

A Discussion of Stability and Homogeneity Issues in Proficiency Testing for Calibration Laboratories

Jeff C. Gust

Abstract: This paper presents the technical issues concerning the homogeneity and stability of artifacts that are used primarily for proficiency testing of calibration laboratories. Additionally, one of the key performance metrics in proficiency testing, the normalized error formula, is analyzed and inadequacies in the formula are identified. This paper makes consensus recommendations concerning the resolution of these issues. These recommendations are intended for use as guidance for the application of a proficiency testing scheme that meets the requirements of *ISO/IEC Guide 43-1:1997*, *ILAC G13:2000* and possibly the standard in development, *ISO 17043*.

1. Introduction and Traceability Paths

This paper has been developed to be a supplement to the Paper “ILAC Discussion Paper on Homogeneity and Stability Testing” presented by Dan Tholen at the second meeting of the ILAC Proficiency Testing Consultative Group in May 2006. [1] The original paper covered the issues of stability and homogeneity in a general manner for proficiency testing of testing laboratories. Tholen’s presentation was a draft paper, which is expected to be revised over time as consensus positions and practical experience in proficiency testing develops. While Tholen’s original draft addressed proficiency testing (PT) from the perspective of test laboratories, and in particular microbiology laboratories, this paper is intended to discuss homogeneity and stability for the sector specific application of proficiency tests for calibration laboratories.

2. Discussion

In order to determine whether or not a participating laboratory is proficient for a particular measurement discipline, an evaluation of the laboratory’s performance must be conducted. While

many methods of evaluation exist, the most commonly used method for determining the performance of an individual calibration laboratory is the normalized error, E_n , formula. [2] The E_n performance statistic is found in *ISO/IEC Guide 43-1:1997* [2], *ISO 15528:2005*, “Statistical Methods for use in Proficiency Testing by Interlaboratory Comparisons,” [3] the A2LA document “R103 – General Requirements: Proficiency Testing for ISO/IEC 17025 Laboratories” [4] and in other documents. The E_n formula is given in equation (1):

$$E_n = \frac{x - X}{\sqrt{U_{lab}^2 + U_{ref}^2}}, \quad (1)$$

where:

- E_n = Normalized error
- x = Participant’s measurement result
- X = Assigned value of the artifact
- U_{lab} = Uncertainty of the participant’s measurement results
- U_{ref} = Uncertainty of the reference laboratory’s assigned value

If the normalized error is between -1.0 to $+1.0$, this is an indication that the difference between the participant’s measurement results and those of the reference laboratory are less than the combined uncertainties of the two organizations. Conversely, if the normalized error is greater than one in magnitude (posi-

Jeff C. Gust

MeasurePT, Inc.
17045 Governors Drive
Columbia City, IN 46725 USA
Email: gust@measurept.com

tive or negative) this would communicate that the difference in measurements was greater than the combined uncertainties, and would be an unacceptable result. Repercussions of an unacceptable normalized error result from a proficiency test usually range from requiring corrective action from the participant laboratory up to removal of accreditation or recognition for the measurement parameter.

Due to the significance associated with proficiency test results, it is important to understand all components of the equation. The focus of this discussion is: How is U_{ref} determined? In other words, what factors should be included in the calculation of U_{ref} ? By the definition given above, U_{ref} is the uncertainty associated with the reference laboratory's assigned value; this could imply that only uncertainty components that the reference laboratory took into account should be included in the calculation of this quantity.¹ However, when the artifact is used as part of a proficiency test scheme, there are other sources of uncertainty that need to be considered and included if applicable. It is important to note that the normalized error formula definitions imply that the reference value and uncertainty is determined by a laboratory. These values may in fact also be determined through a consensus value or weighted mean of participant measurements. However, this paper will focus on reference values established principally by a laboratory.

For example, the National Institute of Standards and Technology (NIST) *Technical Note 1297* [5] is the principal document governing how NIST evaluates and expresses the uncertainty of their measurement results. Section 7.6 of this document states:

“It follows from subsection 7.5 that for standards sent by customers to NIST for calibration, the quoted uncertainty should not normally include estimates of the uncertainties that may be introduced by the return of the standard to the customer’s laboratory or by its use there as a reference standard for other measurements. Such uncertainties are due, for example, to effects arising from transportation of the standard to the customer’s laboratory, including mechanical damage; the passage of time; and differences between the environmental conditions at the customer’s laboratory and at NIST. A caution may be added to the reported uncertainty if any such effects are likely to be significant and an additional uncertainty for them may be estimated and quoted. If, for the convenience of the customer, this additional uncertainty is combined with the uncertainty obtained at NIST, a clear statement should be included explaining that this has been done.”

From this statement, it is clear that there are additional uncertainty factors that a NIST customer needs to evaluate, in addition

to the NIST assigned calibration uncertainty. The statement also indicates that the uncertainty quoted by NIST in the calibration report is valid only at the time of calibration by NIST and only when the standard is located in the NIST laboratory.

