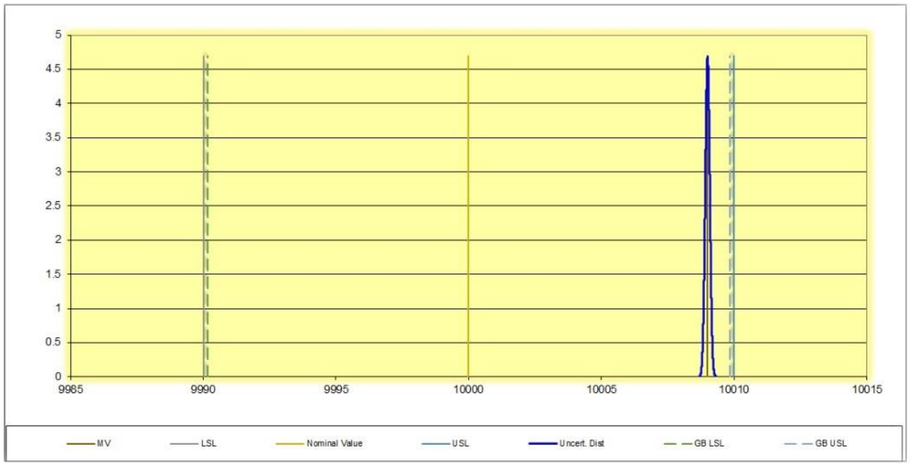

Figure 1: Graph Showing 10 009.0 as the measured value with a 58.789:1 TUR, which is achieved by using a lab with low uncertainties (Morehouse actual example)

Making a conformity assessment of "In Tolerance" is all about location, location, location of the Measurement. It's also about the Uncertainty of the Measurement because anything other than a nominal measurement will significantly raise the risk associated with the Probability of False Accept (PFA).

The probability of a false accept is the likelihood of a lab calling a measurement "In Tolerance" when it is not. PFA is also commonly referred to as ***consumers risk (β: Type II Error)***.

The measurement location we are referring to is how close the Measurement is to the nominal value. If the nominal value is 10 000.0 N and the instrument reads 10 009.0 N, the instrument bias is 9.0 N, as shown in **Figure 1**. The bias is 0.09 % of the measured value or 90 % of the overall tolerance.

The higher the measurement bias from the nominal, the higher the Measurement Uncertainty of subsequent measurements unless the measurement bias is corrected. In **Figure 1**, if the unit under test becomes the reference standard, and the measurement bias is not corrected, future measurements made with this Reference Standard will introduce additional Measurement Risk that is not accounted for in the reported Measurement Uncertainty.

Note: NIST SOP 29 has additional information on bias and gives further examples of how to account for any measurement bias in an uncertainty budget

Nominal Value	10000.0
Lower specification Limit	9990.0
Upper Specification Limit	10010.0
Measured Value	10000.0
Measurement Error	0.0
Std. Uncert. (k=1)	0.501
Total Risk	0.00%
Upper Limit Risk	0.000%
Lower Limit Risk	0.000%
TUR =	9.983373912
Cpk=	10.18730343
TAR=	10
Simple Guard Band (Subtract Uncertainty)	
Guard Band LSL	9991.002
Guard Band USL	10008.9983
Percent of Spec	89.98%
Guard Band Limits for Risk of	2.500%
Guard Band LSL	9990.982
Guard Band USL	10009.018
Percent of Spec	90.18%

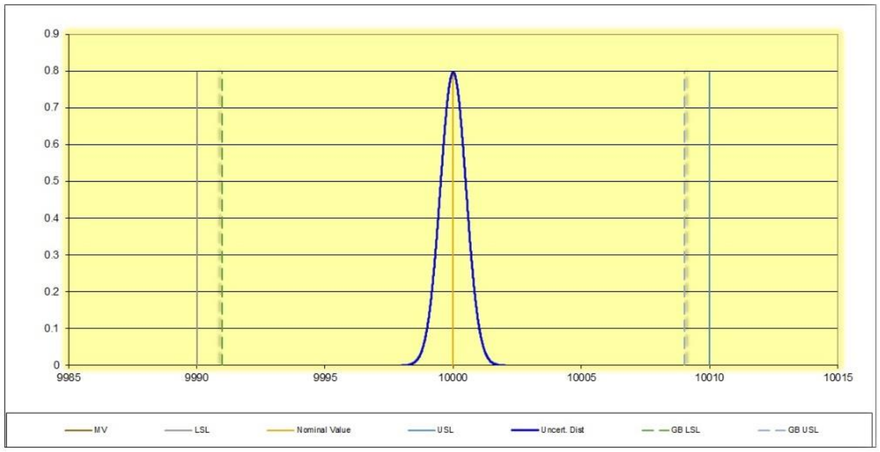

Figure 2: Graph Showing 10 000.0 as the measured value with a 9.98:1 TUR and a Centered Measurement

