Experience Gained from Offering Accredited 3rd Party Proficiency Testing

Speaker/Author: Dr. Henrik S. Nielsen
HN Metrology Consulting, Inc.
HN Proficiency Testing, Inc.
10219 Coral Reef Way
Indianapolis, Indiana, USA
Phone: (317) 849 9577; Fax: (317) 849 9578
Email: hsnielsen@HN-Metrology.com

Abstract

HN Proficiency Testing has been offering third-party proficiency testing for about three years and has been accredited for the last two years. This paper discusses some of the experience gained and the lessons learned in the process. Accreditation bodies generally require accredited laboratories to participate in proficiency testing. Since most laboratories see this as nothing but an inconvenience and an added expense of maintaining accreditation, very few non-accredited laboratories participate. Consequently, the opportunity to analyze third-party proficiency test results provides a unique insight particularly into the state of the accredited laboratories that constitute the backbone of the US metrology infrastructure. As it turns out, technical insight into the measurement processes analyzed is at least as important for the proficiency testing provider as is a thorough understanding of the statistics behind the common proficiency testing metrics. The paper discusses some of the general trends that can be identified from this vantage point as well as some specific examples of where proficiency testing turned out to be more than just an expensive inconvenience for the participating laboratory. In accordance with the rules for accredited proficiency testing providers, the anonymity of all participating laboratories, innocent or otherwise, will be protected throughout the paper.

1. Introduction

The proficiency tests offered by HN Proficiency Testing are aimed at calibration laboratories and dimensional inspection laboratories. These types of laboratories are always required to estimate their measurement uncertainty if they are accredited and providing accredited calibrations. Since these laboratories also tend to use very diverse methods and therefore have very diverse uncertainties, the E_n value is the only meaningful metric for evaluating the performance of these laboratories. Consequently, this paper evaluates test results and laboratory performance based on E_n values. The E_n value is calculated as follows:

$$E_n = \frac{V_{Lab} - V_{Ref}}{\sqrt{U_{Lab}^2 + U_{Ref}^2}}$$

Where

V_{Lab} is the value reported by the laboratory
V_{Ref} is the reference value
U_{Lab} is the laboratory's reported uncertainty at 95% coverage level
U_{Ref} is the uncertainty of the reference value at 95% coverage level

An E_n value in the range of –1 to +1 indicates that the laboratory value and the reference value agree with each other within their respective uncertainties and is considered a success. An E_n value outside this range is considered a failure. If all the usual statistical assumptions (normal distributions, uncorrelated results, etc.) hold true, a failure rate of 5 % should be expected, since the uncertainties are reported at a 95% coverage level.

There are two main reasons that the failure rate can exceed the expected level. Either the quoted uncertainty is too low or there is something atypically wrong with the measurement. Of course, the two problems can be present in the same measurement.

If the failure rate is lower than they expected level, it usually means that the laboratory is conservative in its uncertainty estimates. However, a natural bias in the selection of reference laboratories will also influence the results. Since any conscientious proficiency testing provider will tend to select reference laboratories that are able to live up to their claimed uncertainty for the vast majority of the measurements, the reference uncertainty will tend to be a high estimate, thus reducing the failure rate.