An example of how NIST has implemented this practice can be found in the calibration of standard resistors [6] as documented in *NIST Technical Note 1458*, as well as NIST calibration reports supplied to NIST customers for standard resistors. Section 9 of *NIST Technical Note 1458* states:

“The reported expanded uncertainty contains no allowances for the long-term drift of the resistor under test, for the possible effects of transporting the standard resistor between laboratories, nor for measurement uncertainties in the user’s laboratory.”

While both of these documents make very important statements about the components of uncertainty that are not taken into account by the calibration laboratory, they provide very little guidance to the user on how to estimate the uncertainties due to the passage of time, differences in environment, or transportation effects.

In order for the proficiency test to be technically valid, U_{ref} must contain all uncertainty components which are of importance in the given situation. If U_{ref} is underestimated, then it lowers the value of the denominator of equation (1), which increases the value for E_n . This may cause a false unacceptable result for the proficiency test. When a proficiency test is designed, often an artifact is selected and sent to a reference laboratory, such as NIST, for the assignment of the reference value. It is then the responsibility of the proficiency test provider to determine the sources and magnitudes of uncertainty associated with stability and homogeneity. These additional uncertainties are then combined with the reference laboratory's reported measurement uncertainty for the artifact in order to estimate the uncertainty (U_{ref}) to be assigned to the artifact for the proficiency test and used in equation (1).

In the sections to follow, issues of homogeneity and stability of artifacts will be discussed for different types of proficiency test schemes used for calibration laboratories, together with supporting examples.

2.1 Artifact Homogeneity

In PT schemes for test laboratories, the artifacts may consist of a sample drawn from a large batch of material. In contrast, PT schemes for calibration laboratories usually involve each laboratory measuring the same artifact. Since the calibration PT scheme does not draw from a sample, often the consideration of homogeneity is dismissed entirely. However, when considering the definition of homogeneity, it can be shown that this concept must still be considered even if all participants measure the same artifact.

The term homogeneity is not formally defined in *ISO Guide 43-1:1997* or *ISO 15528:2005*, nor the International Vocabulary of Metrology (VIM). [7] However, the term is defined in *ISO Guide 30*, “Terms and definitions used in connection with reference materials,” [8] as: “Condition of being uniform struc-

¹ Note: Although it is not discussed in any of the given references for the normalized error formula, it is assumed that the uncertainties in the formula are expressed as an expanded uncertainty. For the proficiency test, the coverage factor and its associated level of confidence should be the same and should be agreed to before the proficiency test is initiated. The coverage factor $k=2$, representing an approximately 95% level of confidence, is generally used.

ture or composition with respect to one or more specified properties.” If a property exists for a material that will result in a variance in measurement results, then the homogeneity of the material has not been completely realized. When an artifact is to be selected for a proficiency test, designers of the proficiency test must carefully define the measurand so that all properties of the artifact that can cause variability of the results are appropriately addressed. *ISO Guide 43-1997* does provide some guidance regarding to what degree homogeneity must be addressed in section 5.6.2 where it states “the degree of homogeneity should be such that differences between test items will not significantly affect the evaluation of a participant’s result.” The next sections present specific examples of situations where these homogeneity factors need to be addressed.

2.1.1 PT Scheme – All Participants Measure the Same Artifact Under the Same Conditions

The majority of proficiency test schemes for calibration laboratories involve the measurement of one or more artifacts under appropriately defined measurement conditions. One such example is when the proficiency test artifacts are two standard resistors. Since all participants are measuring the same artifacts, additional measurements or statistical tests of the homogeneity of the artifacts are not required as per the definition cited above. However, in order to ensure that all participants are measuring the same property (i.e., resistance) under the same conditions, the measurand must be appropriately and explicitly defined for the environmental effects that were discussed in section 2.0.

For the case of a PT involving electrical resistance, a complete definition of the measurement process would have to include the temperature of the resistor, because the electrical conductivity of materials comprising the standard resistor change with temperature. The resistance vs. temperature relationship for the artifacts must be known to the PT scheme developer before the scheme is initiated. The relationship is generally expressed as a second degree polynomial. [9] In addition to temperature effects, some resistors are known to have a resistance value dependent upon the head pressure at its terminals. If the participating laboratories are at various elevations, the head pressure at the terminals of the resistor must be quantified (a combination of the local air pressure and bath fluid pressure over the terminals) and accounted for in the reported measurement results.

Additionally for a proficiency test for electrical resistance, it is imperative that the test current used in the measurement is defined. If the participants apply too much current, the measured value for resistance will change as a result of self heating. If the current is too low, then the floor noise of the resistance measurement system causes excessive variability for the measurement. The test current should be set at a point so that the adverse effects of self heating and noise are both optimized. A well designed proficiency test will have some flexibility in the measurement process in order to allow the participant to treat the artifacts as if they were performing routine tests. [10] Since any variability of the measurement result should be accounted for by the participant in the estimate of uncertainty of their measurement, the value set for the measurement current is



Figure 1. Photograph of PT artifacts for Rockwell hardness.

usually a maximum, which means that it should not exceed a specific value.