Introduction to Statistics in Metrology addresses bias (measurement bias) in section 5.2 by stating, *"There are important assumptions associated with using TUR as a metric and the requirement of a TUR of 4 or 10.* ***Using a TUR assumes that all measurement biases have been removed from the measurement process and the measurements involved follow a normal distribution. If there are significant biases that cannot be removed, the TUR will not account for the increased risk.****"* [1]

When the process distribution is centered between the specification limits and does not overstate or understate the nominal value of the Measurement, ***higher TURs*** produce wider acceptance limits. In comparison, ***lower TURs***, such as 1:1, will reduce acceptance limits.

When the measurement bias is corrected, these limits can easily be calculated as a percentage of the specification when the Measurement Uncertainty is known. Acceptance Limits (with the appropriate guard band) based on decision rule applied are covered in detail later. The Metrology Handbook, 3rd edition, Chapter 30 covers the topic on Decision Rules. [2]

When the reference standard measurement value is centered (nominal value), the calibration laboratory can still say the device being tested is within tolerance. A laboratory's scope of accreditation indicates its best capability to call an instrument in tolerance when any measurement bias is observed in the measurand (quality being measured).

Note: The scope of accreditation does not take into account the measurement uncertainty contribution of the equipment submitted for calibration. The laboratory's scope of accreditation only includes the contribution *from the best existing device to be calibrated and may not be what is used for the customer's device submitted for calibration.*

In **Figure 2**, the measured value is centered (nominal value). With the measured value at the nominal, assuming a PFA of 2.5% (based on the decision rule employed), the measurement result is considered to be in conformance ("Pass") as long as it is within the acceptance limits. Please note that the acceptance limits are calculated, taking measurement uncertainty into account and implementing the appropriate decision rule.

What happens when we switch calibration providers?

What if we switched calibration providers, for whatever reason, to someone with a higher calibration and measurement capability uncertainty parameter?

Switching calibration providers may make sense for several reasons. However, if one does not understand the relationship between measurement uncertainty, decision rules, and acceptance limits, shopping on price alone might mean more failed measurements.

More failed Measurements often result in an overall higher cost and increased risk to companies and their customers. These decisions should not be made without properly evaluating the supplier's capabilities and reputation. The recommendation for overall risk reduction is to use accredited calibration suppliers with low uncertainties appropriate to the risk tolerance.

An example of understanding overall risk happened when Morehouse had a customer that would send in two bolt testers periodically for calibration. One was always centered (low to no bias), similar to **Figure 2**. The other slowly approached the acceptance limits (not specification limits), similar to **Figure 1,** showing a high bias.

Eventually, the bolt tester with the non-centered measurements (high bias) failed calibration. Morehouse was informed that this out-of-tolerance Measurement resulted in a one-million-dollar plus recall. The Measurement was approaching the acceptance limit and eventually failed. If the customer had corrected for the bias, the one-million-dollar plus recall would have been avoided. It is important to note that selecting a provider with a larger uncertainty would have resulted in no hope or potential to correct the problem before it worsened, resulting in increased risk.

Low-Risk Scenario

Nominal Value	10000.0
Lower specification Limit	9990.0
Upper Specification Limit	10010.0
Measured Value	10009.0
Measurement Error	9.0
Std. Uncert. (k=1)	0.085
Total Risk	0.00%
Upper Limit Risk	0.000%
Lower Limit Risk	0.000%
TUR =	58.78943644
Cpk=	5.999032319
TAR=	62.5
Simple Guard Band (Subtract Uncertainty)	
Guard Band LSL	9990.170
Guard Band USL	10009.8299
Percent of Spec	98.30%
Guard Band Limits for Risk of	2.500%
Guard Band LSL	9990.167
Guard Band USL	10009.833
Percent of Spec	98.33%

High-Risk Scenario

Nominal Value	10000.0
Lower specification Limit	9990.0
Upper Specification Limit	10010.0
Measured Value	10009.0
Measurement Error	9.0
Std. Uncert. (k=1)	1.250
Total Risk	21.19%
Upper Limit Risk	21.192%
Lower Limit Risk	0.000%
TUR =	3.998933697
Cpk=	0.408061957
TAR=	4
Simple Guard Band (Subtract Uncertainty)	
Guard Band LSL	9992.501
Guard Band USL	10007.4993
Percent of Spec	74.99%
Guard Band Limits for Risk of	2.500%
Guard Band LSL	9992.451
Guard Band USL	10007.549
Percent of Spec	75.49%

