Determining Consensus Values in Interlaboratory Comparisons and 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

An important part of interlaboratory comparisons and proficiency testing is the determination of the reference value of the measurand and the associated uncertainty. It is desirable to have reference values with low uncertainty, but it is crucial that these values are reliable, i.e. they are correct within their stated uncertainty. In some cases it is possible to obtain reference values from laboratories that reliably can produce values with significantly lower uncertainty than the proficiency testing participants, but in many cases this is not possible for economical or practical reasons. In these cases a consensus value can be used as the best estimate of the measurand. A consensus value has the advantage that it often has a lower uncertainty than the value reported by the reference laboratory. There are well known and statistically sound methods available for combining results with different uncertainties, but these methods assume that the stated uncertainty of the results is correct, which is not a given. In fact, the very purpose of proficiency testing is to establish whether the participants can measure within their claimed uncertainty. The paper explores a number of methods for determining preliminary consensus values used to determine which participant values should be deemed reliable and therefore included in the calculation of the final consensus value and its uncertainty. Some values are based on impressive equations and others have curious names. The relative merits of these methods in various scenarios are discussed.

2. The Purpose of Proficiency Testing

As part of their quality assurance programs, accreditation bodies normally require accredited laboratories to participate in proficiency testing. There are two broad categories of proficiency testing or interlaboratory comparisons, one where a set of artifacts is sent around to all participating laboratories and one where a specimen or sample is split up and one piece is sent to each participating laboratory. HN Proficiency Testing offers the former kind of proficiency tests and this paper is based on experiences with this kind of tests. However, the techniques described herein and the conclusions are valid for either kind of tests.

The purpose of proficiency testing is to ensure that the participating laboratories can make reliable measurements. A measurement can be unreliable in two different ways. It can contain a blunder or other serious error, which makes it atypical for what would be expected from the laboratory, or the uncertainty of the measurement can be underestimated, such that the error in a correctly performed measurement is larger than the uncertainty stated by the laboratory.

A measure of the quality of the design of a proficiency test is how well it can distinguish between reliable and unreliable measurements. The width of the gray zone between reliable and unreliable measurements as judged by the proficiency test can be thought of as the uncertainty of the proficiency test. To minimize the uncertainty of the proficiency test it is necessary to have a reference value that is reliable (per the definition above) and has a low uncertainty.

Unless the reference laboratory that can measure with a much smaller uncertainty than the test participants can be identified, a (weighted) average value will generally be more reliable, i.e. be less likely to be affected by the influence of mistakes or blunders, than a value produced by an individual laboratory. Such a value will also have a lower uncertainty than the least uncertain measurement included in the average.

Since the very purpose of proficiency testing is to identify reliable results it is unreasonable to make a priori assumptions about the reliability of the participants' results. Therefore it is necessary to have a robust algorithm for determining which results to consider reliable and therefore include in the average and which to disregard.

3. Algorithms for Identifying Reliable Results

For a variety of reasons HN Proficiency Testing has been using what in our opinion are reliable, accredited, commercial laboratories as reference laboratories. These laboratories do not necessarily offer an uncertainty that is significantly lower than that of the test participants. We have therefore been using weighted averages as reference values for our tests whenever we have considered it prudent to do so based on the particular measurements and the number of participants.

We have been using a weighted average rather than an arithmetic average in recognition of the fact that there is a significant variation in the capabilities of the participating laboratories, thus giving higher weights to the laboratories that claim a lower uncertainty, as long as we deem their results to be reliable.

Another advantage of using a weighted average is that the uncertainty of the weighted average is easy to calculate, if one assumes that the measurements are independent, which is generally a reasonable assumption in these types of tests, as long as there are not a large number of results from one particular laboratory or from laboratories that are related e.g. by using the same, wrong procedure.

3.1 The Median Algorithm

Our first algorithm for determining reliability was fairly simple: We determined the median of all the measurement results in a test and all the results that contained the median value within their uncertainty range were deemed to be reliable and were thus included in the weighted average. We used the median rather than the mean value, as the median is more robust towards outliers.

