Adding Value with Proficiency Testing

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

1. Abstract

Accreditation bodies are increasingly using proficiency testing as a tool to ensure the credibility of their accreditation programs by requiring the laboratories they accredit to demonstrate that they can live up to their uncertainty claims in interlaboratory comparisons. Accredited laboratories mostly see proficiency testing as an added expense they are forced to incur which adds little or no value. However, when used appropriately, proficiency testing can reduce a laboratory's risk of producing incorrect measuring results. While focusing on the En (normalized error) approach, the paper explores the underlying assumptions and associated limitations in various reporting methods traditionally used in proficiency testing. It discusses the important steps that are necessary to ensure that correct conclusions are drawn from a proficiency test and the exposure and potential unnecessary cost participating laboratories are subject to, if these steps are not taken. Additionally, the paper covers some personal experiences, where the author has gained valuable knowledge of measuring processes and their limitations as a participant in interlaboratory comparisons.

2. WECC M 13

My first introduction to proficiency testing or interlaboratory comparison was the WECC M 13 European intercomparison for surface finish conducted in the early 1980s. I learned some valuable lessons on keeping good records of the measurements involved and using the results for improvement purposes. A typical set of results are shown in figure 1.

In the field of surface finish, at least at that time, the German National Laboratory, PTB was a recognized leader.

It was found in a majority of the results that the Institute for Process Technology & Institute for Product Development (PI/IPU), with which I was affiliated, and PTB had much better agreement, than what would be expected, given our respective stated uncertainties. This was especially interesting, since the equipment used was of different brands and quite different in design. The equipment used at PI/IPU is described by De Chiffre and Strøbæk Nielsen [1].

It was also found that in some cases our measurements diverged greatly, see figure 2.

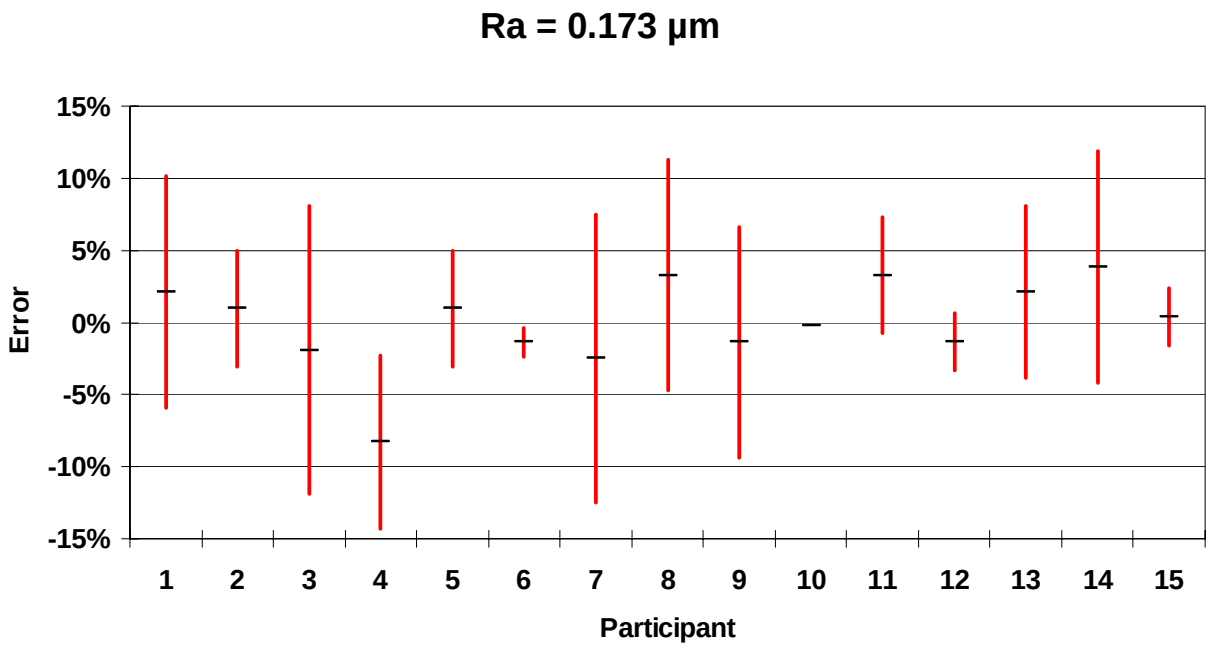

Figure 1: *Typical results from the WECC M 13 surface finish measurements sponsored by the European Commission in the early 1980s. Participant 1 is the German National Laboratory PTB and Participant 8 is PI/IPU, the organization the author was affiliated with at the time.*