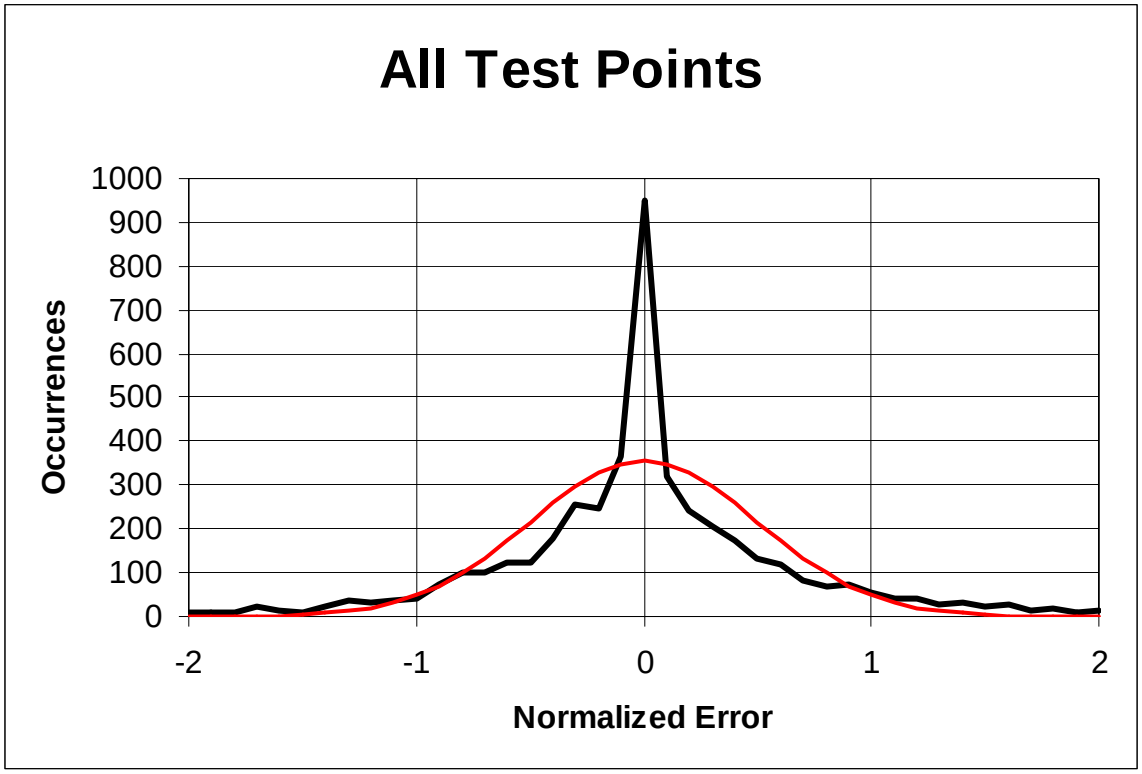

Figure 1: Normalized error (E_n value) distribution for all test results (Black Line) and expected distribution (Red Line).

2. Overall results

HN Proficiency Testing has offered commercial proficiency testing over the last three years. The majority of these tests have been in that dimensional field, but some electrical tests and lately some physical tests have also been offered. The tests that form the basis of this paper are

all either dimensional or electrical. Over 4,000 test results form the basis of this paper. Figure 1 shows the E_n value distribution for all the test results.

The first thing to note about figure 1 is that the results do not appear to follow a normal distribution. The peak around $E_n = 0$ is taller than would be expected of a normal distribution and the "wings" of the distribution are too wide.

One possible explanation for the large peak around $E_n = 0$ is that a substantial portion of the data are from calibration of hand gages, where the resolution of the hand gage is the dominant uncertainty contributor and where more often than not the participating laboratory's measured value is identical to the reference value. Figure 2 shows only the results from the hand gage calibrations and in figure 3 the results from the hand gage calibrations have been removed from the overall data set.

As can be seen from figure 2, the vast majority of the hand gage calibration results show perfect agreement between the participant values and the reference values due to the relative coarseness of the resolution of the hand gages compared to the capability of the overall calibration process.

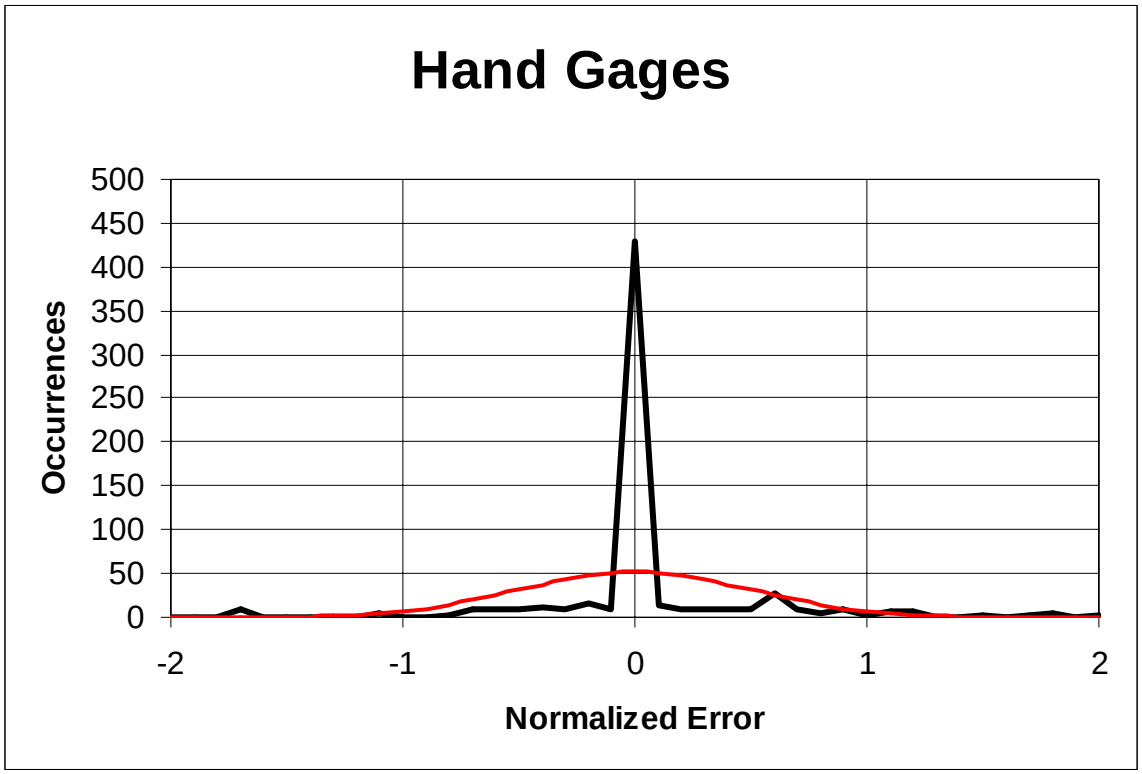

Figure 2: Normalized error (E_n value) distribution for all hand gage calibration test results (Black Line) and expected distribution (Red Line).

