Proficiency Testing Program for U.S. State Weights and Measures Laboratories

Elizabeth J. Gentry, Georgia L. Harris and Val R. Miller

Abstract: The National Institute of Standards and Technology (NIST) Office of Weights and Measures (OWM) manages a State Laboratory Program for weights and measures laboratories that includes: 1) Laboratory recognition using *ISO/IEC 17025:2005* and sponsorship of accreditation through the National Voluntary Laboratory Accreditation Program (NVLAP); 2) Hands-on training courses held at NIST, regional measurement assurance program (RMAPs) training held annually throughout the United States, and a number of web-based short courses; and 3) Formal proficiency testing and interlaboratory comparisons (PT/ILC). The main objective of the State Laboratory Program is to ensure nationally consistent measurement results, acceptable accuracy and metrological traceability, and the credibility and acceptance of state laboratory measurements. This paper presents the key features of the PT/ILC program: measures of success; collaboration challenges; the use of template tools; and continual improvement efforts. Note that while most interlaboratory comparisons are also proficiency tests, some are not. For simplicity, however, this paper will refer to the PT/ILC effort as the PT program.

1. Introduction

The NIST Office of Weights and Measures (OWM) State Laboratory Program [1] has operated interlaboratory comparisons since the early 1980's as a key component of a measurement assurance program and currently completes between 10 and 15 PTs each year. These PTs are conducted on a national or regional basis through six formal Regional Measurement Assurance Program (RMAP) groups as shown in the map in Fig. 1. The RMAPS include the Caribbean Measurement Assurance Program (CaMAP), the Southwestern Assurance Program (SWAP), the Southeastern Measurement Assurance Program (SEMAP), the Northeastern Measurement Assurance Program (NEMAP), the MidAmerica Measurement Assurance Program (MidMAP), and the Western Regional Assurance Program (WRAP). Participants include weights and measures laboratories and other government (e.g., the U.S. Department of Agriculture, Grain Inspection Packers and Stockyards Administration) and industry laboratories as associate members. OWM formalized proficiency testing and interlaboratory comparisons in 2004 by adopting a PT policy [2] to support laboratory recognition and accreditation and to demonstrate metrologist competency in state laboratories. The PT program was further refined in 2005 by publishing and adopting a quality manual [3] for operating and participating in the PT program that follows *ILAC G13* [4] and *ISO/IEC Guide 43* [5]. Changes that are being made now will ensure future compliance with *ISO/IEC 17043:2010* [6] and use of *ISO 13528:2005* [7].

The PT program is primarily operated by NIST staff and by volunteers who are experienced metrologists; have participated in prior PTs, and who are initially mentored as coordinators or analysts. A formalization of voluntary collaboration designations will be implemented with participants, coordinators, and analysts functioning collaboratively as a Technical Advisory Group to NIST as a result of the new *ISO/IEC 17043* standard which requires greater formality of program roles and functions. NIST assistance and oversight is a critical part of the entire process and includes: needs assessment, PT planning, review of the data entry and analysis, and final approval of PT reports. Implementing a quality management system among volunteer collaborators with NIST oversight presents a number of unique management challenges. Each RMAP group accepts a level of ownership and responsibility for program operation to ensure that each laboratory has completed essential proficiency tests. Methods to respond to these challenges include using database tracking tools, standardized analysis methods, and templates for analysis and reporting. The feedback provided by the participants and the NIST staff is used to continually improve the program.

2. Procedure

OWM is part of the NIST Physical Measurement Laboratory (PML). One of our strategic goals beginning in the 1990's was to implement the Baldrige quality framework into our operations as a method of achieving performance excellence according to OWM customers. The Baldrige criteria (http://www.nist.gov/baldridge) include seven key categories of implementation and assessment. The categories are 1) leadership, 2) strategic planning, 3) customer and market focus, 4) measurement, analysis, and knowledge management, 5) human resource focus, 6) process management, and 7) business results.

As a part of the strategic planning (category 2) efforts, the quality and effectiveness of our PT program is regularly evaluated through a Plan-Do-Check-Act process (Deming or Shewart Cycle) and through strength, weakness, opportunity, and threat (SWOT) assessments. Actions are selected to take advantage of opportunities and strengths and to minimize weaknesses and threats. During the assessments, a

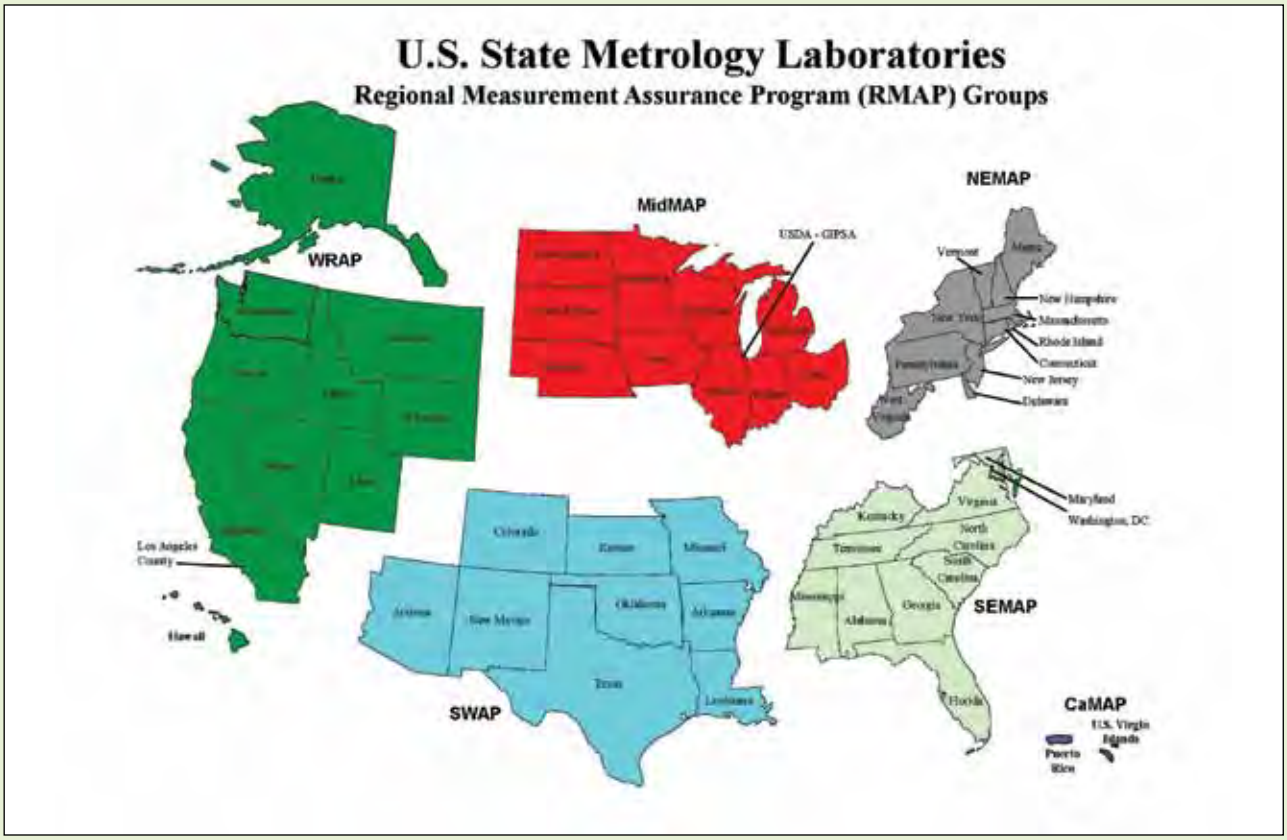

Figure 1. U.S. State Regional Measurement Assurance Programs (RMAPs).