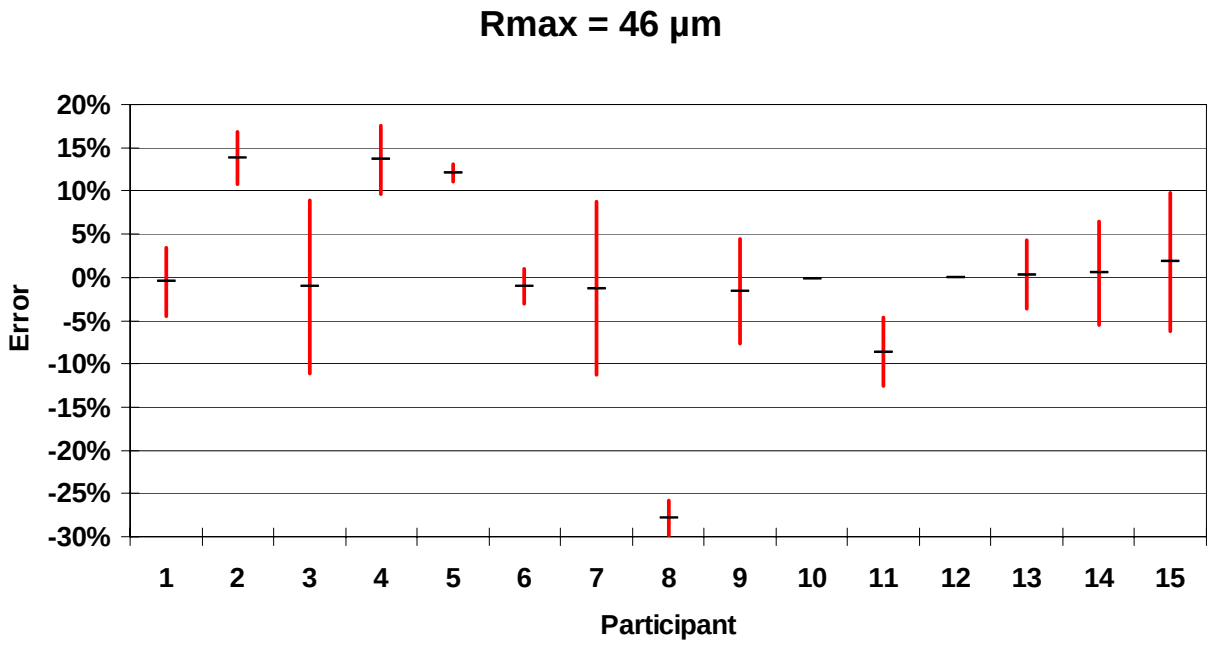

Figure 2: *Results from the WECC M 13 intercomparison. These results showing a large discrepancy by PI/IPU (Participant 8) compared to other participants including PTB (Participant 1).*

The divergent result in figure 2 turned out to be a blunder – a wrong filter setting had been used. When this was corrected, these results agreed well with the other participants' results.

Subsequent research showed that it was essential to keep key measurement parameters, e.g. probe tip radius, traversing speed and sample density constant if one is to obtain consistent results.

The purpose of relating this experience is to emphasize that it is essential to document the measurements, to gain the maximum benefit from participating in intercomparisons. I did not personally perform the measurements, but because of the good records, I was able to reconstruct the measurements, make systematic changes, measure the results of the changes and finally develop a procedure which would ensure consistent results, if followed.

Subsequently, I changed jobs, but even with the system of my new employer, Cummins Engine Co., which was different from what I had worked with before, I was able to make successful intercomparisons with PTB, by applying the procedure developed. Eventually the Cummins Corporate Standards Laboratory was accredited for surface finish measurement by DKD, the German Calibration Service with a lower best measurement uncertainty than anybody else had been granted by DKD at the time.

