NIST Office of Weights and Measures (OWM)
Proficiency Test (PT) Supplemental Report

1. Purpose

The purpose of this document is to assist OWM laboratory participants, PT coordinators, PT analysts, laboratory management, laboratory recognition and accreditation bodies, and assessors to interpret and analyze OWM PT reports. This supplement is an integral part of each PT report but is not copied and integrated into each report to simplify and minimize extra documentation that is generic and duplicative in each report. Portions of the PT Plan Template and the PT Analyses Template spreadsheets provide the foundation of the PT report. Each unique PT report includes components from the planning, organization, PT artifact identifications and purpose(s), participant identification, operations, as well as the draft and final analyses, along with associated summaries, data, charts, and graphs unique to each PT.

2. OWM Policies and Quality System[1]

The Office of Weights and Measures (OWM) is not an accredited PT provider. However, the OWM PT Program seeks to comply with well-designed quality systems, laboratory, and accreditation body needs, ILAC PT policies, as well as ISO/IEC 17043 and ISO 13528 (latest versions where applicable).

2.1. NISTIR 7082, Proficiency Test Policy Plan", January 2018 (Draft March 1, 2023)

This publication provides the policies and plans for the PT Program of the NIST Office of Weights and Measures. This Office of Weights and Measures (OWM) Proficiency Testing (PT) policy and plan has been updated to ensure compliance with the latest applicable documentary standards and policies of the International Laboratory Accreditation Cooperation (ILAC).

The OWM PT program has been operating since the early 1980s as a core part of the support to State weights and measures laboratories, with most operations taking place by coordinating among the through regional measurement assurance programs (RMAPs). Original interlaboratory comparisons activities were conducted as “round robins” in support of ongoing measurement assurance activities related to support for State laws with requirements for metrological traceability to national and international standards. Since the early 1990’s the program has evolved to operate primarily as a proficiency testing (PT) program, with the first Quality System put in place in 2005. Current PT Program efforts

[1] All NIST references noted in this section are publicly available and posted on the NIST website at: https://www.nist.gov/pml/weights-and-measures/laboratory-metrology/proficiency-testing.

provide support for interlaboratory comparisons, method validation, and support laboratory recognition through the NIST OWM, and support for laboratory accreditation efforts, where measurement results are assessed against specific pass/fail criteria.

2.2. NISTIR 7214, Weights and Measures Division Quality Manual for Proficiency Testing and Interlaboratory Comparisons, March 2005 (Draft March 1, 2023)

NISTIR 7214 is the OWM Quality Manual for Proficiency Testing and Interlaboratory Comparisons. This document provides the quality system to ensure that all Proficiency Testing and Interlaboratory Comparison activities within OWM are compliant with ISO/IEC 17043 and ISO 13528 to the extent possible. The quality manual specifies program requirements to ensure that OWM and technical advisory groups, PT coordinators, PT analysts, and participants are technically competent to provide specific types of proficiency testing schemes as required by NISTIR 7082, Proficiency Test Policy and Plan (for State Weights and Measures Laboratories). (NOTE: as written in 2005, the original document was designed to comply with ISO Guide 43 has been updated to comply with the latest versions of ISO/IEC 17043 and ISO 13528).

2.3. NIST OWM Standard Operating Procedures and Resources

2.3.1. *SOP for PTs: Standard Operating Procedure for Office of Weights and Measures Proficiency Tests (OWM PT)*

This procedure is used by the OWM Proficiency Testing (PT) Program to support the State legal metrology laboratories and other laboratories who are members of the RMAPs. This procedure is part of the OWM PT Quality System which includes NISTIR 7214 "Weights and Measures Division Quality Manual for Proficiency Testing and Interlaboratory Comparisons", NISTIR 7082 "Proficiency Test Policy Plan", and associated PT Tools. This rigorous procedure describes how to implement a PT in the OWM program from planning through to final reporting. Specific instructions are provided in the SOP for all stages of proficiency testing to ensure compliance with the OWM Quality System, policies, ISO/IEC 17043, and ISO 13528 to provide rigorous and exceptional quality for participants and to meet minimum requirements of the OWM recognition program and accreditation bodies who are ILAC signatories.

Table 1. PT Phase and Tools Used.

PT Phase	Resource to be Used
Planning	OWM PT Plan Template (Excel file, sections P1, P2, P3, and P4)
Operating	OWM PT Plan Template (Excel file, sections O1 and O2)

Analyzing	OWM PT Analysis Template (Excel file)
Reporting	*Incorporated into PT Plan (Section R1) and PT Analysis Templates*
Follow-Up Actions	GLP for PT Follow Up and Associated form (required by State weights and measures laboratories during annual reviews per NIST Handbook 143, Program Handbook.)

2.3.2. *SOP for Mini-MAPs:*

This procedure is for the operation of a small interlaboratory comparison, typically for two or three laboratories where a full proficiency test among a regional group or a national assessment is not readily available to meet the needs of the laboratory or where there are a very small number of laboratories with similar capabilities. Due to the small number of data points, additional rigorous evaluation of internal laboratory statistics is required. Integrated "measurement assurance" assessments are a key part of conducting small proficiency tests, hence the idea for calling them Mini-Measurement Assurance Programs, or "mini-MAPs". Given the constraints in the usual small number of participants for a Mini-MAP, additional assessments in addition to the proficiency testing (PT) components are essential for providing data validity; assessments include additional evaluations of supporting evidence related to calibration history, traceability, measurement assurance, and uncertainty analysis for each participant.