number of questions are asked to address PT program effectiveness. These include, but are not limited to, the following kinds of questions associated with the Baldrige categories:

- Category 3, Customer and Market Focus: What does success look like? For us? For our customers? For their customers?
- Category 4, Measurement Analysis and Knowledge Management: What measures are meaningful assessments of PT program impact (output and outcome)?
- Category 5, Human Resource Focus: How do we develop staff and delegate expertise in coordinating and analyzing PTs, especially when working with collaborative partners?
- Category 6, Process Management: How can we regularly improve the processes associated with the PT program?
- Category 7, Business Results: Are the PT program objectives of ensuring uniformity and acceptance of measurement results made by state laboratories being met? Are measurement procedures developed by NIST implemented successfully after training is conducted? Does implementation lead to improved measurement consistency? Are corrective actions that are identified in laboratory assessments, training, and proficiency testing effectively completed? Do the data from the PT program provide credibility to laboratory recognition/accreditation efforts?

By asking these types of questions over a number of years, the outputs and outcomes of the PT program have been identified. These measures attempt to quantify program effectiveness and impact. Outputs are defined as numerical measures used to quantify various PT program aspects such as how many PTs are conducted each year, how many measurement results pass, and how many laboratories conduct follow-up actions (corrective, preventive, and continuous improvement) as a result of PT participation.

Output measures have been tracked for more than twenty years and include numbers of completed PTs as shown in Fig. 2, and the number of parameters covered each year as shown in Fig. 3. Results from the State Laboratory Program workload surveys [9] are used to help target the areas of greatest need, resulting in a distribution of PTs as shown in Fig. 3. These output measures provide guidance on the level of effort needed at both the NIST and laboratory levels to manage the PTs and to identify whether all measurement parameters are adequately covered, but they do not necessarily measure PT program outcomes. For example, 90 % of state laboratory

Authors

Elizabeth J. Gentry
elizabeth.gentry@nist.gov

Georgia L. Harris
gharris@nist.gov

Val R. Miller
val.miller@nist.gov

National Institute of Standards and Technology
Office of Weights and Measures
100 Bureau Drive, MS 2600
Gaithersburg, MD 20899

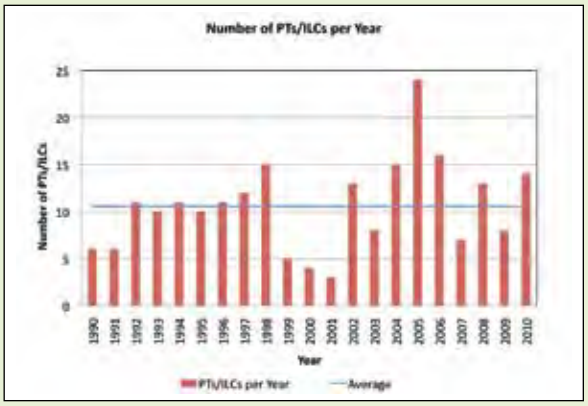

Figure 2. Number of PTs per year.

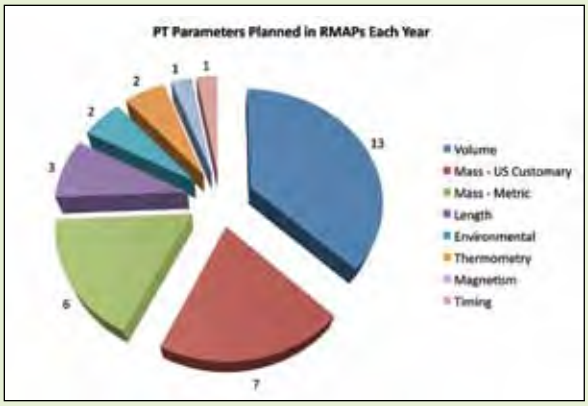

Figure 3. Planned proficiency tests by parameter (typical).

calibrations involve mass parameters, but volume measurements (only 2 % of the laboratory workload) have a significant economic effect on the petroleum industry and are tracked as well. Uniformity in volume measurements to support the petroleum infrastructure is increasingly important. OWM is often asked about coordinating proficiency tests that are outside of the typical parameters needed to support legal metrology. The OWM program is operated to support the needs of U.S. State weights and measures laboratories, and its range of available PTs is fairly narrow when compared to the full spectrum of what an accredited PT provider might have available.

PT program outcomes are defined as big picture, qualitative system-level improvements that can be somewhat conceptual, such as more accepted measurement results, improved validity and acceptance of traceable measurement results from state laboratories, and full implementation of NIST procedures and training. The goal of outcome measures is to assess program impact. One of the metrics we use to determine how well laboratories implement NIST procedures and training is to assess how well metrologists perform on proficiency tests.

OWM recognition and the National Voluntary Laboratory Accreditation Program (NVLAP) accreditation use *ISO/IEC 17025:2005* [1, 8]; additional technical guidelines include criteria that specify laboratory environments, reference standards, instrument quality, and management operations. Because compliance with *ISO/IEC 17025* and the technical criteria are essential for consistently good measurement results, successful PTs provide an outcome measure of the overall laboratory operation and demonstrate individual competence. Failed PTs identify areas that require correction or improvement.

The NIST OWM PT program began capturing data for additional outcome measures in 2006, as shown in Fig. 4 and 5. Critical outcome measures identified for evaluating PT program effectiveness include the percentage of PTs that are passed by the laboratories, as shown in Fig. 4. The Pass/Fail criteria will be discussed in detail in Section 5.

Another outcome measure is increasing the percentage of completed and effective corrective, preventive, and improvement actions, regardless of whether the PT officially failed, as shown in Fig. 5. Group discussions are held at each annual RMAP meeting to review PT results. The discussion is essentially a training exercise, and the laboratories who successfully pass the PT often share their best practices with the laboratories who fail. Informal root cause analysis is conducted by the group during the presentation, and results are evaluated by asking questions about why a laboratory failed or how they passed. Laboratories also conduct independent follow-up assessments to identify potential problems and trends that might not have caused them to fail a PT, but that still indicate a need for improvements or preventive actions. Examples of follow-up actions reported between 2007 and 2010 are shown in Fig. 5.

