The Interrelationship of Proficiency Testing, Interlaboratory Statistics and Laboratory Quality Assurance Programs

Thomas V. Coyner
Analytical Products Group, Inc.

The fundamental concepts of Quality Control and Quality Assurance are understood by most laboratory analysts; however, many times they struggle with the actual implementation of a Quality Control system to support the technical defensibility of their data. How do matrix spikes, matrix duplicates, laboratory control standards, and calibration fit together to give you defensible results? Most approved environmental methods have an extensive Quality Control section, but the pieces still do not seem to fit.

It is important to realize that environmental analysis is essentially a two-step series of processes. The basic process considers the sample as the input and sample data as the outline. In this sense, the matrix spike and matrix duplicate are quality control on this sample processing system. The second process is the test method itself, and the laboratory quality control standard should monitor this process.

It is the second process, the test method, and it's Quality Control which are the focus of this discussion. Analytical Products Group, Inc. has developed a simple graphic to demonstrate the interrelation of calibration, quality control standards, and proficiency testing (PT).

The relative size and location of each of the elements in the triangle correspond to the importance and function of each element. The most important element in an analytical system is the calibration. Without a sound calibration, there is nothing you can do to improve the quality of the data. Therefore, great care should be given to verifying the accuracy and linearity of the calibration. It must be noted that the standard referred to as the Continuing Calibration Verification (CCV) standard is not a verification sample. It simply records the "drift" in the method calibration over time. Calibration is also the area where you can see the greatest return on investment. A simple change from undocumented standards to NIST traceable standards will go a long way to ensure data defensibility. Unfortunately, virtually all of the organic standards commercially available lack traceability, so it is critically important to verify organic calibrations against a second source, or better yet, a quality control standard with documented performance.

Many EPA methods include a requirement for a laboratory control standard (LCS). The requirement usually is that the standard contains every analyte in the method, be prepared within the calibration range, be run on a set frequency, and be processed as a sample. The purposes of the LCS are to verify the calibration and to establish working quality control limits for the laboratory. The most important of these requirements is that the LCS be run as a sample, and therefore, go through the entire sample preparation process. Conversely, the requirement to contain all of the analytes is less important but forces the laboratory to prepare the LCS from a calibration standard. Unfortunately, an LCS prepared from a calibration standard only has a documented true value. Opposed to this approach, a Quality Control Standard certified in an interlaboratory PT program contains more useful information but typically lacks all of the analytes in the method.

An interlaboratory certified Quality Control Standard has a certified true value much like a calibration material. However, because it has been certified in an interlaboratory program where it was run as a sample, it has established typical laboratory performance ranges established in the study. This gives a laboratory a verification of its internal quality control limits and avoids failures on PT samples. Since most inorganic Quality Control Standards are also NIST traceable, this would make every analytical run a NIST traceable analysis which provides better verification of the calibration and provides greater value to the data user.

Similarly, an interlaboratory certified Quality Control Standard is an ideal tool for method validation or analyst training verification. In both cases, the standard has sufficient documentation to support the data needs of these analyses. Finally, this type of Quality Control Standard is also an excellent tool to verify the effectiveness of a corrective action because successful performance within the defined laboratory performance range is a demonstration of performance comparable to other laboratories running the same sample.

Proficiency Testing for Quality Improvement:

Proficiency testing programs, sometimes referred to as round-robins or interlaboratory programs, have been used for many years and for various purposes. ASTM and AOAC use these programs to verify and document the performance of new methods. For the past twenty years, USEPA has used proficiency testing to monitor the quality of environmental laboratory data through the WP, WS, and DMRQA PT programs. During the 1990s, the International Standard Organization (ISO) recognized the need to develop a uniform protocol for proficiency testing of all analytical laboratories and established a workgroup for this purpose. The final Harmonized Protocol for the Proficiency Testing of Analytical Laboratories is generally accepted as the international guide for the operation of these programs. The Protocol recognizes the utility of proficiency testing programs as a quality improvement tool as well as a means of establishing and monitoring laboratory performance.

However, both USEPA and, more recently, the National Environmental Laboratory Accreditation Program (NELAP) have chosen to focus their PT programs on laboratory control rather than monitoring or quality improvement. They both do this by requiring that laboratories be evaluated based upon acceptance limits calculated from the Assigned Value of the samples and previous PT studies rather than using the data developed in the actual study under evaluation. By using this evaluation scheme, both groups lose the important information gathered by the study, fail to understand changes in laboratory performance, lose valuable method performance data, and develop unrealistic performance expectations for their methods and the laboratories that run them.