2.4. ILAC PT Policies

2.4.1. *ILAC-P9:06/2014, ILAC Policy for Participation in Proficiency Testing Activities*

OWM is not an accreditation body nor an ILAC signatory; however, OWM seeks to comply with the policies described in this ILAC policy document. The following items are paraphrased from the ILAC policy, section 4, and includes Notes regarding OWM applications.

1. Accreditation bodies must verify competency of accredited labs; one way may be through proficiency testing. Note: OWM recognizes laboratories based on published criteria in NIST Handbook 143 which directly uses and references ISO/IEC 17025 and requires demonstrated competency through specified training, assigned Laboratory Auditing Program (LAP) problems, onsite observations, and formal PTs and Mini-MAPs.
2. Minimum PT activities related to the laboratory Scope includes a) successful PT participation prior to recognition or accreditation (where available and appropriate) and b) ongoing PT activities consistent with

a documented PT Plan. Note: OWM requires PTs in all measurement areas prior to Recognition, where *reasonably* available according to NISTIR 7082, Policy and Plan, and requires laboratories to maintain a PT Plan (generally through the Regional Measurement Assurance Program, RMAP groups).

3. Accreditation bodies shall have documented policies and may provide additional resources for laboratories regarding PTs and interlaboratory comparisons used for purposes other than PTs. NOTE: NISTIR 7082 Policy and Plan and NIST Handbook 143, Program Handbook addresses OWM implementation of ILAC policy requirements.
4. Some measurement areas in legal metrology may not be practical or readily allow for PTs as demonstration of competence. These include items such as large mass standards above 500 lb, LPG provers, weight carts, railroad test cars, balances, and scales, etc. In those cases, training and on-site observations are substituted as suitable demonstration and evidence of competence.

3. Technical Analysis

3.1. Statistical Concepts and Analyses

Where possible, artifacts with stable historical reference values are chosen for PTs. During the planning process, clear objectives are chosen, and artifacts selected to meet designated PT objectives. Data is evaluated throughout the operation phase of each PT by the PT Coordinator, PT Analyst, and/or OWM staff to provide immediate (as feasible) feedback to each laboratory regarding potential failures or need for corrective actions. Interim E_n or P_n values may have been provided to the laboratory, but reference values of individual standards should *not* be provided during the operating phase of the PT. Because statistics are not finalized during the Operating Phase, interim E_n or P_n values may not match (often do not match) with final PT Reports due to the final PT Analysis and final approved selection of the reference values for each PT artifact. Interim values may occasionally change from final reports, and minor adjustments in the PT plan may also be made (and documented in the final report). A detailed assessment of all data is conducted during the preparation of draft and final reports.

3.1.1. *Official Values Identified for Each Laboratory*

All data is assessed and reported in the final PT report (i.e., no data is omitted). However, to avoid having a mean value that is unduly influenced by multiple participants from a given laboratory, the statistical evaluation represents and uses the data of only one participant from each laboratory for calculations. The data from these designees are referred to as the "official laboratory values". These official values must be designated by the laboratory when submitting PT results to the PT Coordinator or Analyst.

3.1.2. *Initial Data Reviews: Outliers, Blunders, Trends (Drift and Shifts), and Corrective Actions During PT and Draft Reviews (Prior to PT Completion)*

"Initial Statistics" for each PT are calculated using *all* official values. All data is visually evaluated to look for excessive variability, trends/drift, and/or major changes to the measurement results during the PT and after all measurements are completed. Closing values from the starting laboratory may be necessary when there are questions about the stability of the artifact. Data is reviewed for obvious blunders (such as typographical mistakes), unexplained outliers (values outside of three standard deviations of all participant results), uncertainties that are significantly different from peer laboratories with similar scopes, and any data that has widely fluctuating results that may represent artifact instability or poor handling of the standards, and where potential corrective actions may be needed or which can be completed prior to completing the PT round and final PT analysis. New amended values submitted for a laboratory are entered as additional values in the analysis and report and may include selection of amended submissions in lieu of original data as official laboratory values. Notes regarding amendments and corrective actions will be included in the final report.

3.1.3. *Adjusted Statistics (Trimmed Mean, Trimmed Standard Deviation)*

The "Adjusted Statistics" are determined after official values that fail certain criteria are omitted. Values that are outside two standard deviations of the PT official values are flagged in the PT Analysis spreadsheet as "High" or "Low" and then deselected for subsequent calculations of the reference values and uncertainties. The choices for deselection are determined by reviewing all official values and deselecting values in one iteration. This adjustment also identifies any values with gross errors or possible outlying values. When the PT Analysis is done during the course of the scheme, the laboratory may be given an opportunity for immediate corrective actions. In addition, immediate feedback can help the laboratory ensure that problematic measurement results initiate corrective action so that further results are not reported to customers.

In a second step, extreme values that fail the E_n and P_n calculations *may* be omitted to assess the impact on final statistics that will be used for subsequent analyses. The final adjusted mean and adjusted standard deviation are used when evaluating and determining the assigned reference value(s). Values that are deselected are not used in evaluating possible reference value(s).

As noted earlier, all values are retained in the PT Analysis spreadsheet and PT Report with assessment and pass/fail results provided to participants for all submitted values.

3.1.4. *PT Standard Deviation*

The standard deviation of the official values, after any adjustments, is used as the PT standard deviation. This value is used in the Z-score calculation and may be used in ongoing analysis of expected PT variability and estimating future expected PT variability during the Planning phase.