The work inspired by the results of the WECC M13 also had a great influence on the current generation of ISO surface finish standards, published in the early 1990s. In these standards all the important measuring parameters are identified and rules for setting these parameters are given.

3. Proficiency Test Design Considerations

A proficiency test or laboratory intercomparison has to be designed correctly in order for it to yield the maximum amount of information and in order to enable the participants to analyze their results and use them for improvement purposes.

The term proficiency testing is used to cover two very different processes. One is generally used in the testing community, where a set of similar samples are prepared and one or more samples are sent to each participating laboratory. In this process one of the key considerations is to ensure the homogeneity of the samples - that all samples are essentially the same.

The other process is primarily used in the calibration community where one (set of) artifact(s) is circulated amongst the participating laboratories. In this process the key concern is that the artifacts are stable throughout a testing round. This process is often referred to as interlaboratory comparison.

The remainder of this paper will focus on the second process.

3.1 Artifacts

A primary concern is to ensure the stability of the artifacts. If the artifacts change during a testing round, the results from the participating laboratories will not be comparable.

It must also be decided whether to use "perfect" artifacts that generally are easier to measure and will allow laboratories to measure close to their best capability, or to use "normal" artifacts that are harder to measure and require the laboratories to recognize artifact imperfections and take them into account when quoting their uncertainty.

Since "perfect" artifacts also tend to be more stable and present more well defined measurands, these tend to be preferred for interlaboratory comparisons. However, it must be recognized that in using such artifacts, the ability for a laboratory to correctly account for the shortcomings of "normal" artifacts is not tested. Consequently, a laboratory that indiscriminately quotes its best measuring capability for all calibrations may come out well in such an intercomparison, but grossly understate its uncertainty in day to day calibration work.

3.2 Instructions

The next consideration is the instructions to the participating laboratories. The instructions have to very clearly define what is to be measured i.e. the measurands of the intercomparison. For example in the case of a gage being the object of the intercomparison it must be clear whether the participants are to supply a known input to the gage and read its indication or whether they are to provide the input that gives a specified indication. Generally speaking the sign of the deviation changes between these two scenarios.

Beyond clear definitions of the measurands, the more detailed and prescriptive the instructions, the better the chance that participating laboratories get similar results.

However, with very prescriptive instructions, the laboratories' normal procedures are not tested and – as with the perfect artifacts – the intercomparison may not give a true picture of the participating laboratories abilities to generate consistent results.

Consequently, to get a good picture of how well the participating laboratories agree when applying their day to day procedures, the instructions should give clear definitions of the measurands but leave the details of how the measurements are performed up to the participants.

3.3 Reference Values

Reference values (the "true" values) and uncertainties can be determined in two different manners. One is to employ a reference laboratory; the other is to use a consensus value, derived from a weighed average of the participants' results.

The ideal situation is where a reference laboratory is available, which can provide values with a low, but verifiable uncertainty. The advantage of using a reference laboratory is that each participant can receive instant feedback on their results without having to wait for the

intercomparison round to finish. The exposure is that even good, reliable reference laboratories make mistakes sometimes and some participants may falsely be deemed to have failed the intercomparison.

Using a consensus value has the problem that all the participants may agree, but they may all be wrong. For internal intercomparisons, e.g. between laboratories within a company or between laboratories using the same procedures or which have received their training from the same source, this is a significant risk. For comparisons between unrelated laboratories, this risk is smaller.

HN Proficiency Testing uses a combination of these two approaches to be able to provide instant feedback to participating laboratories, while having checks in place to ensure that the reference laboratory did not make any mistakes. First the individual participants' results are compared to those of a reference laboratory. At the end of each round a consensus value is calculated based on a weighed average of what is considered "reliable" participant results. A reliable result is defined as one that contains the median value of all reported results within its uncertainty range. The weight assigned to each result is based on the reported uncertainty. Reliable results with a low stated uncertainty receive a higher weight than those with a high stated uncertainty.

The consensus value generally will have a lower uncertainty than the reference laboratory value, so it often represents a more stringent criterion than the comparison to the reference laboratory's value.

Finally, the reference laboratory's value is compared to the consensus value to detect any problems with the reference laboratory value.

