Designation: ~~D7372 – 17~~ D7372 – 21

An American National Standard

Standard Guide for Analysis and Interpretation of Proficiency Test Program Results[1]

This standard is issued under the fixed designation D7372; the number immediately following the designation indicates the year of original adoption or, in the case of revision, the year of last revision. A number in parentheses indicates the year of last reapproval. A superscript epsilon (ε) indicates an editorial change since the last revision or reapproval.

1. Scope*

1.1 This guide covers the evaluation and interpretation of proficiency test program (PTP) results. For proficiency test program participants, this guide describes procedures for assessing participants' results relative to the collective PT program results and potentially improving the laboratory's testing performance based on the assessment of findings and insights. For the committees responsible for the test methods included in PT programs, this guide describes procedures for assessing industry's ability to perform test methods and for potentially identifying opportunities for improvements.

1.2 *This standard does not purport to address all of the safety concerns, if any, associated with its use. It is the responsibility of the user of this standard to establish appropriate safety, health, and environmental practices and determine the applicability of regulatory limitations prior to use.*

1.3 *This international standard was developed in accordance with internationally recognized principles on standardization established in the Decision on Principles for the Development of International Standards, Guides and Recommendations issued by the World Trade Organization Technical Barriers to Trade (TBT) Committee.*

2. Referenced Documents

2.1 *ASTM Standards:*[2]

D4175 Terminology Relating to Petroleum Products, Liquid Fuels, and Lubricants
D6259 Practice for Determination of a Pooled Limit of Quantitation for a Test Method
D6299 Practice for Applying Statistical Quality Assurance and Control Charting Techniques to Evaluate Analytical Measurement System Performance
D6617 Practice for Laboratory Bias Detection Using Single Test Result from Standard Material
D6792 Practice for Quality Management Systems in Petroleum Products, Liquid Fuels, and Lubricants Testing Laboratories
D7915 Practice for Application of Generalized Extreme Studentized Deviate (GESD) Technique to Simultaneously Identify Multiple Outliers in a Data Set
E177 Practice for Use of the Terms Precision and Bias in ASTM Test Methods
E456 Terminology Relating to Quality and Statistics
E2586 Practice for Calculating and Using Basic Statistics
E2655 Guide for Reporting Uncertainty of Test Results and Use of the Term Measurement Uncertainty in ASTM Test Methods

2.2 ASTM standards used only in Appendix X3 are also listed in X3.1.

[1] This guide is under the jurisdiction of ASTM Committee D02 on Petroleum Products, Liquid Fuels, and Lubricants and is the direct responsibility of Subcommittee D02.94 on Coordinating Subcommittee on Quality Assurance and Statistics.

Current edition approved ~~Oct. 1, 2017~~May 1, 2021. Published ~~October 2017~~September 2021. Originally approved in 2007. Last previous edition approved in ~~2012~~2017 as ~~D7372 – 12.~~D7372 – 17. DOI: ~~10.1520/D7372-17.~~10.1520/D7372-21.

[2] For referenced ASTM standards, visit the ASTM website, www.astm.org, or contact ASTM Customer Service at service@astm.org. For *Annual Book of ASTM Standards* volume information, refer to the standard's Document Summary page on the ASTM website.

***A Summary of Changes section appears at the end of this standard**

3. Terminology

3.1 *Definitions:*

3.1.1 More extensive lists of terms related to quality, statistics, and related terms are found in Terminology D4175.

3.1.1.1 In the event of disagreement between the quoted text and the latest in the referenced standard, the latter supersedes the text in this standard.

3.1.2 *accuracy, n*—closeness of agreement between an observed value and an accepted reference value. **E177, E456**

3.1.2.1 *Discussion*—
The term accuracy, when applied to a set of test results, involves a combination of a random component and of a common systematic error or bias component. **E177**

3.1.3 *analytical measurement system, n*—a collection of one or more components or subsystems, such as sample handling and preparation, test equipment, instrumentation, display devices, data handlers, printouts or output transmitters, that are used to determine a quantitative value of a specific property for an unknown sample in accordance with a standard test method.