3.2. Determining the Assigned Reference Value and Its Uncertainty

3.2.1. *Metrological Traceability Required for Participants and Assigned Reference Value*

The OWM PT policy and plan (NISTIR 7082) requires all OWM PT participant laboratories to have demonstrated metrological traceability, either through OWM laboratory Recognition, Accreditation through an Accreditation Body that is an ILAC Signatory, or through an assessed process that is compliant with ISO/IEC 17025:2017 and OWM Good Measurement Practice (GMP) 13 (NISTIR 6969). Because metrological traceability is a requirement of all participants, any laboratory, group of laboratories, or all official values (one per laboratory) could conceivably be used during the assessment of results when selecting a suitable reference value, provided that all uncertainty values are comparable. Figure 1 provides an example traceability hierarchy that demonstrates the concept of metrological traceability as a characteristic of each participant laboratory, keeping in mind that each successive level down usually, though not always, has a larger uncertainty. This Figure is also used in the discussion of selecting the assigned reference value.

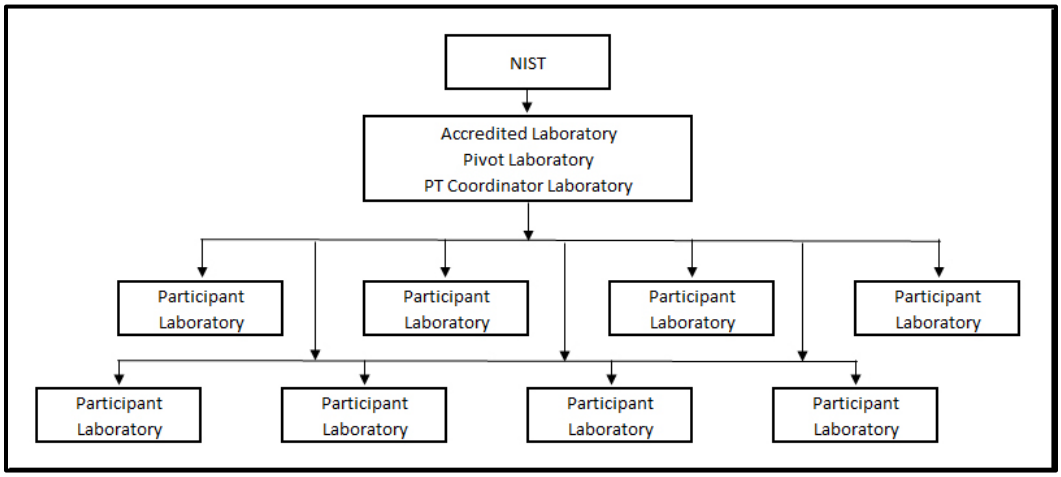

Figure 1. Metrological Traceability Hierarchy.

3.2.2. *Technical Analysis Required for Selecting Assigned Reference Value(s)*

After careful review of all PT data and initial (and adjusted) statistics are determined, suitable reference values and corresponding uncertainties are considered for each item following the documented process described here and in the PT SOP. The hierarchy of selecting an assigned reference value is shown below and preferences are prioritized for selecting an assigned reference value, but the technical assessment of the data by the PT Analyst and OWM staff is required when reviewing options for each standard used in the PT. Even when a higher-level (smaller uncertainty or higher on the hierarchy list) reference value is desired or was used to begin the PT, it may not be the best reference value once all data are reviewed. For example, a precision mass standard may have been calibrated by NIST, but once the item is circulated, its measured mass value might not remain stable throughout the round when compared to the original NIST value, even though it might be stable during the course of the PT. In that case, a consensus value or adjusted consensus value may be the most technically correct choice to use as the assigned reference value. In some cases, ideal reference sources may not provide the smallest or most suitable reference values. It is always appropriate to check the validity of an assigned value against the data from each round of a PT scheme.

As noted in ISO 13528, section 7.1.2. *"Alternative methods for determining the assigned value and its uncertainty may be used provided that they have a sound statistical basis and that the method used is described in the documented plan for the proficiency testing scheme, and fully described to participants. Regardless of the method used to determine the assigned value, it is always appropriate to check the validity of the assigned value for that round of a proficiency testing scheme."*

The OWM PT Analysis spreadsheet allows the PT Analyst and OWM staff to select alternative reference values to determine the most appropriate value for each PT standard artifact.

Choices in the PT Analysis spreadsheet include the following options in a drop down cell for selecting reference values:

- Adjusted mean (will be identical to mean value if no data is deselected; this approach helps to identify values that might have gross errors or are outlier values);
- Adjusted mean with μ_b (uncertainty of bias must be reported and entered during analysis; this will be used in the reference value uncertainty) – rarely used;
- Calibration Source from one laboratory (this selection value defaults to zero in the selection list if no values are entered in the PT Analysis spreadsheet);
- Mean of accredited laboratories and average uncertainty (a formula must be entered to calculate the mean of specific accredited laboratory values and also the mean of their uncertainties) – rarely used;

- Weighted mean and trimmed average uncertainty (set as the default) – this value weights the selection of the reference value based on the reported uncertainty and uses the mean of uncertainties nearest the median uncertainty;

Several of these choices might be used to evaluate the impact on the final PT analysis for all participants; however, the final selection of assigned reference values is evaluated and approved by OWM prior to release of a final report. Criteria used by PT Analysts and OWM staff have been evaluated by NIST statisticians to ensure appropriate values are selected for standards used in each PT. All choices should be considered and compared to other options in the spreadsheet as a part of ensuring the validity of the selected reference value.

3.2.3. *Reference Value and Uncertainty from a Single Laboratory (Externally Derived Criteria) – ISO 13528, section 7.5. ISO/IEC 17043, B.3.1. item c.*

A reference value from a single laboratory may be one from an NMI, such as one from NIST. This might be considered an ideal reference value to use when there is also evidence of stability, and the uncertainties are sufficiently small relative to the participant values. This source is not always an option due to the high cost and the time associated with obtaining this value. Stability of the standard may also make this value less desirable due to the lack of long-term stability of reference values in some measurement areas. In some cases, where standards have demonstrated stability over a long period of time, these values may be used. (Examples: 100 gal prover, 500 lb reference standards). The uncertainty associated a single-laboratory reference value is taken from the calibration certificate. It is critical that this value be compared to all other reference options to ensure validity of the reference value.