3. Collaborative Coordination

State and industry laboratory participants partner in the NIST OWM PT program by planning the PTs and by coordinating and then analyzing the initial data. This is followed by NIST reviews of draft and final reports. Metrology laboratories that participate in a PT activity must agree to quality policies and conditions [2], including training at NIST and annual attendance at RMAP training and participation in PT discussions. Participants can be excluded from future PT program participation when quality policies and procedures are not followed. In the early years, the PT results were coded to protect the participant's identity. However, by the end of each RMAP training session, the participants had all shared their coded identity with each other. Thus, one of the program policies is that participants must agree to openly report their PT results; anonymity is not implied nor guaranteed. We have found that open discussions provide better training and improvement opportunities, and help all participants to perform better measurements.

Operating collaboratively within a quality management system is a challenge, though most laboratories are willing to support the PT program. Participants must reconcile issues that arise from complying to a quality management system that is not their own. The *NISTIR 7214* quality manual [3] uses several forms, templates, and spreadsheets that are not a part of individual laboratory document control policies or subject to their software validation criteria. PT analysts are required to submit training documentation that might duplicate the technical training records maintained in their quality management system. PT follow-up corrective action forms must also be completed in addition to internal corrective action process or measurement assurance programs as a part of the laboratory recognition effort. The recognition program allows alternative follow-up forms that have been integrated into the laboratory's quality management system [1]. NIST reviews the data entry, analysis, and final reports for every PT. Additional effort is required when overseeing a new coordinator or analyst, and more

Figure 4. PT success measures.

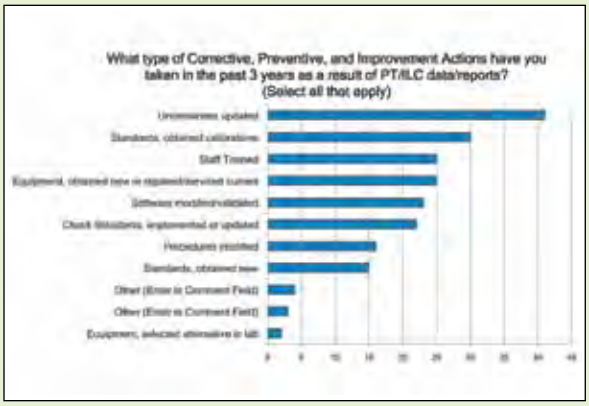

Figure 5. PT follow-up actions (improvement, preventive, and corrective).

experienced volunteers often mentor newer ones. Each participant in an RMAP could be expected to coordinate or analyze a proficiency test and generate a draft final report.

The PT program participants have diverse experiences as coordinators and analysts. The average state laboratory metrology experience level is just over nine years, with staff ranging from new hires (no experience) to master metrologists (35 years of experience) [9]. The long-term success of the PT program depends on strategically recruiting new analysts in the various RMAP regions, providing formal training, and relying on extensive mentoring and guidance from NIST and experienced RMAP participants, thus addressing the human resource and leadership concerns associated with Baldrige categories 1 and 5.

4. Planning and Participation

The PT program requires each laboratory to complete a minimum level of proficiency testing that is established to meet international accreditation requirements [2]. Laboratories must participate in at least one PT for each measurement parameter included in the laboratory's scope of recognition and/or accreditation during each four-year period. A greater frequency may be required for some parameters. For instance, nearly 90 % of the state laboratory workload is in mass calibrations [9]. Therefore, more PTs are required in mass than in a discipline such as length, which typically accounts for less than 1 % of a laboratories' workload. Based on the scope of laboratory recognition or accreditation, the local participation requirements will vary. The RMAP groups develop a plan for the current and future years to meet the requirements for all participating laboratories.

For recognition and accreditation purposes, the staff members who are authorized to sign calibration reports for specific tests must each be able to demonstrate proficiency. The laboratory manager is responsible for ensuring that all personnel who are authorized to perform a particular type of measurement participate in PTs when they are offered. Full participation by all trained staff increases the number of PT participants quite dramatically for common measurement parameters.

If there are too few laboratories in a region to run a comparison in a given measurement area, the RMAP must have an alternative plan. For example, a national PT may be necessary to meet laboratory recognition or accreditation requirements. In addition to ensuring that enough PTs are available to cover the entire scope of a laboratory, NIST also ensures that the ranges of measurements are covered as well as possible and often coordinates the national proficiency tests. PTs for specialized or lower workload areas (e.g., Mass Echelon I, Rigid Rule) are often coordinated on a national basis.

Certain challenges are associated with transporting large, bulky or massive calibration standards across the United States (such as 100 gal[1] volumetric standards, 500 lb mass standards, and even some of the precision mass standards where possible impacts to artifact stability need to be controlled or minimized). Participants often cooperate by travelling to meet each other at designated locations to transfer custody of the PT artifact. The national PTs are often coordinated through one RMAP region, where the artifact is calibrated at each laboratory with the appropriate scope. The report for that region is reviewed by NIST and published as an interim report. The artifact is then sent to another RMAP region where the process is repeated. However, when the data are compiled, they are added to the prior RMAP report(s) to create a national report. This strategy allows NIST to monitor the national measurement system and the capabilities of all laboratories that participate.

Allowing each laboratory to participate in all PTs at their best measurement capability (versus similar capability) was once common in many RMAPs. However, analyzing data with three groups of uncertainty and capability was found to be ineffective and made PT analysis difficult for several reasons. For example, *NIST Handbook 143* describes Mass Echelon I, II, and III which correspond to the International Organization for Legal Metrology (OIML) recommendation *R111* [10] E, F, and M classes of weights—each with differing specifications, calibration requirements, tolerances, and uncertainties. One graph scale was once used to display three distinctively different uncertainty levels. The small variations typically observed among laboratories operating at the Mass Echelon I level could not be effectively graphed with Mass Echelon II calibrations which have larger uncertainties. This made it nearly impossible to effectively determine the statistical elements (e.g., standard deviation, mean, median) and reference values when comparing the results from laboratories with different calibration capabilities, or to show these results on the same graph. For these reasons, nationally coordinated

[1] Artifacts are tested in both the SI and U.S. customary system and measurement unit systems are selected to reflect that of the laboratory workload.

PTs designed to consider specific calibration levels and grouping of uncertainties, such as at the Mass Echelon I level, are now preferred. This also presents challenges due to the infrequency of national meetings where results can be presented and discussed. However, the infrequency of national meetings can now be supplemented by the use of web meetings.