For a discussion of these algorithms, see Nielsen [2]

3.4 Calculations

ISO Guide 43-1 [3] gives two basic measures for evaluating the results of proficiency testing, the E_n-value and the z-score. The E_n-value approach requires each laboratory to report an uncertainty. This is the most suitable measure for interlaboratory comparisons.

The z-score approach does not require a reported uncertainty from each participant, but is based on the assumption that all the measured values are part of the same population – that they have the same uncertainty. Consequently, the z-score tends to be most useful where all participants use similar methods and where uncertainty estimation is difficult or impossible. In practice the z-score approach is most useful and meaningful for chemical or biological analysis.

Related to the z-score is the Youden plot, which is also based on the assumption that all the measured values are part of the same population and also does not require the participants to explicitly state an uncertainty of the measured results. The Youden plots are used to attempt to identify significant bias (and thus opportunities for improvement) within participating laboratories. While the basic approach is very sound, there are significant assumptions and

conditions that have to be validated and fulfilled, before meaningful conclusions can be drawn, see Youden [4] and [5].

3.4.1 z-score

The formula for the z-score is:

$$z = \frac{Value_{Lab} - Average}{s_{Population}}$$

Where:

$Value_{Lab}$ is the value reported by the individual laboratory
Average is the average of all participants' values

and

$s_{Population}$ is the standard deviation of all participants' values

Results are judged as follows:

$|z| \leq 2$ is satisfactory
$2 < |z| \leq 3$ is questionable
$|z| > 3$ is unsatisfactory

3.4.2 E_n-value

The formula for the normalized error or E_n-value is:

$$E_n = \frac{Value(Lab) - Value(Ref)}{\sqrt{U(Lab)_{95}^2 + U(Ref)_{95}^2}}$$

Where:

Value(Lab) is the value reported by the participating laboratory
Value(Ref) is the reference value for the measurand
$U(Lab)_{95}$ is the uncertainty reported by the participating laboratory
$U(Ref)_{95}$ is the uncertainty of the reference value

Results are judged as follows:

$-1 \leq E_n \leq 1$ is satisfactory
$E_n > 1$ or $E_n < -1$ is unsatisfactory

3.5 Quality Assurance

There are many elements to quality assurance of the results of proficiency testing. Most of these are common to all data based activities, such as review of data transfers etc, but some elements are unique to interlaboratory comparisons. These elements are aimed at ensuring that the reference value is correct and that the artifact(s) is stable.

Typically the reference laboratory measures the artifact(s) before and in some cases after each testing round. The artifacts travel in a logical circular pattern from the reference laboratory to the first participant, then to the second participant and so on, until it is finally returned to the reference laboratory.

If there is some concern that the artifact(s) are not stable, a star pattern can be used. In the star pattern the artifact(s) is returned to the reference laboratory between each (or every few) participant. Obviously, the star pattern is more time consuming and more expensive to employ than the circular pattern, but it provides a better assurance of detection of artifact instabilities.

HN Proficiency Testing uses a circular pattern for all tests currently offered, but uses some quality assurance techniques to ensure that the artifact(s) have remained stable and that the reference value is correct within its stated uncertainty. Figure 3 shows data from an imaginary testing round, where the majority of the participants appear to have failed the test. In order to ensure the anonymity of the participants, the sequence of the participants in the official final report is randomized.