3.2.4. *Accredited Laboratory, Pivot Laboratory, PT Coordinator Laboratory or Groups of Expert Laboratories Initial Reference Value and Uncertainty (Externally Derived Criteria) – ISO 13528, sections 7.5, 7.6 Consensus value from expert laboratories; ISO/IEC 17043, B.3.1. items c or d.*

As a part of the PT Plan, OWM and the PT Administrative Team and PT participants may have discussed using an initial reference value and an ending value from an Accredited laboratory, a Pivot Laboratory, or a PT Coordinator laboratory measurement result and uncertainty. This is the next level in a hierarchy from the NIST or NMI value as shown in the hierarchy in Figure 1. Unless measurement results at this level have uncertainties that are significantly smaller than other laboratories in the group, exceptional care must be taken to ensure suitable agreement in the final measurement results to avoid conflict among participants and disagreements about assigned reference value(s). This avoidance of perceived conflict is

especially important given that OWM PTs are not anonymous, and participants are familiar with the other laboratories and capabilities.

Using a single laboratory (often called a Pivot Laboratory) with a "better procedure" is sometimes chosen and may be used to monitor for trends/drift during the scheme, often with before and after measurements, however the evaluation and selection of the reference value must include assessment against other options. The risk with this option may include challenges or appeals to the pivot laboratory value(s) where a laboratory that fails one or more of the statistics used in the analysis. This approach must be compared to other options in the spreadsheet as a part of ensuring the validity of the reference value.

This approach may be suitable in some instances, for example:

- Where more than one level of calibration will be performed in the PT, with some laboratories performing a higher-level procedure (lower uncertainty) and the remaining laboratories performing a lower-level procedure, a mean value from these laboratories may be used to select a best assigned reference value. Calculations of the mean values and uncertainties of the better subgroup of procedures could be used.
- Where the standard to be used in the PT belongs to one of the participants and significant history of calibrations and stability is available the "owner" may be selected to provide initial and closing measurement results and the value from that laboratory used as the initial assigned reference value.

3.2.5. *Historical Reference Value and Uncertainty (Externally Derived Criteria)*

An historical reference value can be an individual value or a collection of values from a variety of sources including past NMI calibrations, past RMAP calibrations, or past accredited lab calibrations. The uncertainty is often a mean of the uncertainty of the selected values (average uncertainty from contributing values). These values can often be used to assess stability of the standard artifacts over time.

3.2.6. *Mean of "Official" Participants and Uncertainty (Consensus Value) (Comparison Derived Criteria) – ISO 13528, 7.7 Consensus value from participant results; ISO/IEC 17043, B.3.1. item e.*

When all official values agree with no need for omitting data as part of the analysis, and when the associated uncertainty is acceptable for the assessment needed, the mean value of all participant results may be used. This value is most often used when there is no other good alternative, or when the tolerances are sufficiently large that the use of this value no

significant negative impact on the analysis. The uncertainty is from the standard deviation of values used, multiplied by *k* as a coverage factor. OWM PTs are coordinated among laboratories that all have demonstrated traceability to the International System of Units (SI) and any or all of the laboratory values could conceivably be used in demonstrating traceability for the reference value (provided robust statistics support the selection decisions). This approach must be compared to other options in the spreadsheet as a part of ensuring the validity of the reference value.

3.2.7. *Weighted Mean and Average Trimmed Uncertainty – ISO 13528, 7.7 Consensus value from participant results, ISO/IEC 17043, B.3.1. item e.*

This is the default method selected in the OWM PT Analysis spreadsheet. After the initial data is reviewed and initial failures are flagged and removed from the reference value analysis, the remaining values and statistics are considered the adjusted, trimmed, or Winsorized mean and include an associated uncertainty. The weighted mean and average trimmed uncertainty are then used to ensure that laboratories with smaller uncertainties contribute a greater proportion of the assigned values. This approach may overestimate the uncertainty of the reference value when the PT standard deviation might be smaller and could impact the normalized error calculations. Again, this approach must be compared to other options in the spreadsheet as a part of ensuring the validity of the reference value.

3.2.8. *Simulations and Monte Carlo Assessments*

Although not widely used for OWM PT analyses, this tool generates simulated values based on an inputted distribution and variables for your data set. Simulation iterations can run in the tens of thousands, hundreds of thousands, or more depending on the computing capabilities. When this analysis is conducted, the values are often entered as additional participant data points for reference in reviewing the graphs and the selection of reference values. This approach has been considered in a number of PTs although it is not explicitly referenced in either ISO/IEC 17043 or ISO 13528, though the standards do reference alternative rigorous statistical approaches that must be documented.

3.2.9. *Multiple assigned reference values.*

Selection of different reference values may be required for each standard within a set of standards circulated for a given PT. Typically this approach is reserved for problem artifacts that seem to be trending in a consistent pattern or direction. Problematic data could include situations where standards are cleaned or damaged in some way and an obvious shift in the data occurred. Combinations of other reference value and uncertainty options may be used for each subgrouping of data. Use of alternative

methods by participants is normally assessed according to method without mixing results for analysis. The summary data chart for standard/artifact in the PT designates the value that was used and selected as the assigned reference value. All other statistics and uncertainties performed for that standard will then be based on the selected reference value.

3.2.10. *Summary of Methods (from PT SOP)*

Table 2. Selection Choices for Reference Values and Uncertainties.