OWM also coordinates PTs with NVLAP to ensure that sufficient tests are completed to demonstrate proficiency. The accreditation status of the state laboratories is one of the metrics OWM uses to determine PT program success, so we work closely with NVLAP, the RMAPs, and the individual laboratories to ensure that enough PTs are regularly conducted. In the case of dedicated NVLAP PTs, the confidentiality of the individual participants is maintained as described in *ISO/IEC 17043:2010*, section 4.10, unless all participants agree to an open format.

The OWM PT program shares operational costs through collaboration. NIST generally pays for shipping costs to and from state laboratories, and for the cost of calibrations performed at NIST. Industry participants generally pay their own shipping expenses. State and industrial laboratories coordinate PTs and analyze the initial data, followed by NIST reviews. The coordination time for a typical PT can range from 8 to 24 hours. The simpler PTs might only require from 5 to 8 hours for analysis and report preparation. However, the more complex PTs might require 30 to 40 hours of analysis and preparation time. The PT program encourages laboratories to share resources and demonstrates the economic benefits of collaboration.

5. Template Tools

A number of tools are readily available on the NIST website for planning, coordinating, and analyzing PTs. The tools are standardized and are an integral part of the PT quality management system. The tools include a planning form, reporting checklist, reporting form, analysis guidance (including standardized statistics), standardized terminology, and analysis spreadsheets. The development of these template tools has led to many positive results.

5.1 Planning Form

The planning form is used to standardize the planning process for each PT. This form includes entries that define the purpose of the PT, identify artifacts, clarify the measurand in question, identify all participants and refine their purpose for participation. Measurement system requirements and the measurement process are also specified. Additionally, the expected measurement results are clarified, as is guidance about the expected form of the reported measurement results and the analysis that will be performed. Using the planning form also ensures that all participants agree to the plan and have access to all pertinent information. Prior to the use of this form, many of these issues were addressed informally and the results of the PT cycle did not necessarily meet the stated objectives.

5.2 Reporting Checklist and Template Report

A reporting checklist was originally used to help volunteer analysts ensure that all information needed in the PT report was included. A template report was later developed to replace the reporting checklist. All of the components on the template report have improved uniformity, which is especially helpful when reviewing draft and final PT reports. The sections of the template report correspond with the PT planning checklist, providing consistent formatting. This consistency makes the final report easier to read and makes it possible for participants and assessors to quickly locate information.

Guidance is provided to the volunteer analyst regarding the information that should be entered in each section of the template report. This guidance makes it less time consuming to complete the report. Each report must be reviewed and approved by NIST staff prior to publication. The NIST staff has a national perspective and views many PT reports each year, making it possible to evaluate data with a broader focus. In some cases, this national perspective makes it possible to identify measurement issues that might be systematic or undetectable by a regional or accreditation body.

5.3 Analysis Guidance

An analysis guidance document is provided to PT analysts to ensure the consistent treatment of data. This includes selecting the best reference value and associated uncertainty, handling data with excessive offsets from the mean or reference values (outliers), and handling data from laboratories whose uncertainty values are significantly different from the rest of the group.

Each analysis begins by first identifying the official reported values for each participating laboratory. This step is necessary because some laboratories submit values from each staff member who is approved for performing calibrations in a given parameter. If a PT has 15 reported values and five of the values came from one laboratory, that laboratory has the potential to unduly influence the mean or median value, which may be used to determine or validate the accepted reference value. To prevent one laboratory from biasing the mean, all laboratories are allowed to equally contribute to the reference value, providing that their results are not outliers. All submitted data are shown in the analysis, but only one official value from each laboratory is allowed to contribute to the reference value.

The initial mean and standard deviation of all the reported official data are used in every PT analysis as the next level of analysis. Any submitted value that is outside two standard deviations of the median is discarded before calculating an adjusted mean (trimmed mean) and adjusted standard deviation. The adjusted mean and standard deviation may be used to validate the reference value or to establish a consensus reference value and uncertainty when a better reference value is not available. Additional guidance is provided to participants on how to:

- assess the initial data;
- look for artifact and data stability;
- use statistical tests and tools to evaluate shifts, drifts, and general instability where appropriate;
- review the data to see how well all possible potential reference values might agree;
- determine if and when any shifts in the data occurred and if the suspected change exceeds the standard deviation of the data set; and
- determine whether drift occurred and if the suspected drift on a trend line exceeds the standard deviation among the laboratories.

Many of these additional considerations may require more rigorous statistical treatment than most volunteer analysts will be able to perform. A hierarchy for how to select the best reference value is

Selection Level	Value	Criteria	Associated Uncertainty
1	National Measurement Institute (NMI) value.	Depends on measurement area, and the level of the PT (e.g., an NMI calibration may be excessive), and date of original calibration. The value needs to be compared to the mean and median values and may not be current or valid depending on artifact stability.	Reported by the NMI for the calibration.
2	Mean, adjusted mean, or median.	One point from each laboratory is preferred (the official value), but all results may be used if there are not an excessive number of participants from any single laboratory.	Standard deviation of the mean or adjusted standard deviation of the mean of those points used to determine the reference value times a coverage factor (e.g., Student's *t*-distribution based on degrees of freedom) or average uncertainty of the points used to determine the reference value.
3	Mean or median of some designated laboratory values.	For example: 1) labs working at the lowest uncertainty levels; 2) only accredited labs; 3) only labs who all have recent calibrations of their standards and values agree well.	Standard deviation of those points used to determine the reference value times a coverage factor (e.g., Student's *t*-distribution based on degrees of freedom) or average uncertainty of the points used to determine the reference value.
4	Pivot lab values.	Designated during the planning phase.	Pivot lab reported uncertainty.
5	Value provided.	For example: 1) past data for the artifact; 2) mean of prior high-level calibrations.	Uncertainty provided, or associated with the past data or other calibrations.

Table 1. Reference value selection hierarchy.

provided as part of the analysis guidance. This reference document describes the recommended hierarchy to be used in determining the reference value and uncertainty of a PT. The selection process builds on training efforts and discussions that have spanned many years and represents a consensus reached between OWM and the NIST Statistical Engineering Division for purposes of proficiency testing. The hierarchy generally follows the levels in Table 1.