In order to effectively eliminate inhomogeneity in the measurement result for this resistance example, a complete definition of the measurand is required such as the following:

“The measurand of the PT is the electrical resistance of the artifacts at a temperature of 25.00 °C and a pressure of 101 325 Pa. The test current is not to exceed 10 mA for the 100 ohm artifact and 150 μA for the 19 000 ohm artifact.”

Even with this definition of the measurand, there will still be some uncertainty associated with realizing the measurement at the defined conditions. The uncertainty in the measurement of temperature, pressure and applied current must be considered and included in the proficiency test uncertainty analysis if they are significant. In some cases, the participants may not be able to perform measurements at the defined parameters. In this case, either the PT provider or the participant will have to apply a correction to the measured value in order to correct for the known offset. Thus the uncertainty associated with the measurement of environmental parameters, and uncertainty of any correction, must be estimated and included in the overall uncertainty analysis if they are significant.

In this example where all participants measured the same artifact(s), a complete definition of the measurand may effectively eliminate any uncertainty due to variation in homogeneity or make it insignificant to the overall proficiency test uncertainty. In some cases the limitations in the measurement process (i.e., magnitude of uncertainty for temperature or pressure measurement) may require inclusion of uncertainty due to homogeneity in an expanded uncertainty statement for the proficiency test despite the thorough definition of the measurand.

2.1.2 PT Scheme – All Participants Measure Different Locations on the Same Artifact

As in the previous example, a PT scheme can be conducted for metallic hardness where all participants measure the same PT item. However, the measurement of metallic hardness is essen-

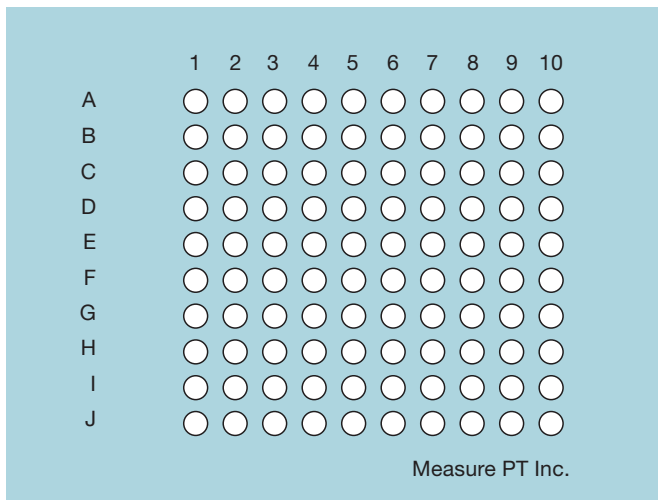


Figure 2. Schematic of PT artifact for Rockwell hardness.

tially a destructive test [11], in that, in performing the measurement, the hardness tester makes a small indentation into the material. Since the artifact is permanently indented, other measurements must be taken on different locations of the artifact. During the determination of the reference value for the artifact, the reference laboratory takes a series of measurements at random locations across the artifact. The artifacts also return to the reference laboratory after completion of participant measurements for another set of measurements by the reference laboratory. The standard deviation of the reference laboratory measurements (for both opening and closing the round) may be used as an estimate of homogeneity for the artifact.

An example of this type of proficiency test is when the artifact is configured as in Figs. 1 and 2. To establish the reference value of the artifact, the reference laboratory makes a preliminary indentation on the artifact to seat it onto the anvil of the hardness measurement machine within the circled area at a matrix location such as row B, column 3, or B3. The reference laboratory then makes measurements at various locations selected randomly across the artifact at locations such as A9, E7, I2, B1, and E5. The average of the five measurements determines the initial assigned value and standard deviation of the artifact by the reference laboratory.

Each PT participant also performs a preliminary indentation in the artifact and five additional measurements on the artifact at locations as directed by the PT scheme provider.

At the end of the PT round, the artifact is returned to the reference laboratory for closing measurements where a preliminary indentation and five additional measurements are taken. The homogeneity of the artifact may be determined by computing the standard deviation of the ten reference laboratory measurements across the artifact. The standard deviation of the ten measurements by the reference laboratory then becomes the standard uncertainty associated with homogeneity of the artifact which accounts for variations in hardness across the artifact.

The PT scheme provider should also evaluate each participant’s data to determine if there is any detectible trend identification that could be related to the homogeneity of the artifacts. However it is strongly cautioned that any review of participant data must result in a very clear trend and must show strong cor-

respondence with the reference laboratory’s opening and closing data before any conclusions about the homogeneity of the artifact can be made. In most cases, the participants may not possess equipment capable of the same accuracy or resolution and therefore their assigned uncertainties will vary significantly. In addition, the participating laboratory staff may not have the same level of technical expertise as the reference laboratory staff. Therefore, all participant data should be considered suspect unless an obvious trend is observed.

While the measurement examples presented in Sections 2.1.1 and 2.1.2 do not address all the issues associated with the homogeneity for proficiency tests designed for calibration laboratories, the majority of proficiency test schemes currently being performed fall into either design 2.1.1 or 2.1.2. In any case, the homogeneity of the artifact and the measurement process always needs to be considered, and if appropriate, it should be clearly defined, measured, quantified, and taken into account before initiating the PT round.