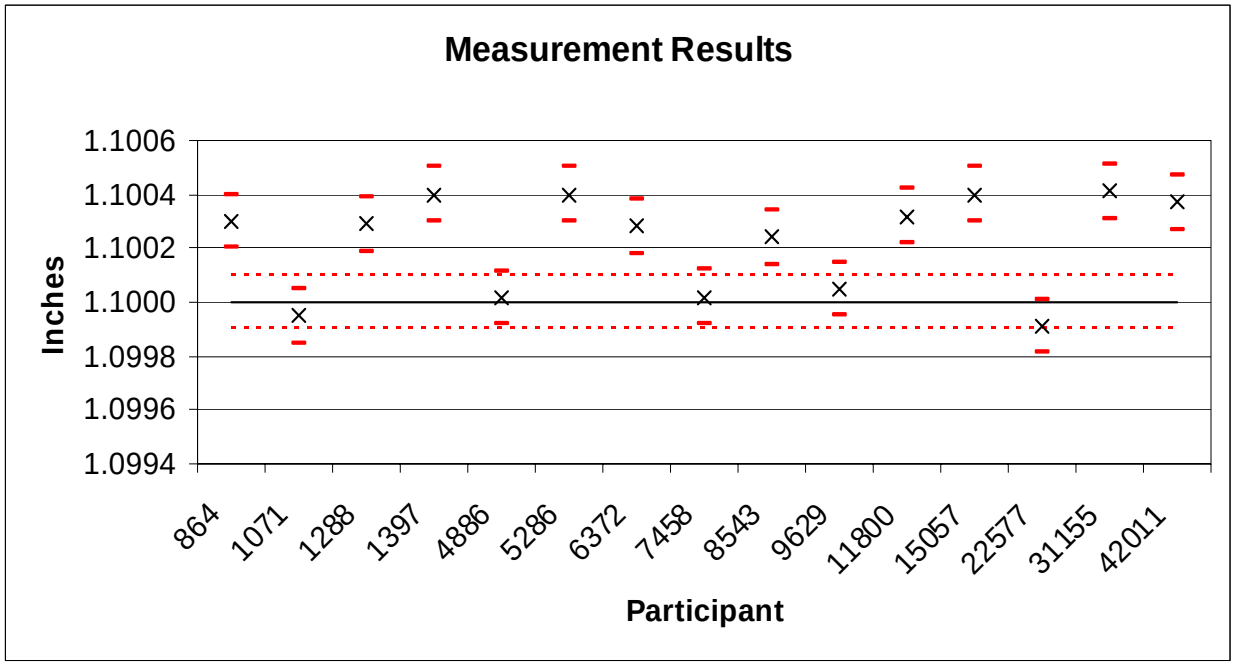

Figure 3: *Plot of results of an imaginary testing round. The X'es indicate each participants' measured value and the associated horizontal lines indicate the reported uncertainty. The dashed lines represent the uncertainty interval for the reference value. Data is sorted by participant ID, which is a randomly generated number, to ensure the anonymity of the participants. 10 participants appear to have failed the test.*

For quality assurance purposes HN Proficiency Testing generates another plot, see figure 4, where the data is sorted in chronological order. This report is only used internally, as it can compromise the anonymity of the participants. Figure 4 clearly shows a shift in values between the 5th and the 6th participant. In this case corrective action would have to be initiated.

However, if the quality assurance plot did not show such a shift, see figure 5, then it cannot be concluded that the problem is a change in the artifact. In this case, the analysis of the results against a weighed average may show that the reference laboratory was not correct, see figure 6.

It should be pointed out that these imaginary results are simplified somewhat for illustrative purposes, as it appears that there are two distinct measured values possible. If this happened in a real testing round, it would raise the suspicion that there are two different techniques possible for measuring the artifact and that these yield different results.