Item	Source	Value	Uncertainty	Comments
3.2.3	NIST or other National Metrology Institute (NMI) Value (demonstrated appropriate through CIPM MRA database review)	From calibration certificate	From calibration certificate	If values are stable and sufficiently small uncertainty
3.2.4	Accredited Laboratory, Pivot Laboratory, Small Subset of Participants working at higher level (Groups of Expert Laboratories)	Value or mean of values calibration certificate(s)	Value or mean of values from calibration certificate(s)	If values are stable and sufficiently small uncertainty
3.2.5	Historically Stable Reference Values	Value used in prior group or mean of values	Uncertainty from value used in prior group or mean uncertainty	E.g., other RMAP regions If values are stable and sufficiently small uncertainty
3.2.6	Mean/Median Value – Consensus	Adjusted Mean or Median value	Adjusted uncertainty	E.g., one value per lab

Item	Source	Value	Uncertainty	Comments
3.2.7	Weighted Mean and Adjusted Trimmed Uncertainty	Adjusted and Weighted Mean or Median value (each value contributes a proportion of the contribution based on uncertainties)	Weighted uncertainty	Must be enough remaining data after adjustments to be valid; only one official value per laboratory is used
3.2.8	Monte Carlo Simulation Values	Special statistics	Special statistics	

3.3. Performance Statistics in the PT Report

The PT Final Report presents the reported measurement results and associated uncertainties. All participants and official participant results from each laboratory are identified and assessed. According to the OWM policy and waiver agreements, there is no assurance of confidentiality in OWM PTs. Laboratories who participate in OWM PTs are notified during planning that they must waive anonymity to participate.

Items that are included in the PT Analysis include:

- Tabulations of data submitted and the baseline analysis for each standard/artifact that was calibrated in the PT. Tables contain the laboratory identification, participant initials, date of calibration, measurement results and uncertainties, initial and adjusted statistics, bias (offsets), E_n, P_n, and Z-scores, status of in/out of two standard deviation limits, and selection criteria for values that were not used in selecting the assigned reference values.
- Summary tables of Pass/Fail statistics showing E_n and P_n values with a total number of failed results for each person.
- Graphs showing measurement results and uncertainties with associated reference values for each standard/artifact in the PT.
- Graphs of E_n and P_n values for each standard/artifact in the PT. In OWM reports, the E_n is graphed with the P_n value. Unlike most PT providers, OWM uses an absolute value for the E_n value so that it can be easily graphed with the P_n statistic on the same chart. To determine consistent directionality of measurement offsets from reference values over time

when evaluating uncertainties associated with minor biases in laboratory results, the laboratory can review the Z-score values.

3.3.1. *Difference or Bias from the Reference Value (Offset), 17043, B.4.1.3., item a, Eqn. B.1.*

The difference, bias, or offset (however named) of each reported value from the selected reference value is calculated and reported as part of the PT analysis data using Eqn. 1. This value is not used as a pass/fail statistic but is used in the initial assessment of data by the PT Analyst and by OWM to review the overall data for obvious blunders and outliers. The laboratory may use this value as a part of its follow-up assessments of laboratory bias, accuracy assessment, and evaluations of recalibration intervals. E.g., for precision calibrations, a laboratory might want to set recalibration goals such that whenever the bias/offset exceeds some ratio of its reported uncertainty, a recalibration or interim assessment of metrological traceability is conducted. Historical OWM PT statistics (no longer used) included an assessment of this offset as shown in Eqn. 2 with a modification of the *Z*-score that was based on laboratory uncertainties rather than the PT statistics. The laboratory may still wish to conduct this assessment for internal evaluations, but it is no longer reported in the OWM PT Reports.

$$x_{lab} - X_{ref} \qquad \text{Eqn. (1)}$$

$$OWM_{historical}\ Z = \frac{x_{lab} - X_{ref}}{U_{lab}} \qquad \text{Eqn. (2)}$$

3.3.2. *Normalized Error, E_n , 17043, B.4.1.3., item e, Eqn. B.6*

Normalized Error, E_n, is defined in ISO/IEC 17043 as the ratio of the difference between the reference value and the reported value compared to the root sum square of associated expanded uncertainties. The normalized error is an indicator of accuracy/inaccuracy as compared to an assigned reference value with respect to the associated uncertainties. Conceptually, the normalized error asks whether the bias is less than the expanded uncertainties of the laboratory and reference value combined in root sum square as shown in Eqn. 2.

$$E_n \text{ assessment: Is } \left(x_{lab} - X_{ref}\right) < \sqrt{U_{lab}^2 + U_{ref}^2}? \qquad \text{Eqn. (3)}$$

OWM uses the absolute value of the calculated E_n results in order to graph multiple statistics on the same charts and to have a simple pass/fail criteria. Using the absolute value, the value of E_n must be less than one to pass. Values of E_n between 0.7 and 1 are highlighted on the charts to alert

laboratories of the possible need to investigate bias with respect to the combined expanded uncertainties.

$$E_n = \left| \frac{x_{lab} - X_{ref}}{\sqrt{U_{lab}^2 + U_{ref}^2}} \right| \quad \text{Result must be} < 1 \text{ to pass.} \qquad \text{Eqn. (4)}$$

A visual assessment of example (unitless) E_n results are shown in Figure 1. Assuming that the assigned reference value of 1.25 with a corresponding expanded uncertainty of 0.5 is correct and acceptable, and that submitted laboratory values vary in a normal distribution, laboratories A, B, and C were selected to illustrate the normalized error concept. In general, the E_n assessment determines the degree to which the measurement results and associated uncertainties overlap each other.