Proficiency testing programs are fundamentally a peer review process. They allow the comparison of laboratories on the same samples, run at the same time. However, the idea of a peer review process requires that all laboratories be treated fairly and equally. Therefore, the samples must be properly prepared and evaluated for homogeneity prior to the study and for stability after the study. Similarly, the samples must be designed to “challenge” the method; that is, they must not be too easy or be so difficult that they cannot be successfully run by the appropriate methods. Given these precautions, a PT program can provide valuable information for both the laboratory and the program operator.

What can be gained from a PT program is dependent upon the final data analysis. If the goal of the program is to evaluate the performance of the laboratories against the True Value of the samples, then percent recovery or En value, which measure deviation from the reference value, can be used. If the purpose is to establish typical laboratory performance then z Scores or Deviation from the Mean are appropriate. If data is analyzed by both method and laboratory, then method comparison data can be developed. However, inappropriate data analysis can easily lead to improper conclusions and misrepresentation of laboratory performance.

If we look again at the USEPA procedure for proficiency testing which evaluates laboratories based upon regression equations to develop an “expected” Mean and “expected” Standard Deviation, we can see that all of the actual data in the study is lost. Since 1999, when the USEPA privatized its PT programs, the limits have been calculated from data developed in EPA operated programs throughout the 1990s. During that period, the Agency used a complex and robust statistical procedure to estimate laboratory Means and Standard Deviations; however, the procedure tended to underestimate the Standard Deviation, and a significant number of laboratories failed even though the Agency was using limits set at Mean±3SD for WP and Mean±2SD for WS studies. This gave, and continues to give, the impression that general laboratory quality is poor. In fact, many of the regression equations used by EPA to set limits had correlation coefficients of less than 0.7 and were questionable at best.

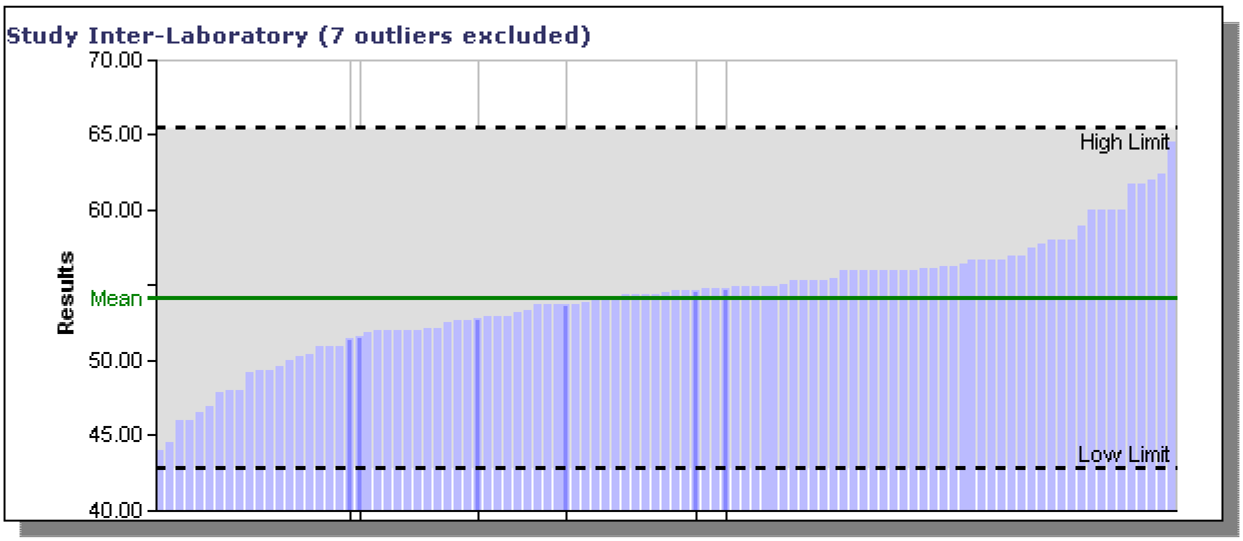
Comparison of Actual vs. EPA Estimated Standard Deviations from DMRQA Study 20