The normalized error (E_n) evaluation is used in all PTs to compare each laboratory's reported result (x_{lab}) and its expanded uncertainty (U_{lab}) to the reference value ($X_{reference}$) and its uncertainty ($U_{reference}$). This is used as the first of the Pass/Fail statistics, Eq. (1), and is also used as a means of identifying outliers when selecting data to determine the reference value. Official laboratory values failing the E_n test are usually excluded from the data when calculating the adjusted mean and adjusted median to avoid influencing the value by excess offset. An E_n value of less than 1 indicates that the PT passes**.** The absolute value of E_n is used to enable graphing multiple values on a single chart.

$$E_n = \left| \frac{\left(x_{lab} - X_{reference}\right)}{\sqrt{\left(U_{lab}^2 + U_{reference}^2\right)}} \right| \quad (1)$$

The acceptability of the reported uncertainty for each reported measurement result is further evaluated against tolerance requirements (when applicable) using a normalized precision test (P_n), as shown in Eq. (2). In this test, the uncertainty is compared to criteria established by the applicable documentary standard for the artifact (for example, some documentary standards require the uncertainty of the calibration to be less than one-third, one-fourth, or even one-tenth of the applicable tolerance). A P_n result of less than 1 indicates the PT passes. Official laboratory values failing the P_n test may be excluded from the reference value selection due to excessive uncertainties. Laboratory values with large uncertainties may pass the E_n test while having large bias; if included these values would negatively influence the reference value selection.

$$P_n = \frac{Unc_{lab}}{\frac{1}{3}Tolerance} \quad (2)$$

5.4 Standardized Terminology

The terminology document contains sample text that can be used when preparing a PT report. There are basic explanations of the statistical tests as well as sample terminology to describe the methods used for selecting the reference values and uncertainties. The wording of the sections may need to be altered in order to make sense for the PT being evaluated. Sample text is also provided for situations where corrective action will be recommended. Coordinators and analysts are part of the Technical Advisory Group and may provide suggested corrective actions, mentoring and consultation in the report. However,

Artifact Counter:					1		2		3	
Tab name:					'Data (1)'!		'Data (2)'!		'Data (3)'!	
These cells are referenced in the other worksheets to minimize data entry effort.										
Enter Nominal Denomination ID on this line:					1 kg		1 kg *		500 g	
Date of Test	#	Official Value	SOP	Participant ID	Reported Value	Reported Unc	Reported Value	Reported Unc	Reported Value	Reported Unc
1/2/15	1	*	4	1	1.29	0.12	1.65	0.12		
1/3/15	2	*	5	2	1.27	0.12	1.58	0.12		
1/4/15	3	*	5	3	1.2836	0.063	1.669	0.063		
1/5/15	4	*	5	4	1.278	0.066	1.607	0.066		
1/6/15	5	*	5	5	1.278	0.066	1.607	0.066		
1/7/15	6	*	4	6	1.29	0.21	1.65	0.21		
1/8/15	7	*	4	7	1.29	0.2	1.65	0.2		
1/9/15	8	*	4	8	1.267	0.051	1.653	0.051		
1/10/15	9	*	4	9	1.286	0.11	1.59	0.11		
1/11/15	10	*	4	10	1.27031	0.026	1.59904	0.026		
1/12/15	11	*	28	11	1.25	0.11	1.69	0.11		
1/13/15	12	*	5	12	1.28	0.11	1.67	0.11		
1/14/15	13	*	5	13	1.279	0.061	1.6835	0.061		
1/15/15	14	*	4	14	1.24	0.038	1.387	0.038		
1/16/15	15	*	3	15	1.307	0.038	1.678	0.038		
1/17/15	16	*	3	16	1.223	0.21	1.7	0.21		
1/18/15	17	*	28	17	1.225	0.16	1.5	0.16		
1/19/15	18	*	?	18	1.31	0.05	1.35	0.05		
1/20/15	19	*	*	19	1.305	0.074	1.665	0.074		
1/21/15	20	*	*	20	1.289	0.01947	1.701	0.01947		

Figure 6. Data entry worksheet.

as many volunteer coordinators or analysts do not want to provide corrective action guidance to other laboratories, the standardized terminology is helpful and provides consistency among reports.

5.5 Template Spreadsheets

Template spreadsheets are provided to each analyst and are readily available on the NIST website (see references). Figures 6 through 10 illustrate the data entry format, a summary results worksheet, a data analysis worksheet where values are presented and evaluated, and examples of the standardized graphs that are generated with the template spreadsheets. The spreadsheet is arranged so that many of the calculations are performed automatically as the submitted data are entered on the data entry worksheet.

The data analysis worksheet for each artifact is where the core of the individual analysis must be performed. The initial data are evaluated to determine which points are outside of two standard deviations of the reported laboratory results. This assessment is used during the process of selecting the best reference value. The previous hierarchy for selecting the best reference value can be chosen from a drop down list once any outside reference values have been entered. All submitted values are included on all tables and graphs, even if the point is not an official laboratory value, or if was not used to select the reference value.

After the reference value and its associated uncertainty are established, numeric values for the normalized error and normalized precision tests are used to determine the Pass/Fail status for each point. The Pass/Fail statistics associated with the E_n and P_n tests are color coded. Failed results ($E_n > 1.0$ and $P_n > 1$) are highlighted in a red background while marginal values, between 0.7 and 1.0, are highlighted in yellow as a warning limit. The warning limits were originally set at 0.5, although numerous recommendations were submitted to raise this value since actions are not required until there is a failure [10].

The results calculated on the data analysis worksheet automatically populate the PT summary worksheet (Fig. 7) where the users of the PT report can see the results for multiple artifacts at a glance. Throughout the spreadsheets, only cells with a medium yellow background are not locked. This simplifies data entry and ensures the integrity of the automatic calculations.

There are now several versions of the spreadsheet. The one chosen depends upon the number of participants and artifacts contained in a PT. The core of the most basic version is updated with any changes and all other versions are developed after validation is completed. The versions in use at this time automatically create the titles and other header data for graphs and presentation of the results.

Many of the PTs coordinated through this program use identical artifacts (e.g., a 5 kg to 1 mg weight kit, a 5 gal test measure, a single stopwatch). Therefore, it is possible to set up standard spreadsheets for each type of proficiency test. Tasks such as adding pages and creating associated graphs have already been completed, so the analyst only needs to enter the data and make decisions about what values should be included in the 'adjusted mean' and 'adjusted standard deviation' for each artifact in the set and select the best reference value. Standardized spreadsheets help ensure consistent and successful PT analysis.

The use of template tools has dramatically improved the quality of the PT data analysis and reporting process. Volunteers know what information to expect, and where the information will be located in each report. Additionally, the standardized PT reports provide assessors with a quick summary of the PT results for each participant.

