# A DISCUSSION ON ISSUES OF STABILITY AND HOMOGENEITY IN PROFICIENCY TESTING FOR CALIBRATION LABORATORIES

Jeff C. Gust
Quametec Proficiency Testing Services
501 W. Van Buren St.
Unit A
Columbia City, Indiana 46725
260 244-7450 260 244-7905 (fax) gust@quametec-pt.com

**Abstract:** This discussion is intended to present the technical issues concerning the homogeneity and stability of artifacts that are used primarily for proficiency testing of calibration laboratories. It is also intended to make 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 and ILAC G13:2000.

## 1.0 INTRODUCTION

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. 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 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.

## 2.0 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 may be found in ISO/IEC Guide 43-1:1997, ISO 13528:2005 (Statistical Methods for use in proficiency testing by Interlaboratory comparisons), the A2LA Proficiency Testing Requirements for Accredited Testing and Calibration Laboratories and in other documents. The $E_n$ formula is defined in equation (1) as:

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

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

The focus of this discussion is: What is $U_{ref}$ comprised of? By its definition, it 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 reported in this quantity.

NIST Technical Note 1297[3] is the principal document for how NIST evaluates and expresses measurement results. In section 7.6 of this document, it 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."

An example of how NIST has implemented this practice can be found in the calibration of standard resistors[4] in NIST Technical Note 1458 and NIST calibration reports for standard resistors. Section 9 of this document 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 they did not take into account, they provide no further guidance how to estimate the uncertainties due to the passage of time, differences in environment, or transportation effects. The uncertainties associated with passage of time, and transportation effects are excellent examples of issues of stability, while the uncertainty associated with differences in environment is a good example of issues involving homogeneity.

In order for the proficiency test to be valid, $U_{ref}$ must contain all components which are of importance in the given situation. If the uncertainty associated with $U_{ref}$ is underestimated, it lowers the value for the denominator of equation (1), which would in turn increase the returned value for $E_n$. This may cause a false failure 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 up to the proficiency test provider to determine the types and magnitudes of uncertainty associated with stability and homogeneity, and combine these estimates with the reference laboratory's measurement uncertainty for the artifact in order for the estimate of uncertainty to be complete.

In the proceeding sections, issues of homogeneity and stability will be discussed for different types of Proficiency Test (PT) schemes used in proficiency testing for calibration laboratories, with the support of examples.

### 2.1 Homogeneity

The term Homogeneity is not defined in ISO Guide 43-1:1997, ISO 13528:2005, nor the VIM. The term is defined in ISO Guide 30, Terms and definitions used in connection with reference materials[5]. ISO Guide 30 defines Homogeneity as "Condition of being uniform structure or composition with respect to one or more specified properties." While this definition of homogeneity may serve for a reference material, it may not be complete as a metrological definition. If a property exists for a material that will cause a variance in measurement results, then homogeneity has not been completely defined. If an artifact is to be selected for a proficiency test, designers of the proficiency test must define the measurand so that all properties of the artifact that can cause variability of the results have been 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 homogeneity are not required as per the definition cited above. However, in order for it to hold that all participants are measuring the same property (i.e. resistance), the measurand must be appropriately and explicitly defined.

For the case of electrical resistance in a proficiency test or interlaboratory comparison for National Measurement Institutes (NMI's) or industrial laboratories with uncertainties within an order of magnitude of NMI's, a complete definition would have to include the temperature of the resistor, because the conductivity of materials comprising the standard resistor changes as a function of temperature. The resistance to 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[6].

Additionally for a proficiency test for electrical resistance at this level, it is imperative that the test current be defined. If the participants apply too much current, the measured value for resistance will increase 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 defined at an optimal point so that the effects of self heating and noise are minimized. A well designed proficiency test will have some leeway to allow the participant to treat the artifacts as if they were performing routine tests[7], and since the variability of measurement should be accounted for by the participant in their estimate of uncertainty of measurement, the definition of current is usually single sided, or not to exceed a certain amount of current.

In order to effectively eliminate homogeneity for these examples, a complete definition of the measurand is required such as:

The measurand is electrical resistance of the artifacts at 23 degrees Celsius. The test current is not to exceed 10 mA for the 100 ohm artifact and 150 μA for the 19,000 ohm artifact.

### 2.1.2 PT Scheme – All Participants Measure an Artifact at Different Locations

The measurement of metallic hardness is essentially a destructive test[8], in that, in order to take a 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 on 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 measurement run 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 Figure 1. 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 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 of the artifact.