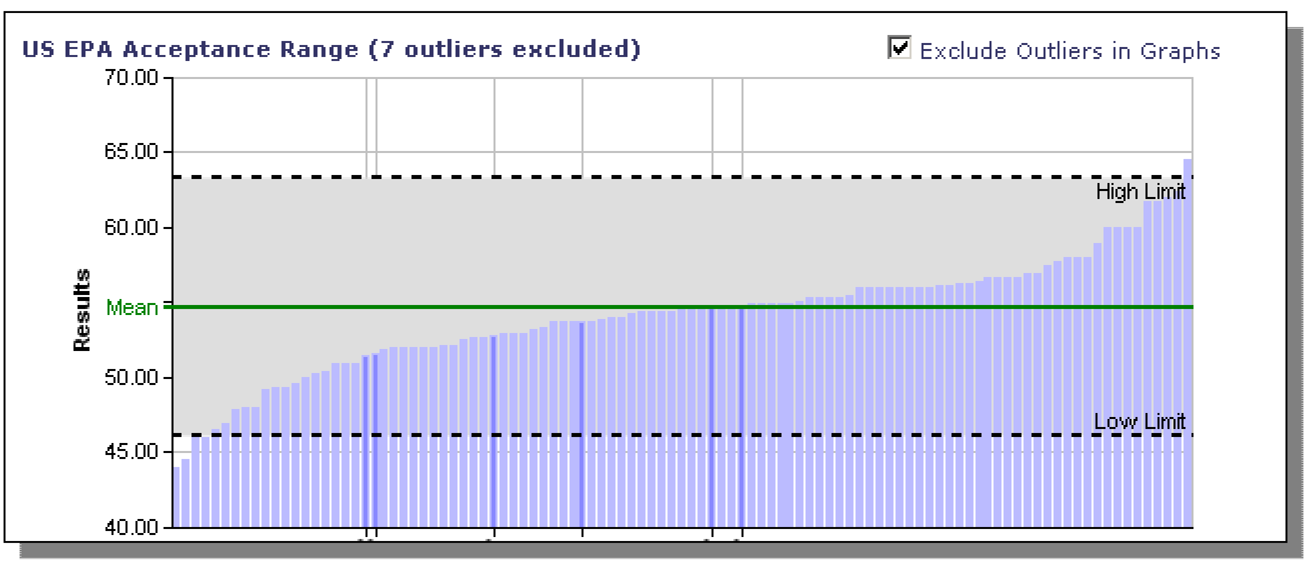
(Data set includes more than 2,000 laboratories)

	Standard Deviation			
Parameter	***EPA Estimate***	***Actual Study***	***% Difference***	***Lab Pass Rate***
Biochemical Oxygen Demand	17.6	28.4	61	91.99
Carbonaceous BOD	16.8	25.1	49	94.44
Chemical Oxygen Demand	10.5	38.9	270	87.63
Orthophosphate as P	0.17	0.37	118	85.32
Total Kjeldahl Nitrogen	1.37	3.88	183	86.09
Aluminum	40.2	58.8	46	90.73
Arsenic	6.17	10.6	72	89.08
Chromium	5.48	7.38	35	89.38
Lead	15.3	20.8	36	90.21
Mercury	1.12	2.29	104	87.16
Vanadium	39.4	52.1	32	92.05
Zinc	16.8	20.3	21	89.29
Total Cyanide	0.057	0.098	72	85.99
Total Residual Chlorine	0.14	0.29	107	84.04
Oil & Grease	2.27	4.02	77	83.45

Unfortunately, NELAP uses a similar regression-based evaluation process. The NELAP regression equations have recently been revised to be based upon data developed since the 1999 privatization and have more reasonable limits for correlation coefficients. These new limits take effect on June 1, 2005, and it is expected that they will improve the laboratory pass percentages for most WS and WP parameters. However, even with the new limits, the NELAP method of laboratory evaluation against estimated parameter statistics results in the loss of the real data that is available in the study.

