

Managing Calibration Confidence in the Real World

David Deaver
Fluke Corporation
Everett, Washington

ABSTRACT:

Previous papers have investigated the risk of making false test decisions as a function of Test Uncertainty Ratio, Confidence Interval, and Guardbands for fairly idealized distributions. This paper continues that pursuit by assessing the impact less than ideal distributions have on the Consumer Risk and Producer Risk. Specifically, the increase in risk associated with systematic bias in the unit under test and the measurement standard is investigated.

INTRODUCTION:

Standards that govern the operations of a calibration or standards lab are written by experts in the calibration and the academic communities to control and improve the quality of the calibration process and to provide a means of audit or critique of that process. Of particular concern is the adequacy, maintenance, and traceability of standards which are used to calibrate subject units under test. One of the tests for adequacy is the extent to which the accuracy of the calibration standard (STD) exceeds that of the unit under test (UUT). Many trees have been sacrificed to record the efforts of those who have labored to analyze the risks associated with making an imperfect determination of the state of UUT. The impact of declaring a defective unit to be performing within its specifications (false accept) must be weighed against the costs of performing the calibrations themselves and the cost of declaring a properly performing instrument as exceeding its specifications (false reject). It is a difficult task for the authors of these documents to anticipate the impact of false test decisions in every application of calibration.

THE ADEQUACY REQUIREMENT:

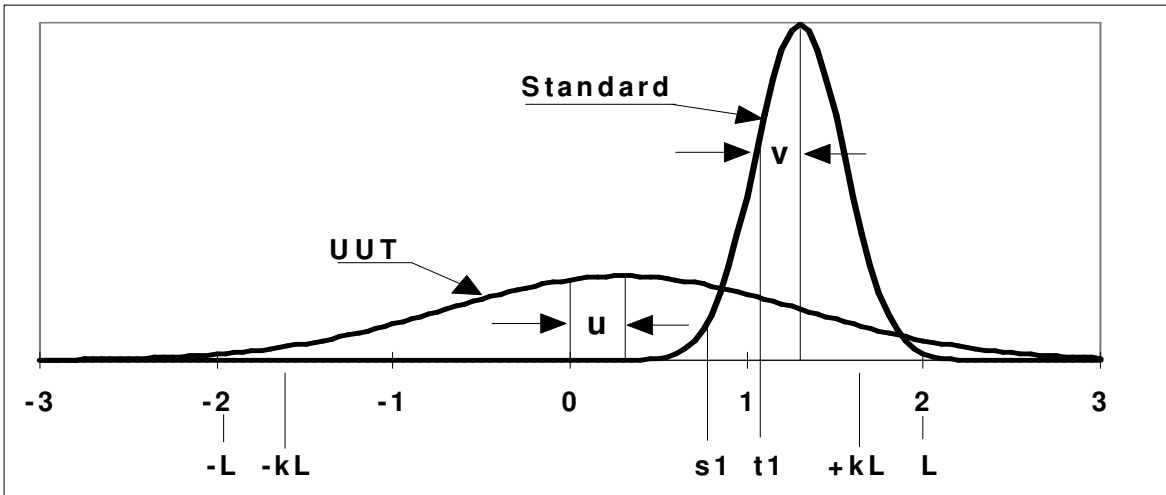
The adequacy clause in Z540-1 [1] is found in section 10.2.b and 10.3:

The laboratory shall ensure that the calibration uncertainties are sufficiently small so that the adequacy of the measurement is not affected. Well defined and documented measurement assurance techniques or uncertainty analyses may be used to verify the adequacy of a measurement process. If such techniques are not used, then the collective uncertainty of the measurement standards shall not exceed 25% of the acceptable tolerance (e.g., manufacturer's specification) for each characteristic of the measuring and test equipment being calibrated or verified.

Where methods are not specified, the laboratory shall, wherever practical, select methods that have been published in international or national standards, those published by reputable technical organizations or in relevant scientific texts or journals.

Similar requirements are found in other standards. It is the observation of the author that, in the majority of the calibration and standards laboratories, that the test uncertainty ratio (TUR) provision of these standards (...shall not exceed 25% of the acceptable tolerance...) receives far more attention than the sections that surround it. This has resulted in the widespread presumption that the risks associated with a TUR of 4:1 are acceptable and those less than 4:1 are unacceptable. Though the statisticians and mathematicians have determined that a test uncertainty ratio of 4:1 (or 10:1 or 3:1) provides adequacy in most instances, the responsibility still rests with the laboratory to determine that the associated risks implied by the procedure are acceptable. Metrology costs must be weighed against the consequences of false test decisions. In some critical applications, huge false reject losses may be acceptable in order to reduce the probability of a false accept. In others, where the cost of false test decisions is not particularly high, considerable savings in calibration costs may be realized by using lower accuracy standards.

Numerous technical papers [2-10] have been written to allow for risk assessment to be performed as a function of many more parameters than just TUR. Previous papers by this author [11,12] provided charts, figures, and tables to allow risk to be estimated and managed as a function of confidence interval and guardband factor as well as TUR. This paper adds charts and figures to show how systematic bias in the UUT and the STD affects the risk of false test decisions. This analysis was undertaken in response to concerns that the risk assessments and guardbanding recommendations presented in those papers may be too idealized (assuming normal distributions with means centered within the specification limits, etc.) and do not reflect the actual characteristics of the standards and units under test likely to be encountered. It is interesting that the same concern does not seem to be directed to the similar assumptions that surround the determination of adequacy based on a TUR of 4:1.



- UUT: The distribution of possible values for the unit under test.
- STD: The distribution of possible values for the Standard.
- t: Local variable for the UUT distribution
- t1: A possible value of the UUT
- s: Local variable for the Standard
- s1: A possible value of the Standard with the UUT at t1
- L: Specification Limit
- kL: Test Limit [Guardband Factor (k) times the Specification Limit (L)]
- u: Offset bias of the UUT
- v: Offset bias of the Standard
- R: Test Uncertainty Ratio
- CR: Consumer Risk or False Accept Risk
- PR: Producer Risk or False Reject Risk

Figure 1: Nomenclature

IMPLIED RISK:

Figure 2 summarizes some of the work done in the previous papers [11,12] and shows the level of risk implied by Z540, MIL-STD-45662A, and other standards which use 4:1 as the recommended minimum TUR if we assume no offset bias in the UUT nor the standard and normal distributions with confidence intervals of $\pm 2\sigma$. With these assumptions, the risk of false accepts, or consumer risk is 0.8%. As the chart shows, however, that level of risk can be maintained at lower test uncertainty ratios if the confidence interval is higher (more conservative specs) or through the use of guardband techniques (dashed lines). The curves were generated using MathCAD[®] [13] to solve:

$$Eq. 1 \quad CR = \frac{1}{\pi} \cdot \int_L^{\infty} \frac{-R(t - k \cdot L) - \frac{s^2 + t^2}{2}}{-R(t + k \cdot L)} ds dt$$

CONSUMER RISK WITH UUT BIAS:

The number of UUTs out of tolerance (OOT) rises significantly with UUT bias. Figure 3 shows OOT probabilities from 1σ to 6σ with offset biases of 0%, 25%, 50%, 75%, and 100%. The Motorola 6σ paper [14] concurs that the OOT probability is only 2 parts per billion if the confidence interval is $\pm 6\sigma$ with no offset bias but indicates that a more reasonable expectation is that there would be some bias. They picked 1.5σ of bias (25%) which yields an OOT probability of 3.4ppm. At a $\pm 2\sigma$ confidence interval, the OOT probability increases from 4.6% with no bias to 7.3% with 25% bias and 16% with 50% bias. At 100% bias, the mean of the distribution is aligned with the specification limit and half the units will be out of tolerance as can be seen in the figure. The equation to calculate the OOT condition of a normal distribution with bias is:

$$\text{Eq. 2} \quad OOT = 1 - \frac{1}{\sqrt{2 \cdot \pi}} \cdot \int_{-L}^L e^{-\frac{(t-u)^2}{2}} dt$$

The significant increase in the number of units out of tolerance due to UUT bias will increase the probability of false test decisions since there is a higher probability of units being at or near the specification limit. Figure 4 shows the effects of the UUT bias on the probability of accepting defective units (Consumer Risk) at the $\pm 2\sigma$ confidence level. Equation 3 was evaluated with UUT bias (u) to generate the curves. Charts are shown for TURs of 4:1 and 2:1 and with guardbands. First of all, it can be seen that for small biases, there is very little effect on the risk. All the curves have zero slope at the zero bias point because the slopes of the normal distribution at the specification limits are equal; the increase in OOT risk at one specification limit is canceled by a decrease in OOT risk at the other. As the amount of bias increases, however, the slopes are unequal and the increase in risk with bias at one limit is more than the reduction in risk at the other. Note also that guardbanding seems to still work with bias; the lines for the different guardband factors (k) are reasonably parallel. Figure 5 plots each of the $\pm 2\sigma$ curves relative to its value with no bias. Here it can be seen fairly easily that the risk is reasonably tolerant of modest biases, doubling for 40% bias. We can see that the increase in risk with bias is a little higher for lower TURs but that guardbanding (k values less than 1) reduces the risk slightly.