3.1.4 *Anderson-Darling Resolution Sensitive Statistic, ADrs, n*—a goodness-of-fit statistical tool used to objectively test for normality of proficiency testing data.

3.1.4.1 *Discussion*—
ADrs is a modified version of the Anderson-Darling Statistic (see D6299) and was developed specifically for use in assessing normality in proficiency test program data. The ADrs statistic assesses normality regardless of the adequacy of data measurement resolution relative to the overall variation in the dataset.

3.1.5 *assignable cause, n*—factor that contributes to variation and that is feasible to detect and identify. **E456**

3.1.6 *bias, n*—systematic error that contributes to the difference between a population mean of the measurements or test results and an accepted reference or true value. **D6299, E177**

3.1.6.1 *Discussion*—
Bias is the total systematic error as contrasted to random error. There may be one or more systematic error components contributing to the bias. A larger systematic difference from the accepted reference value is reflected by a larger bias value. **E177, ~~E456~~**

3.1.7 *common (chance, random) cause, n*—for quality assurance programs, one of generally numerous factors, individually of relatively small importance, that contributes to variation, and that is not feasible to detect and identify. **D6299**

3.1.8 *control limits, n*—limits on a control chart that are used as criteria for signaling the need for action or for judging whether a set of data does or does not indicate a state of statistical ~~control.~~control based on a prescribed degree of risk. **E456**

3.1.9 *in-statistical-control, adj*—process, analytical measurement system, or function that exhibits variations that can only be attributable to common cause. **D6299**

3.1.10 *median, x̃, n*—the 50th percentile in a population or sample.

3.1.10.1 *Discussion*—
The sample median is the [(n + 1)/2] order statistic if the sample size n is odd and is the average of the [n/2] and [n/2 + 1] order statistics if n is even. **E2586**

3.1.11 *median absolute deviation (MAD), n*—a robust measure of the variability of a data set.

3.1.11.1 *Discussion*—
MAD is a measure of statistical dispersion that is more resilient to outliers than the standard deviation. MAD is calculated as the median of the absolute deviations of individual results from the median.

3.1.12 *modified Z-score (M_i), n*—a standardized and dimensionless measure of the difference between an individual result in a data set and the sample median re-expressed in units of median absolute deviation of the dataset.

3.1.12.1 *Discussion*—

M_i is a robust statistic that is calculated as the difference between individual result minus the median divided by the MAD and then multiplied by the constant 0.6745 to approximate the standard deviation.

3.1.13 *out-of-statistical-control, adj*—a process, analytical measurement system, or function that exhibits variations in addition to those that can be attributable to common cause and the magnitude of these additional variations exceeds specified limits. **D6299**

3.1.14 *proficiency testing, n*—determination of a ~~laboratory's~~laboratory's testing capability by ~~participation in an interlaboratory proficiency test program~~evaluating its test results in interlaboratory exchange testing or crosscheck programs.

3.1.14.1 *Discussion*—

One example is the ASTM D02 committee's proficiency testing programs in a wide variety of petroleum products and lubricants, many of which may involve more than a hundred laboratories. ~~D6299~~**D6792**

3.1.15 *proficiency test program (PTP), n*—statistical quality assurance activities that enable laboratories to assess their performance in conducting test methods within their own laboratory when their data are compared against other laboratories that participate in the same program cycle using the same test method.

3.1.15.1 *Discussion*—

Proficiency test programs are also known as crosscheck programs and check schemes. The term Interlaboratory Crosscheck Program (ILCP) was previously used by ASTM for its PTP with Committee D02.

3.1.16 *site precision (R')*—the value which the absolute difference between two individual test results obtained under site precision conditions is expected to exceed about 5 % of the time (one case in 20 in the long run) in the normal and correct operation of the test method.

3.1.16.1 *Discussion*—

It is defined as 2.77 times $\sigma_{R'}$, the standard deviation of results obtained under site precision conditions. **D6299**

3.1.17 *site precision conditions, n*—conditions under which test results are obtained by one or more operators in a single site location practicing the same test method on a single measurement system which may comprise multiple instruments, using test specimens taken at random from the same sample of material, over an extended period of time spanning at least a 15 day interval.