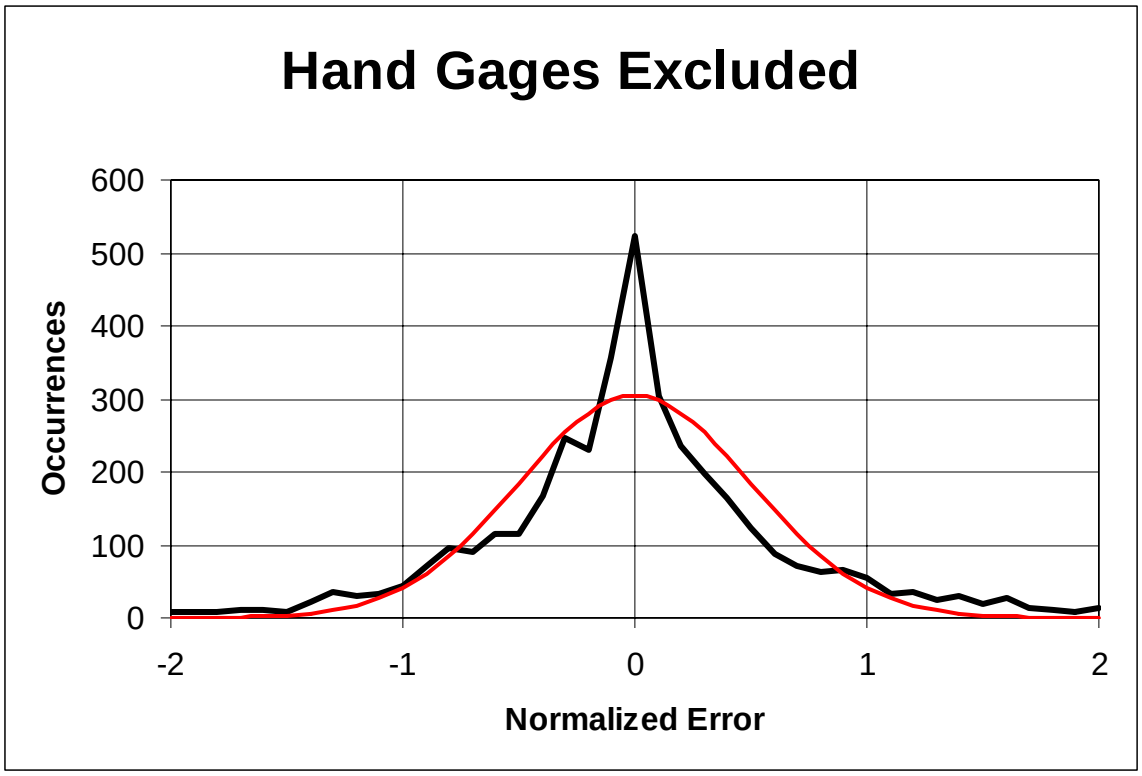

Figure 3: Normalized error (E_n value) distribution for all test results, except those for hand gage calibration (Black Line) and expected distribution (Red Line).

Excluding the hand gage results has reduced the central peak somewhat, but it is still quite pronounced. One explanation for the distribution, which fits anecdotal evidence, is that it is really a convolution of two distributions: a narrow distribution from laboratories that significantly overestimate their uncertainty on top of a much wider distribution from laboratories that underestimate their uncertainty.

3. Discipline specific results

In the following we shall see that laboratories' ability to estimate measurement uncertainty very significantly from one field of measurement to another. However, in all cases we shall find that there is a number of conservative laboratories that overestimate their uncertainty.

3.1 Electrical calibrations

Figure 4 shows the results from the calibration of an Agilent 34401a digital multimeter. The test points include voltage and current, both DC and AC, as well as resistance measurements. The participants sourced the test point values and read the deviation on the multimeter.

These results are the only ones to show our failure rate significantly lower than the theoretically expected one. This suggests that either the electrical laboratories are conservative in their claims of measurement uncertainty or that the manufacturers of electrical calibration equipment are conservative in the specifications they give for their products.