For example, if a calibration were being made with a 4:1 TUR at the $\pm 2\sigma$ confidence level and an offset bias of 80% were encountered, from the relative curves of Figure 4, we see that the risk of falsely accepting an out of tolerance unit is 4.2 times greater than with no bias. (The absolute risks of 3.4% vs. 0.8% can be seen on the upper charts of Figure 4.) What if we had only a 2:1 TUR and but were using a guardbands to maintain the risk of 0.8%? From the previous papers [11,12] we found a guardband factor (k) of 0.91 would be required. With an 80% bias, the risk would be 4.8 times greater than with no bias (TUR=2:1, $k=0.9$ curve at 80% bias). Thus, with 80% bias, using guardbands and a

2:1 TUR increased the risk only slightly ($4.75/4.24 = 12\%$) more than the factor of 4.2 that would have been experienced with a 4:1 TUR.

Figure 5 also contains plots consumer risk with UUT bias but at the $\pm 3\sigma$ confidence level. Their shapes are similar to those of Figure 4, the $\pm 2\sigma$ relative curves, but the risk increases by a much larger factor than for the $\pm 2\sigma$ case. With 40% bias the $\pm 2\sigma$ risk doubled, but the same 40% bias causes a ten-fold increase in the risk in the $\pm 3\sigma$ case. Comparing the absolute risk charts, we see that $\pm 3\sigma$ confidence level gives much more improvement at low biases than at high biases. Indeed we saw in Figure 3, in the extreme case of 100% bias, that 50% of the units are out of tolerance regardless of the confidence interval. Though the high confidence units continue to out-perform the lower less conservatively specified units, the tremendous improvement in risk level is diminished in the presence of bias.

Equation 3 was used to calculate the consumer risk for UUT bias ($u \neq 0$) and for bias in the standard ($v \neq 0$):

$$CR = \frac{1}{2 \cdot \pi} \cdot \int_{-\infty}^{-L} \int_{-R(t+k \cdot L)}^{-R(t-k \cdot L)} e^{-\frac{(t-u)^2 + (s-v)^2}{2}} ds dt$$

Eq. 3

$$+ \frac{1}{2 \cdot \pi} \cdot \int_L^{\infty} \int_{-R(t+k \cdot L)}^{-R(t-k \cdot L)} e^{-\frac{(t-u)^2 + (s-v)^2}{2}} ds dt$$

CONSUMER RISK WITH STANDARD BIAS:

By varying “v”, in Equation 3, we can investigate the effect that bias in the standard has on the consumer risk. Figure 6 shows the consumer risk for confidence intervals of $\pm 2\sigma$ and Figure 7 for $\pm 3\sigma$, but this time as a function of offset bias in the standard. These figures show somewhat different behavior for standard bias vs. UUT bias. Whereas we saw larger relative increases in the TUR=2:1 curves for UUT bias, for standard bias, the larger relative increases are in the TUR=4:1 curves. The relative increase is worse with guardbanding instead of less as we saw in the UUT bias curves. In fact if we use aggressive guardbanding with TUR=4:1, the tremendous reductions in risk are much less if we have bias in the standard. For example, the risk at 80% bias with TUR=4:1 increases to 0.13% from 0.006% with no bias, a factor of about 20. For the $\pm 3\sigma$ curves the increase is a whopping factor of 550 but only to a risk of 0.005%.

For the areas of practical metrology, however, the results are again quite reasonable. If we again consider our strategy of using a $k=0.91$ guardband with a TUR of 2, with 80% bias we could expect the consumer risk to twice the no-bias risk vs. increase a factor of 1.7 if we were using a 4:1 TUR.

PRODUCER RISK WITH BIAS:

Equation 4 was used to evaluate producer risk with UUT bias ($u \neq 0$) and standard bias ($v \neq 0$) similarly to the way Equation 3 was used to evaluate the consumer risk.

$$PR = \frac{1}{2 \cdot \pi} \cdot \int_{-L}^L \int_{-\infty}^{-R(t+k \cdot L)} e^{-\frac{(t-u)^2 + (s-v)^2}{2}} ds dt$$

Eq. 4

$$+ \frac{1}{2 \cdot \pi} \cdot \int_{-L}^L \int_{-R(t-k \cdot L)}^{\infty} e^{-\frac{(t-u)^2 + (s-v)^2}{2}} ds dt$$

Figures 8 & 9 document the absolute and relative producer risks associated with UUT bias and Figures 10 & 11, standard bias, for the same conditions considered for consumer risk. For both UUT and standard bias, the relative increase in risk is lower as guardbands are applied.

With UUT bias (Figures 8 & 9), the relative increase for a TUR=2:1 is lower than with the TUR=4:1, for the same guardband factor, meaning there is no producer risk “UUT bias penalty” when using guardbands for risk management.

With standard bias, however, the TUR=2:1 exhibit the higher relative increases in risk. For our practical situation of TUR=2:1 and $k=0.9$, however, we still have a lower increase in risk than the TUR=4:1 and no guardband for the $\pm 2\sigma$ case (Figure 10) and only slightly higher at $\pm 3\sigma$ (Figure 11).

SUMMARY:

The table below summarizes the results of offset bias for the example of managing calibration risks using a 2:1 TUR and a guardband factor (k) of 0.9. The incidence of false test decisions will increase with bias regardless of TUR. The guardbanding method of maintaining the same consumer risk as a 4:1 TUR nearly does so in the presence of bias. For conditions where bias is suspected, the last column of the table (k=0.8) shows that a little more guardbanding can reduce the risk below that of a 4:1 TUR.

		TUR k	4:1	2:1		
			1.0	1.0	0.9	0.8
No Bias	CR		0.80%	1.2%	0.76%	0.42%
	PR		1.5%	4.1%	7.0%	11%
80% UUT Bias	CR		3.4%	6.1%	3.6%	1.9%
	PR		3.8%	7.7%	12%	18%
80% STD Bias	CR		1.3%	1.8%	1.5%	1.2%
	PR		3.8%	12%	17%	22%

Table 1: Consumer Risk and Producer Risk with bias for $L=\pm 2\sigma$

CONCLUSION:

An analysis of the degradation in the ability to make quality calibrations in the presence of offset bias in the UUT and the standard shows that the guardbanding techniques continue to perform quite well as an means of managing the risk of accepting units which are not performing to their specifications when it is difficult to maintain the desired 4:1 test uncertainty ratio. With the tools available to the metrologist today, the ability exists to use the “justifications” and “other methods” allowed in the standards to significantly improve the quality and reduce the cost of calibrations performed in calibration and standards laboratories.

REFERENCES:

- [1] ANSI/NCSL Z540-1-1994, National Conference of Standards Laboratories, August, 1994
- [2] Capell, Frank, "How Good is Your TUR?", Evaluation Engineering, January, 1991, pp. 80-84
- [3] Castrup, Howard, "Evaluation of Customer and Manufacturer Risk vs. Acceptance Test Instrument In-Tolerance Level", NCSL Workshop & Symposium, 1980.
- [4] Eagle, Alan R., "A Method for Handling Errors in Testing and Measuring", Industrial Quality Control, March, 1954, pp. 10-15
- [5] Ferling, John A., "The Role of Accuracy Ratios in Test and Measurement Processes", Measurement Science Conference, 1984
- [6] Grubbs, Frank E. and Coons, Helen J., "On setting Test Limits Relative to Specification Limits", Industrial Quality Control, March, 1954, pp. 15-20
- [7] Hayes, Jerry L., "Calibration and Maintenance of Test and Measuring Equipment", Encyclopedia of Applied Physics, Vol 3, VCH Publishers, Inc., 1-56081-062-9/92, 1992
- [8] Hutchinson, Bill, "Setting Guardband Test Limits to Satisfy MIL-STD-45662A Requirements", 1991 NCSL Workshop & Symposium, pp. 305-309
- [9] Schumacher, Rolf B. F., "Decision Making and Customizing Decision Limits in Measurement Assurance", 1992 NCSL Workshop & Symposium, pp. 301-310
- [10] Weber, Stephen F. and Hillstrom, Anne P., "Economic Model of Calibration Improvements for Automatic Test Equipment", NBS Special Publication 673, 1984
- [11] Deaver, David, "Maintaining Your Confidence (In a World of Declining Test Uncertainty Ratios)", 1993 NCSL Workshop & Symposium, pp. 133-153
- [12] Deaver, David, "Guardbanding with Confidence", 1994 NCSL Workshop & Symposium, pp. 383-394
- [13] MathCAD is a registered trademark of MathSoft, Inc. The author has no interest nor affiliation with MathSoft except that of a satisfied user of the MathCAD software program.
- [14] Harry, Mikel J., "The Nature of Six Sigma Quality", Motorola Inc., Government Electronics Group