3.1.17.1 *Discussion*—

Site precision conditions should include all sources of variation that are typically encountered during normal, long term operation of the measurement system. Thus, all operators who are involved in the routine use of the measurement system should contribute results to the site precision determination. In situations of high usage of a test method where multiple QC results are obtained within a 24 h period, then only results separated by at least 4 h to 8 h, depending on the absence of auto-correlation in the data, the nature of the test method/instrument, site requirements, or regulations, should be used in site precision calculations to reflect the longer term variation in the system. **D6299**

3.1.18 *test performance index—industry (TPI_{IND}), n*—an approximate measure of a PT program's testing capability for a specific test method, defined as the ratio of the ASTM reproducibility (R_{ASTM}) to *these data reproducibility* ($R_{these\ data}$).

3.1.18.1 *Discussion*—

TPI_{IND} is like the TPI used in D6792 except that the $R_{these\ data}$ is substituted for the site precision (R').

3.1.19 *these data, n*—term used by the ASTM International D02 PT program to identify statistical results calculated from the data submitted by program participants.

3.1.20 *uncertainty, n*—an indication of the magnitude of error associated with a value that takes into account both systematic errors and random errors associated with the measurement or test process. **E2655**

3.1.21 *Z-score, n*—standardized and dimensionless measure of the difference between an individual result in a data set and the arithmetic mean of the dataset, re-expressed in units of standard deviation of the dataset (by dividing the actual difference from the mean by the standard deviation for the data set). ~~D6299~~

3.1.21.1 *Discussion*—

The Z-score term described here is equivalent to Eq. A1.3 in Practice D6299.

3.1.22 *Z'-score, n*—~~measure similar to~~ standardized and dimensionless measure of the ~~Z-score~~difference ~~except that the PT program standard deviation is replaced with one that takes into account the site precision of the laboratory. Z' is a valid approach when the laboratory's~~ between an individual result in a data set and the arithmetic mean of the dataset, re-expressed in units of

the individual laboratory site precision standard deviation ~~is less than that for the PT program (that is,~~ of the dataset.~~these data standard deviation) or stated otherwise when the TPI > 1.~~

$$Z' = \frac{\left(X_i - \bar{X}\right)}{\sqrt{\left((s')^2 + \left(\frac{s^2_{these\ data}}{n}\right)\right)}}$$

where:

Z' = site precision adjusted *Z-Score*,
X_i = laboratory's result,
X^- = PT average value,
s' = site precision standard deviation estimate,
$s_{these\ data}$ = PT Program standard deviation estimate, and
n = number of non-outlier data.

3.1.22.1 *Discussion—*
This measure is like the Z-score except that the PT program standard deviation is replaced with one that takes into account the laboratory's site precision.

3.1.22.2 *Discussion—*
Z' is a valid approach when the laboratory's site precision standard deviation is less than that for the PT program (that is, *these data standard deviation*) or stated otherwise when the TPI > 1.

3.1.22.3 *Discussion—*
Z'-score described here is equivalent to Eq. 2 in Practice D6299 for pre-treated results, when the "standard error of ARV" is expressed as "standard deviation of ARV/ √n."

~~3.2 *Definitions of Terms Specific to This Standard:*~~

~~3.2.1 *common (chance, random) cause, n*—for quality assurance programs, one of generally numerous factors, individually of relatively small importance, that contributes to variation, and that is not feasible to detect or control. **D6299**~~

~~3.2.2 *site precision (R'), n*—value below which the absolute difference between two individual test results obtained under site precision conditions may be expected to occur with a probability of approximately 0.95 (95 %). It is calculated as 2.77 times the standard deviation of results obtained under site precision conditions. **D6299**~~

~~3.2.3 *site precision conditions, n*—conditions under which test results are obtained by one or more operators in a single site location practicing the same test method on a single measurement system which may comprise multiple instruments, using test specimens taken at random from the same sample of material, over an extended period of time spanning at least a 15 day interval. **D6299**~~

~~3.2.4 *these data, n*—term used by the ASTM International D02 PT program to identify statistical results calculated from the data submitted by program participants.~~

3.2 *Symbols:*

3.2.1 *ADrs*—Anderson-Darling Resolution Sensitive Statistic.