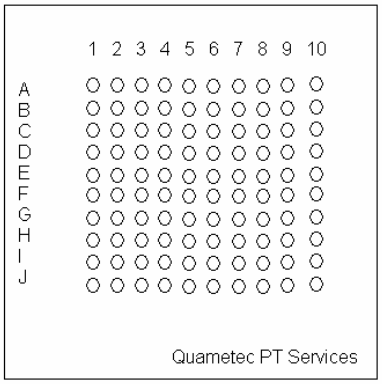

***Fig. 1*** Schematic of PT Artifact for Rockwell Hardness

The participants also perform a preliminary indentation in the artifact, and five additional measurements on the artifact as directed by the PT scheme provider.

At the end of the round, the artifacts are returned to the reference laboratory for measurement 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 then becomes the standard uncertainty associated with homogeneity of the artifact.

The PT Scheme provider should also evaluate the participant data to see if there is any detectible trend identification regarding the homogeneity of the artifacts. However it is strongly cautioned that any review of participant data must have a very clear trend and must show strong correspondence with reference laboratory data before any conclusions about the homogeneity about the artifact can be made. In most cases, the participants may not possess equipment of the same accuracy or resolution and therefore their uncertainty is larger. The participating laboratory staff may not have the technical expertise of the reference laboratory. All participant data should be considered suspect unless an obvious trend is observed.

While the measurement examples of 2.1.1 and 2.1.2 do not address every given issue of homogeneity for proficiency tests designed for calibration laboratories, the majority of proficiency test schemes performed today fall into the design of either 2.1.1 or 2.1.2. In any case, the homogeneity of the artifact needs to be considered, and if appropriate, measured, before initiating the PT round.

### 2.2 Stability

As it was in the case for homogeneity, the term stability is not defined in the documents ISO Guide 43-1:1997, ISO 13528:2005, nor the VIM. 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." Once again, 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 may not be able to be specified or known, such as the change of the assigned value of the artifact with respect to time. These types of issues of stability may be very significant with respect to the uncertainty estimated by the participant and reference laboratory. Since it is not possible to always accurately estimate the uncertainty due to a particular condition of stability, the only alternative is to measure and evaluate uncertainty due to 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 expanded uncertainty, false failure results will occur for participants.

During the design phase of the PT, stability should be considered and estimated in order to develop an estimated PT expanded uncertainty. An appropriate estimate of the PT expanded uncertainty is used by the PT scheme provider and participants to determine if the PT scheme is suitable for the participant. The design phase for the PT should also consider an appropriate model for the measurement of artifact stability. When stability is measured and analyzed (upon completion of the round), the PT scheme provider should also compare the measured stability versus the estimated stability, so that if the artifact stability exceeds pre-established limits and becomes unsuitable, 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 through the use of a Petal[9] or Modified Petal design, in which the artifact is measured before and after shipments (to the participant laboratory) by a pivot laboratory (PL). This type of design allows stability to be artifact to be determined with the smallest uncertainty and is most applicable when participants are NMI's or industrial laboratories with uncertainty 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 compared to the initial estimate of stability uncertainty. A drawback to the petal design is that it has a high operational cost due to sending the artifact back to the reference laboratory after each participant. Figure 3 below shows a Modified Petal design, often referred to the Quametec Petal[10]. In a formal Petal design, the only difference is that the reference laboratory performs the before/after or pivot measurements in addition to establishing the reference value for the artifact. In the Quametec Petal, the PT scheme provider has the ability measure the artifacts with sufficient resolution and sensitivity. The PT scheme provider does not establish the absolute value of the artifact, but instead leaves this to a competent, accredited calibration laboratory. The artifacts are measured before and after sending the artifact to either the reference laboratory or the participants. In either the Petal or Quametec Petal design, it allows the PT scheme provider to measure the short term stability of the artifact which would capture any change in the reference value of the artifact due to effects of transportation, measurement of the artifact by the participant, and any environmental changes.