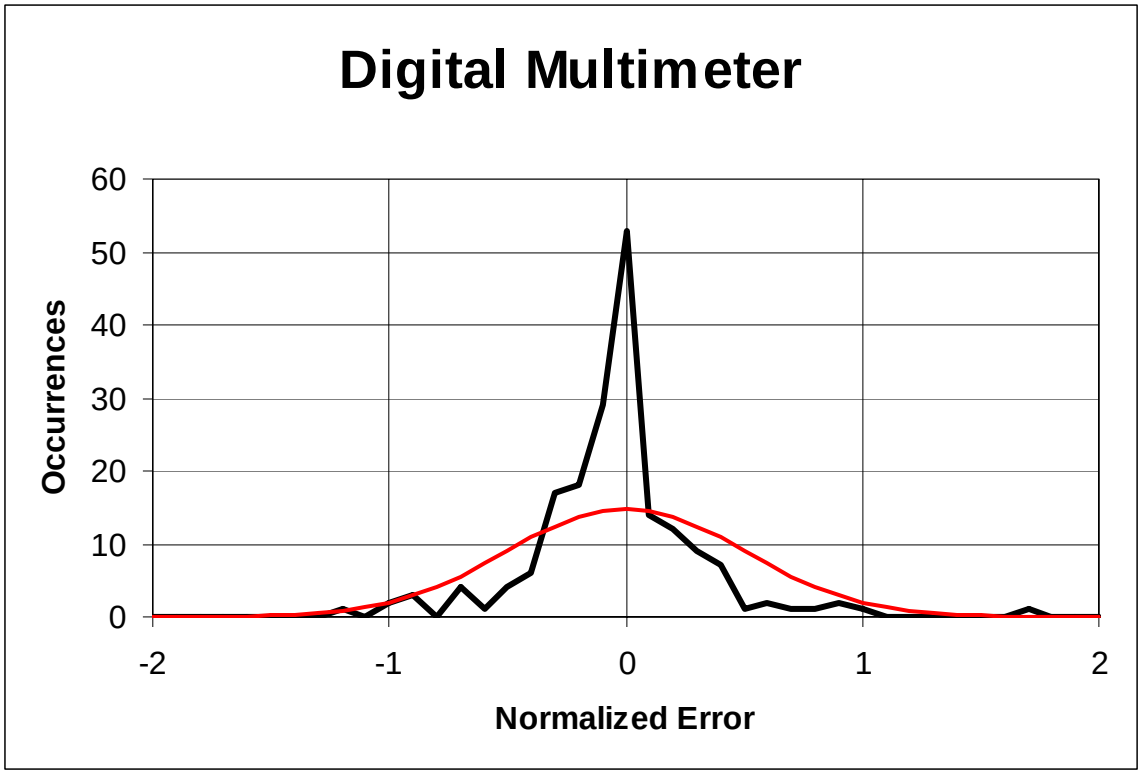

Figure 4: Normalized error (E_n value) distribution for all digital multimeter calibration test results (Black Line) and expected distribution (Red Line).

Over the last several years there has been a substantial and well-published debate in the accreditation community concerning how to take the specifications of electrical calibration equipment into account in uncertainty budgets, since the manufacturers on one hand claim that the specifications are 99% limits, but typical calibration procedures treat them as the hard limits of a rectangular distribution. Assuming that the majority of the participants have taken the rectangular distribution route in their uncertainty estimates, it appears that this is a conservative assumption.

Another electrical proficiency test is the temperature simulation test based on an Omega CL 26 readout unit for thermocouples and RTDs. The measurands and uncertainties are given in °C, but are really Volts and Ohms in disguise. From the underlying results it is clear that many participants express the uncertainty as a simple percentage of the measured value. However, this is intuitively problematic for a scale such as the temperature scale that contains an arbitrary zero. The same measurement reported in °C and °F will have a different uncertainty, which cannot logically be the case.

Figure 5 shows the results of the temperature simulation test. As the figure shows the results are considerably worse than what one would expect having seen accredited laboratories ability to measure electrical quantities in the DMM calibration test in figure 4. This would suggest that the problem lies not in the ability to measure electrical quantities, but in the ability to convert the uncertainty of the electrical measurement to a corresponding amount of °C.