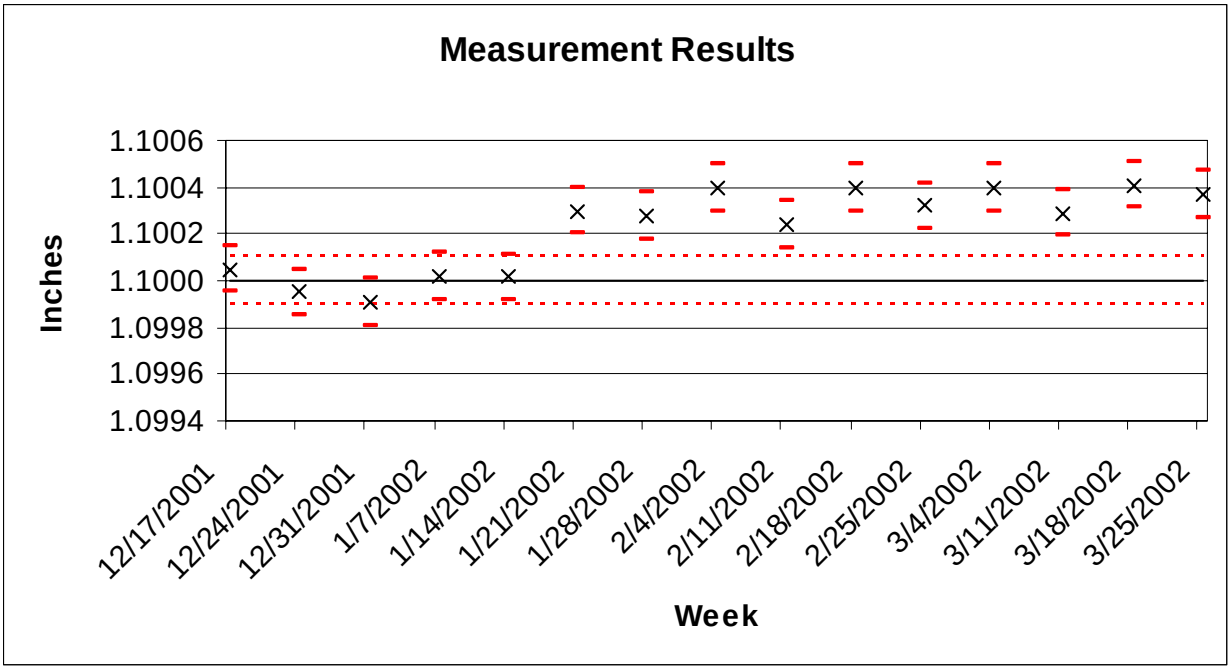

Figure 4: *The same data as given in figure 3, but plotted in chronological order. Scenario 1: The artifact has changed during the testing round.*

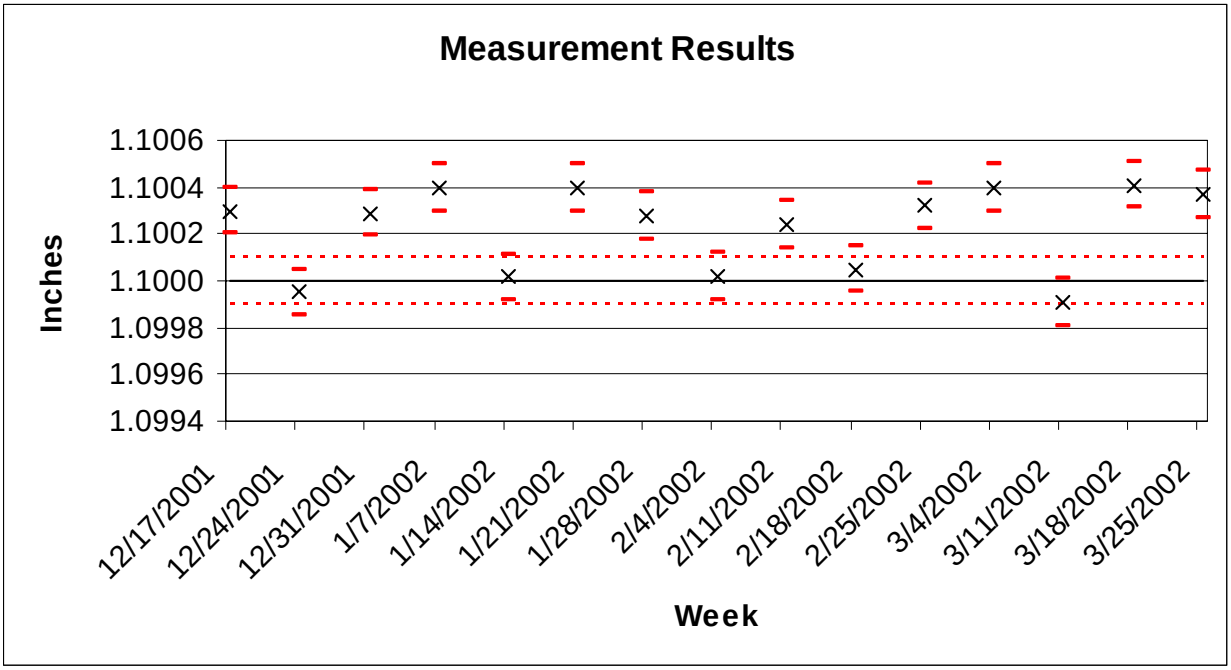

Figure 5: *The same data as given in figure 3, but plotted in chronological order. Scenario 2: The artifact has not changed during the testing round.*