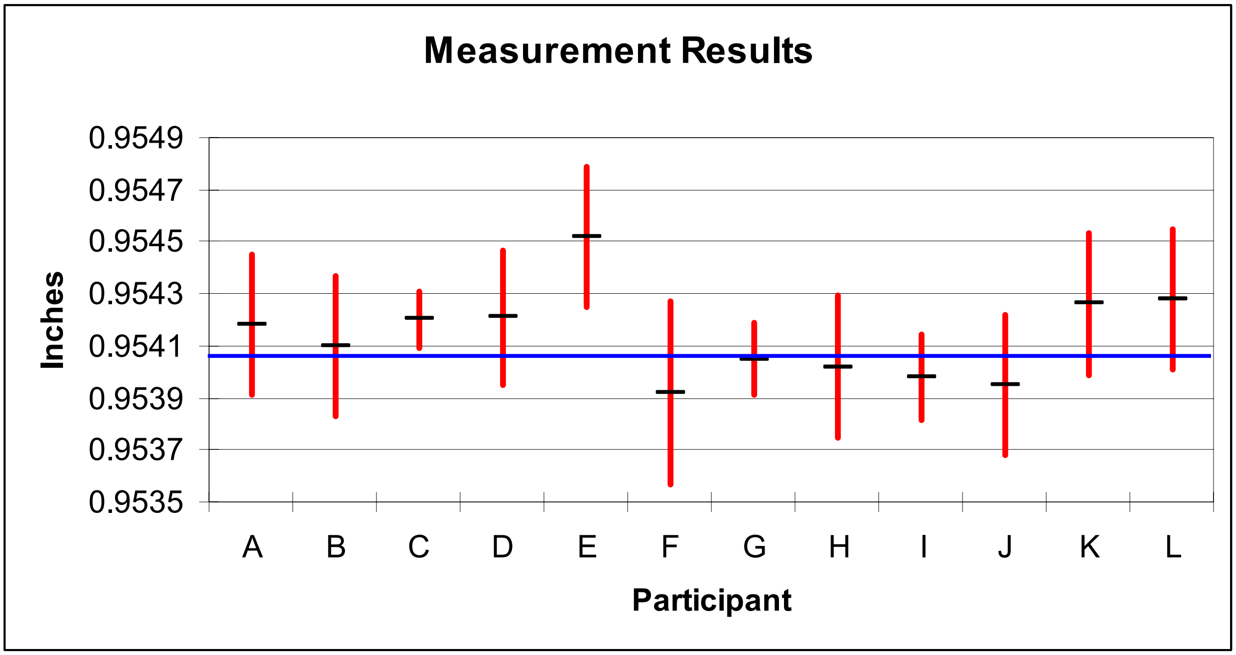

Figure 1: *Measurements from 12 laboratories. The vertical lines indicate uncertainty ranges. The bold horizontal line indicates the median value. The values from participants C and E are not considered reliable for the purposes of calculating the weighted average value, as their uncertainty ranges do not include the median.*

Figure 1 shows a typical set of measurements and their median value. As can be seen from the figure, the uncertainty ranges of participants C and E do not include the median value (the bold line). Therefore, the values from these to participants are not included in the weighted average.

It can be argued, that the results from participant C is probably reliable, but the criterion is designed to eliminate rather than include the results that are only probably reliable. As it turns out when the weighted average value and its associated uncertainty are calculated, the result from participant C is indeed acceptable based on the E_n calculation, but that is a different issue.

The algorithm works well on most data sets, but it turned out to have some weaknesses, see figure 2. The range of uncertainties quoted by the participants in this test is very wide with participants A, C and G quoting low but – as it turns out – realistic uncertainties and participants B, E, F and I quoting fairly high but – based on their results – also realistic uncertainties for this particular measurement. Note that the values of participants A, C and G correspond quite well to each other, given their uncertainties, but since they are in the minority in this case, their uncertainty ranges do not include the median value.

The weighted average of the values that include the median in their uncertainty range is 1.774 and the uncertainty of the weighted average is 0.00051. These particular measurements were in inches, but this is irrelevant to the analysis. When the results of the individual participants are compared to this weighted average, the results of participants A, C and G are found to have an En value outside +/-1 and they are thus deemed unacceptable.