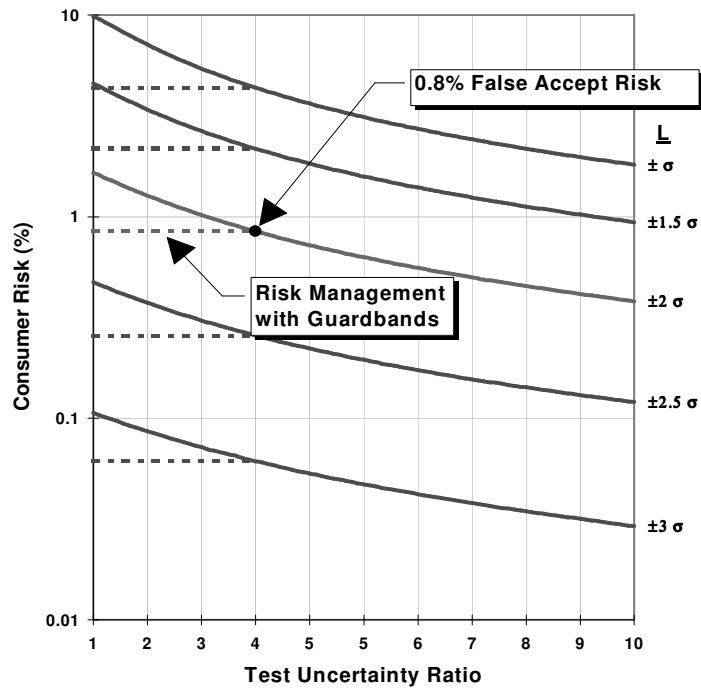


Figure 2: False Accept Risk as a Function of TUR, Confidence Interval and Guardbands

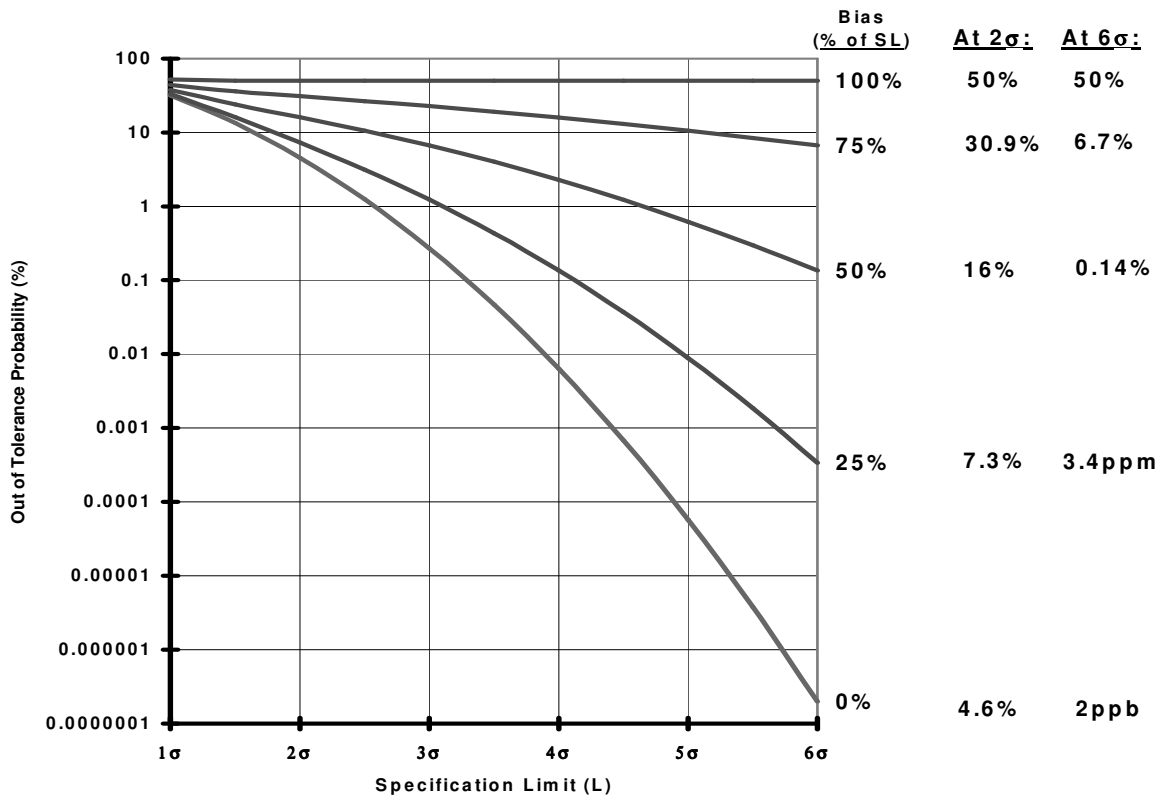


Figure 3: Out of Tolerance Probability of a Normal Distribution with Offset Bias.

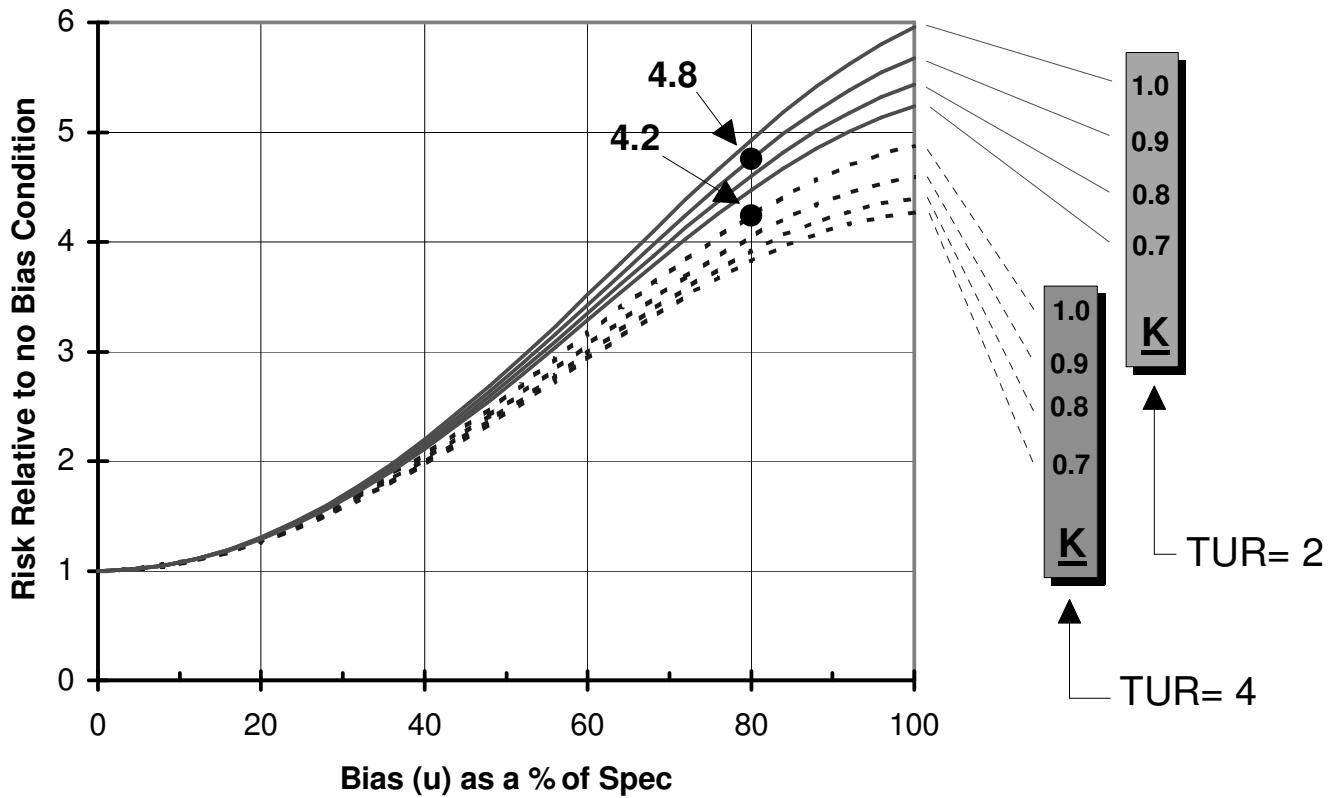
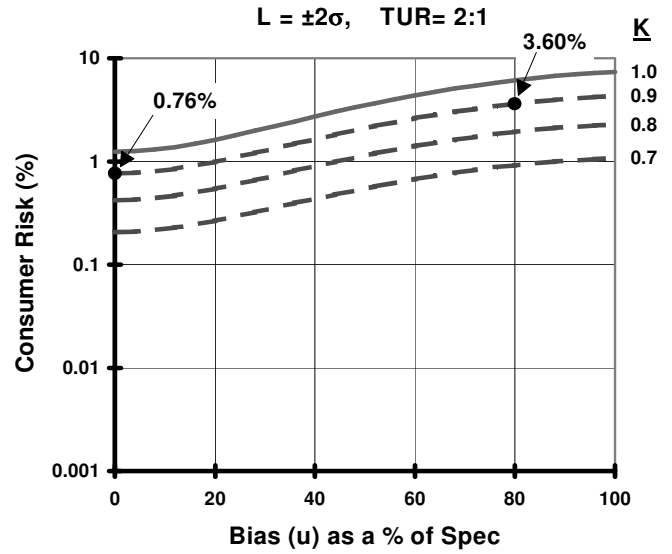
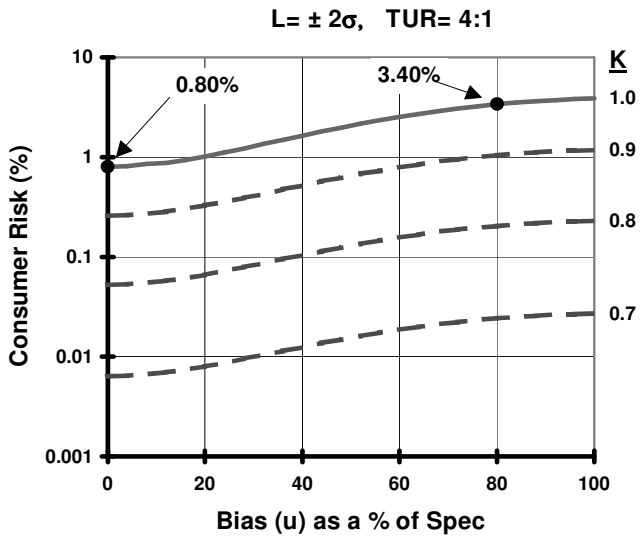