3.2.2 *I*—individual observation (as in *I*-chart).

3.2.3 ~~*PTP M_i or PT program*—~~ proficiency test program.Modified Z-score.

3.2.4 ~~*QC*—~~R_{ASTM} ~~quality control.~~published ASTM reproducibility.

3.2.5 *R'*—site precision.

3.2.6 $R_{\text{these data}}$—reproducibility determined in PT program.

3.2.7 ~~*r*~~$\tilde{x}_{\text{these data}}$ ~~repeatability determined in PT program.~~median.

~~3.3.7 R_{ASTM}—published ASTM reproducibility.~~

3.3 *Acronyms:*

3.3.1 *MAD*—median absolute deviation

3.3.2 *PTP or PTP program*—proficiency test program

3.3.3 *QC*—quality control

3.3.4 TPI_{IND}—test performance index (industry)

4. Summary of Guide

4.1 Petroleum product, liquid fuel, and lubricant samples are regularly analyzed by specified standard test methods as part of a proficiency test program. This guide provides a laboratory with the tools and procedures for evaluating their results from a PT program. Techniques are presented to screen, plot, and interpret test results in accordance with industry-accepted practices.

5. Significance and Use

5.1 This guide can be used to evaluate the performance of a laboratory or group of laboratories participating in a proficiency test (PT) program involving petroleum and petroleum products.

5.2 Data accrued, using the techniques included in this guide, provide the ability to monitor analytical measurement system precision and bias. These data are useful for updating standard test methods, as well as for indicating areas of potential measurement system improvement for action by the laboratory. This guide serves both the individual participating laboratory and the responsible standards development group as follows:

5.2.1 Tools and Approaches for Participating Laboratories.
- Administrative Reviews
- Flagged Data and Investigations
- Data Normality Checks
- QQ Plots
- Histograms
- Bias (Deviation from Mean)
- *Run-Sum*
- *Z-Scores, Z'-Scores Trends*
- *Precision Performance—TPI_{IND}, F-test*
- Comparison of PTP and Individual Laboratory Site Precision

5.2.2 Tools and Approaches for Responsible Standards Development Groups.
- TPI and precision trends
- Bias and precision comparisons via box & whisker plots
- Normality evaluations
- Relative standard deviations
- Uncontrolled variables

5.3 Reference is made in this guide to the ASTM International Proficiency Test Program on Petroleum Products, Liquid Fuels, and Lubricants, version PTP 2.0 implemented in 2016–2017. Program reports containing similarly displayed results and statistical treatments may be available in other PT programs. Appendix X2 summarizes the statistical tools referenced in this guide and Appendix X3 is a collection of examples covering ~~many of the approaches~~ QQ plots, histograms, and Run-Sum described in this guide.

6. Procedure—Evaluation and Interpretation by Participating Laboratories

6.1 *Administrative Reviews*—Laboratories should review the results published for each proficiency test program and for each test

method or parameter for which the laboratory submitted data. The following cover the evaluations that the laboratory should consider during their review of proficiency test results.

6.1.1 *Reported versus Submitted Data*—Verify that the values ascribed to the laboratory in the proficiency test (PT) report agree with the values recorded by the laboratory in its PT records. Report discrepancies to the respective PT program contacts. Investigate, as appropriate, to determine the root cause of the problem.

6.1.2 *Units for Results*—Verify that the units for the data reported by your laboratory are the same as that requested by the PT program. Report discrepancies to the respective PT program contacts. Investigate, as appropriate, to determine the root cause of the problem.

6.1.3 *Missing Data*—If data and corresponding results are not present when they are clearly expected, then investigate to determine the cause. In some cases it could be an error within the PT program data entry system, or it could be an omission on the part of the laboratory.

6.2 *Flagged Data and Investigations:*

6.2.1 *Rejected Data*—Perform an investigation for each instance where laboratory data are rejected by the PT program data treatment processes. Investigations should consider the entire analytical measurement system and not focus just on the instruments used by the test method. Attempt to determine the root cause and take corrective actions as needed. Document all such investigations and outcomes. Causes should be shared with the laboratory staff performing the testing. Guidelines on conducting these types of investigations are available in Appendix X1.