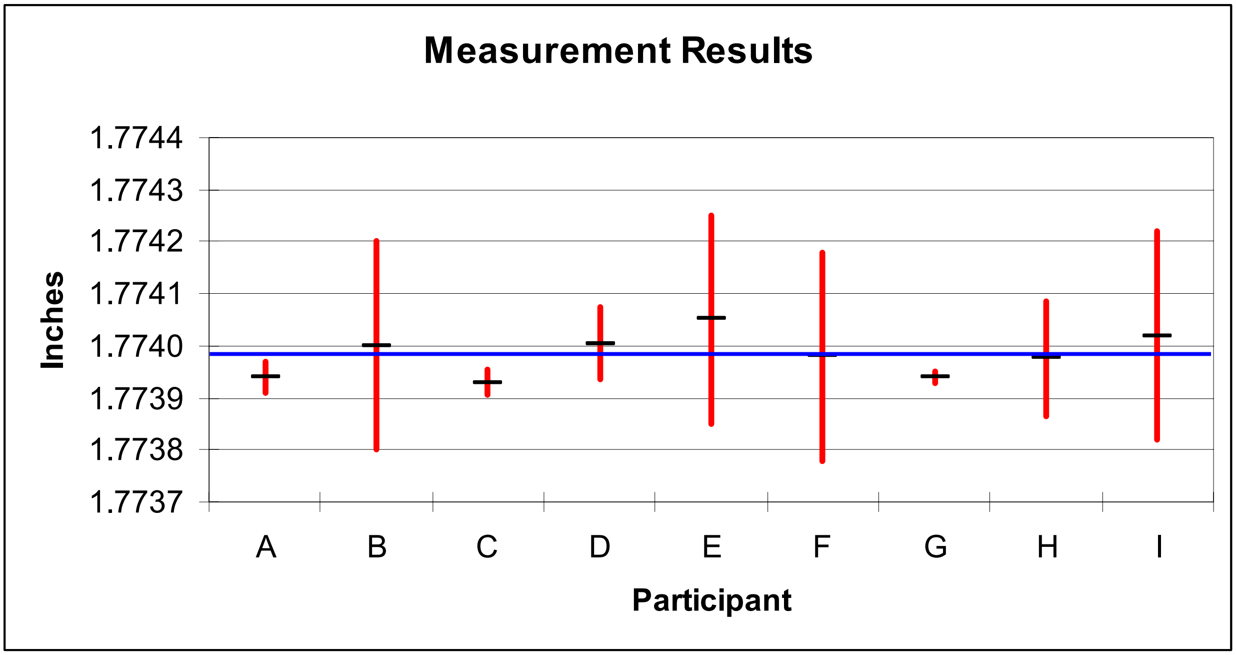

Figure 2: *Measurements from 9 laboratories. The vertical lines indicate uncertainty ranges. The bold horizontal line indicates the median value. The values from participants A, C and G are not considered reliable for the purposes of calculating the weighted average value, as their uncertainty ranges do not include the median. However, although they are in the minority, these 3 results agree well with each other and their stated uncertainty is within what is reasonable for the measurement in question.*

The median algorithm works well for eliminating results with unrealistically low uncertainty claims from the weighted average. It does not work well for data sets that include a few results with low, but realistic uncertainty claims amongst a majority of results was significantly higher uncertainty claims.

3.2 The Cumulative Probability Algorithm

Having identified this weakness in the median algorithm, it became clear that it would be necessary to develop a more robust algorithm. The median algorithm assigns the same weight to all results when determining the first estimate of the reference value (the median value). To correct for the behavior described above it would be necessary to assign a higher weight to results with low uncertainty claims when determining the first estimate of the reference value.

The cumulative probability algorithm does this by modeling each result as a normal distribution with a standard deviation equal to one half of the claimed uncertainty. This distribution represents the probability distribution for the true value, given the measured value and the claimed uncertainty. This is consistent with the assumptions in and the theory behind the ISO Guide [1], as participants are required to report their uncertainty as an expanded uncertainty using k=2.