Figure 4: False Accept (Consumer) Risk with UUT Bias for $L = \pm 2\sigma$

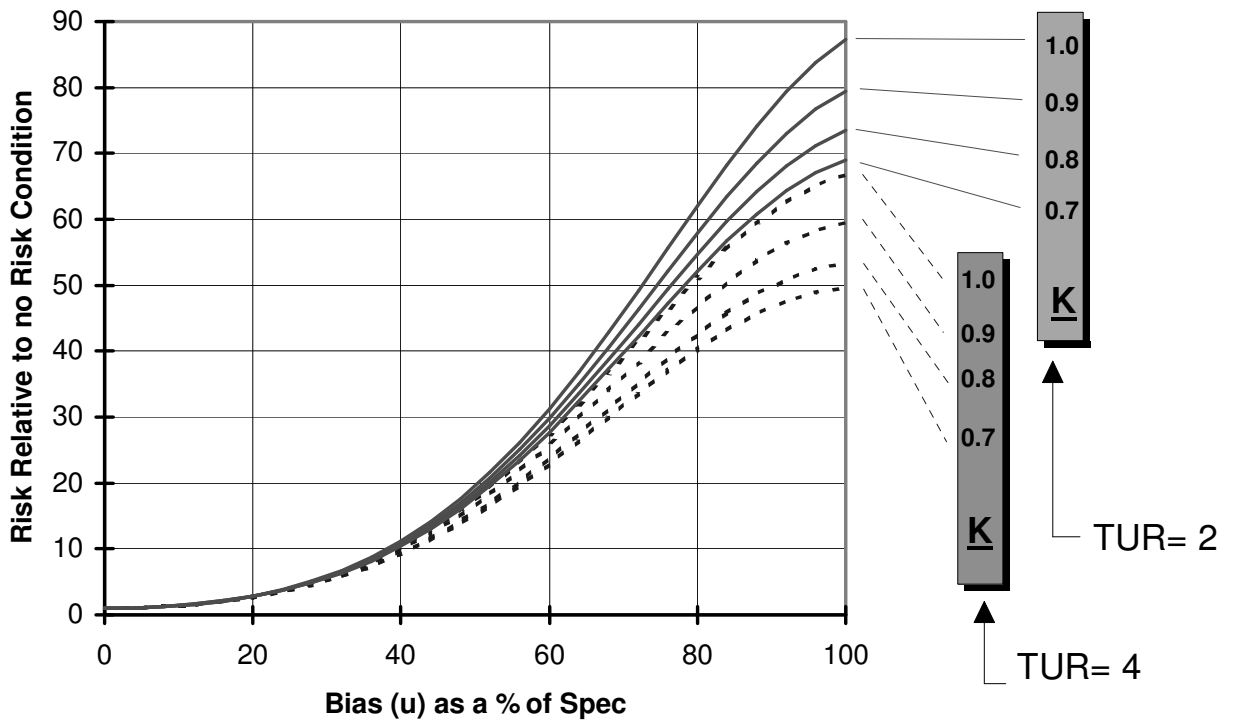
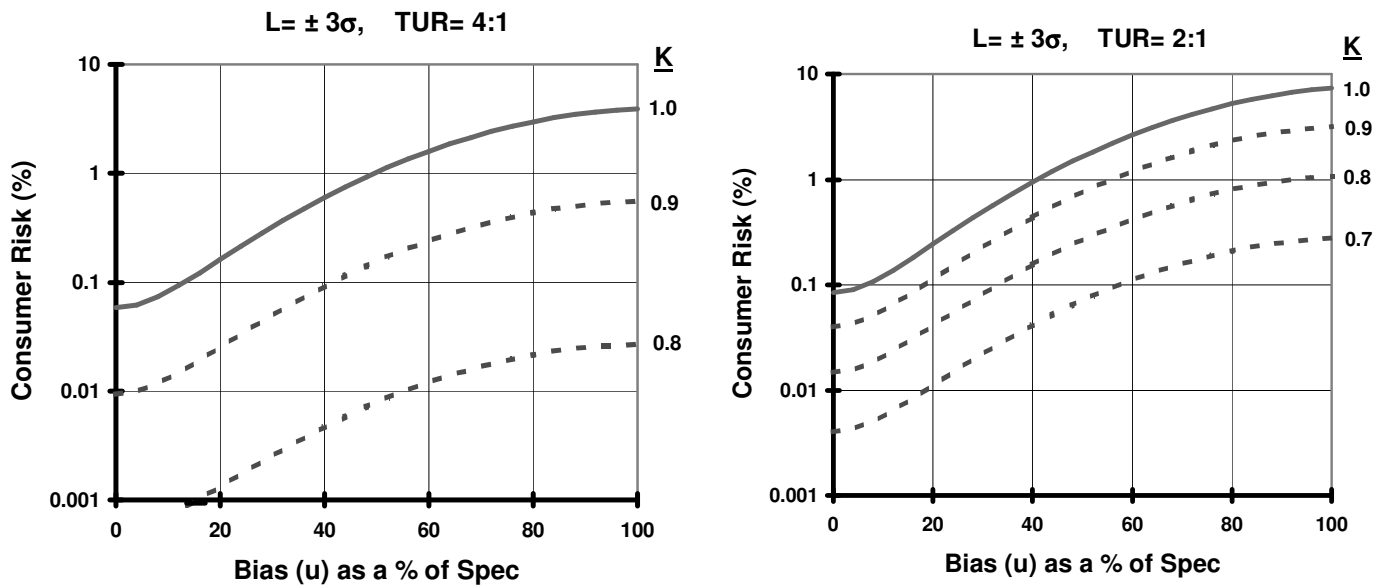