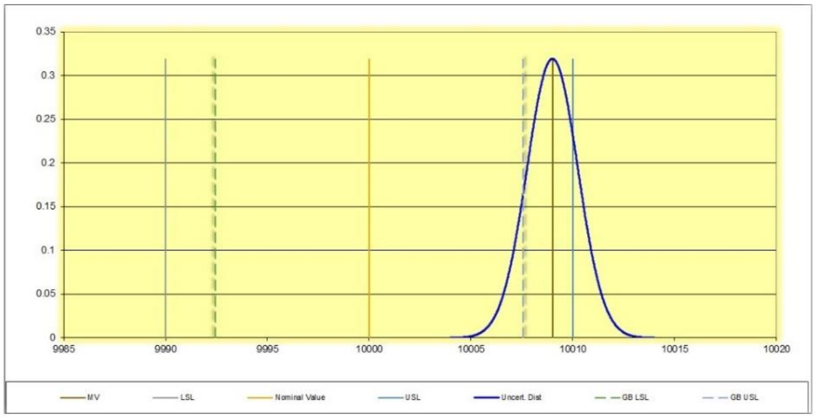

Figure 3: Graph Showing 10 009.0 as the measured value comparing two different TUR values

Figure 3 shows the comparison between two different suppliers, resulting in different TUR values. The bottom graph shows the higher risk level using a different supplier. The new provider has a higher Measurement Uncertainty of 0.025 % than shown in **Figure 1**, where the calibration provider had a 0.0016 % Measurement Uncertainty. Everything else has remained the same. However, the overall measurement risk is now 21.19 %.

The assumption is that the measurement bias is known (+ 9 N). Although the risk is 21.19 %, the bias can usually be corrected (adjusting the measuring system) or incorporated in a measurement model as a correction. Using the high-risk scenario, we will discuss what happens when bias is not corrected.

What happens when we do not correct the bias?

Let us look at the **high-risk** scenario in **Figure 3**. When 10 000.0 N of force ± 2.50 N was applied, the measured value was 10 009.0 N.

The right thing for the end-user to do is to load the device to 10 009.0 N to apply 10 000.0 N of force. Let us assume they do not do that and use this device to calibrate another 10,000 N instrument.

If we look at the minimum Measurement Uncertainty for the device that read 10 009.0, assuming the bias is corrected, the Measurement Uncertainty would have to be greater than that of the Measurement Uncertainty used for the calibration of the device, which was ± 2.50 N.

The Measurement Uncertainty for this device would be ±2.5 N plus additional Measurement Uncertainty contributors for repeatability, reproducibility, resolution, environmental, stability between calibrations, and other error sources. Likely our measurement uncertainty assuming stability of 0.02 % as the second highest contributor would become around 5.178 N. For more information on how to calculate Measurement Uncertainty for Force, see A2LA Guidance Document G126 [3]

Nominal Value	**10000.0**
Lower specification Limit	9990.0
Upper Specification Limit	10010.0
Measured Value	**9996.0**
Measurement Error	-4.0
Std. Uncert. (k=1)	**2.589**
Total Risk	**1.02%**
Upper Limit Risk	0.000%
Lower Limit Risk	1.024%
TUR =	**1.931223436**
Cpk=	**1.182403422**
TAR=	**3.99840064**
Simple Guard Band (Subtract Uncertainty)	
Guard Band LSL	**9995.178**
Guard Band USL	**10004.8219**
Percent of Spec	**48.22%**
Guard Band Limits for Risk of	**2.500%**
Guard Band LSL	**9995.074**
Guard Band USL	**10004.926**
Percent of Spec	**49.26%**

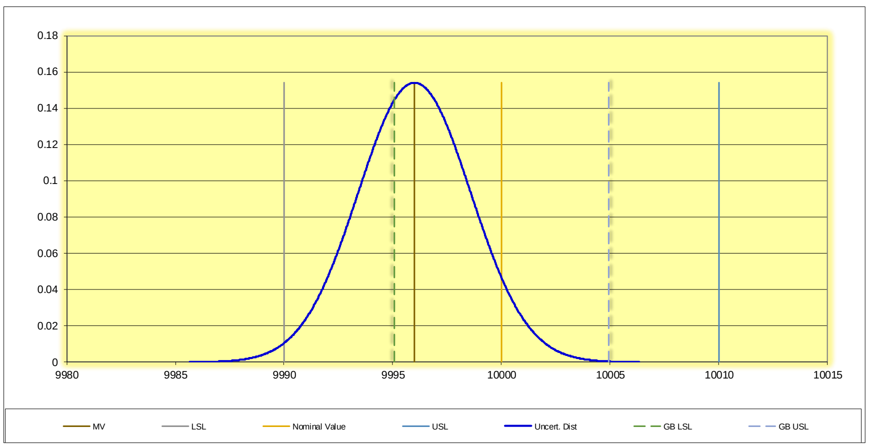

Figure 4: Graph Showing 9 996.0 as the measured value with a 1.93:1 TUR

Scenario 1: Bias is corrected by loading the reference standard to 10 009.0 N to apply 10 000.0 N.