2.2 Artifact Stability

As in the case for homogeneity, the term stability is not formally defined in the documents *ISO Guide 43-1:1997* or *ISO 13528:2005*. However, the current version of the VIM [7] does define stability as “property of a measuring instrument, whereby its metrological properties remain constant in time” *ISO Guide 30* defines stability as: “Ability of a reference material, when stored under specified conditions, to maintain a stated property value within specified limits for a specified period of time.” When developing proficiency tests for calibration laboratories, both of these definitions are somewhat incomplete. With respect to the VIM definition, there may not be a measuring instrument that maintains metrological properties constant with time. Our only alternative is to determine how unstable the PT artifact may be during the proficiency test process. As for the *ISO Guide 30* definition, while this may be suitable for defining stability of a reference material, it may be somewhat incomplete when discussing stability of a metrological artifact used for a proficiency test of a calibration laboratory. There are some conditions that cannot be specified or are not known, such as the change of the assigned value of the artifact with respect to time. These issues of stability may be very significant with respect to the uncertainty estimated by the participant and reference laboratory. Since it is not always possible to accurately estimate the uncertainty due to a particular condition of stability, the only alternative is to measure and evaluate the artifact’s stability during the PT round.

The evaluation of stability is extremely important when conducting proficiency tests for calibration laboratories. Often, the reference laboratory can provide a measured value and uncertainty of measurement for the artifact that is extremely small. Despite the best efforts of the PT scheme developer, effects of transportation will make the uncertainty associated with the stability of the artifact several times larger than the uncertainty associated with the reference laboratory measurement. Unless the uncertainty associated with artifact stability is appropriately accounted for in the PT, false unsatisfactory results will occur for participants.

During the design phase of the PT, the stability of the artifact should be considered and estimated in order to develop an understanding of its magnitude relative to the other components of uncertainty. An appropriate estimate of the uncertainty associated with the artifact’s stability is used by the PT scheme provider and each participant to determine if the PT scheme is suitable for validating each participant’s calibration capabilities. The design phase for the PT should also consider an appropriate measurement model for describing the artifact’s stability. When the artifact’s stability is measured and analyzed at the end of the PT process, the PT scheme provider should compare the measured stability with the initially estimated stability and the model. If the artifact’s measured stability exceeds pre-established limits, and thereby indicates that it has become unsuitable for use in the PT scheme, a nonconformance investigation may be initiated by the PT scheme provider.

2.2.1 Short Term Stability – Petal or Modified Petal Design

The most conservative PT design for measuring PT artifact stability is a Petal [12] or Modified Petal design, in which the artifact is measured by the pivot laboratory (PL) before and after each shipment to the participant laboratory. This type of design allows the stability of the artifact to be determined with the greatest confidence and smallest uncertainty. Therefore the petal design is most applicable when participants are national metrology institute’s (NMI’s) or industrial/governmental laboratories with uncertainties within an order of magnitude of NMI’s. It is also a sound design when there are concerns about the stability of the artifact as indicated by a large initial estimate of the stability uncertainty. A disadvantage of the petal design is that it has a high operational cost due to sending the artifact back to the pivot laboratory after each participant; this also increases the time to complete the PT.

Figure 3 below shows a Modified Petal design. In a formal Petal design, the reference laboratory also performs as the pivot laboratory so it performs the before/after measurements in addition to establishing the reference value for the artifact. In this particular Modified Petal design, the PT scheme provider has the ability to measure the artifacts with sufficient resolution and sensitivity to detect any significant change in the artifact. Thus the PT scheme provider acts as the pivot laboratory. However, the PT scheme provider does not establish the reference value of the artifact, but instead leaves this to a competent, accredited calibration laboratory (reference laboratory). The artifacts are measured by the PT scheme provider both before and after they are sent to either the reference laboratory or the participants. In either the Petal or this Modified Petal design, the PT scheme provider determines the short term stability of the artifact which includes changes in the reference value of the artifact due to effects of transportation, handling by the participants, aging of the artifact or changes due to environment.

One major benefit of a Petal Design is that it isolates the measurement data associated with each participant. Thus if the artifact becomes damaged at a point in the PT process, some of the participant data can be rescued and reported. This method also allows a PT provider to supply a final report to each participant in a much shorter timeframe, under the condition that the

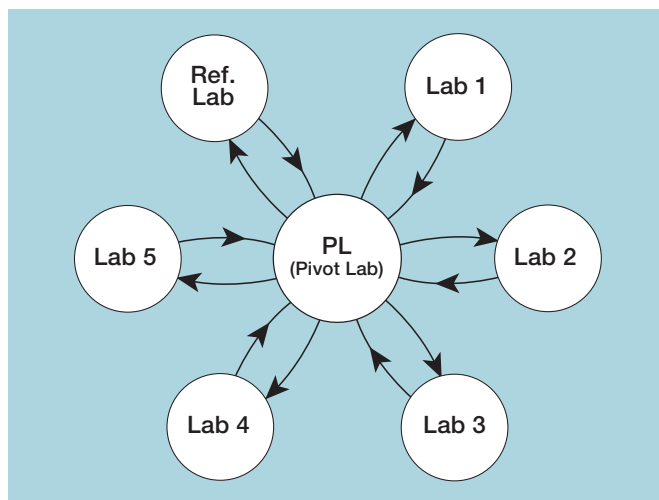


Figure 3. Schematic of a modified petal design.