Figure 5: False Accept (Consumer) Risk with UUT Bias for L = ±3σ

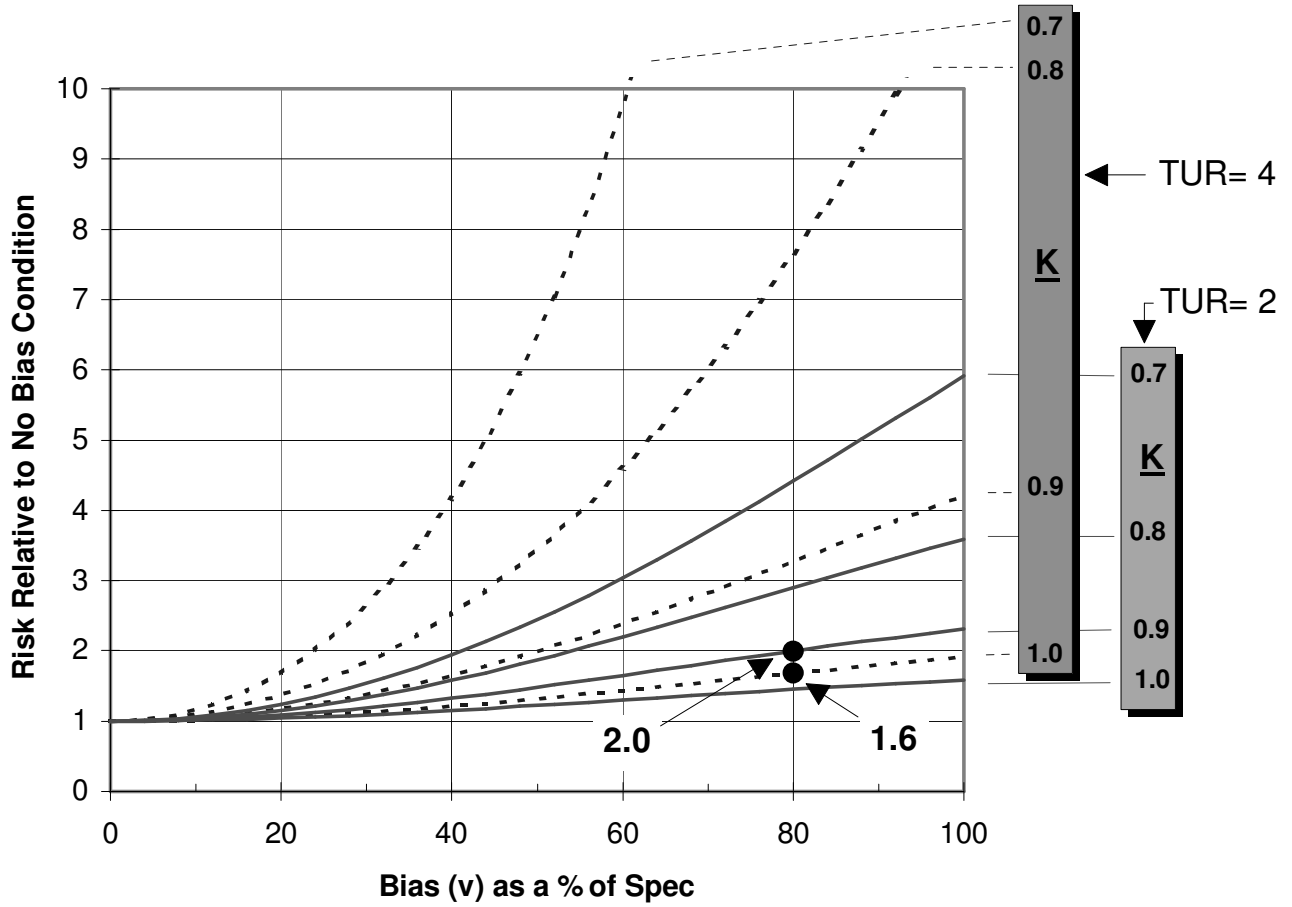
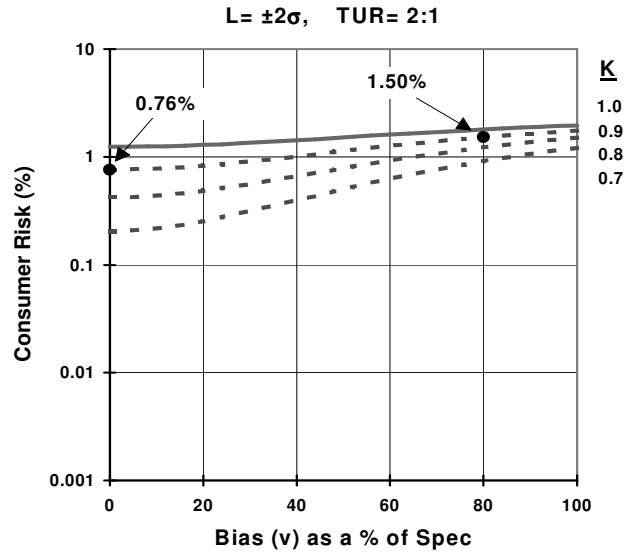
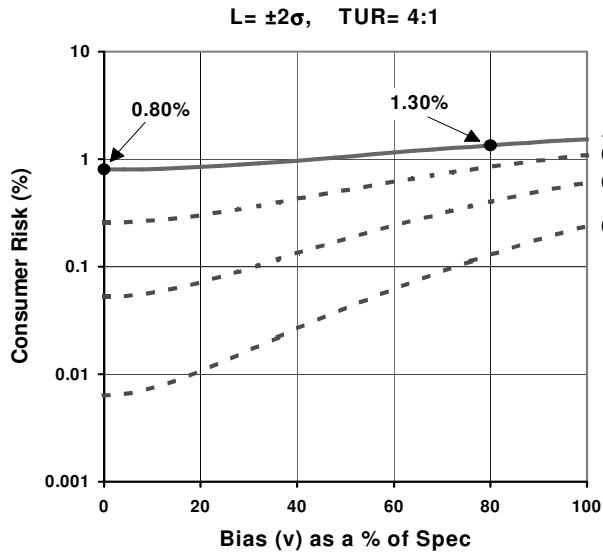


Figure 6: False Accept (Consumer) Risk with STD Bias for L = $\pm 2\sigma$

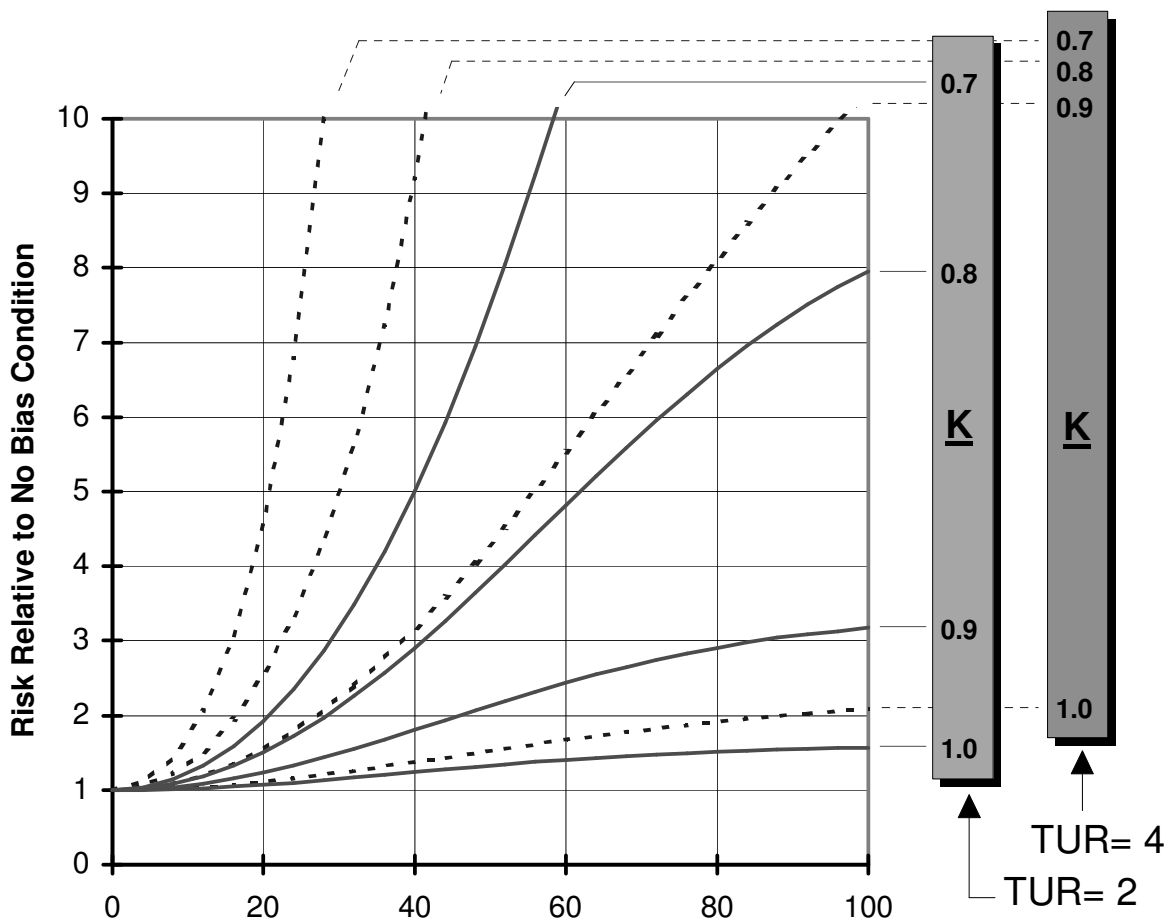
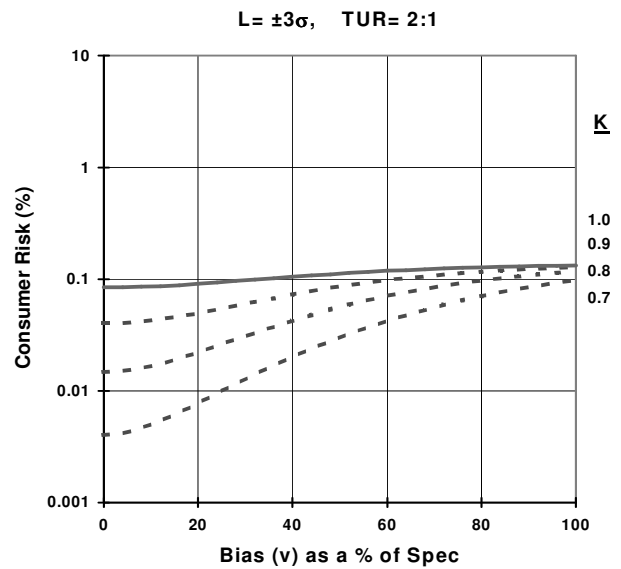
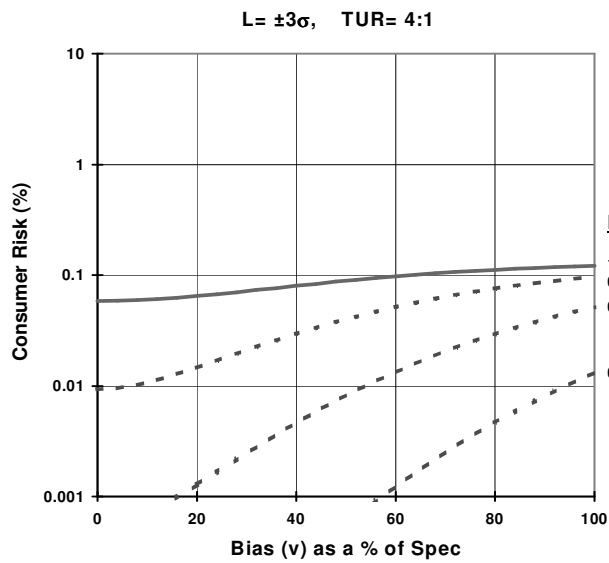