Region:	Your Region									
Reporting Year:	2011									
Round Robin:	Two 1 kg standards									
PT Identification:	XXX-11-MII-M-01									
Eval pts	80									
Failing Eval pts	11									
PT Summary										
Date of Test	**Participant ID * - Official Value**	**1 kg**		**1 kg ***		**500 g**		**200 g**		
	Red has failures	En	Pn	En	Pn	En	Pn	En	Pn	
1/1/2011	1 *	0.08	0.72	0.18	0.72	---	---	---	---	
1/2/2011	2 *	0.08	0.72	0.39	0.72	---	---	---	---	
1/3/2011	3 *	0.05	0.38	0.58	0.38	---	---	---	---	
1/4/2011	4 *	0.03	0.40	0.28	0.40	---	---	---	---	
1/5/2011	5 *	0.03	0.40	0.28	0.40	---	---	---	---	
1/6/2011	6 *	0.05	1.26	0.10	1.26	---	---	---	---	
1/7/2011	7 *	0.05	1.20	0.11	1.20	---	---	---	---	
1/8/2011	8 *	0.25	0.31	0.41	0.31	---	---	---	---	
1/9/2011	9 *	0.05	0.66	0.33	0.66	---	---	---	---	
1/10/2011	10 *	0.35	0.16	0.70	0.16	---	---	---	---	
1/11/2011	11 *	0.27	0.66	0.54	0.66	---	---	---	---	
1/12/2011	12 *	0.00	0.66	0.37	0.66	---	---	---	---	
1/13/2011	13 *	0.02	0.37	0.80	0.37	---	---	---	---	
1/14/2011	14 *	1.02	0.23	4.82	0.23	---	---	---	---	
1/15/2011	15 *	0.68	0.23	1.00	0.23	---	---	---	---	
1/16/2011	16 *	0.27	1.26	0.34	1.26	---	---	---	---	
1/17/2011	17 *	0.34	0.96	0.78	0.96	---	---	---	---	
1/18/2011	18 *	0.58	0.30	4.66	0.30	---	---	---	---	
1/19/2011	19 *	0.33	0.44	0.46	0.44	---	---	---	---	
1/20/2011	20 *	0.40	0.12	1.93	0.12	---	---	---	---	
---	---	---	---	---	---	---	---	---	---	

Figure 7. PT summary worksheet.

Region:	Your Region	SOP (From Planning Checklist):	4	Units:	mg					
Reporting Year:	2011	Accepted Reference Value:	1.280181875	Ref. source:	Adjusted Mean					
Round Robin:	Two 1 kg standards	U of Ref. Value(k...*u...):	0.010068446	Tolerance:	0.5	mg				
Artifact ID:	1kg			Tol. Source:	OIML R111					
PT Identification:	XXX-II-MII-M-01	P(n) of U...:	0.05							
		Description:	1kg OIML E1 mass standard							
DATA ANALYSIS										
Initial Statistics:		Do not change the Initial Statistics calculations! Initial Statistics are used to assess the group distribution and evaluate initial choices for reference values.								
Mean:	1.2755455	E(n): Normalized Error; evaluates bias vs combined uncertainties.								
Median:	1.2795	P(n): Normalized Precision; evaluates uncertainty vs tolerance.								
Max:	1.31	Typically, Adjusted Statistics include only those values that are designated the Official Laboratory Value, pass E(n), pass P(n), and are within 2 standard deviations of the Initial Statistics Mean value.								
Min:	1.223									
Range:	0.087	Adjusted Statistics are used for assessing and selecting a PT/LC Reference value.								
Standard Deviation:	0.02453033	16 values were included in the Adjusted Statistics:								
SD of Initial Mean:	0.005485149	4 values were NOT included in the Adjusted Statistics:								
Adjusted Statistics:		Official values not included in the Adjusted Statistics:								
Mean:	1.280181875	6 , 7 , 16 , 17 ,								
Median:	1.2795									
Max:	1.31									
Min:	1.24									
Range:	0.07									
Standard Deviation:	0.01889502									
SD of Adjusted Mean:	0.004723755									
k [adjusted] [15 df]=	2.131449536									
DATA		Values highlighted in Red (dark shading) FAIL. Values highlighted in Yellow (light shading) are values of concern.								
Date of Test	**#**	**Participant ID * - Official Value**	**SOP Used**	**Reported Value**	**Uncertainty**	**Bias from Reference value**	**E(n) Pass < 1**	**P(n) Pass < 1**	**Z value (bias/U)**	
1/1/2011	1	1 *	4	1.28	0.12	0.0098181	0.08	0.72	0.08	
1/2/2011	2	2 *	5	1.27	0.12	-0.0101819	0.08	0.72	-0.08	
1/3/2011	3	3 *	5	1.2836	0.063	0.0034181	0.05	0.38	0.05	
1/4/2011	4	4 *	5	1.278	0.066	-0.0021819	0.03	0.40	-0.03	
1/5/2011	5	5 *	5	1.278	0.066	-0.0021819	0.03	0.40	-0.03	
1/6/2011	6	6 *	4	1.29	0.21	0.0098181	0.05	1.26	0.05	
1/7/2011	7	7 *	4	1.29	0.2	0.0098181	0.05	1.20	0.05	
1/8/2011	8	8 *	4	1.267	0.051	-0.0131819	0.25	0.31	-0.26	
1/9/2011	9	9 *	4	1.286	0.11	0.0058181	0.05	0.66	0.05	
1/10/2011	10	10 *	4	1.27031	0.026	-0.0090719	0.35	0.16	-0.35	
1/11/2011	11	11 *	28	1.25	0.11	-0.0301819	0.27	0.66	-0.27	
1/12/2011	12	12 *	5	1.28	0.11	-0.0001819	0.00	0.66	0.00	
1/13/2011	13	13 *	5	1.279	0.061	-0.0011819	0.02	0.37	-0.02	
1/14/2011	14	14 *	4	1.24	0.038	-0.0401819	0.68	0.23	-1.06	
1/15/2011	15	15 *	3	1.307	0.038	0.0268181	0.68	0.23	0.71	

Figure 8. Sample data table with PT analysis and results.