Figure 4 above shows a subsequent measurement being made with the calibrated device that read 10 009.0 N when 10 000.0 N ± 2.5 N was applied. This device is now used as a reference standard to calibrate other devices (UUT).

The graph represents correcting the reference standard for the + 9 N bias and using it to calibrate another device (UUT). The measured value of the Unit Under Test reads 9,996 N.

The reference standard is being loaded to 10 009.0 N to apply 10 000.0 N ± 5.178 N. The UUT reads 9996.0 with a Total Risk of 1.02 %.

Nominal Value	**10000.0**
Lower specification Limit	9990.0
Upper Specification Limit	10010.0
Measured Value	**9987.0**
Measurement Error	-13.0
Std. Uncert. (k=1)	**2.589**
Total Risk	**87.67%**
Upper Limit Risk	0.000%
Lower Limit Risk	87.672%
TUR =	**1.931223436**
Cpk=	**-0.59120171**
TAR=	**3.99840064**
Simple Guard Band (Subtract Uncertainty)	
Guard Band LSL	**9995.178**
Guard Band USL	**10004.8219**
Percent of Spec	**48.22%**
Guard Band Limits for Risk of	**2.500%**
Guard Band LSL	**9995.074**
Guard Band USL	**10004.926**
Percent of Spec	**49.26%**

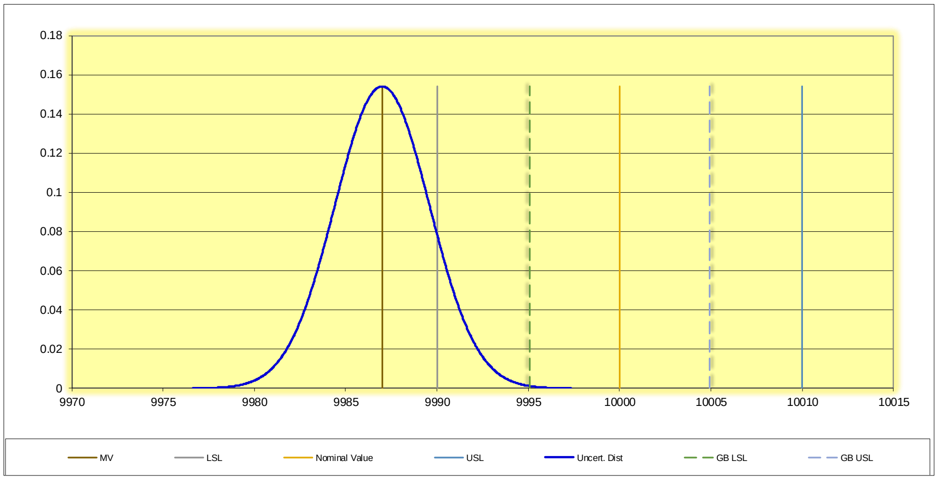

Figure 5: Graph Showing what happens if we do not correct for the + 9 N bias (1.93:1 TUR, which stays the same)

Scenario 2: The reference standard is not loaded to 10 009.0 N to apply 10 000.0 N. Instead, the device is loaded to 10 000.0 N, which means only 9 991.0 N is applied (10 000.0 – 9.0 = 9 991.0)

We show not correcting for this +9 N bias graphically by subtracting 9 N (9 996.0 – 9.0 = 9 987.0) from the measured value. The UUT reads 9 987.0 N, which could result in the lab failing the instrument and deciding to adjust the device within the acceptance limits (the measured value of this calibration is now off by 9 N and transferred to the UUT).

The result of not correcting for the +9 N bias is a failed instrument that has been adjusted using a reference standard with a high bias and a measurement risk above 87 %.