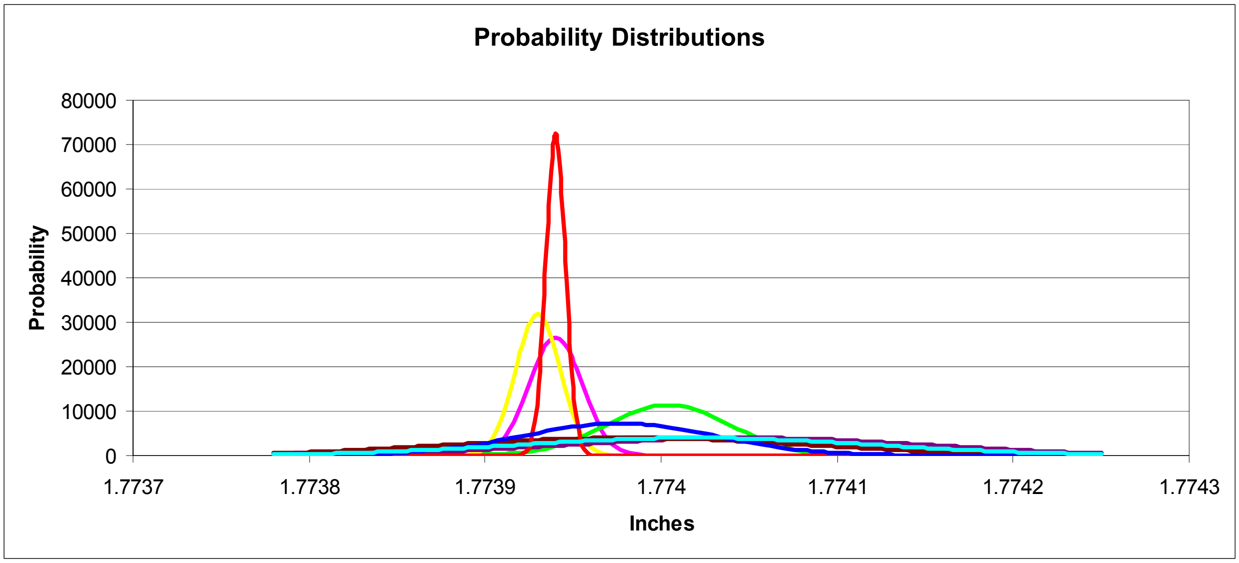

Figure 3: *The same measurements as in figure 2, represented by normal distributions with a standard deviation equal to their respective combined standard uncertainty (half their expanded uncertainty). The amplitude of the distributions are such that the area under each curve is unity.*

The cumulative probability distribution based on all the measured values is calculated using the following formula:

$$f(x) = \frac{\sum_{i=1}^{n} \left(\frac{1}{\sqrt{2\pi}\sigma_i} e^{-\left(\frac{(x-\mu_i)^2}{2\sigma_i^2} \right)} \right)}{n}$$

where:

f(x) is the cumulative probability distribution
n is the number of measurements/participants
μ_i is the value measured by the i'th participant
σ_i is the standard deviation (one half of the expanded uncertainty) associated with the value measured by the i'th participant