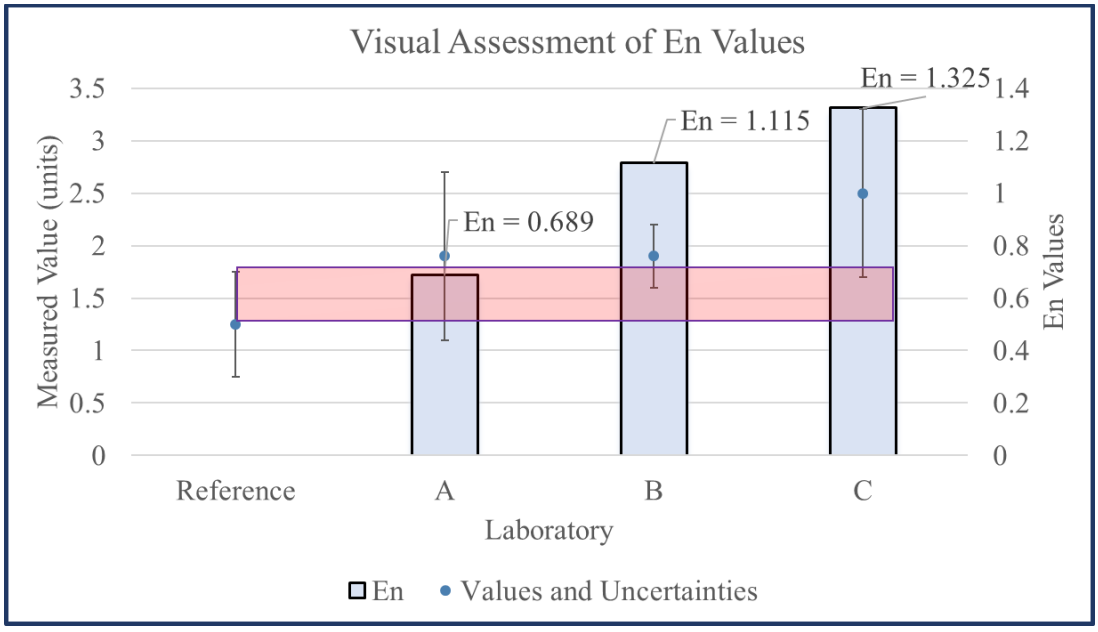

Figure 2. Visual Assessment of E_n Values.

- A: The value submitted by laboratory A is outside the uncertainty of the reference value although its uncertainty overlaps the reference value. Visually, there is a good amount of overlap of the uncertainty bars. The calculated E_n value of 0.689 is less than 1 and passes this assessment. However, an E_n value of 0.689 *might* still warrant further assessment of the laboratory accuracy by determining if the difference or bias that is shown has been consistent in previous PTs or is observed in a laboratory control chart. Further evaluation depends on the applicable tolerances or required measurement limits for the application and the desired level of accuracy needed by the laboratory or its customers.

- B: The value submitted by laboratory B is identical to the value from laboratory A, thus the Bias (Difference) calculated from Eqn. 1 is identical. However, the uncertainty for laboratory B is smaller than the laboratory A uncertainty, it is also smaller than the reference value uncertainty. While the uncertainty values still overlap slightly, the laboratory uncertainty does not overlap the reference value, the uncertainty of the reference value does not overlap the submitted laboratory B value, and the E_n value of 1.115 fails the assessment. As noted, the Bias (Difference) for both laboratories A and B are identical, but the uncertainty for laboratory B does not support this level of bias. Either the uncertainty is too small if all other laboratories performed the using a similar procedure and submitted uncertainties comparable to Laboratories A and C (likely) or the laboratory needs to identify the root cause of this failure (e.g., a systematic error of some type or the need for recalibration of standards to bring values closer to the reference value). In this case, the laboratory might question the choice of reference values, reinforcing the importance of rigorous analysis of reference values.
- C: The value submitted by laboratory C is not inside reference value uncertainty and its uncertainty is the same as that of laboratory A. In this case, there is very minor overlap of uncertainty values, but the overlap is not enough and the calculated E_n value of 1.325 fails this assessment and corrective action is needed to identify the cause for the bias shown in the results. Some laboratories working with larger tolerances might suggest that an offset of this nature "does not matter" and the failure is "not significant", which is counter to the purpose of PTs. When tolerances are significantly larger than the offset shown in this case, a larger uncertainty to cover the gap and pass the E_n assessment is likely warranted.

Note that the observed biases for all three laboratories A, B, and C do not pass criteria in SOP 29, Standard Operating Procedure for Assignment of Uncertainty (NISTIR 6969) to allow incorporation of the bias into the uncertainty!

3.3.3. *Normalized Precision, P_n*

The Normalized Precision, P_n, is a performance assessment of fitness for purpose (suitability) of the laboratory uncertainty compared to applicable documentary standards and is related to decision rules and conformity assessments as described in ISO/IEC 17025. Where decision rules and conformity limits are provided and reported uncertainty must be considered, the precision assessment, P_n is conducted. The precision assessment asks whether the reported uncertainty is less than the specified limits, as shown in Eqn. 4 where the example is given that uncertainty must be less than one-

third of the maximum permissible error (as is the case in mass calibrations according to OIML R111 and ASTM E 617).

$$P_n \text{ assessment: Is } U_{lab} < \frac{1}{3} m.p.e.? \qquad \text{Eqn. (5)}$$