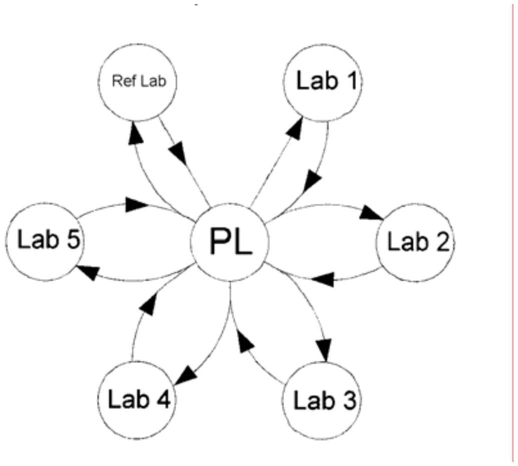

***Fig. 2*** Quametec Petal Design Schematic

The benefits of a Petal Design are that they capture all data associated with each participant, so if the artifact is damaged during subsequent measurements, some of the laboratory data can be rescued and reported upon. This method also allows a PT provider to provide a final report to the participant in a much shorter timeframe, if the estimate of stability is appropriately performed with suitable equipment. When the PT scheme provider performs the pivot measurements, the cost associated with the proficiency test is reduced from having the reference laboratory perform the pivot measurements. In the Quametec Petal, the anonymity of the participant is better maintained, because of the shipments directly from and to the PT scheme provider rather than traveling to the reference laboratory. Additionally, either the Petal or Quametec Petal allow participation at will, that is to say that the number of participants is not restricted as in a one-off type of scheme, so long as the artifact returns to the reference laboratory in an appropriate amount of time.

The estimate of uncertainty due to short term stability is usually determined by considering the measured deviation from pivot measurements completed before and after a participant. The most conservative analysis of this information would be to consider the opening and closing pivot measurements to be opposite ends of a rectangular distribution, so the standard uncertainty due to stability associated with the artifact 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 stability uncertainty may be due to the measuring equipment itself, and therefore the stability measurement distribution may be estimated to be triangular.

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 measured stability is determined by the deviation in the assigned value for the artifact from the two reference laboratory measurements. This longer term stability value should be compared to the stability observed through the shorter term stability measurements performed by the pivot laboratory, in order to assure that the PT expanded uncertainty provided to each participant from pivot measurements was not less than the PT expanded uncertainty estimated from long term data.

### 2.2.2 Short Term Stability – When Stability Can Be Assumed

Some artifacts are inherently stable by design, such as 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 are not required.

### 2.2.3 Long Term Stability – Reference Laboratory Measurements

When the artifacts are understood to be stable, a Petal Design is not required, but a Ring Design is used instead. In a Ring Design, the Reference Laboratory establishes a reference value for the artifact. The stability is assumed be small or insignificant as compared to the uncertainty of Reference Laboratory measurement.

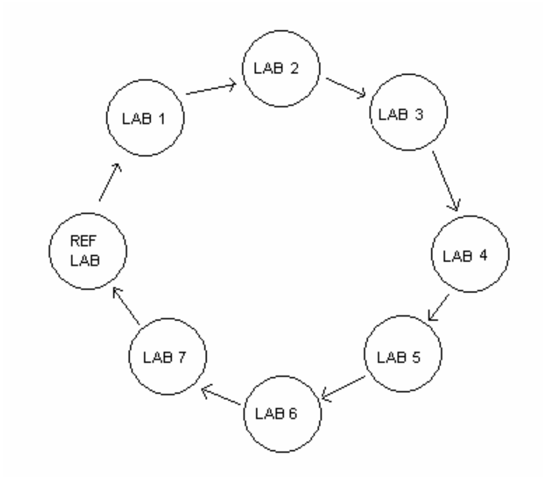

***Fig. 3*** Ring Design Schematic

The advantages for using a Ring Design are that the labor and costs associated with the pivot measurements is eliminated. A PT conducted for the same number of laboratories would conclude earlier with a Ring Design as opposed to a Petal Design, because of the reduced shipping time 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 is lost, and shipping the artifact from one participant laboratory to another compromises some of the confidentiality of the participants.

Regardless of design, in proficiency tests designed for calibration laboratories, the artifacts travel 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 for the assigned value, the most conservative estimate of stability can be considered to be half the deviation between the opening and closing measurements, rectangularly distributed.

In Figure 4 below, the reference laboratory's opening data is represented by the solid blue line, the closing data from the reference laboratory is represented by the solid brown line (uncertainty of reference laboratory measurements was also considered for this graph). The pink dashed line is the average of the opening and closing measurements and 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 ends of a rectangular distribution.