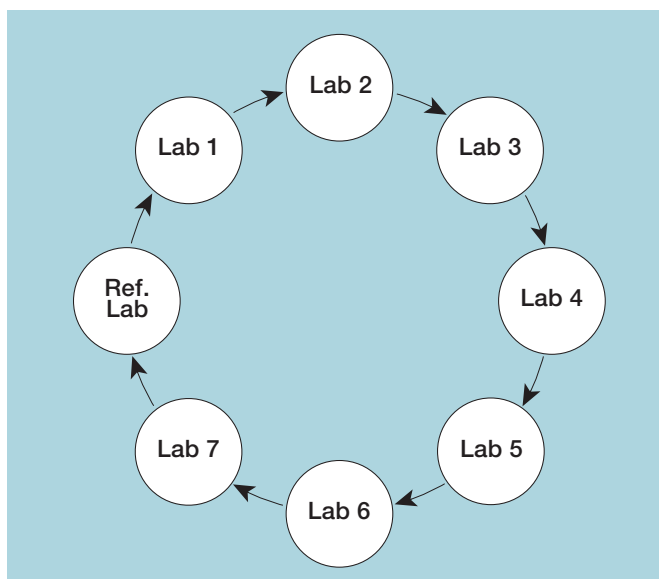


Figure 4. Ring design schematic.

estimated stability of the artifact was appropriately performed using suitable equipment and measurement procedures. When the PT scheme provider performs the pivot measurements, the cost associated with the proficiency test can be reduced compared to having the reference laboratory test perform all of the pivot measurements. In this Modified Petal design, the anonymity of each participant is easier to maintain, because all shipments are sent directly from and to the PT scheme provider rather than having a participant ship to the reference laboratory or to another participant. Additionally, either the Petal or Modified Petal designs allows participation at will, that is to say that the number of participants is not restricted as in a one-off type measurement scheme, so long as the artifact is stable and is returned to the reference laboratory periodically.

The estimate of uncertainty associated with the short term stability of the artifact is usually determined by considering changes in the pivot measurements. The most conservative analysis of this information considers the opening and closing pivot measurements to be opposite ends of a rectangular distribution; this means that the standard uncertainty associated with the artifact’s

stability is the difference between the opening and closing measurement divided by the square root of three. Sometimes if the artifact is known to be sufficiently stable and the majority of the difference between opening and closing measurements is a result of changes in the measuring equipment/process itself, the stability measurement distribution may be estimated to be triangular. When analyzing these stability measurements, it is assumed that the deviations are largely the result of random effects in the measurement process and are not due to a shift in the reference value of the artifact, unless the artifact is known from its design or from previous measurement data to have a drift over time (see Section 2.2.3).

At the completion of a proficiency test round, in which the artifact has traveled from the reference laboratory, through a group of participants, and back to the reference laboratory, the long term stability of the artifact is determined by calculating the change in the reference value assigned to the artifact from the two reference laboratory measurements. This long term stability value should be compared to the short term stability determined through the measurements performed by the pivot laboratory. If the long term stability value is significantly larger than the short term stability value, then the expanded uncertainty assigned to the artifact should be appropriately increased and a revised PT report supplied to each participant.

2.2.2 Short Term Stability – When Stability Can Be Assumed

Some artifacts are inherently stable by design, construction, or materials (as per the VIM definition of stability). These types of artifacts primarily include dimensional standards like gauge blocks, ring gauges, length standards, etc. In cases where the artifact is understood to be inherently stable, short term stability measurements can be minimized. However, changes in the artifact due to use or transportation still need to be considered.