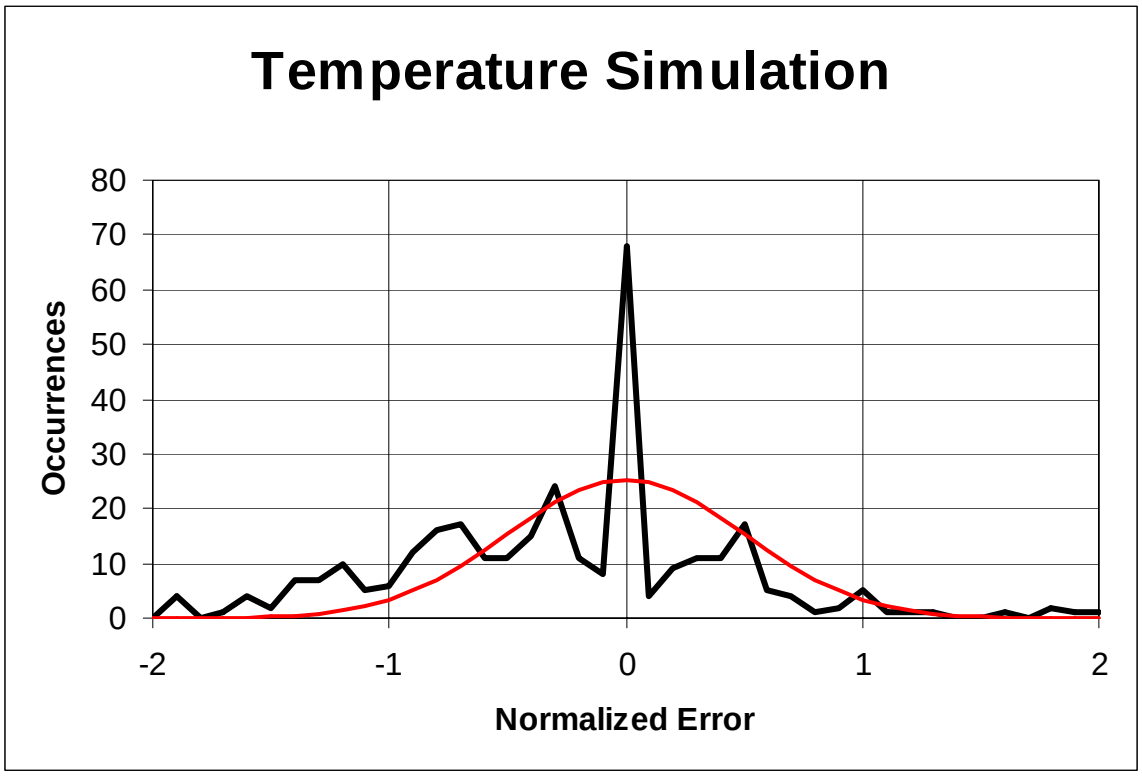

Figure 5: Normalized error (E_n value) distribution for all temperature simulation test results (Black Line) and expected distribution (Red Line).

3.2 Dimensional calibrations

We have already seen the results for calibration of hand gages in figure 2. In this section we will look at the calibration of gage blocks and calibration of hard gages, including plain plug gages, plain ring gages, thread gages and thread wires.

The results of the gage block calibration tests are given in figure 6. This is the only set of tests in addition to the DMM tests that show a failure rate of less than the expected 5%. One explanation may be that while gage block calibration requires a good environment, good standards, and patience, it is still a very simple 1:1 comparison process and the major uncertainty contributor is generally the uncertainty of the master gage blocks used, which the participating laboratory "inherits" from the higher level calibration laboratory in the traceability chain.