Global risk and bias

Global consumer's risk is defined in JCGM 106:2012[4]. The role of CPU in conformity assessment is defined as "the probability that a non-conforming item will be accepted based on a future measurement result." [4]

The acronym CPU is Calibration Process Uncertainty, which is used in the calculation of risk, and a requirement for any ISO/IEC 17025 accredited calibration provider to take into account when making a conformity assessment using a decision rule. Section 3.7 of the ISO/IEC 17025 defines a decision rule as a "rule that describes how measurement uncertainty is accounted for when stating conformity with a specified requirement." [5]

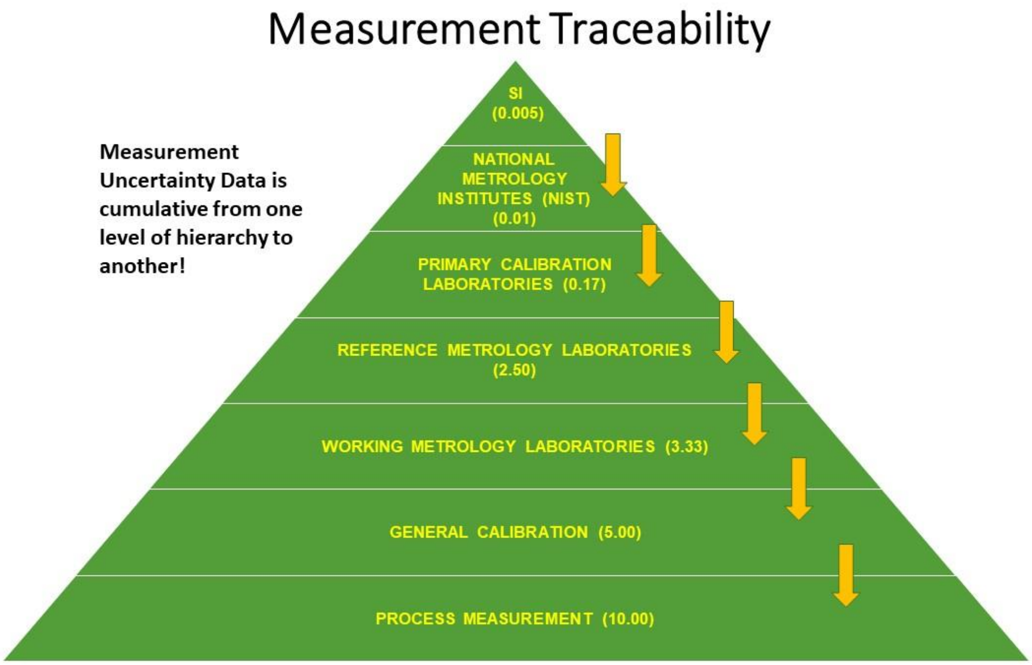

Figure 6: Measurement Traceability Pyramid Used with Measurement Uncertainty

Suppose we follow this logic further, following the progression from the initial calibration at the Primary level through the pyramid, correcting for bias and not correcting at each step. In that case, we can generate random variations due to the measurement uncertainty at each level.

When we do not correct for bias, measurement uncertainty might be underestimated, and therefore may not align with the definition of metrologically traceability, undermining measurement confidence.

Metrological traceability is defined in JCGM 200:2012 as "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." [6]

If we do not correct for the bias and fail to calculate the impact of the bias to our measurement uncertainty, we no longer have metrological traceability as required in ISO/IEC 17025 section 6.5.[5]

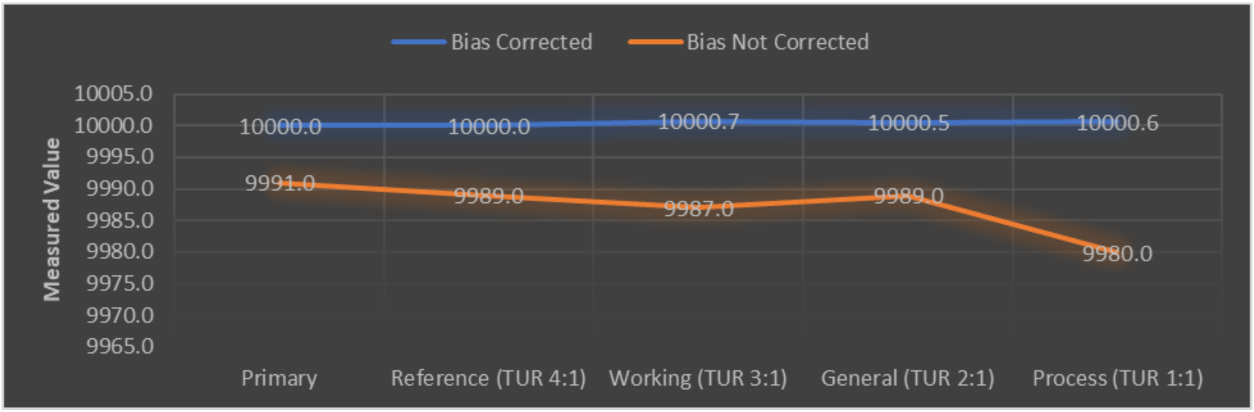

Figure 7: Randomly Generated Differences in correcting for Bias and Not Correcting (Reference through Process Tiers in Figure 7)

Figure 7 above shows what could happen when the reference laboratory does **not correct** for bias and applies 9 991.0 N (10,000.0 – 9.0) versus what could happen when **Bias is Corrected**.