Analytical Products Group, Inc. (APG) as a NIST accredited PT provider operates twelve EPA/NELAP compliant WP studies and six EPA/NELAP compliant WS PT studies a year. However, as a benefit to our participating laboratories, APG reports both EPA/NELAP required limits and the actual interlaboratory statistics based upon the methods of the Harmonized Protocol. This allows a laboratory to better understand of its relative performance compared to other participants in the program. It also allows a laboratory to evaluate the quality of the EPA or NELAP limits based upon actual study data. Interlaboratory z Scores show laboratories quantitatively how their performance compares to other laboratories running the same sample. Since method data is reported for all laboratories and results can be sorted by method they can better understand the limitations of the methods they are using. Similarly, by providing z Score plots of all data in the study for both the EPA/NELAP limits and the actual interlaboratory limits, laboratories can understand when they fail an EPA limit that is inappropriate and avoid unnecessary corrective actions.





WP Trace Metals Copper

Lot Number: 41592-41593 **Authority:** Regressions, NELAP Code 1055
Method: EPA220.2 **Technology:**

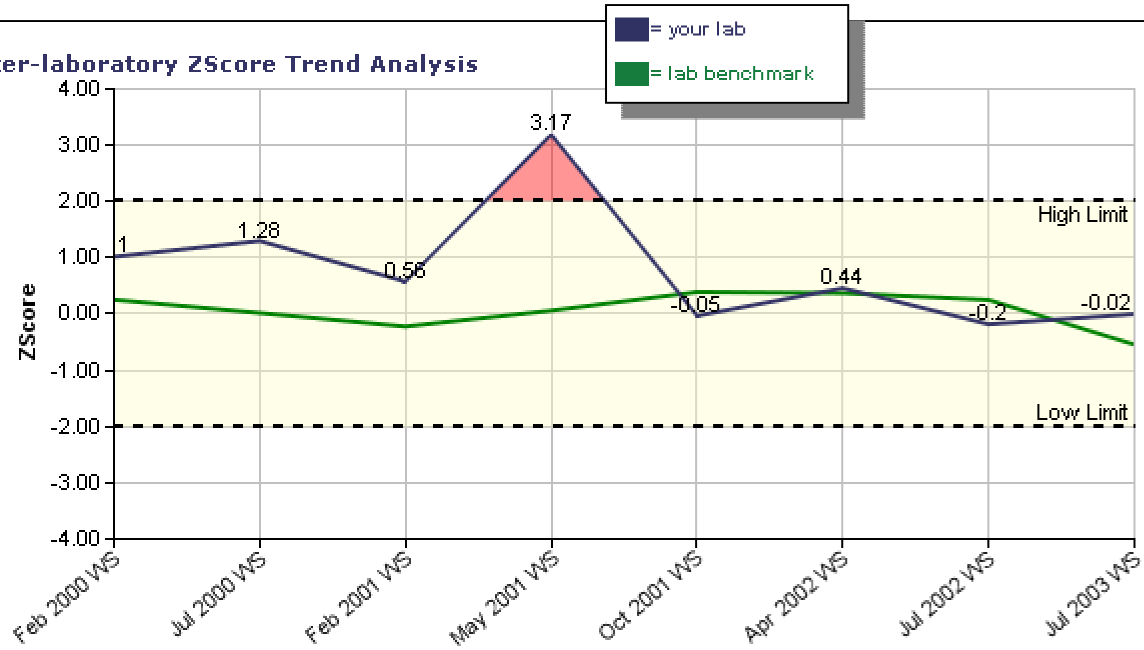
Evaluation	Reported Value	Units	Assigned Value	Percent Recovery	Certified Value
Acceptable	51.5	ug/L	54.6	94.15%	52.3 +/- 0.29

	Study Inter-laboratory	Regressions	Historic Performance
Your ZScore	-0.684	-1.12	
Your Rank		35 of 63	
Total Studies	1	1	23
Total Data Points	111	111	1227
Mean	54.1	54.7	100.37%
Standard Deviation	3.80	2.86	5.15
Acceptance Range (+/- 3 SD)	42.8 - 65.5	46.1 - 63.3	84.9 - 116%
Warning Range (+/- 2 SD)	46.5 - 61.7	49.0 - 60.4	90.1 - 111%
Average Percent Recovery	99.08%	100.18%	100.37%
Relative Standard Deviation	7.02%	5.23%	5.15%
Average Percent Acceptable	Review	Review	95.46%

In this case the EPA regressions under estimate the Standard Deviation by only 6% but it results in an additional 5% of participating laboratories failing the evaluation.