2.2.3 Long Term Stability – Reference Laboratory Measurements

When the artifacts are understood to exhibit appropriately small uncertainty

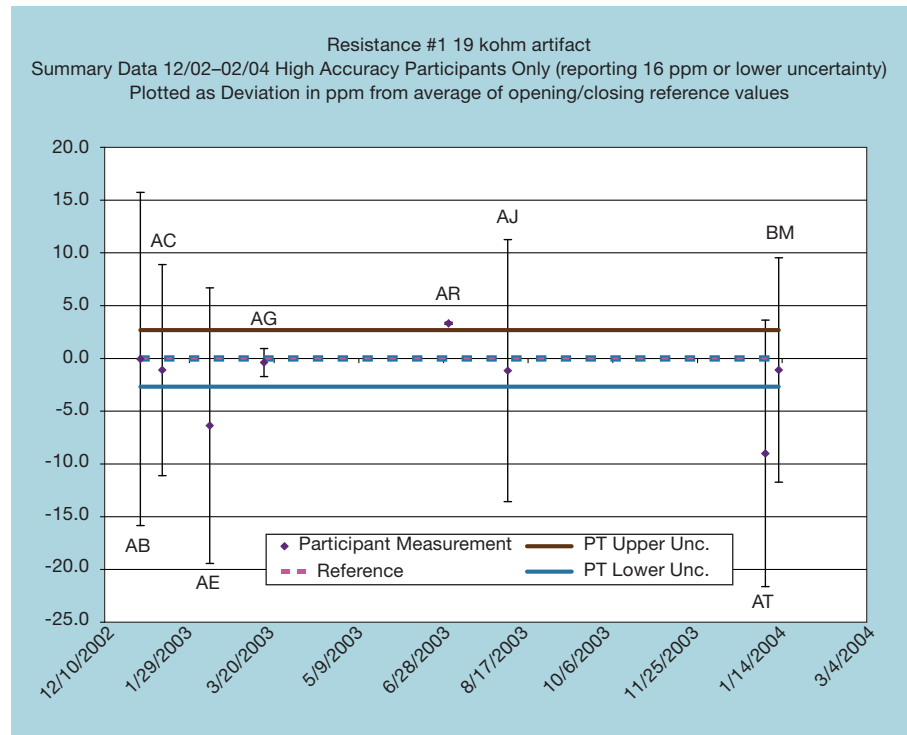


Figure 5. Graph of PT data with long term stability information.

due to short term stability, a Petal Design is not required, but a Ring Design can be used instead. In a Ring Design, the reference laboratory initially establishes a reference value for the artifact. The uncertainty due to stability is assumed to be small or insignificant as compared to the measurement uncertainty of the reference laboratory. In the Ring Design, the artifact is measured successively by each participating laboratory (1 through 7 in Fig. 4).

The advantages of a Ring Design are that the labor and costs associated with the pivot laboratory measurements are eliminated. In addition, the PT conducted for the same number of participating laboratories would conclude earlier with a Ring Design as opposed to a Petal Design, because of the reduced shipping time and elimination of pivot laboratory measurements for the artifact.

The disadvantages of a Ring Design are that if the artifact becomes unstable at any point in the PT, all data for the round are lost, and shipping the artifact from one participant laboratory to another compromises some of the confidentiality of the participants. In addition, because the final results cannot be released until stability of the artifact has

been confirmed by the reference laboratory, it takes longer for participants to receive final reports.

Regardless of design used in proficiency tests for calibration laboratories, the artifact needs to be sent back to the reference laboratory on a periodic basis. In this case, the assigned value of the artifact for the PT round is generally considered to be the average of the opening and closing measurement by the reference laboratory. Since the average of the opening and closing measurements is used to determine the reference value, the most conservative estimate of the uncertainty due to stability can be considered to be half the change between the opening and closing measurements, with a rectangular distribution.

Figure 5 shows PT resistance results for eight participating laboratories. In the figure, the reference laboratory's opening data are represented by the solid blue line and the closing data by the solid brown line. Note that this shows that the reference value increased by about 5 ppm over the time of the PT. The pink dashed line is the average of the opening and closing measurements. One can see that it is reasonable to estimate that the artifact was most likely not lower than

the opening measurement or higher than the closing measurement; therefore the opening/closing measurement may be considered the limits of a rectangular distribution that estimates the uncertainty due to artifact stability.

2.2.4 Long Term Stability – Linear Regression

There are certain types of artifacts that exhibit a uniform change in their value with time. The two most common examples of this behavior are standard resistors and zener voltage reference standards. If the artifacts have a sufficient amount of history, it is possible to use linear regression to predict what the reference value of the artifact should be on any given day. Most spreadsheet programs have accurate worksheet functions to predict the value of a trend line for a given day. However, if a linear regression model is used to assign the reference value to an artifact, then the uncertainty of the regression must be included in the overall uncertainty budget for the proficiency test. If a linear regression model provides a better estimate for the reference value of the artifact, then the uncertainty of the regression should be smaller than the uncertainty derived only from opening/closing measurements.

Figure 6 shows the individual calibration values and uncertainties (less than 1 μΩ/Ω) from the reference laboratory. These values were modeled with a linear regression line (pink), where the blue and yellow lines are the two sigma estimated uncertainty associated with the linear regression.

Figure 7 has the same information as Fig. 6, but the calibration data points were removed from the graph and a point has been plotted on the regression line representing the predicted calibration value for a day seven months in the future. For this case, the estimated uncertainty due to linear regression prediction for the value of this standard resistor is 0.30 μΩ/Ω (k = 2) seven months after the reference laboratory measurement.