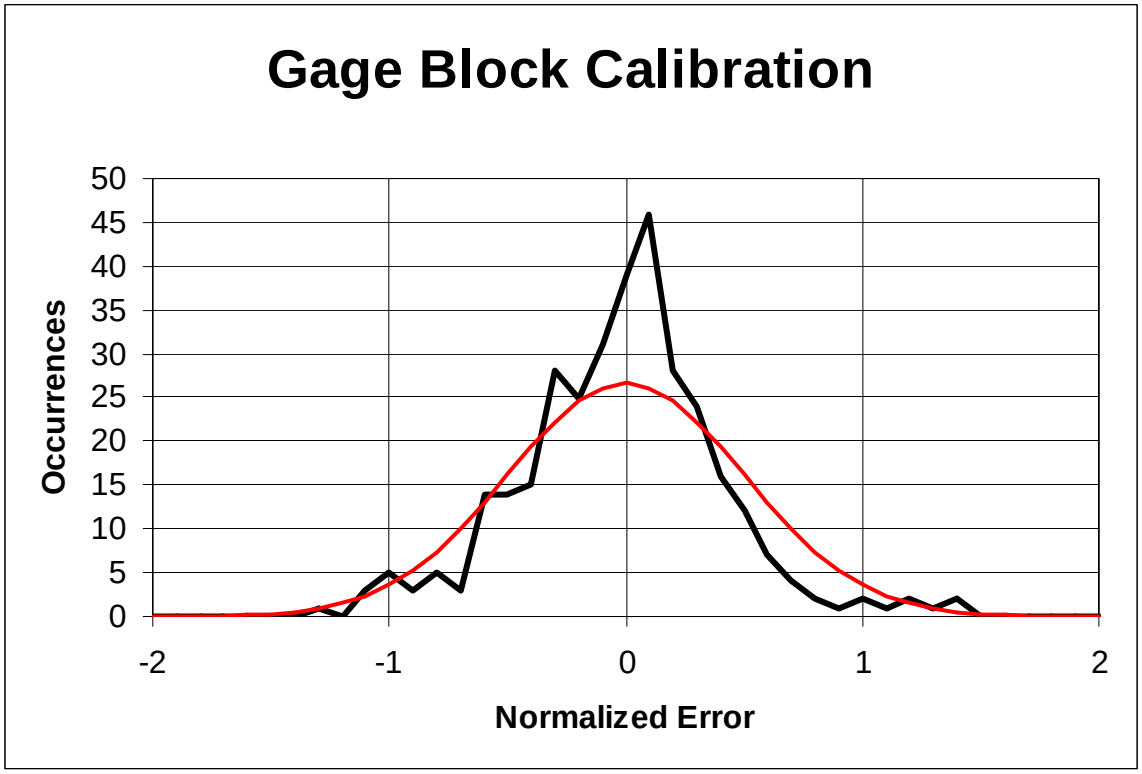

Figure 6: Normalized error (E_n value) distribution for all gage block calibration test results (Black Line) and expected distribution (Red Line).

I have chosen to treat the calibration of hard gages as one discipline since they in many respects resemble each other and is a larger statistical material to ensure the anonymity of the participants. Although not as pronounced as the shift from the DMM calibration results to the temperature simulation results, there is still a significant difference between the proficiency of dimensional laboratories to calibrate gage blocks and their proficiency to calibrate other hard gages, given their claimed uncertainty.

Figure 7 shows the results of the hard gage calibrations. The failure rate is over 15% or three times the expected rate. This suggests that while dimensional calibration laboratories are capable

of accounting for their uncertainty when calibrating gage blocks, they tend to be either optimistic or miss some of the contributors, when the calibration is no longer a 1:1 comparison.

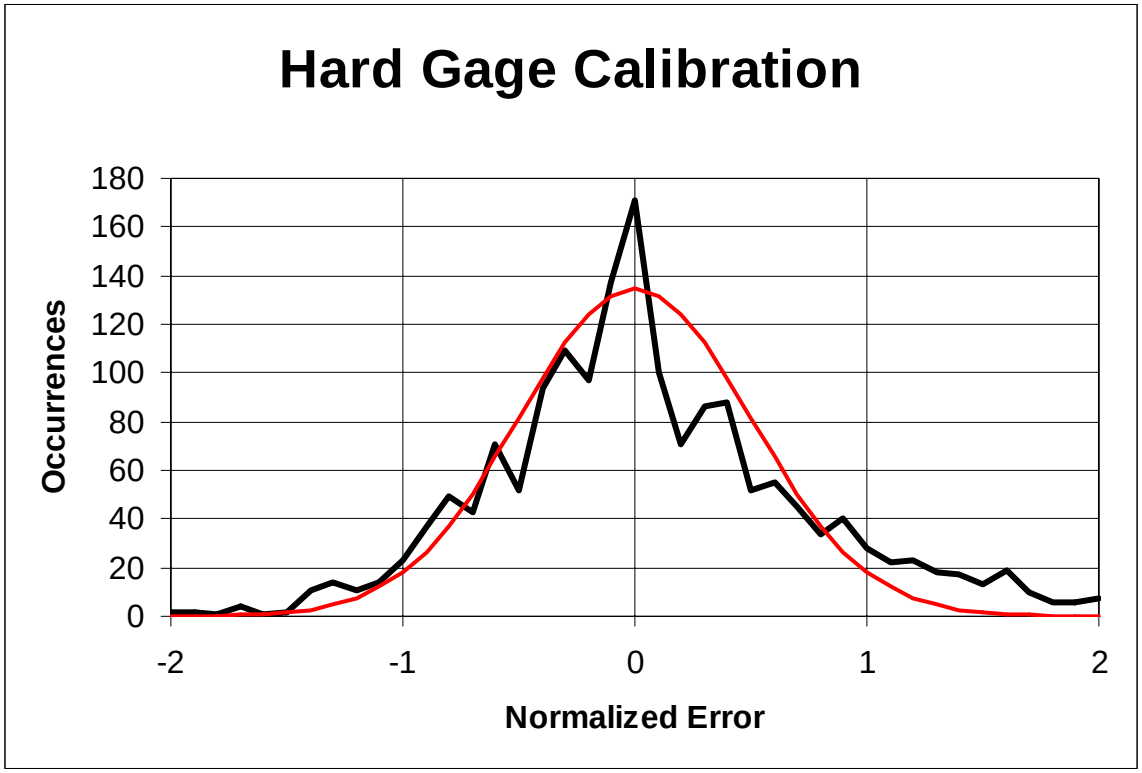

Figure 7: Normalized error (E_n value) distribution for all hard gage calibration test results, including plain plug gages, plain ring gages, thread gages and thread wires (Black Line) and expected distribution (Red Line).

3.3 Dimensional inspection

The third broad category of tests offered by HN Proficiency Testing is tests for dimensional inspection laboratories. I have chosen to pool the results from the optical inspection test, which is designed to use an optical comparator, a measuring microscope, or an optical CMM and the dimensional inspection test, which is designed to use either hand gages or surface plate inspection, while treating the CMM inspection tests separately.