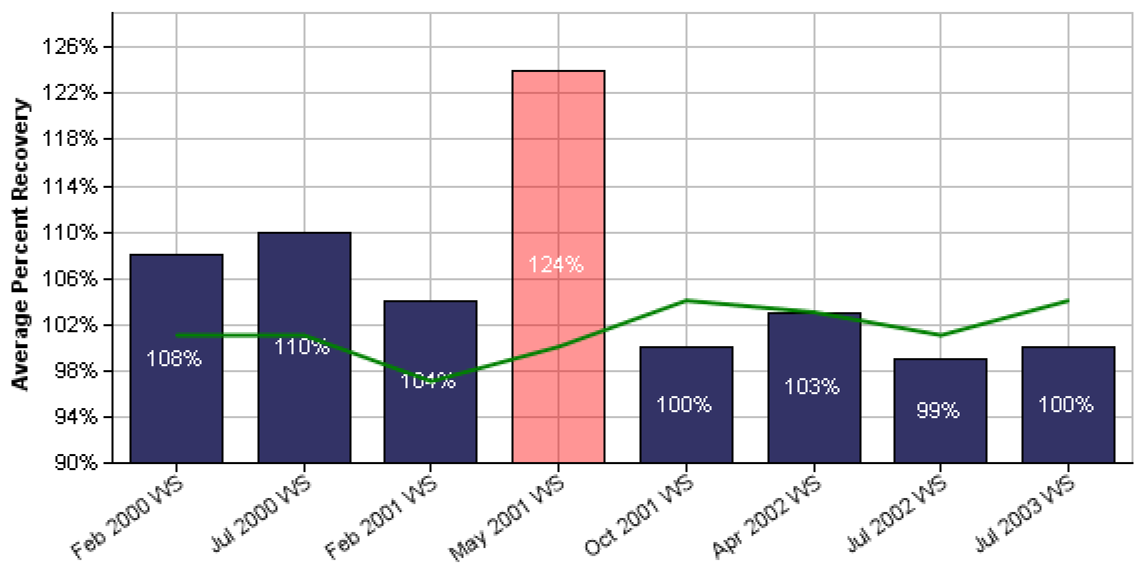
A fundamental failure of PT programs is that, much like a financial statement, they tell you a great deal about a laboratory at a single point in time but nothing about what happened before or after the study. However, by providing participating laboratories with plots of z Scores and percent recoveries over time against the average performance for the same measures, we have found that valuable information regarding trends in performance can be demonstrated. This is important if quality issues are to be found and corrected before they become accreditation failures. These same graphs can be used to demonstrate to an auditor that corrective actions on a particular parameter have been successful and that appropriate quality levels have been restored.

Inter-laboratory ZScore Trend Analysis



Percent Recovery Trend Analysis

Recoveries that differ more than 15% from the inter-laboratory average are red.



Inter-laboratory Benchmark Comparison

	Your Laboratory	Inter-laboratory Benchmark
Total Data Points	8	
Average Percent Acceptable	87.5%	90.39%
Average Percent Recovery	105.77%	101.3%
Absolute Median ZScore	0.5	0.1368

It is clear that the participating laboratories can gain a wealth of information by having the actual interlaboratory statistics from a PT study. It is also true that the regulatory agencies can also find this data valuable. The Agency has used interlaboratory data to demonstrate regulatory method performance for many years. EPA Method Study Number 1 was based on the classical methods of BOD and COD. However, many of these studies were based upon very limited data sets and were operated under tight controls that did not represent typical laboratory performance. For these reasons, the interlaboratory data quoted in many regulatory methods is inconsistent with performance in most environmental laboratories. Unfortunately, many of these limited data sets have been used to establish Quality Control Limits for the methods. The Office of Water recognized the value of the data available in the WS and WP studies, and within the past ten years, summarized the data as typical performance. Unfortunately, the Office used the results from the biased, robust statistical procedure used to set limits rather than re-analyze the original data sets which would have been more useful.

When the EPA privatized its PT program in 1999, it did not require the NIST accredited PT providers to collect method data in a useable format. However, NELAP, who accredits laboratories based upon matrix/method/analyte does collect this information. Since the focus of the NELAP PT programs has been laboratory control rather than environmental research, the data has never been analyzed by method, and to a large extent, the data has been ignored. Since NELAP data is within the public domain, it should be ready available. The NELAP PT Board currently has a project to establish a national database of NELAP PT data which could provide this information in the future. Proper analysis of this database could provide a wealth of information about the actual performance of the regulatory methods and shed new light on the actual Quality of the environmental data that is used to make critical decisions.