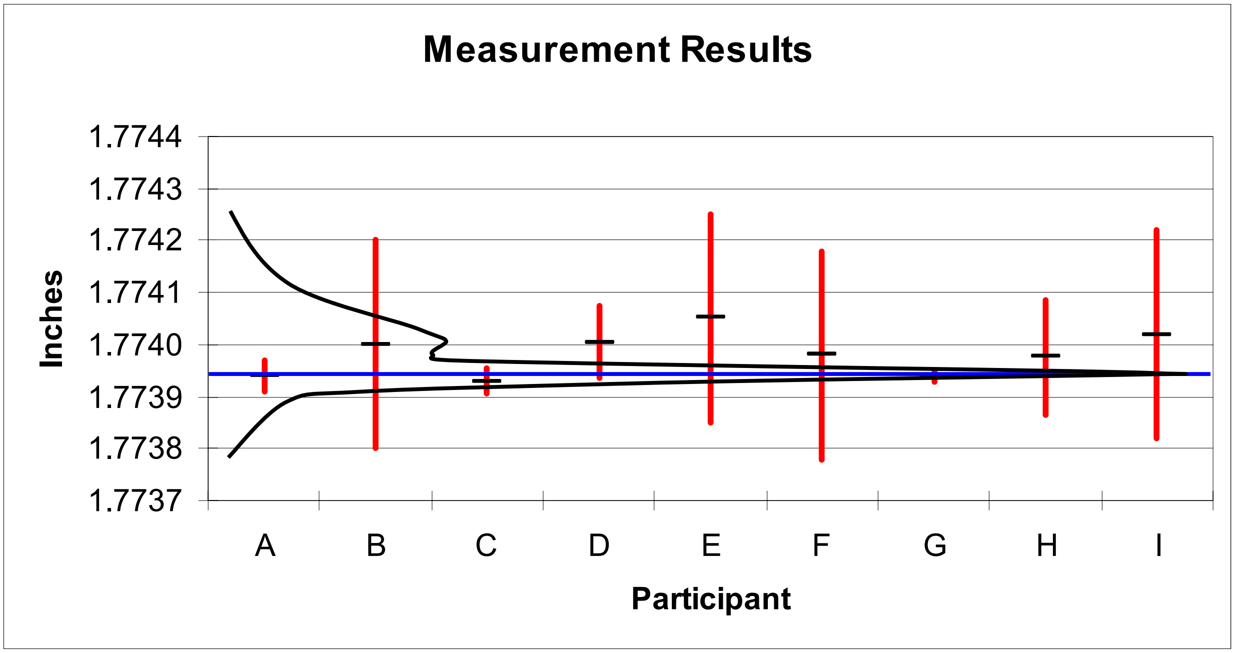

Figure 4: *The same measurements as in figure 2. The curve indicates the cumulative probability distribution and the bold horizontal line indicates the value with the highest cumulative probability, the first estimate of the reference value using this algorithm. As it can be seen, the cumulative probability distribution is dominated by the values with the lowest claimed uncertainties (Participants A, C and G). The value of participant G is obscured by the cumulative probability distribution curve.*

The cumulative probability algorithm works well for the data set used in figures 2, 3 and 4. However, it relies very heavily on the assumption that the measured values and the quoted uncertainties are correct, which is the very issue proficiency testing is intended to validate.

The weakness of the cumulative probability algorithm is illustrated in figure 5. The value of participant G has been changed to simulate an outlier with a very low quoted uncertainty. This is a quite common situation in proficiency testing. Because of the exponentially higher weight given to values are quoting low uncertainty, this outlier is able to "overpower" the rest of the measured values. It is clear from figure 5 that the cumulative probability algorithm is not a robust algorithm.

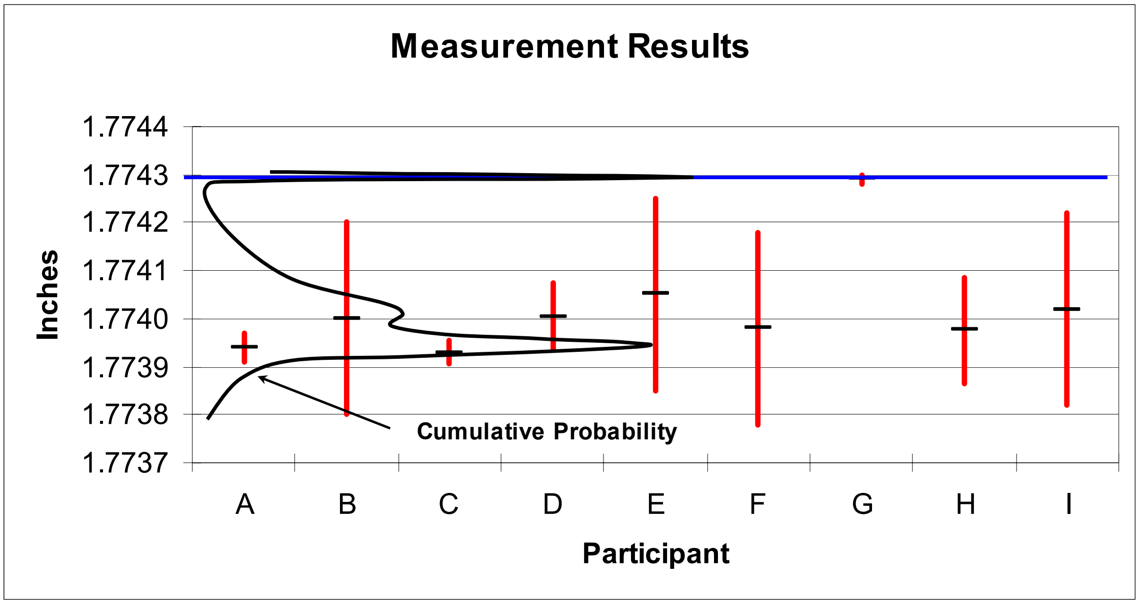

Figure 5: *The same measurements as in figure 4, except the value from participant G had been altered to simulate an outlier. The curve indicates the cumulative probability distribution and the bold horizontal line indicates the value with the highest cumulative probability at about 1.7743 inches. As it can be seen, the cumulative probability distribution is dominated by the values with the lowest claimed uncertainties (Participants A, C and G). The uncertainty claimed by participant G is one half of that claimed by participants A and C, which puts the peak of the cumulative probability curve at the value of participant G. The value of participant G is obscured by the cumulative probability distribution curve and the bold horizontal line.*