The estimate of uncertainty for linear regression is computed using the following equations: [13]

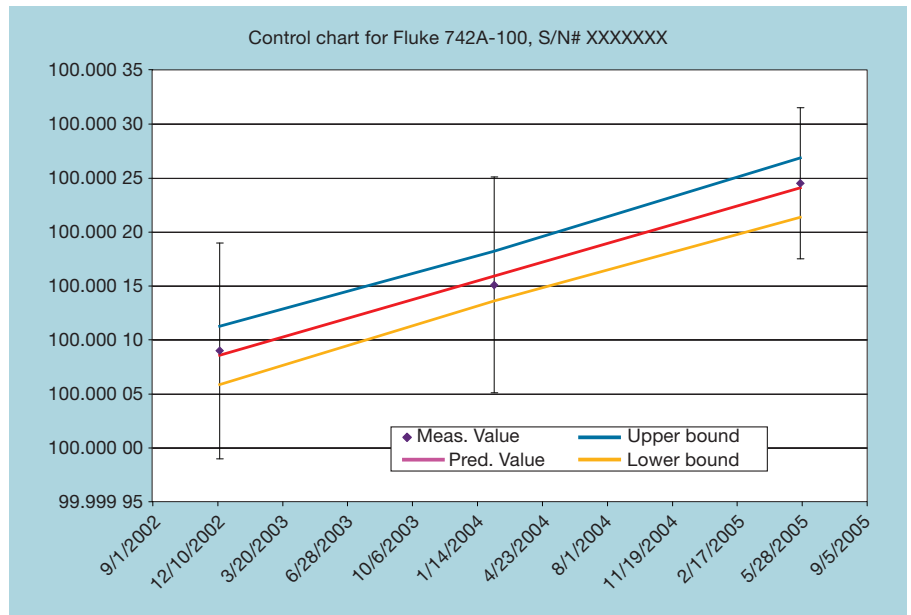


Figure 6. Control chart with linear regression error limits.

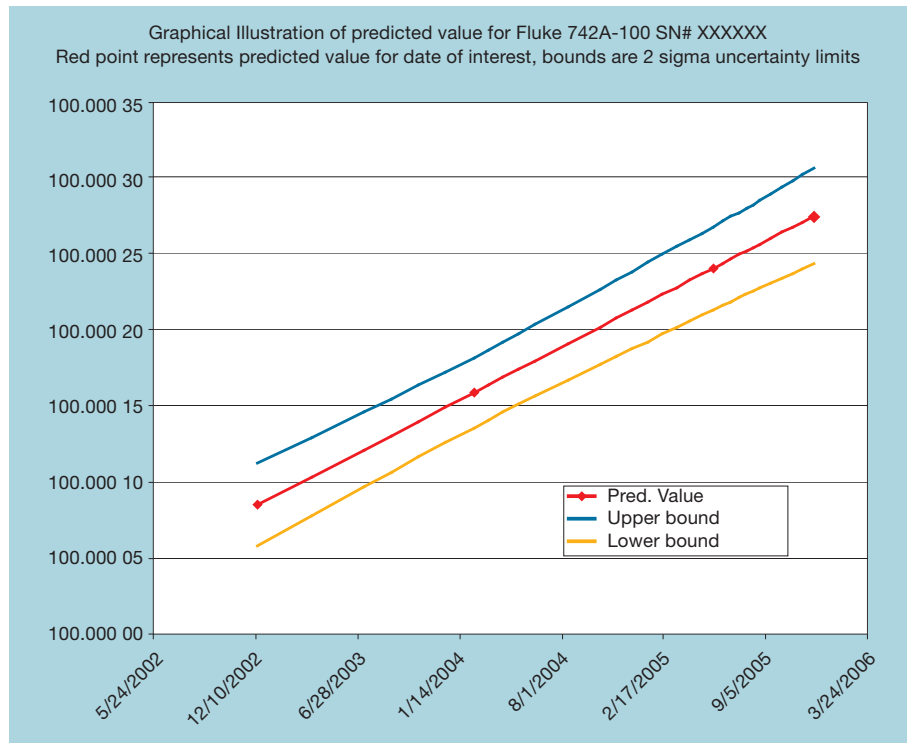


Figure 7. Graphical estimate of a predicted value.

$$S_{reg} = S_{yx} \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}}, \tag{2}$$

where:

- S_{reg} = Uncertainty of the Regression
- S_{yx} = Standard Error of the Regression (Defined Below)
- n = Number of Data Points
- x = Time Data or Calendar Date
- \bar{x} = Mean of Time Data
- S_{xx} = Standard Error of Time Data
- y = Resistance Data

and

$$S_{yx} = \sqrt{\frac{1}{n(n-2)} \left[n \sum y^2 - (\sum y)^2 - \frac{[n \sum xy - (\sum x)(\sum y)]^2}{n \sum x^2 - (\sum x)^2} \right]} \quad (3)$$

$$S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n} \quad (4)$$

It is important to note that Microsoft Excel does not have a function that calculates S_{reg} and S_{xx} [eqs. (2) and (4)]; however, the function STEYX does determine the value of S_{yx} in equation (3).

Although linear regression is a useful tool to reduce the estimate of uncertainty due to stability, the PT scheme provider should still perform interim measurements on the artifact to ensure that it is performing as expected and has not undergone any catastrophic event.