The precision assessment is a ratio of the reported uncertainty versus the decision rule limits. Passing values for the precision assessment are less than one and are graphed with the E_n values. This statistic is unique to OWM assessments but is related to ISO/IEC 17025 decision rules and ISO/IEC 17043 performance assessments. Documentary standards used in legal metrology generally specify appropriate uncertainty to tolerance (or maximum permissible errors, m.p.e.) ratios on which to base decision risks. In this supplemental report, tolerances and m.p.e. terminology is used interchangeably. Documented decision risks and use of uncertainties in making conformity decisions are specified in the ISO/IEC 17025 standard. Many of the OWM published procedures and documentary standards that are referenced for legal metrology include uncertainty to m.p.e. ratios of 1:1 or 1:3, where the uncertainty must be less than the applicable m.p.e. or the uncertainty must be less than one-third of the m.p.e. The 1/3 ratio is common in international legal metrology documentary standards such as those from the International Organization of Legal Metrology (OIML) and a number of the NIST Handbook 150-x series documentary standards. ASTM E617 for mass standards also includes this common ratio of uncertainty to tolerances. In some cases, the ratio will be 1:1, where the uncertainty must simply be smaller than the applicable tolerance or m.p.e.

The P_n value should always be assessed by the laboratory prior to participation in applicable PTs with corrective action taken prior to participation. Failures of the P_n assessment are preventable with appropriate risk mitigation methods and illustrate a failure of complying with the precision requirements and a failure of completing suitable corrective action. Failures of the P_n statistic in a PT always require suitable follow up corrective action and may immediately impact laboratory Recognition and or Accreditation status.

$$P_n = \left| \frac{U_{lab}}{\frac{1}{3} m.p.e.} \right| \text{ Result must be < 1 to pass.}$$

alternative ratios that may be used: Eqn. (6)

$$P_n = \left| \frac{U_{lab}}{m.p.e.} \right|, \; P_n = \left| \frac{U_{lab}}{\text{fraction or \% of } m.p.e.} \right|$$

A visual assessment of example (unitless) P_n results are shown in Five examples are shown to illustrate the relationship between the maximum permissible error (m.p.e.) or tolerances and the uncertainties submitted by the laboratory. In the P_n assessment, the actual values are not what is being assessed. For example, laboratory A is exactly the same as the reference nominal value of zero error, yet the calculated value of its normalized precision is 3, and fails the requirements of being less than one-third of the m.p.e. Also, in the case of laboratories D and E, they have identical passing P_n results even though laboratory D reported a result identical to the reference nominal value and laboratory E is significantly away from the reference value (and would likely fail an E_n assessment).

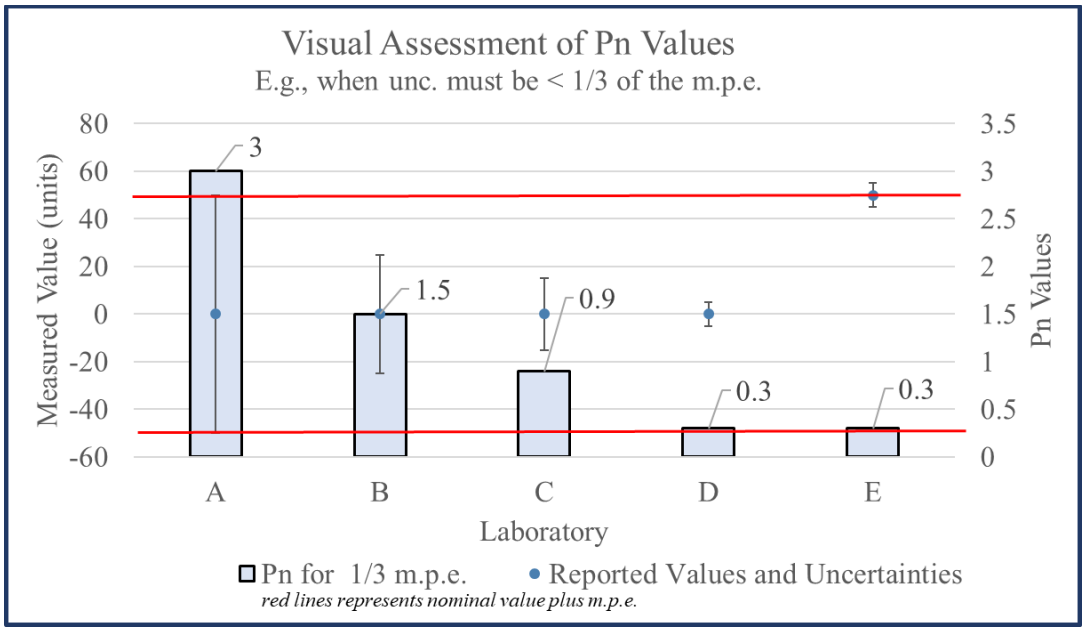

Figure 3. Visual Assessment of P_n Values.

Laboratory B fails the P_n assessment because the uncertainty is one-half of the tolerance instead of one-third. Laboratory C passes this assessment but is very nearly at the limit of 1 and may want to evaluate the uncertainty further.

3.3.4. *Z Score , ISO/IEC 17043, item B.4.1.3. item b, Eqn. B3.*

This statistical evaluation of Z Score comes from ISO/IEC 13528, 3.7 as: *"standardized measure of performance, calculated using the participant result, assigned value and the standard deviation for proficiency assessments"*. The Z-score may be used in combination with the adjustment statistics (trimmed mean and associated uncertainty) described earlier.