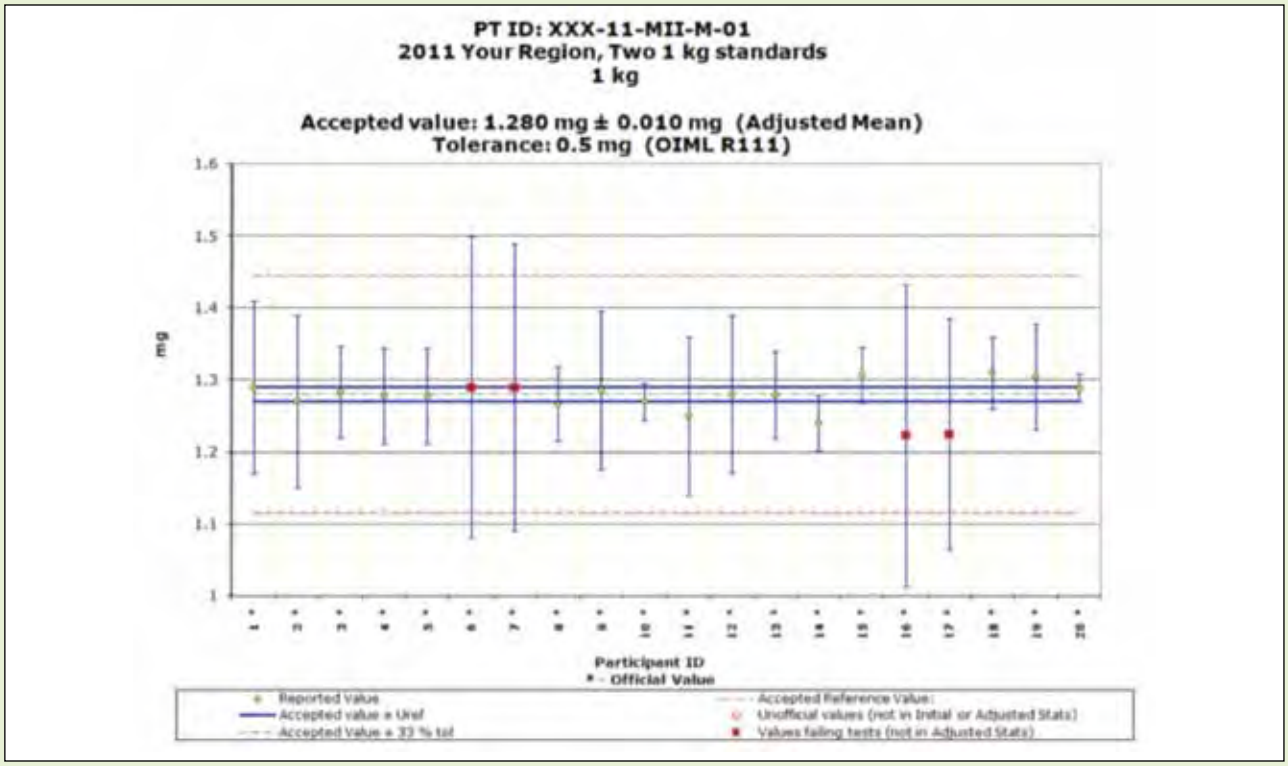

Figure 9. Sample data graph with reported values, uncertainties, and limits.

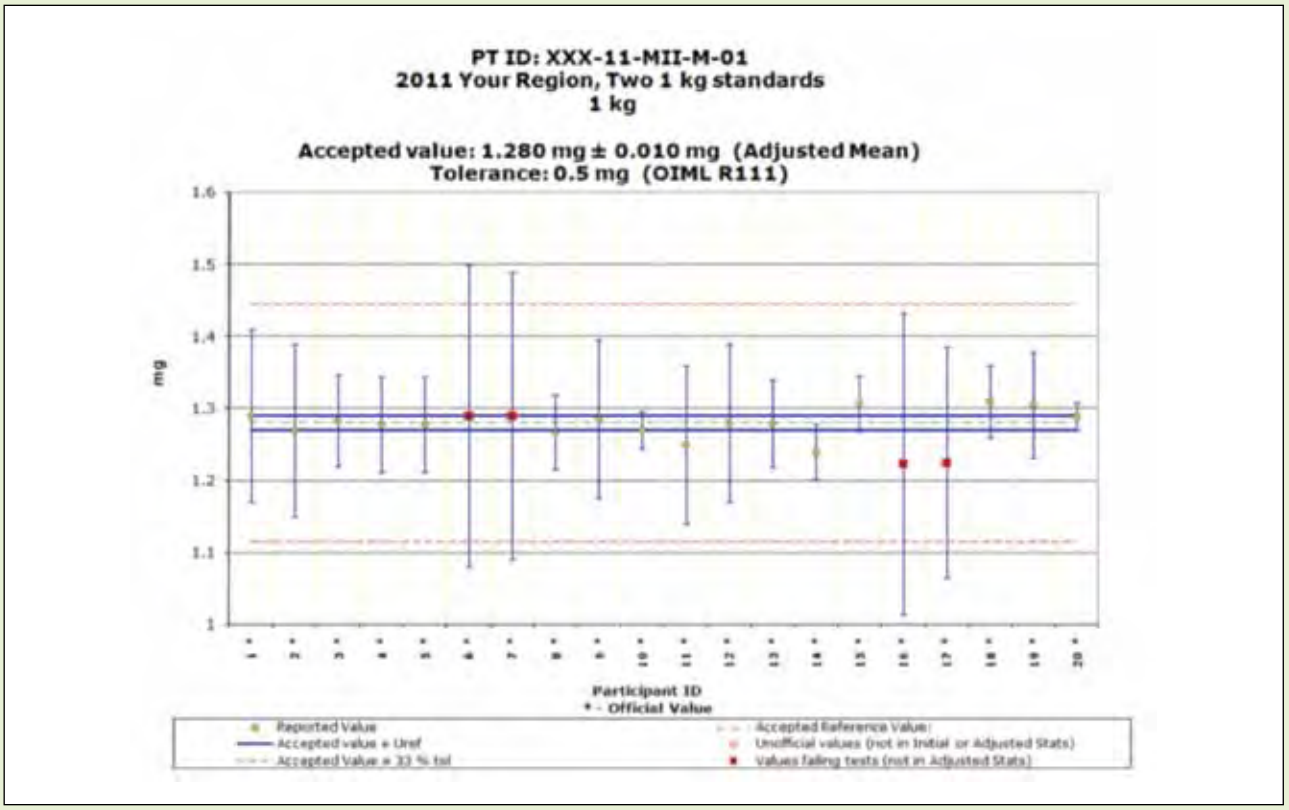

Figure 10. Sample E_n and P_n graph.

Improve Compliance to PT Program Quality Policies
• Work to ensure that the policies of each regional group support, rather than conflict with, the OWM PT policies. For instance, each laboratory must meet minimum training and attendance requirements for regular participation (on going) [10]. • Create a PT plan within each RMAP region to ensure the calibration scope is covered within a four year period and to ensure that all essential PTs are conducted and that unnecessary PTs are not conducted (implemented) [10]. • Complete all reports 30 days prior to the meetings so data can be evaluated by participants for accuracy in data entry and analysis can be reviewed (and approved) by OWM, permitting completion of any follow-up testing or corrective actions by laboratories before the final report is issued (on going). • Conduct web meetings 30 days prior to the RMAP training events to review data with coordinators and analysts and ensure that reports are complete (to be started in 2012). • Update the PT policy and Quality Manual so they are consistent with *ISO/IEC 17043* and *ISO 13528* and ILAC policies (to be started in 2012).
Use Technology to Improve PT Coordination
• NIST should develop an online inventory of possible artifacts and shipping containers that are available for use, including their current RMAP location and anticipated availability (future) [10]. • Provide an online system for PT management that includes scheduling functions, automated uploading of data sheets and reports, and automated email generation to participants when data and reports are not submitted by deadlines (future) [10].
Develop PT Analyst Expertise
• Each state should coordinate a PT once every three years to develop expertise in the coordination and analysis, as well as knowledge, of the specific measurement process (future) [10]. • Conduct PT planning, analysis, and follow-up training via regularly scheduled webinars (implemented). • Hold another PT Workshop to further enhance analyst skills and invite new participants to ensure ongoing succession planning (planned for 2012).
Develop PT Tools to Improve Uniformity and Streamline the Process
• Implement a standardized numbering system that efficiently communicates the report, the RMAP region participating, the level of precision or measurement echelon, and the measurement area (implemented). • Create standardized planning checklists for common artifacts such as standard weight sets and volumetric measures (implemented) [10]. • Create standardized template data sheets that can be downloaded and tailored for each region (future) [10]. • Continuously improve the data analysis spreadsheets: o Provide additional automation in labeling graphs and in selecting reference values(implemented); o Evaluate the current limits for precision test in volume calibrations (on going); o Evaluate the use of "marginal" flags on E_n values and raise the 0.5 limit (raised to 0.7) [10]. Implement internationally recognized methods for data analysis and reference value decision making referencing *ISO 13528* [7]; and o Consider not using absolute values of E_n values to more clearly identify when reported values are greater or less than the reference value (would require separate graphs for E_n and P_n).