Figure 7: False Accept (Consumer) Risk with STD for $L = \pm 3\sigma$

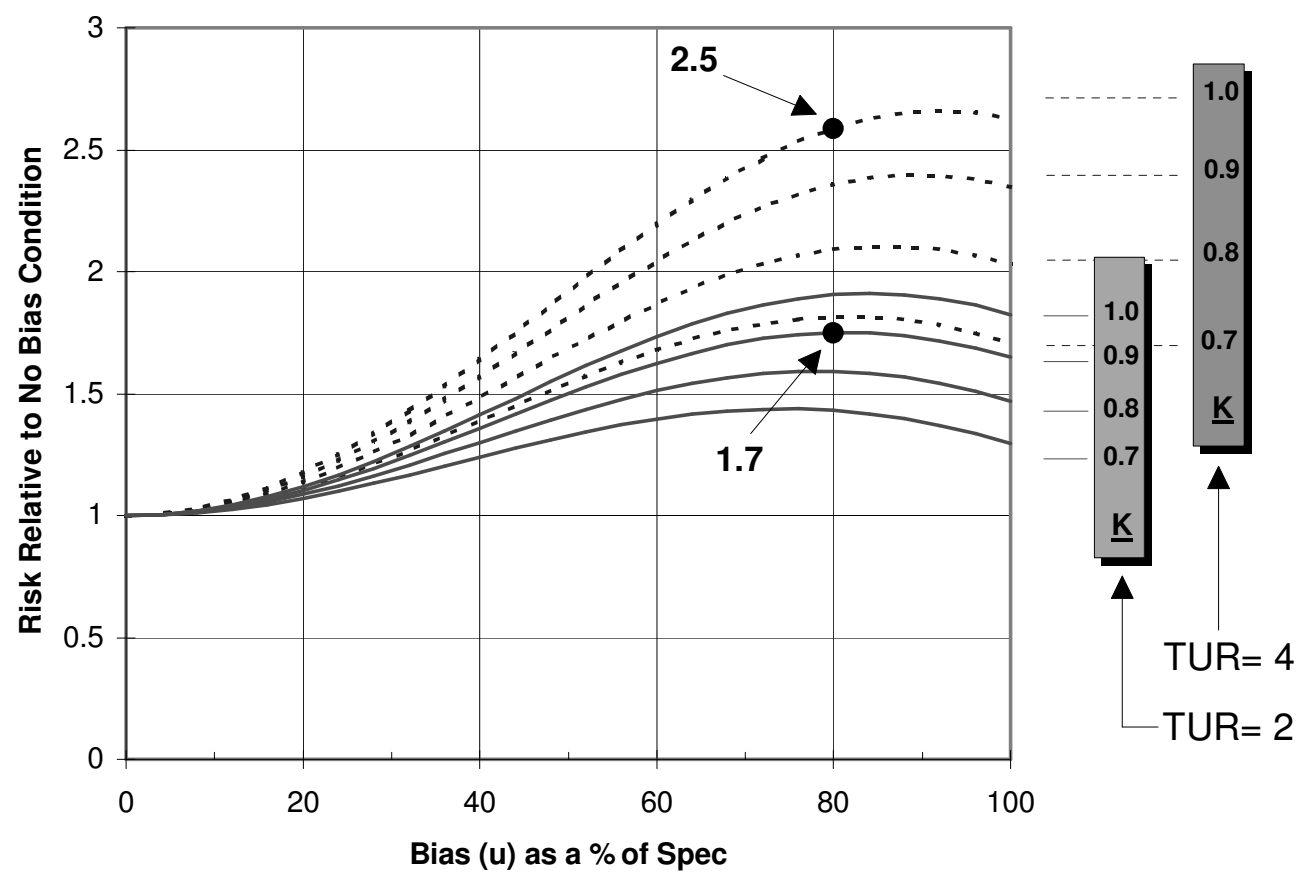
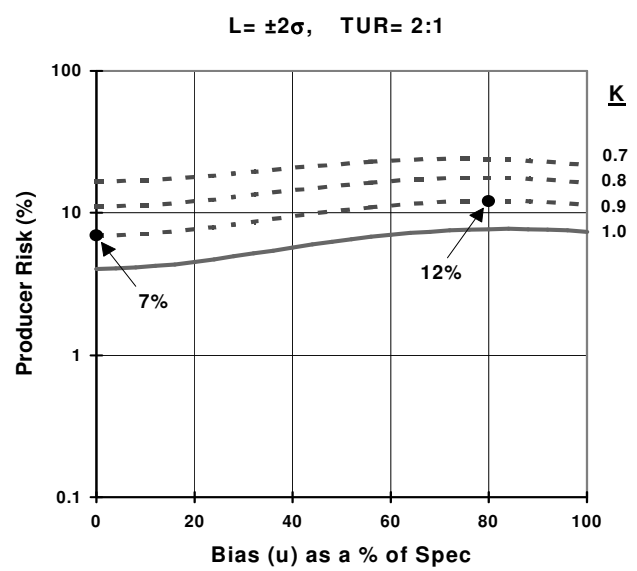
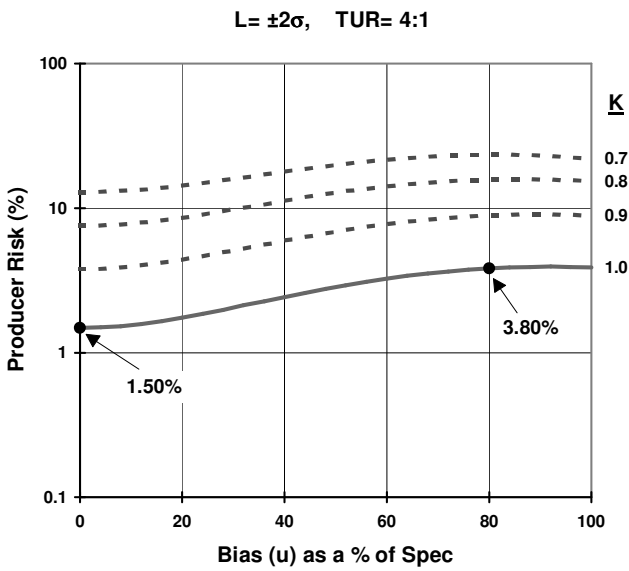