6.2.2 *Data Warnings/Alerts*—The ASTM International PT programs provide comments (that is, Warnings/Alerts 1 to 3 in results tables) that warn participants when their result is:

Warning/Alert
1—Test results outside ±3-sigma range for *these data*
2—Test results outside ±3-sigma range for ASTM reproducibility
3—*Z-score* outside range of –2 to 2

Investigations should be conducted when any of these warning situations occur. The priority for conducting investigations should be for Warning/Alert 1 > 2 > 3. Note that 1 indicates that the laboratory is out-of-statistical-control with respect to the data set (with the rejected data removed), which is a potentially serious situation with respect to the quality control performance of the corresponding standard test method. A similar argument could also be made for Warning/Alert 2. Finally, Warning/Alert 3 is a less severe situation, but should be investigated from a continuous improvement standpoint.

NOTE 1—If the user notices that the majority of ~~the~~ laboratories providing data have been cited with a Warning/Alert 2, then an investigation may not produce any meaningful corrective actions. This occurrence may be the result of the precision statement not accurately reflecting the variability of the test method and should be addressed by the subcommittee responsible for the method. Also, when the ~~Anderson-Darling statistic or the ADrs statistic is >1.3,~~ ADrs statistic signals not normal (6.3.1.1), then the Warning/Alert 2 may not be valid.

6.2.3 *Investigations*—It is important to recognize statistical outliers, but it is even more important to take action to identify assignable causes (factors that contribute to variation and that are feasible to detect and identify). Investigations should continue to identify root cause(s) and to implement corrective and preventative measures. A checklist for investigating the root cause of unsatisfactory analytical performance is provided as Appendix X1.

6.3 *Data Normality Checks:*

6.3.1 Typical statistical evaluations of proficiency test results assume data are from normal distributions, so it is appropriate to evaluate the data for normality. The Anderson-Darling ~~(AD)~~ statistic is a goodness-of-fit test to determine if the data are from a normal distribution. ~~The AD~~ This statistic is sensitive to inadequate data measurement resolution relative to the overall variation in the dataset. ~~Practice D6299 covers the calculation of the Anderson-Darling statistic.~~ The ASTM D02 ~~PTP 2.0~~ PT program uses ADrs, a resolution-sensitive version of the Anderson-Darling ~~statistic referred to as ADrs. The ADrs was developed for the ASTM PT programs. The ADrs~~ statistic. The ADrs is a special case of the ~~AD~~Anderson-Darling statistic for dealing with step normal distributions. ADrs is designed not to signal non-normality when presented with normally distributed data that have poor resolution or are coarsely rounded. The ADrs statistic is designed to assess the normality of datasets regardless of the coarseness of the reporting resolution.

NOTE 2—~~Until the~~See X2.1 ~~approach~~ for calculating ~~ADrs~~ ADrs.~~is~~[3] ~~included in Practice D6299, this approach can be obtained from the ASTM International PTP Office.~~

6.3.1.1 ~~The ASTM PTP 2.0 program uses~~ Use the following guidelines for interpretation of the ~~AD and~~ ADrs ~~statistics.~~statistic. This guide recognizes a range of ~~AD and~~ ADrs values where the data could be ~~consider marginally normal.~~considered normal, marginally normal, and not normal. The critical value for acceptance of normality for the ADrs is 0.752 for alpha = 0.05. The practical upper limit for acceptance of marginal normality is 1.12 for alpha = 0.05.

ADrs Range		Interpretation
~~AD, AD_{RS} <0.75~~	~~Normal~~	~~Data are likely normally distributed and the participants should take action to address all data flags.~~
< 0.75	Normal	Data are likely normally distributed; participants should take action to address all data flags.
~~AD, AD_{RS} 0.75 – 1.3~~	~~Marginally Normal~~	~~Data exhibit near normal behavior, so participants should consider action to address all data flags.~~
0.75 – 1.12	Marginally Normal	Data exhibit near normal behavior; participants should consider action to address all data flags.
~~AD, AD_{RS} >1.3~~	~~No~~	~~There is strong evidence that the data are not distributed normally, so corrective actions for data flags should be considered with some caution.~~
≥ 1.12	Not Normal	There is strong evidence that the data are not distributed normally; corrective actions for data flags should be considered with some caution.