Bias Not Corrected Measured values are generated using Upper and Lower Specification Limits that are modified by the 9.0 N bias taking into the Measurement Uncertainty at each tier.

Remember: When 10 000.0 N was applied, the device read 10,009 N. When the laboratory only loads the device to 10 000.0 N, 9 991.0 is the actual force applied.

In this scenario, **not correcting for bias** can result in making an **incorrect conformity statement** when stating conformity to the Tolerance/specification limit (e.g., pass/fail, in-tolerance/out-of-tolerance)

When measurement bias is not corrected, a conformity statement of "Fail" might result in the calibration laboratory adjusting an instrument that should have passed calibration to the wrong nominal value.

If we continue to generate data randomly with and without measurement bias corrected, we might end up with the table and risk scenarios in the graphs below.

	Measurement Uncertainty k = 2	BIAS Measured Value With Bias	BIAS CORRECTED Measured Value Bias Removed
Primary	0.17	9991.0	10000.0
Reference (TUR 4:1)	2.5	9989.0	10000.0
Working (TUR 3:1)	3.3	9987.0	10000.7
General (TUR 2:1)	5	9989.0	10000.5
Process (TUR 1:1)	10	9980.0	10000.6

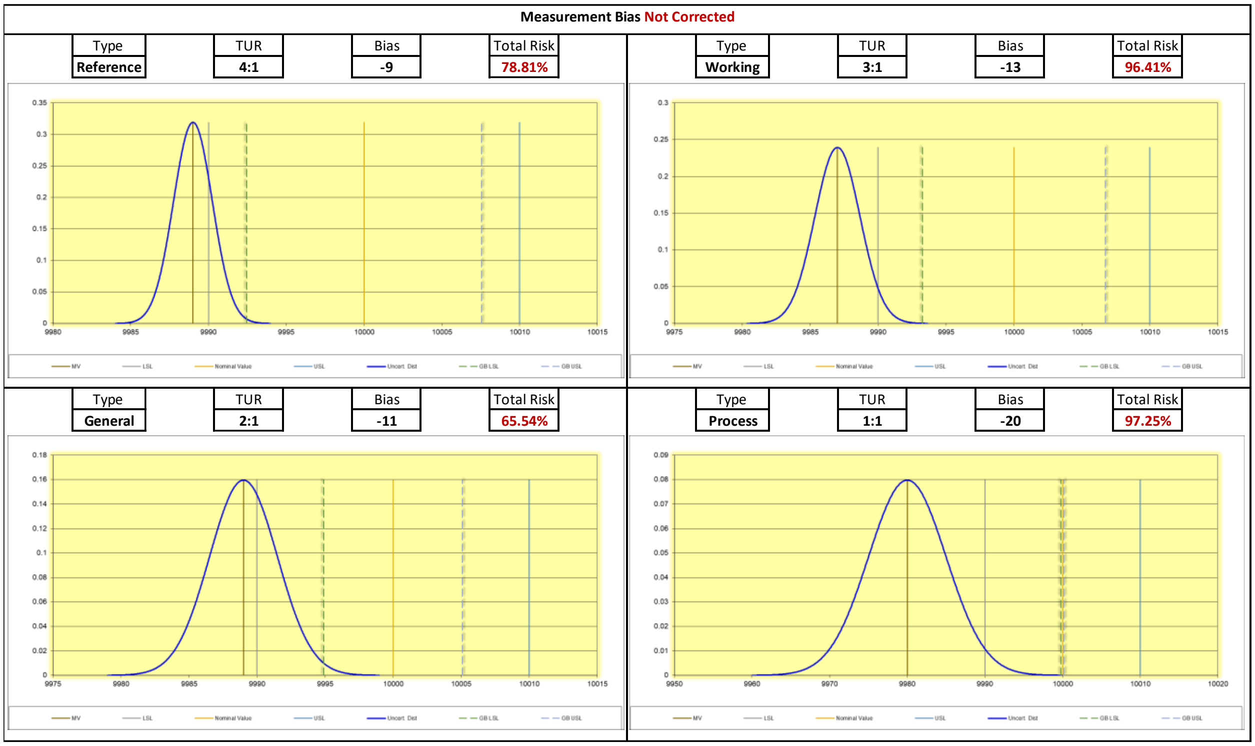

Figure 8: Randomly Generated Differences in not correcting for bias total risk graph

By the time we get to the process measurement, the device might have a bias of -20 N from nominal. In our simulation using the measurement uncertainty at each tier, a starting measured value of 9 991.0, and randomly generating numbers within the tolerance of 0.1 %, we prove that not correcting for bias raises the total risk at each measurement tier.