It is interesting to note that the calculation of uncertainty for CMM inspection has been the subject of great debate and great controversy, while the uncertainty of measurements using e.g. an optical comparator has been considered simple and straightforward. While tasks for the CMM test were chosen on the Goldilocks principle, not to simple, not too hard, but just right, the results seem to indicate that the laboratories have a better general understanding of the uncertainty associated with CMM measurements, than that of traditional dimensional inspection. The CMM measurements have a failure rate of 9%, whereas the optical and dimensional inspection tests have a failure rate of 33%.

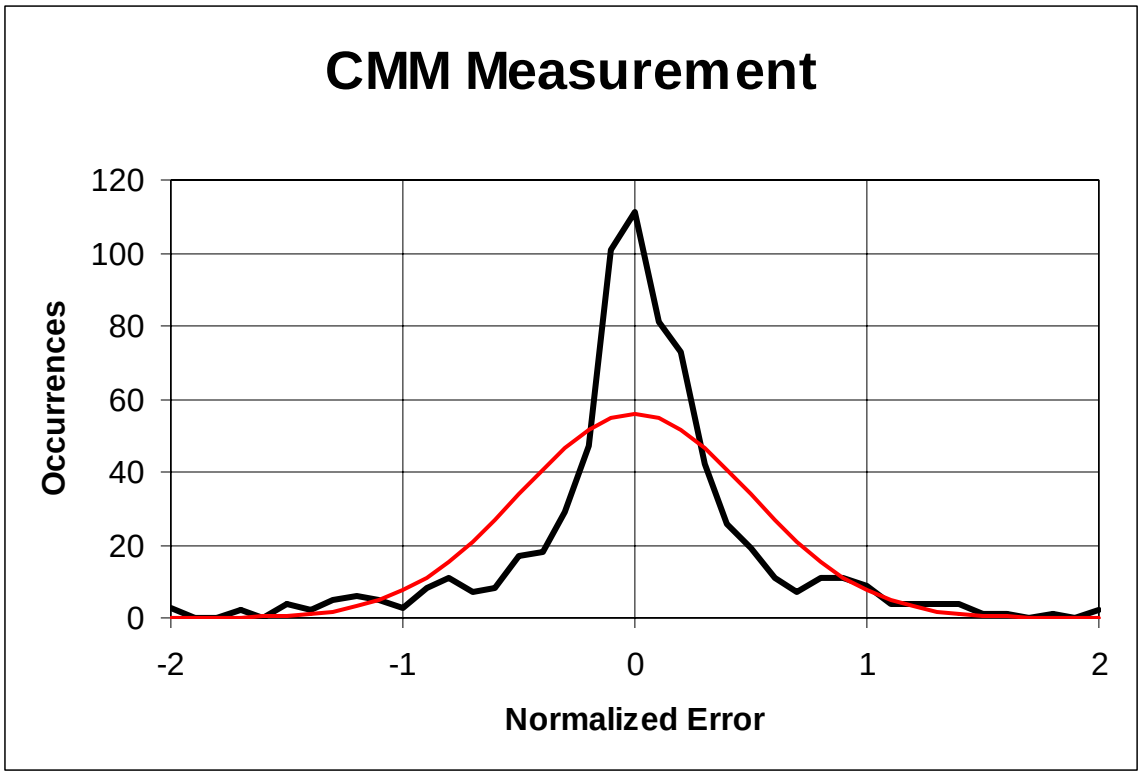

Figure 8: Normalized error (E_n value) distribution for CMM inspection test results (Black Line) and expected distribution (Red Line).