OWM reports this value in the tables of the PT Report but does *not* use this statistic for pass/fail criteria in PTs because the Z score does not include

assessment of the laboratory uncertainty. The value may be used in isolating values outside 2 standard deviations of the accepted reference values. According to ISO/IEC 17043, satisfactory performance is generally indicated as $Z \leq 2$; unsatisfactory performance is indicated as $Z > 3$. and marginal performance is anything between $Z > 2$ and $Z \leq 3$. However, further evaluation of the Z score requires an assessment of the observed bias from the assigned reference value with respect to the reported laboratory uncertainty, such as is provided by the E_n assessment. However, the directionality (positive or negative values) of this statistic can provide additional insights to the laboratory for ongoing evaluation of differences, bias, or offsets in measurement results especially when compared to internal measurement assurance data.

$$Z - score = \frac{x_{lab} - X_{ref}}{s_{PT}} \qquad \text{Eqn. (7)}$$

The Z-Score statistic and analysis is very similar to that of control charts where plus and minus two standard deviations serve as warning limits and plus and minus three standard deviations are the control or action limits. In the case of the PT, however, the standard deviation of the PT is based on the final statistics of the official values when any adjustments (if needed) have been completed. In the graph shown in Figure 3, the Z scores for each laboratory are given on the X-axis with the laboratory identification. It can be seen that the values are sequentially placed on one standard deviation intervals. Again, the assumption must be made that these laboratory values were selected for illustration purposes and the submitted values are all normally distributed around the assigned reference value.

The bias (difference) determined with Eqn. 1 is observable in these values and may impact which values are used in the selection of the assigned reference values, but further evaluation requires consideration of accuracy in conjunction with the E_n assessment, the reported uncertainty, and any applicable tolerances.

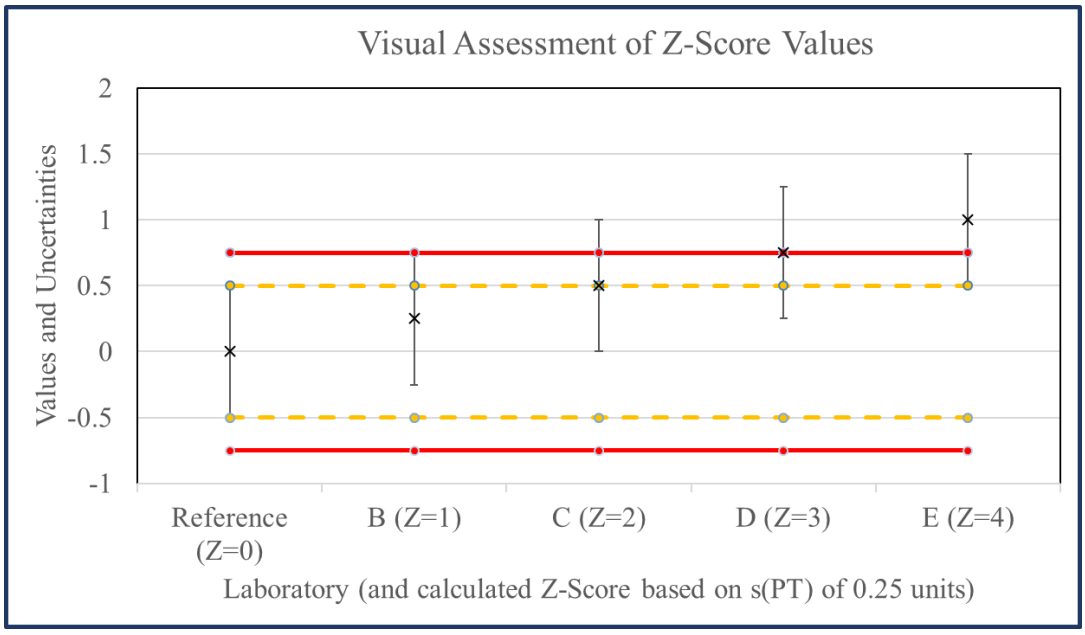

Figure 4. Visual Assessment of Z-Score Values.

4. Non-statistical Pass/Fail Criteria

Some PTs are planned and designed to assess laboratory participation for performance measures unrelated to specific measurement results and associated uncertainties. Additional non-statistical pass/fail criteria might include any or all of the following items that are explained in the PT Report. These criteria should have been selected as a part of the PT Plan during the original planning phase to ensure that all laboratories are aware of the additional assessments:

- Compliance of the certificate to ISO/IEC 17025, Section 7.8;
- Errors on submitted certificates and/or data sheets;
- [Unreasonable] time delays on standard/artifact shipments and/or report submission (e.g., communicating with the coordinator; reports within 2 weeks);
- Improper packaging and shipping (and handling);
- Deviations from the approved and accepted PT Plan (e.g., using a different SOP);
- Switching or substituting standards or PT artifacts with laboratory artifacts;
- Unapproved cleaning, adjustments, or other identified care and handling problems;
- Uncertainty analysis: Detailed uncertainty analysis may be planned as part of the PT Plan. Uncertainty components as specified in the SOP and PT Plan were not included.
- Uncertainty reported on a certificate that is smaller than what is on the published Scope (for Accredited labs).

5. Follow-up Actions (Corrective, Improvement, Tracking)

Pass/fail status of each standard evaluated in the PT is not the only thing a laboratory should consider when participating in a PT. The OWM Good Laboratory Practice for PT Follow-

ups is a procedure designed to enable a thorough follow-up assessment and includes writing an Executive Summary that can be used in a Management Review and guiding the laboratory in performing a thorough assessment of the PT Report and the analysis of their results; this further follow-up assessment is valuable as a communication tool even when all indicators were successful. Ongoing tracking and evaluating of PTs are part of ensuring the validity of measurement results provided by the laboratory and the PT follow up assessment should be integrated into evaluating laboratory measurement assurance data from other sources and include review of data from periodic calibrations, internal evaluations of reference standards, similar past PTs, control charts, repeatability charts, and other statistical analyses. Regular assessment of PT data, even when successful, can mitigate risk and provide opportunities for continual laboratory improvement. See the Good Laboratory Practice for PT Follow Ups and its associated form.