Table 2. PT program continual improvement opportunities.

6. Continual Improvement

The PT quality management system requires regular internal audits and management reviews. It is a challenge to keep up with a separate assessment. Positive feedback and suggestions for improvement are obtained from participants at the annual RMAP meetings, through individual inquiries, from accreditation bodies, and from requests for direct feedback. Input from the feedback results is used to implement corrective and preventive actions and contributes to responding to Baldrige category 3. These internal processes (category 6) and the programmatic measures (category 4) discussed earlier also contribute to continual improvement. The input received through the many sources is considered during the strategic planning sessions and the SWOT analyses.

A PT Workshop was held in November 2007. The participants included RMAP representatives, NIST statisticians, and NVLAP assessors. Many suggestions from this workshop have been implemented. Due to its effectiveness, another workshop is being planned for 2012. A compilation of improvement opportunities is shown in Table 2.

The purpose of the PT workshops is to improve the PT program, by establishing goals and covering a variety of topics.

PT/ILC workshop goals:

- Develop greater skill and number of regionally-based PT coordinators, analysts, and technical reviewers for OWM, NVLAP, and the regional groups;
- Ensure higher quality and consistent proficiency test analysis and reporting;
- Improve processes and templates for planning, reporting, and analysis;
- Obtain customer feedback on the PT quality system to enable improvements;
- Identify improved measures of success/failure for PTs; and
- Update the PT program documents.

PT/ILC workshop topics:

- International PT standards and perspectives;
- Accreditation body and OWM PT policies;
- Planning and coordination requirements and process;
- Statistical design and analysis techniques (including use of standard and Youden template analysis spreadsheets with hands-on case studies);
- Uncertainty considerations and calibration measurement capabilities;
- Selection of reference values;
- Interpretation of PT results;
- Methods for correlating PT results and the laboratory internal measurement assurance data (integrated measurement assurance); and
- Changes due to adoption of *ISO/IEC 17043* [6] and *ISO 13528* [7].

Recommendations from all sources are compiled, evaluated, and implemented as appropriate and as staff and time is available. Coordinating and analyzing proficiency tests must be streamlined as much as possible to enable the volunteers to provide quality data and to ensure that consistent and substantive reports are available for participants to meet *ISO/IEC 17025:2005* accreditation and/or recognition requirements. Participant feedback and suggestions, such as those provided by Van Hyder [10] and others, are critical inputs and continue to improve the PT program. Updates to the quality management system and data analysis will be started in 2012 to ensure compliance of the PT program with *ISO/IEC 17043*. These updates will ensure continued acceptance of our PT reports by accreditation bodies.

7. Conclusion

The OWM PT program has produced a number of unique benefits. Participating laboratories receive the feedback necessary to assess and improve their measurement results. Successful PT results also support laboratory recognition and accreditation requirements. The PT opportunities are regularly available, of high quality, and the costs are shared amongst participants. Participating staff members gain expertise and insight in the use of spreadsheets, in statistical analysis techniques, in identifying problematic data, and in working with other laboratories to troubleshoot measurement errors. As a result, the depth of metrology expertise continues to expand beyond what it would have been if NIST independently coordinated and analyzed the PTs necessary to support the U.S. weights and measures laboratories. In addition to the benefits received by participants, the NIST staff gains insight into failed PT results which helps us to provide future training that is focused on measurement problems.

8. References

[1] "State Weights and Measures Laboratories, Program Handbook," *NIST Handbook 143*, 5th Edition, 2007.

[2] "Proficiency Test Policy and Plan for State Weights & Measures Laboratories," *NISTIR 7082*, January 2004. (Controlled versions of PT management system are posted at: http://www.nist.gov/pml/wmd/labmetrology/roundrobins.cfm)

[3] "Office of Weights and Measures Quality Manual for Proficiency Testing and Interlaboratory Comparisons," *NISTIR 7214*, March 2005. (Controlled versions of PT management system are posted at: http://www.nist.gov/pml/wmd/labmetrology/roundrobins.cfm)

[4] International Laboratory Accreditation Cooperation, "Guidelines for the Requirements for the Competence of Providers of Proficiency Testing Schemes," *ILAC-G13:08/2007*, 2007.

[5] International Organization for Standardization, "Proficiency Testing by Interlaboratory Comparisons – Part 1: Development and Operation of Proficiency Testing Schemes," *ISO/IEC Guide 43-1*, 1997. (superseded by *ISO/IEC 17043:2010*)

[6] International Organization for Standardization, "Conformity Assessment — General Requirements for Proficiency Testing," *ISO/IEC 17043*, 2010.

[7] International Organization for Standardization, "Statistical Methods for Use in Proficiency Testing by Interlaboratory Comparisons," *ISO 13528*, 2005.

[8] International Organization for Standardization, "General Requirements for the Competence of Testing and Calibration Laboratories, International Organization for Standardization," *ISO/IEC 17025*, 2005.

[9] *State Laboratory Program, Workload Surveys (1996 to 2011)*, published collaboratively from 1996 to 2003 by the National Conference on Weights and Measures and from 2005 to 2011 as a product of the NCSL International, Legal Metrology Committee. (http://www.nist.gov/pml/wmd/labmetrology/lab-resources.cfm)

[10] International Organization of Legal Metrology, "International Recommendation, Weights of Classes E1, E2, F1, M1, M1-2, M2-3 and M3 – Part 1: Metrological and Technical Requirements," *OIML R 111-1*, 2004. (http://www.oiml.org/publications/R/R111-1-e04.pdf)

[11] S. V. Hyder, "Proficiency Testing: State Weights & Measures Laboratory Experience Over the Last 10 Years: The Trials and Tribulations of a Nag," *Proceedings of the NCSL International Workshop and Symposium*, St. Paul, Minnesota, 2007.