3.3 The "Value Voted Most Likely To Be Correct" Algorithm

The median algorithm does not consider the uncertainty claimed by the participants in the first estimate of the correct value. While it works well in most cases, this algorithm is not robust for data sets containing a few results with low, realistic uncertainty amongst a higher number of results with higher uncertainty and significant deviation.

The cumulative probability algorithm assigns significant weight to the uncertainty claimed by the participants in the first estimate of the correct value. Therefore this algorithm is not robust against outliers with a low claimed uncertainty.

Given the problems with these two algorithms it would appear that an ideal algorithm would add some weight to the uncertainty claimed by the participants in the first estimate of the correct value, but not as much as the cumulative probability algorithm does. The "Value Voted Most Likely to Be Correct" algorithm is one such algorithm.

Where the cumulative probability algorithm interprets the uncertainty claimed by the participants as a normal distribution, the Value Voted Most Likely to Be Correct algorithm interprets the uncertainty range of each participant as a modified rectangular distribution. The modification is that the distribution has a height of one, regardless of its width. Conceptually this is equivalent to saying that each participant gives one vote to each value within their uncertainty range and zero votes to values outside this range. By tallying the votes one can determine which value, or range of values, the highest number of participants considers a likely correct value.

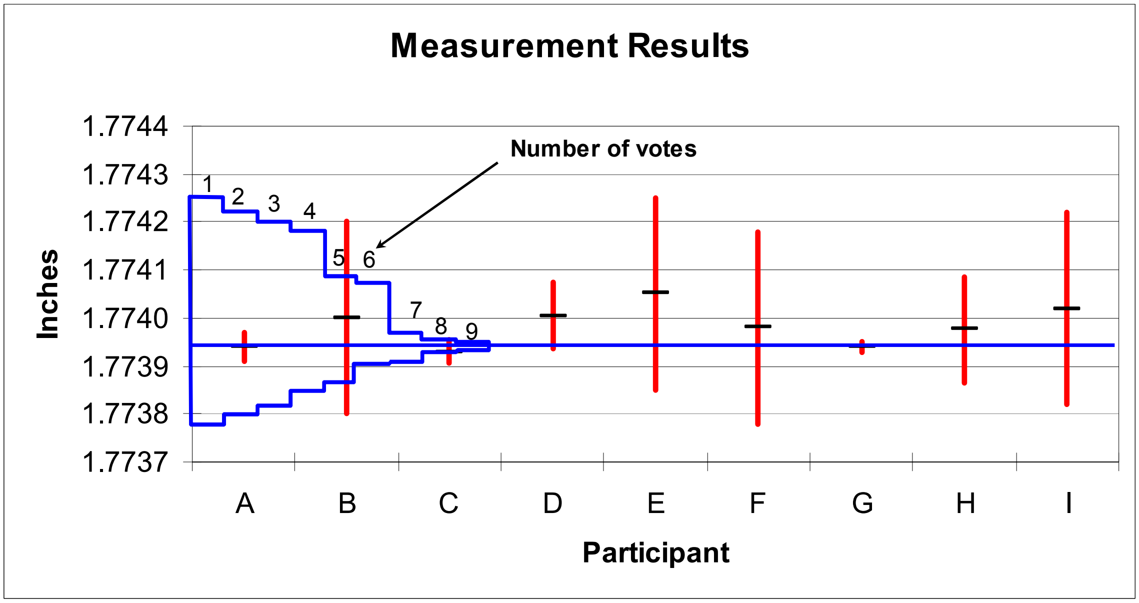

Figure 6: *The same measurements as in figure 2 and 4. The stepped curve indicates number of votes for each range of values and the bold horizontal line indicates the value with most votes, the first estimate of the reference value using this algorithm. As it can be seen, the algorithm finds the value most participants can agree on.*