Figure 8: False Reject (Producer) Risk with UUT Bias for $L = \pm 2\sigma$

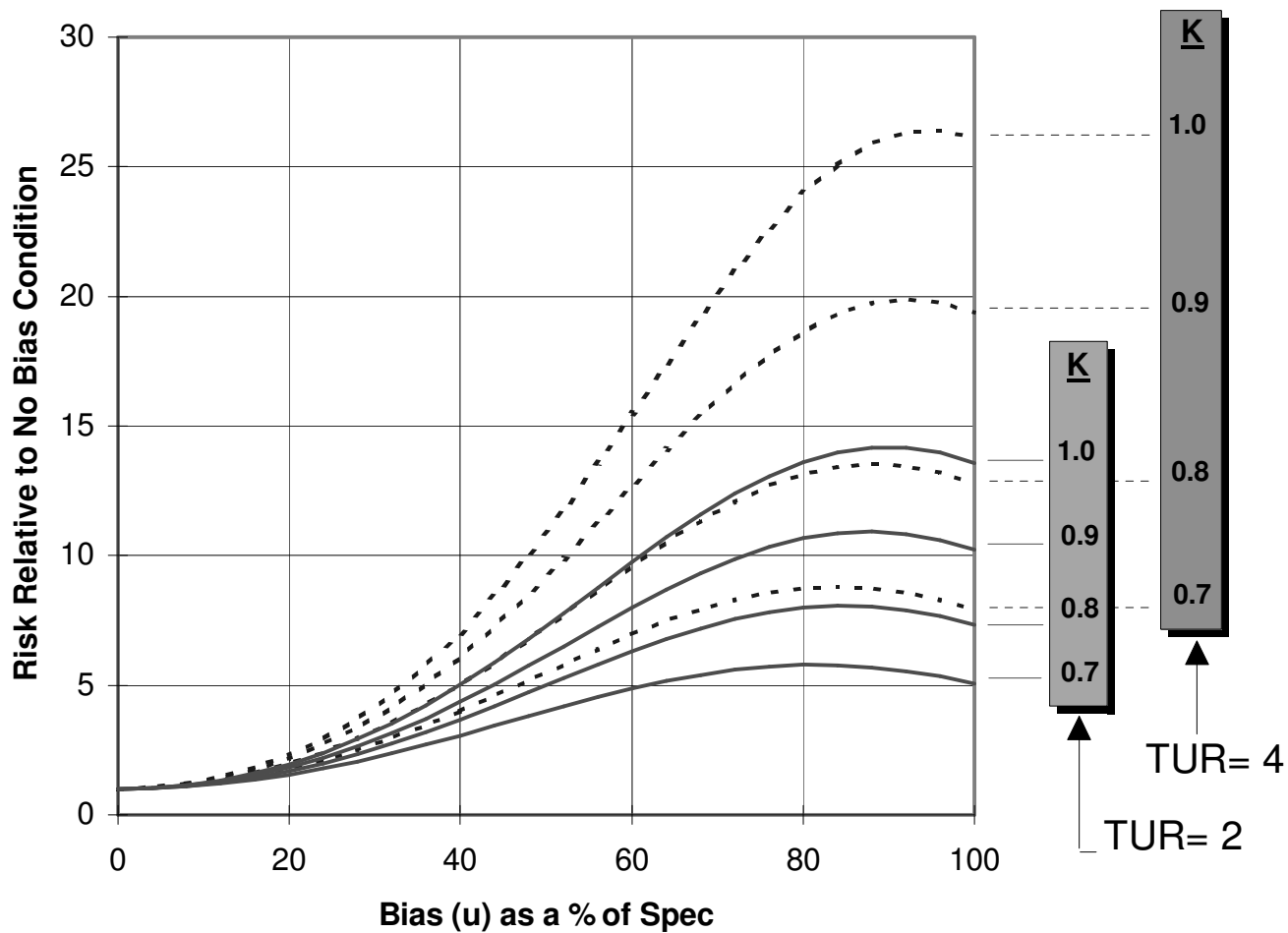
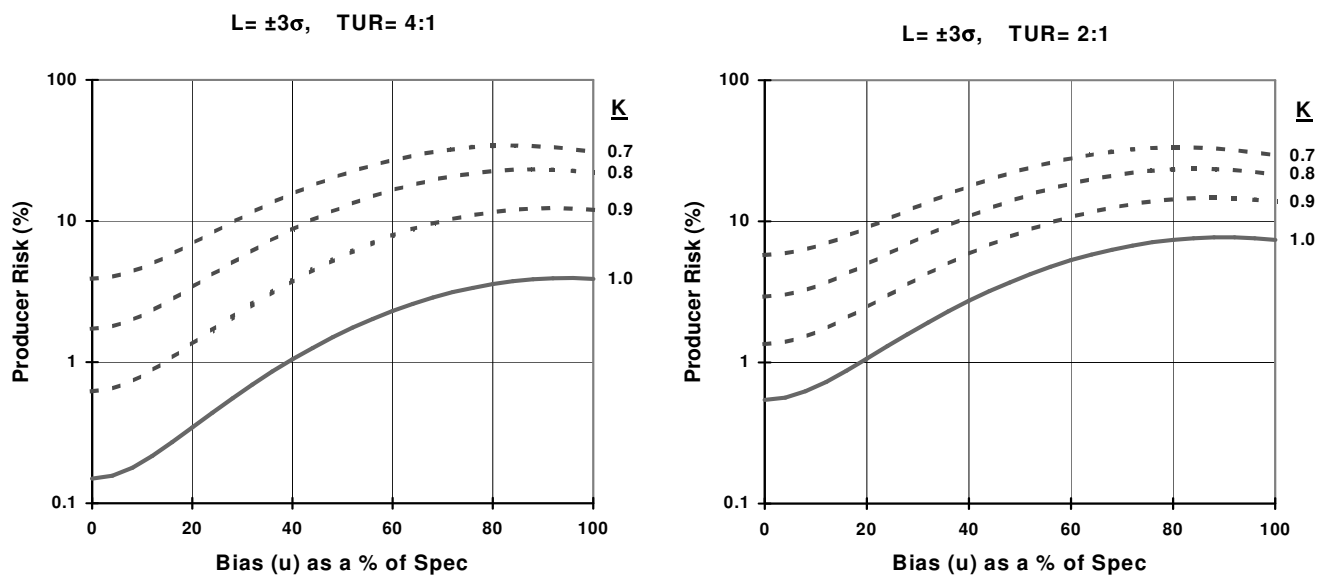


Figure 9: False Reject (Producer) Risk with UUT Bias for $L = \pm 3\sigma$

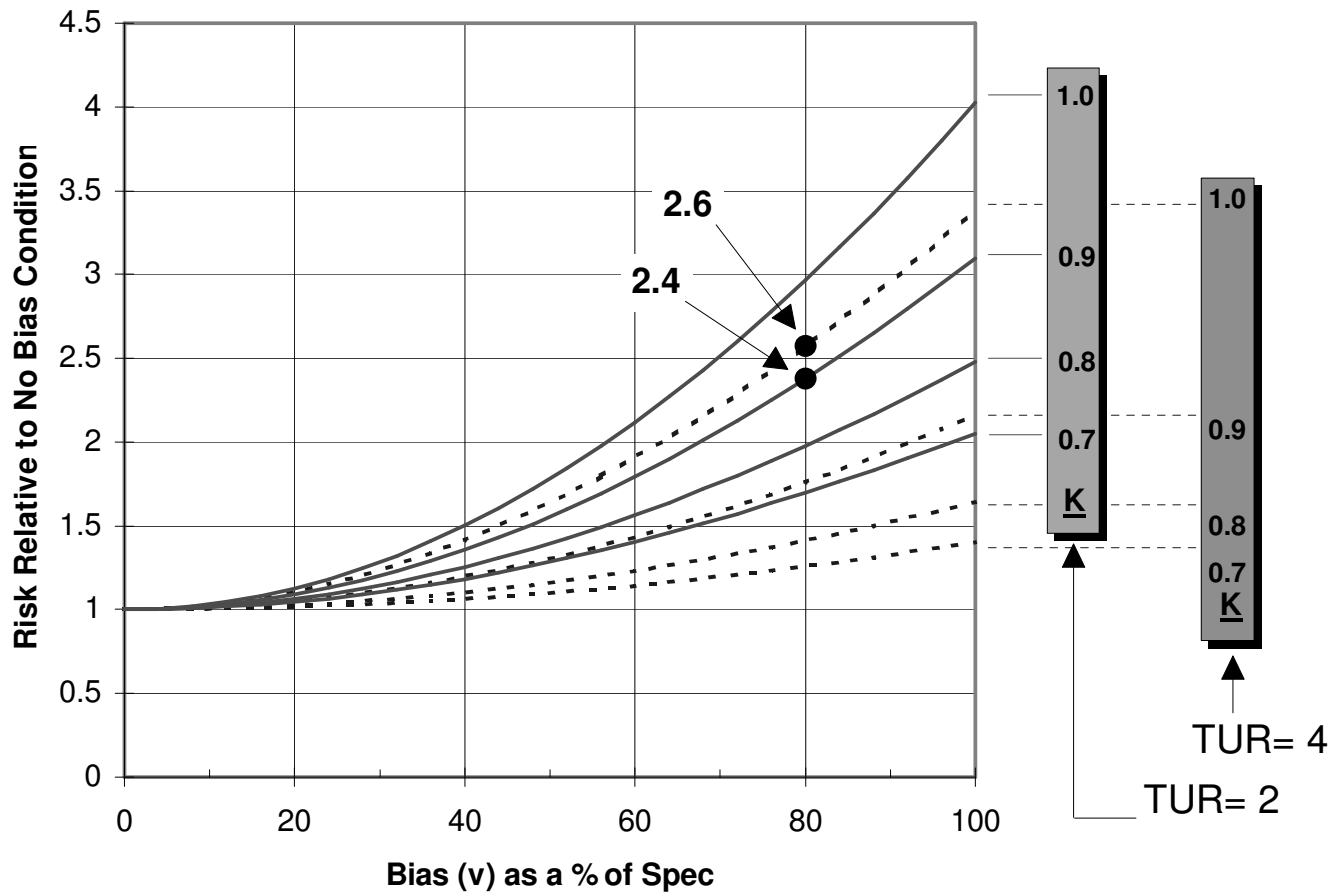
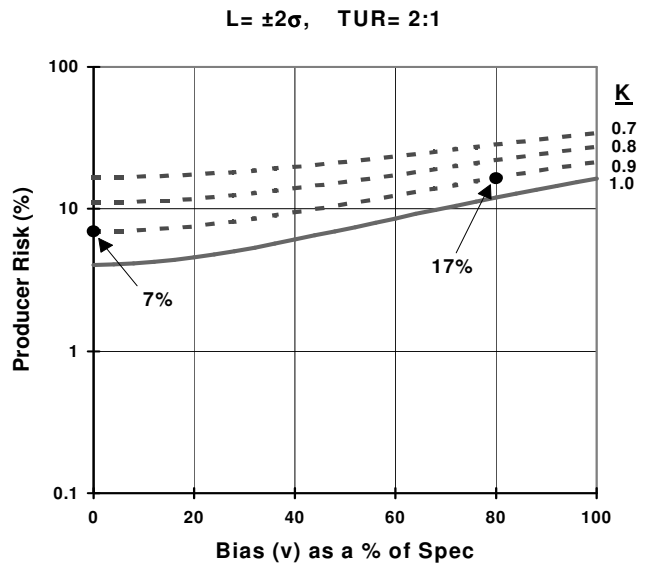
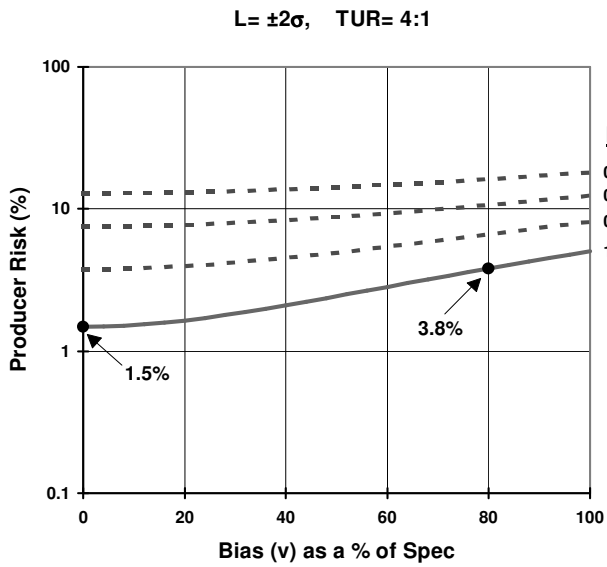