When the bias is not corrected, the starting measured value is 9 991.0; the difference becomes 9 N or 9/10th of the specification limits of $\pm$ 10 we are trying to maintain throughout the process with our TUR ratios. (For these graphs, bias is the difference from the nominal value, measured value minus the nominal value)

If Primary Standards calibrate the Reference with a 58.79:1 TUR (Shown in **Figure 1**), the total risk is 0.0 %. When the next level uses this Reference, if they correct for bias, the risk with a 4:1 TUR is 0.0 %, as shown in **Figure 9**. If they do not correct for the bias, as shown in **Figure 8**, the risk is 78.81 %. Randomly generating numbers and not correcting for bias at a 2:1 TUR, the total risk becomes 65.54 %, compared to 0.0 % when bias is corrected in **Figure 9**.

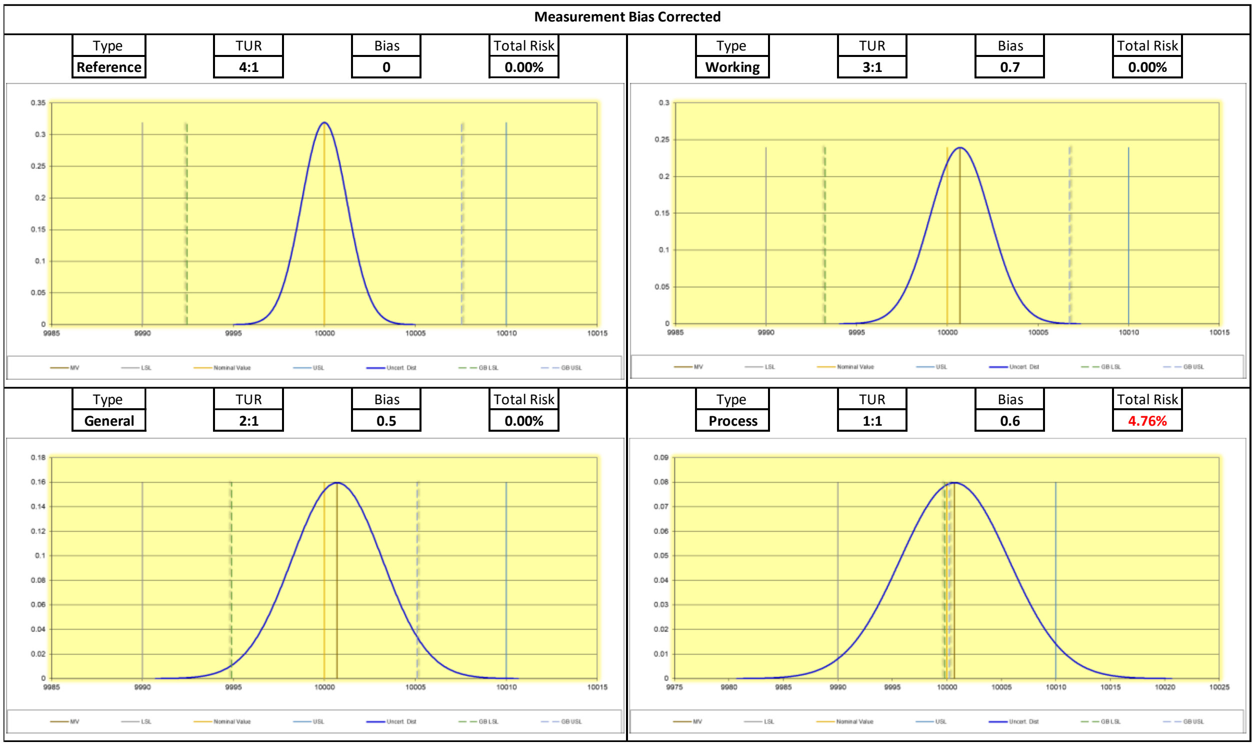

Figure 9: Randomly Generated Differences in correcting for Bias total risk graph

Figure 9 shows randomly generated numbers assuming each tier from the Reference tier to the General calibration tier is correcting for bias. In each scenario, the measurement risk is drastically different.

The larger the measurement uncertainty becomes, the greater the measurement risk. When the bias is corrected, the total risk should follow the percentage of specification in **Figure 10**. Meaning at a 4:1 TUR, if the measured value falls between 9 992.45 and 10 007.55, the total risk will be less than 5 %.