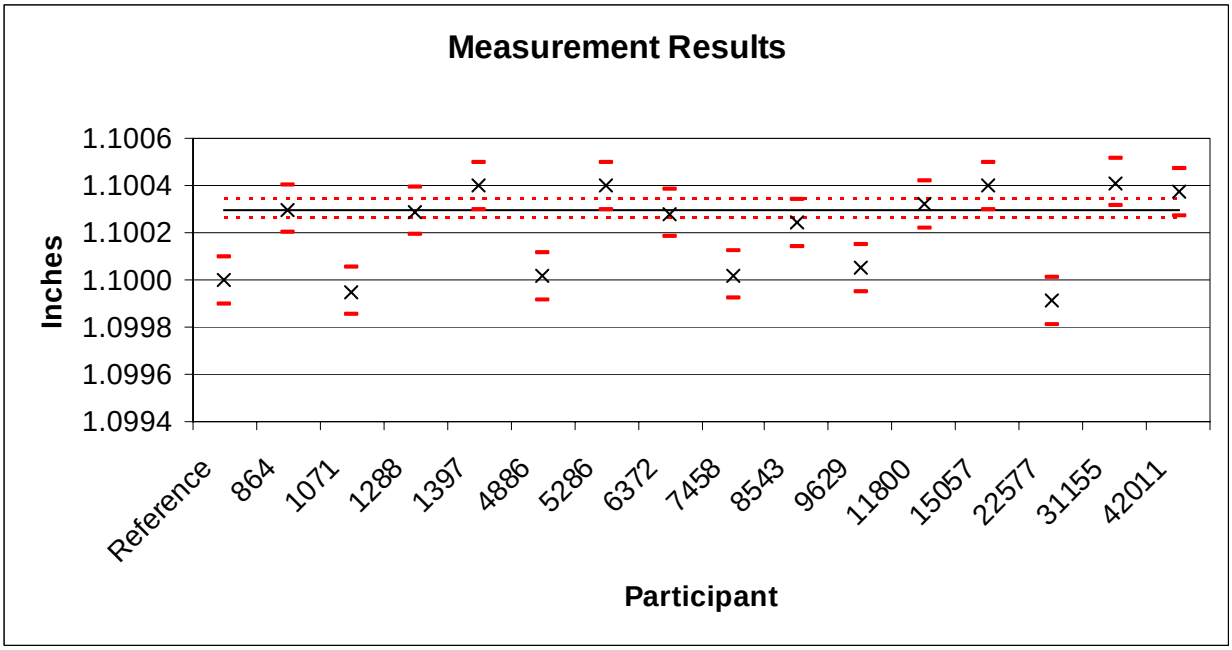

Figure 6: *Plot of results of an imaginary testing round. The X'es indicate each participants' measured value and the associated horizontal lines indicate the reported uncertainty. The dashed lines represent the uncertainty interval for the weighed average of the reliable measurements. Data is sorted by participant ID, which is a randomly generated number, to ensure the anonymity of the participants. The reference laboratory and 5 participants appear to have failed the test.*

4. Evaluating Your Results

Participating laboratories must evaluate their results – both good and bad – to get the maximum benefit out of proficiency testing.

Accreditation bodies generally require accredited laboratories to participate in proficiency testing within their scope and initiate corrective action for any unsatisfactory result.

4.1 Unsatisfactory Results

Unsatisfactory results can be split up into those caused by errors in the measurement and those caused by a too optimistic uncertainty estimate.

Errors in the measurement include blunders, such as adding corrections instead of subtracting them (or vice versa), wrong instrument settings etc. Better procedures with clearer instructions or education and training are the typical corrections for these issues.

A too optimistic uncertainty estimate either means that some uncertainty contributors are underestimated or missed altogether or that there are mathematical errors in the uncertainty estimate.
It generally requires significant knowledge and experience to distinguish between these two error modes. One has to know what a typical and reasonable uncertainty is for the measurements in question and one has to understand the measurements well enough to recognize the symptoms of common mistakes. Having a knowledgeable technical advisor associated with each proficiency testing scheme is essential for this activity.

4.2 Minimum Uncertainty

For failed results, where it has been determined that the cause was not measurement error, the minimum uncertainty that would have yielded a satisfactory result can be calculated by the following formula:

$$U_{\min} = \sqrt{Error^2 - U_{Ref}^2}$$

Where:

U_{min} is the uncertainty that would have resulted in an E_n-value of 1
Error is the difference between the laboratory's value and the reference value
and
U_{Ref} is the uncertainty of the reference value.