Figure 10: False Reject (Producer) Risk with STD Bias for $L = \pm 2\sigma$

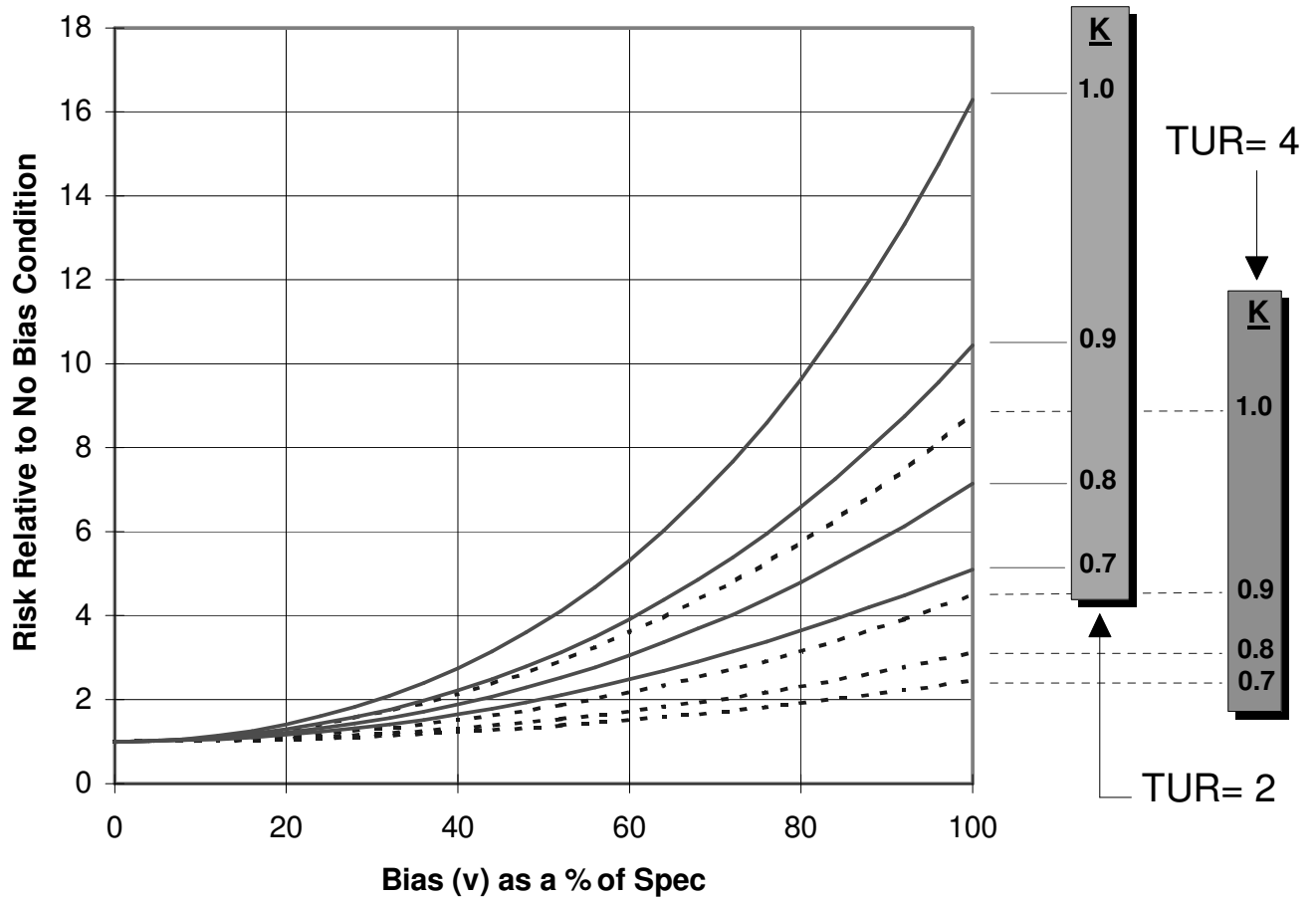
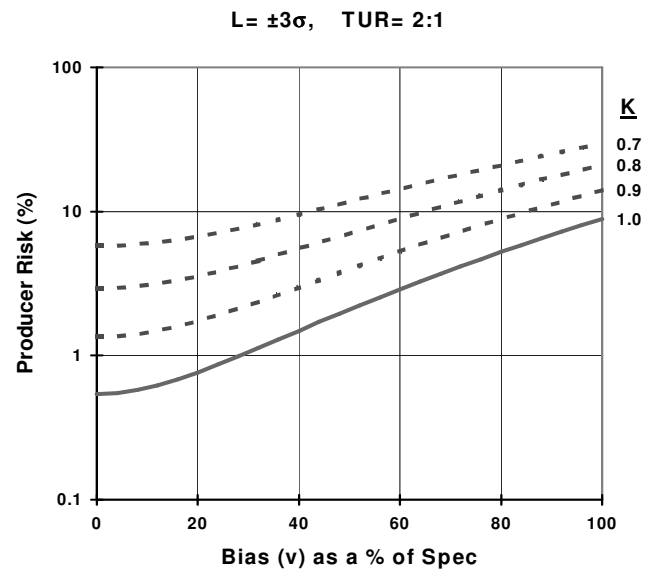
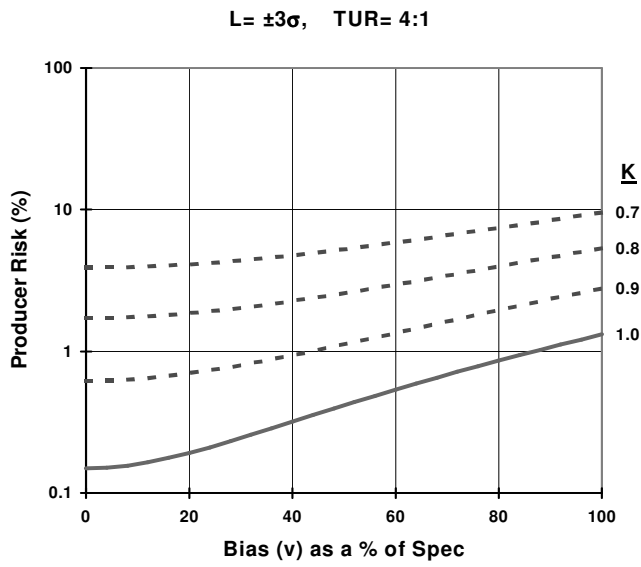


Figure 11: False Reject (Producer) Risk with STD Bias for $L = \pm 3\sigma$