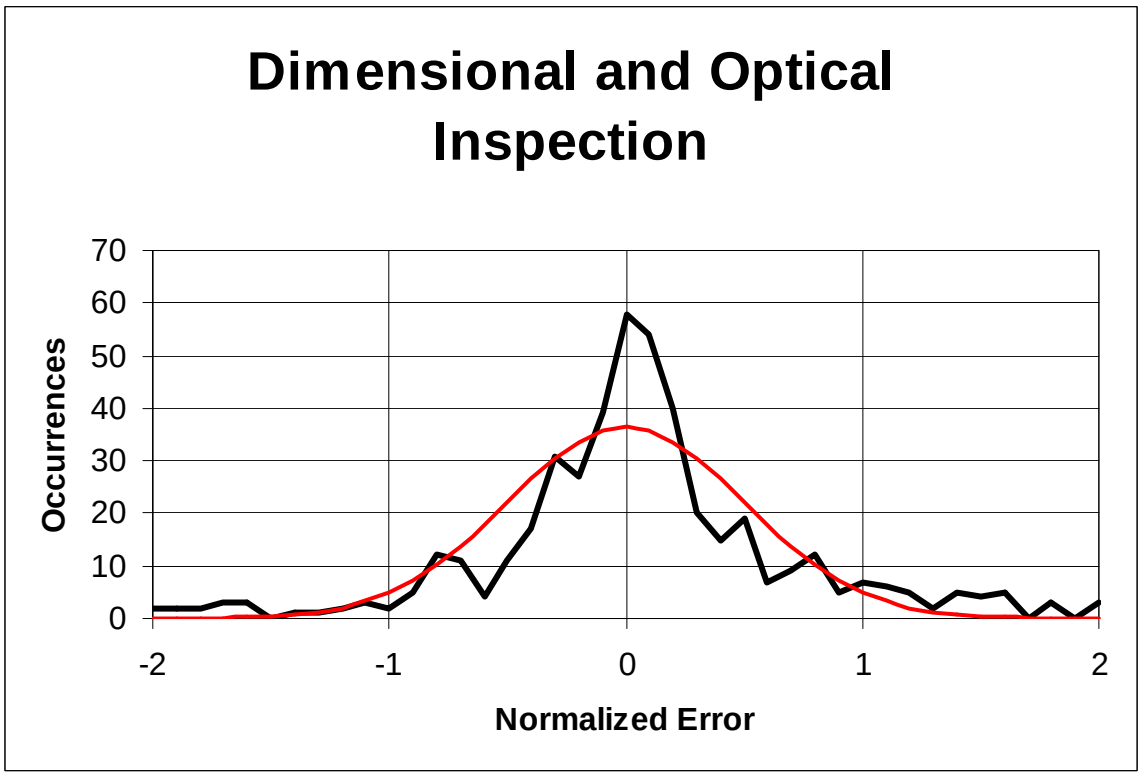

Figure 9: Normalized error (E_n value) distribution for dimensional and optical inspection test results (Black Line) and expected distribution (Red Line).

4. Adding value

The bulk of this paper has focused on statistical evaluation of relatively large data sets. The intent has been to give an overview of the state of accredited laboratories in the United States and their ability to measure up to their claims. The purpose of the requirement imposed by accreditation bodies for accredited laboratories to participate in proficiency testing is to be able to do this kind of analysis on a laboratory by laboratory basis.

However, proficiency testing has other effects and benefits the usually only become apparent to a laboratory after they have participated. It is usually the laboratories that fail a test that benefit the most, because the proficiency testing either unveils a hidden problem in their measurement process or lets them know that there is something missing in their uncertainty analysis. A good proficiency testing provider with a knowledgeable technical advisor assigned to each test can usually held a laboratory troubleshoot failures and covered by the test. Over the last three years HN Proficiency Testing has helped laboratories identify issues such as:

- Wrong measurement pressure
- Unevenly worn anvils
- Reverse polarity of corrections
- Outdated calibration certificates
- Misalignment of comparator optics

- In addition to general uncertainty analysis problems.

What all these issues have in common is that they are virtually impossible to detect, unless you are either specifically looking for them (which laboratories generally do not too) or you are comparing your results to those of another laboratory, which laboratories rarely, if ever, do unless they are in a crisis or participating in a proficiency test. None of the laboratories we have helped thought they had a problem when they first signed up for the proficiency test. It was only because their accreditation body forced them to participate that they signed up for the test in the first place.

However, it is my impression that more and more laboratories realize the value of proficiency testing as an ongoing reality check and an insurance policy against deceiving yourself in your evaluation of your own measurement capabilities.

5. Conclusions

Over the last three years HN Proficiency Testing has collected test data primarily from accredited laboratories. The primary metric for these tests has been the E_n value. Since the uncertainty values used for the calculation of the E_n value are based on 95% coverage, a failure rate of 5% should be expected. The actual failure rates for the tests are summarized in table 1. The overall failure rate is a little more than three times the expected rate, indicating that there is still room for improvement in the evaluation of uncertainty amongst calibration and inspection laboratories.

Test Group	Failure Rate
Digital Multimeter Calibration	1.6 %
Gage Block Calibration	4.2 %
Hand Gage Calibration	7.5 %
CMM Measurement	9.1 %
Hard Gage Calibration	15.6 %
All Tests	16.3 %
All Tests Except Hand Gages	17.5 %
Temperature Simulation	25 %
Dimensional and Optical Inspection	33 %

Table 1: Summary of failure rates for the test groups discussed in this paper.

However, when you look at the results in terms of success rate, the expected rate is 95% in the actual success rate is about 84%. While the still suggests that accredited laboratories on average underestimate their uncertainty by 40%, there is no reason to believe that non-accredited laboratories would have fared any better, or indeed nearly as well, as it is only the accreditation process that forces laboratories to critically evaluate their measurement processes and their uncertainties. Accreditation and proficiency testing has brought the laboratory community a long way towards realistic evaluation of the capabilities of measurement processes. The goal is not reached yet, but we are much closer than we were five years ago.