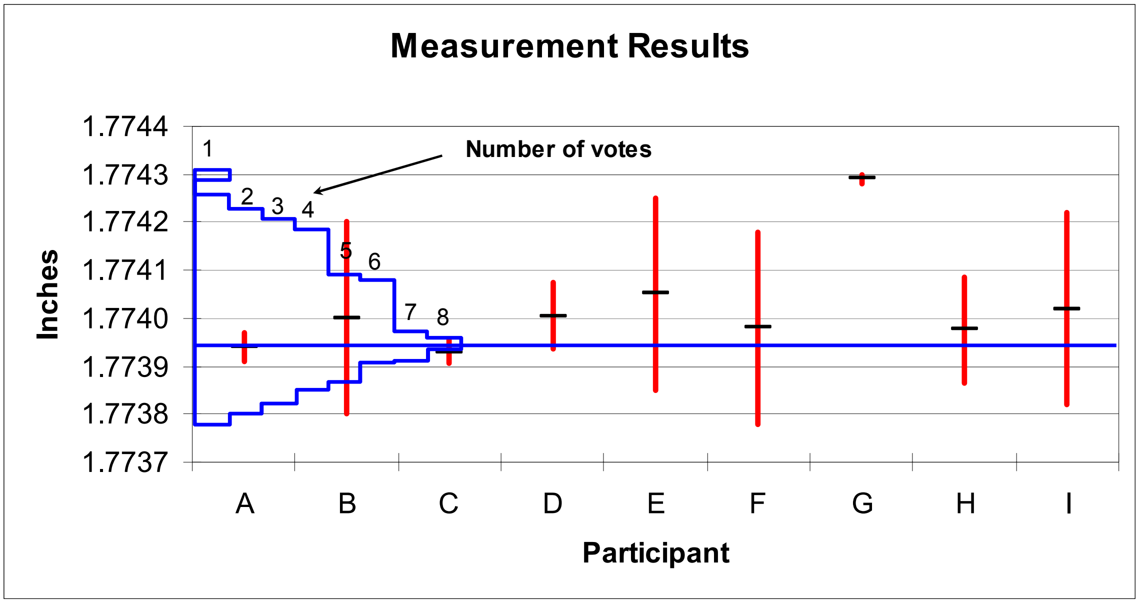

Figure 7: *The same measurements as in figure 5 with the result for participant G modified to simulate an outlier. The stepped curve indicates number of votes for each range of values and the bold horizontal line indicates the value with most votes, the first estimate of the reference value using this algorithm. As it can be seen, the algorithm is robust against outliers.*

Figure 6 shows how this algorithm works on the original problem data set. It finds a first estimate of the reference value that is acceptable not only according to the high uncertainty measurements but also according to the three low uncertainty measurements.

Figure 7 shows how this algorithm, contrary to the cumulative probability algorithm, is robust against outliers. Early results that include several data sets indicate that the Value Voted Most Likely To Be Correct algorithm is not only more robust than the median algorithm but also in most cases identify more participant values as reliable and therefore lead to weighted average values with lower uncertainties than the median algorithm, while at the same time finding fewer participant values unacceptable according to the E_n calculation.

4. Conclusions

A weighted average value is one way to create a reference value in interlaboratory comparisons in proficiency testing. In order for the weighted average value and its associated uncertainty not to be contaminated by incorrect measurements (measurements that yield a wrong value due to a measurement error and measurements with a too low quoted uncertainty) a criterion must be established for which measurements are included in the weighted average.

A good criterion for determining which measurements to include in the weighted average is one that, while excluding all potentially incorrect measurements, includes as many measurements as possible to yield a reliable weighted average value with as low an uncertainty as possible. At the same time a good criterion must be robust against "problem data sets".

Requiring the measurement's uncertainty interval to include the median value of all the measured values is one such criterion. This criterion is statistically well founded and generally works well, but as it is shown in this paper it may come to the wrong conclusion for data sets that include correct measurements with significantly different uncertainties.

While one new algorithm, the cumulative probability algorithm, works well on the data set where the weakness of the median algorithm was discovered, it is shown to be very sensitive to outliers and lack the required robustness.

Another new algorithm, the Value Voted Most Likely To Be Correct algorithm, is presented. It is shown how this new algorithm works better than the median algorithm on data sets with significantly different uncertainties, while at the same time exhibiting good robustness against outliers.

References

1. Guide to the Expression of Uncertainty in Measurement. BIPM, IEC, IFCC, ISO, IUPAC, IUPAP, OIML., 1995