3. Conclusions

Organizations that develop PT schemes are required to demonstrate that the homogeneity and stability of artifacts are quantified and suitable for their intended use. If the PT is to serve the purpose of verifying the measurement capability of the participants and not provides false results, it is essential that the artifacts meet intended estimates of uncertainty associated with homogeneity and stability. In order to completely understand all the issues associated with the stability and homogeneity of artifacts, these terms should be clearly defined in a way that is suited to PT applications. All measurements associated with the homogeneity or stability of an artifact should be used to quantify each source of uncertainty, converted to a standard uncertainty, and combined with the measurement uncertainty associated with the reference laboratory measurement in order to produce an expanded uncertainty for the proficiency test artifacts [14] that is used to judge the proficiency of each participant.

Considering all potential sources of uncertainty is consistent with both the principles of the ISO "Guide to the Expression of Uncertainty in Measurement," as well as NIST *Technical Note 1297*. It has been the personal experience of the author, both as a consultant to NMI's in the development of proficiency tests and as an assessor for calibration laboratory accreditation, that it is a common mistake to not include an uncertainty component related to stability and homogeneity of the artifact in the statement of U_{ref} . Although the E_n formula implicitly states that it should include all sources of uncertainty which are of importance in the give situation, in order to more clearly communicate the appropriate estimate of uncertainty for the proficiency test, the following considerations to the E_n formula are suggested:

First, define the expanded uncertainty for a PT, using equation (5) as follows:

$$U_{PT} = k \sqrt{u_{ref}^2 + u_{stab}^2 + u_{homo}^2} \quad (5)$$

where:

- U_{PT} = Expanded uncertainty of the proficiency test
- u_{ref} = Standard uncertainty of the reference laboratory's assigned value
- u_{stab} = Standard uncertainty of the artifacts due to effects associated with artifact stability
- u_{homo} = Standard uncertainty of artifacts due to the effects associated with artifact homogeneity

k = Coverage factor used to obtain an expanded uncertainty

The E_n formula given in equation (1) could be amended to equation (6) substituting U_{PT} for U_{ref} , giving:

$$E_n = \frac{x - X}{\sqrt{U_{lab}^2 + U_{PT}^2}} \quad (6)$$

Note that U_{lab} and U_{PT} must use the same value for k .

This paper provides specific guidance for addressing issues of artifact stability and homogeneity for PT schemes used to assess calibration laboratories. It is the author's hope and expectation that this paper will be amended with additional information as the body of knowledge surrounding PT schemes for calibration laboratories grows.

4. References

- [1] D.A. Tholen, et al., "ILAC Discussion Paper on Homogeneity and Stability Testing," Presented at the *ILAC Proficiency Testing Consultative Group Meeting*, Madrid, Spain, May 12 & 13, 2006.
- [2] "Proficiency testing by interlaboratory comparisons – Part 1: Development and operation of proficiency testing schemes," *ISO/IEC Guide 43-1:1997*, Appendix A.2, International Organization of Standardization, 1997.
- [3] "Statistical Methods for use in Proficiency Testing by Interlaboratory Comparisons," *ISO 13528:2005*, International Organization of Standardization, 2005.
- [4] "General Requirements: Proficiency Testing for ISO/IEC 17025 Laboratories," *A2LA R103*, American Association for Laboratory Accreditation, April 2008. [Available from the A2LA web site: http://a2la.org/requirements/A2LA_General_Requirements_for_Proficiency_Testing.pdf]
- [5] B.N. Taylor and C.E. Kuyatt, "Guidelines for evaluating and expressing the uncertainty of NIST measurement results," *NIST Technical Note 1297*, 1994 Edition.
- [6] R.E. Elmquist, et al, "NIST Measurement Service for DC Standard Resistors," *NIST Technical Note 1458*, December 2003.
- [7] "International Vocabulary of Metrology – Basic and General Concepts and Associated Terms," (VIM) 3rd Edition, *Joint Committee for Guides in Metrology*, JCGM 200:2008.
- [8] "Terms and definitions used in connection with reference material," *ISO Guide 30*, International Standards Organization (ISO), 1992.
- [9] R.F. Dziuba, "Resistors," *Encyclopedia of Applied Physics*, vol. 16, VCH Publishers, Inc., pp. 423–435, 1996.
- [10] "Proficiency testing by interlaboratory comparisons – Part 1: Development and operation of proficiency testing schemes," *ISO/IEC Guide 43-1:1997*, Section 6.2.4, International Organization of Standardization, 1997.
- [11] S.R. Low, "Rockwell Hardness Measurement of Metallic Materials," *NIST SP 960-5*, January 2001.
- [12] "Guide for Interlaboratory Comparisons," *NCSL RP-15*, NCSL International, Boulder, CO, March 1999.
- [13] R.A. Johnson, "Miller & Freund's Probability and Statistics for Engineers," 5th edition, Prentice Hall, New Jersey, pgs 341–342, 1994.
- [14] "Guide to the Expression of Uncertainty in Measurement," International Organization of Standardization, 1995.