Since the uncertainties used in the E_n-value calculations are expressed at a 95% coverage level, E_n-values outside +/-1 should be expected in 5% of the cases with underestimation of the uncertainty up to on the order of 20 % – 30 %.

4.3 Uncertainty Estimates

When sufficient data has been collected, by participation in a number of interlaboratory comparisons, a laboratory can evaluate the general validity of its uncertainty statements.

If it is assumed that:

- Reference laboratories quote realistic uncertainties
- Reference laboratories' errors follow a normal distribution
- The participating laboratory's errors follow a normal distribution
- The reference laboratory's measurement and the participating laboratory's measurement are independent

Then the long term average absolute En value will be about 0.4, if the laboratory quotes the correct uncertainty for their measurements.

5. Lessons Learned

From my own experiences participating in interlaboratory comparisons and my experience from offering proficiency testing on a commercial basis, I have learned several lessons, the main one being that the more information you have available, the easier it is to establish what happened, if the results are not satisfactory.

Therefore it is important to keep very good notes on how the measurements were performed. Video can be a great memory aid in this case. It is also important to keep all data, including raw data and to document each step and each calculation that take place between the raw data and the reported result.

Without this information it is impossible to go back and check instrument settings and calculations, which are the root cause of most of the unsatisfactory results I encounter.

6. Conclusions

This paper has explored some of the design considerations for proficiency tests and interlaboratory comparisons, which should be employed to ensure that maximum value can be gained from the schemes. It should be clear from this discussion that the design of the test, both in terms of artifacts and instructions to the participants are paramount for the overall usefulness of the testing scheme. The only way to ensure that these considerations are made during the design phase is to involve somebody with a significant expertise in the measurements in question.

Even a well designed proficiency testing scheme is of little or no value, if the results are not reliable. Therefore it is important to use a reference laboratory with good reliability (if a reference laboratory is used), which means that the reference laboratory has to quote a realistic uncertainty for its measurements and have a low occurrence of blunders in its measurements. Data analysis techniques should be implemented to ensure that the reference values are indeed correct and that the test artifacts have remained stable throughout the testing round.

The design of the test and the reliability of the reference values are the two key items for ensuring that no participant results are falsely deemed unsatisfactory. The effort, expense and aggravation involved in implementing unnecessary corrective action far outweighs the cost of participating in proficiency testing in the first place.

Curiously enough, proficiency testing often adds most value when it identifies unsatisfactory results, than when all results are satisfactory, as was seen by my experience with WECC M 13.

In order to benefit from proficiency testing in this case by diagnosing what went wrong so procedures can be corrected and future measurements improved, it is imperative that participating laboratories keep good notes of their measurements, record all instrument settings, raw data etc. and document all calculations and data transfers.

At this stage it will again be beneficial to have somebody with expertise in the measurements associated with the testing scheme. Often such a person can diagnose a problem with a glance at the data presented, where it may look like "random error" to somebody less familiar with the particular measurements.

Proficiency testing is often the quickest and least expensive way to ensure that the results generated by a laboratory are correct and, if not, to diagnose what must be changed to make the results correct in the future. This is how competently administered proficiency testing adds value to the measuring and calibration community.

7. References

1. De Chiffre, L.; Strøbæk Nielsen, H.: A digital system for surface roughness analysis of plane and cylindrical parts. Precision Engineering. 1987, 9 (2), 59

2. Nielsen, H.S.: Determining Consensus Values in Interlaboratory Comparisons and Proficiency Testing. Winner of "Best Paper on Theoretical Metrology" - Proceedings of the 2003 NCSL-I Workshop and Symposium.

3. ISO Guide 43-1:1997 Proficiency testing by interlaboratory comparisons – Part 1: Development and operation of proficiency testing schemes.

4. Youden, W. J.: Graphical Diagnosis of Interlaboratory Test Results. Industrial Quality Control, Vol. XV, No. 11, May 1959.

5. Youden, W. J.: The Sample, The Procedure, and The Laboratory. Analytical Chemistry, Vol. 32, No. 13, December 1960.