Table for 95 % Confidence Interval 5 % Total Risk						
Std Unc	Std Unc %	TUR	Percent of Spec	In Engineering Units ±	GB LSL	GB USL
0.085 048	0.001%	58.79	98.33%	9.833	9990.167	10009.833
1.250 000	0.013%	4.00	75.50%	7.550	9992.450	10007.550
1.501 502	0.015%	3.33	70.57%	7.057	9992.943	10007.057
2.500 000	0.025%	2.00	51.00%	5.100	9994.900	10005.100
5.000 000	0.050%	1.00	2.00%	0.200	9999.800	10000.200

Figure 10: Guard Banded Acceptance Limiting Risk to Total Risk 5 %

Earlier, we mentioned how knowing the TUR makes it easy to calculate acceptance limits. In **Figure 10**, we use the ILAC G8 guard banding method that allows for a maximum of 5 % total risk. [7]

The assumption here is that the Measurement is centered. At our 58.79:1 TUR we achieved at Morehouse, we know our deadweight machine's uncertainty is ± 0.0016 %, including any bias added to the uncertainty in the calibration by NIST on our weights.

The TUR formulas work for us; however, if the end-user has not considered (and corrected) for the contribution of the effect of bias in their evaluation of Measurement Uncertainty, the acceptance criteria may be skewed. Skewed acceptance criteria can increase measurement risk at all tiers, starting from the first tier of the pyramid that did not correct for bias correctly.

Conclusion

Using the manufacturer's accuracy specification and not correcting for bias can further increase Measurement Risk. Morehouse did the sampling by varying the TUR and using randomly generated values after the initial calibration by correcting for bias and then by not correcting for bias, which showed a significant difference in Measurement Risk.

Not correcting for bias seems to be a problem many in the calibration deal with, and their unsuspecting customers are likely getting calibrations that carry too much overall Measurement Risk.

The risk of not correcting for this offset (Bias) should concern anyone making measurements.

Furthermore, the habit of insisting on a 4:1 TUR, shown in **Figure 10**, only works if the measurement process is centered (measurement bias is corrected).

In all cases, paying attention to the location of the Measurement and calculating Measurement Risk is imperative to making accurate measurements.

Anyone wanting more accurate measurements (measurements with less Measurement Uncertainty) should have a defined process to account for and correct bias. They should also examine their calibration providers' practices on how they handle and correct their measurement biases.

Figure 11: Morehouse 4215 Plus that Uses Coefficients to Reduce Bias

Morehouse has many options with our force calibrations systems that use coefficients generated at the time of calibration. Our 4215 plus and C705P use coefficients that are programmed into the indicator to help correct and minimize measurement bias.

The reason this is important is JCGM 106 references that when a measuring system is used in conformity assessment that, the measuring system has been corrected for all recognized significant systematic errors (Bias) [8]

When bias is not corrected, the risk of making a measurement that does not properly account for bias can result in an underestimation of measurement uncertainty and therefore disagrees with the metrologically traceability definition and undermines measurement confidence

References

1. Introduction to Statistics in Metrology Section 5.2
2. The Metrology Handbook, 3rd edition, Chapter 30
3. A2LA G126 Guidance on Uncertainty Budgets for Force Measuring Devices
4. JCGM 106:2012_E clause 3.3.15 "Evaluation of measurement data – The role of measurement uncertainty in conformity assessment."
5. ISO/IEC 17025:2017 "General requirements for the competence of testing and calibration laboratories."
6. JCGM 200:2012 "International vocabulary of metrology – Basic and general concepts and associated terms (VIM) 3rd edition
7. ILAC-G8:09/2019 "Guidelines on Decision Rules and Statements of Conformity
8. JCGM 106:2012_E clause A.4.3.3 "Evaluation of measurement data – The role of measurement uncertainty in conformity assessment."

Additional Information Summary and Notes

For **Figures 7 – 9,** random numbers were generated.

The Excel function to generate numbers randomly is NORM.INV(RAND(),Measured Value, Measurement Uncertainty at $k = 1$)

After the initial calibration, the measured values after Tier 1 (Primary) were generated randomly using =RANDBETWEEN(USL, LSL). Both the USL (10 010.0 and LSL 9 990.0) were adjusted to the 9 N difference at tier 2 (Reference) and then, at each tier, adjusted using the measured value from the previous tier. Thus tier 2 used USL of 10 001.0 and LSL of 9 981.0)

The term bias in this paper is the nominal value minus the measured value. It is a known systematic error where a correction can be applied to compensate for this error, such as adding $\pm$ 9 N to the 10 000.0 N measurement to generate 10 000.0 N.

Of course, no correction would need to occur if the device is loaded to 10 009.0 N to apply 10 000.0 N as that is the value needed to generate 10 000.0 N in our example.

The associated Measurement Uncertainty of the system, including resolution, repeatability, reproducibility, Reference standard uncertainty, reference standard stability, environmental factors, and other error sources would still need to be calculated.