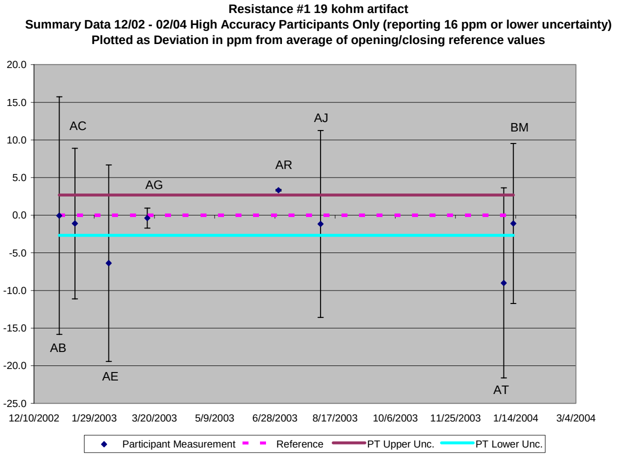

***Fig. 4*** Graph of PT Data with Long Term Stability Information

## 3.0 CONCLUSION:

Organizations that develop PT schemes are required to demonstrate that the homogeneity and stability of artifacts used are quantified and suitable for use. It is essential if the PT is to serve its purpose of verifying the measurement capability of the participants, and not provides false PT results, that the artifacts meet intended estimates of uncertainty for homogeneity and stability. In order to completely understand issues of stability and homogeneity, perhaps these terms should be defined so that they are better suited for metrological applications. Any measurement of homogeneity or stability should be treated as a source of uncertainty, converted to a standard uncertainty, and combined with the estimate of uncertainty associated with the reference laboratory measurement to produce an expanded uncertainty for the proficiency test artifacts[11] for use in judging proficiency of the participant. Considering all potential sources of uncertainty meet with both the principles of the ISO Guide to the Expression of Uncertainty in Measurement, as well as those previously mentioned in NIST Technical Note 1297. It is been the personal experience of the author both as a consultant to NMI's in the development of proficiency tests, and as an accreditation assessor for calibration laboratories, that it is a common mistake to not include uncertainty due to stability and homogeneity in the statement of $U_{ref}$. Although the Normalized Error ($E_n$) formula implicitly states that it should include all sources of uncertainty which is 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 is suggested:

Define equation (2) as follows:

$$U_{PT} = \sqrt{U_{ref}^2 + U_{stab}^2 + U_{homo}^2} \qquad (2)$$

Where:

$U_{pt}$ = expanded uncertainty of the proficiency test
$U_{ref}$ = uncertainty of the reference laboratory's assigned value
$U_{stab}$ = uncertainty of the artifacts due to effects associated with artifact stability
$U_{homo}$= uncertainty of artifacts due to the effects associated with artifact homogeneity

And the $E_n$ formula (1) could be amended to equation (3) substituting $U_{pt}$ for $U_{ref}$, giving us:

$$E_n = \frac{x - X}{\sqrt{U_{lab}^2 + U_{PT}^2}} \qquad (3)$$

This paper is meant to provide specific guidance for addressing issues of stability and uniformity for PT Schemes for calibration laboratories. This is not an all inclusive discussion of the subject, and it is expected that this appendix will be amended with additional information as the body of knowledge grows.

## REFERENCES

[1] Tholen, D.A. 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] ISO/IEC Guide 43-1:1997, Proficiency testing by interlaboratory comparisons – Part 1: Development and operation of proficiency testing schemes, Appendix A.2. International Organization of Standardization, 1997

[3] Taylor, B.N. and Kuyatt, C.E. NIST Technical Note 1297, Guidelines for evaluating and expressing the uncertainty of NIST measurement results. NIST. 1994 Edition

[4] Elmquist, R.E. et al NIST Technical Note 1458, NIST Measurement Service for DC Standard Resistors. NIST. December 2003.

[5] ISO Guide 30, Terms and definitions used in connection with reference materials. ISO 1992.

[6] Dzuiba, Ronald F. "Resistors" An entry in Encyclopedia of Applied Physics, Volume 16. VCH Publishers, Inc. 1996. Pages 423 – 435.

[7] ISO/IEC Guide 43-1:1997, Proficiency testing by interlaboratory comparisons – Part 1: Development and operation of proficiency testing schemes, Section 6.2.4. International Organization of Standardization, 1997

[8] Low, Samuel R. NIST SP 960-5, Rockwell Hardness Measurement of Metallic Materials. National Institute of Standards and Technology, January 2001.

[9] NCSL RP-15, Guide for Interlaboratory Comparisons. NCSL International. March 1999

[10] Gust, J.C. Final Results from an Accredited Proficiency Test. Published in the proceedings of the 2005 Measurement Science Conference.

[11] GUIDE TO THE EXPRESSION OF UNCERTAINTY IN MEASUREMENT, International Organization of Standardization, 1995