6.3.2 *Median-based Approach When Data are Not Normally Distributed*—When ADrs > 1.12 and the proficiency testing data are thus not normally distributed, the usual data flags (see 6.2.2) should be used with caution and may not apply. In these cases, median-based statistics can be used to identify data that need investigation. This approach uses the median-based counterparts to the mean and standard deviation, namely the median ($\tilde{x}$) and the median absolute deviation (MAD). Using these statistics, a Modified Z-score (M_i)[4] can be determined for each result, $M_i = 0.6745 (X_i - \tilde{x})/MAD$. Data are flagged for investigation when the corresponding $|M_i|$ exceeds a critical value, D. A critical value of 3.5 has been shown to flag results that would correspond to exceeding a 3-sigma limit. See X2.2 for computation of median, MAD and M_i.

NOTE 3—The ASTM International D02 Proficiency Test Program is considering implementing this approach to report median, MAD, and Mi statistics along with corresponding flagged results.

6.4 *QQ Plots*—In addition, graphical tools are available for evaluating normality. For example, the ASTM PTP 2.0 uses a normal probability or a QQ plot (an equivalent plot to the normal probability plot) to visually assess the validity of the normality assumption and to identify data that are on the extremes of the distribution. Refer to Practice D6299 for guidance regarding the preparation and interpretation of normal probability plots. If data are normally distributed, the normal probability plot should be approximately linear. Major deviations from linearity are an indication of non-normal distributions. The appearance of a series of steps in the plotted data rather than a smooth line is an indication that the data (or measurement) resolution is too coarse relative to the precision of the test method. A few examples of these normal probability plots are shown in parallel with histograms in X3.2.

6.5 *Histograms:*

6.5.1 Histograms are a useful graphical tool for viewing data distribution and variability. The ASTM PT programs generate histograms for all data sets where n > 20; and includes the mean and the 1st and 99th percentile limits on the histogram for data sets with n > 100. These limits are based on "median ± 2.33 · Standard Deviation," where ±2.33 are respectively the first and 99th percentiles of the standard normal distribution.

6.5.2 PT program participants should review histograms when available and note unusual data distributions. Participants should locate where their result falls within the histogram bins. Depending on the histogram, the location of data in certain bins could indicate a potential issue such as bias. Consider reviewing the histogram in parallel with corresponding statistics such as the Z-score, AD statistic, TPI (Industry), and the normal probability (or deviate) plot. See X3.2 for examples.

[3] Supporting data have been filed at ASTM International Headquarters and may be obtained by requesting Research Report RR:D02-2023. Contact ASTM Customer Service at service@astm.org.

[4] Boris Iglewicz and David Hoaglin (1993), "*Volume 16: How to Detect and Handle Outliers*," The ASQC Basic References in Quality Control: Statistical Techniques, Edward F. Mykytka, Ph.D., Editor.

6.6 *Single Laboratory Bias (Deviation from Mean):*

6.6.1 As mentioned in Practice D6299, ~~subsection 7.6,~~ it is appropriate for PTP partcipants to evaluate proficiency test results by plotting the signed deviations from the mean for each result for each test cycle. Practice D6299 suggests plotting the signed deviations on control charts. Laboratories would then apply the strategies outlined in that standard to identify outliers and other issues such as long-term biases. The recommended control chart is a chart of individual observations (called an I-Chart) with an exponentially weighted moving average (EWMA) overlaid on the data. ~~See X3.3 for examples.~~

6.6.2 ~~Another~~A graphical approach for monitoring bias involves use of box and whisker graphs. As is the case for reviewing histograms, laboratories should use the box and whisker graphs to observe where their particular result lies in the graph relative to the general distribution of results for the test method they used. Consider investigating any data outside the whisker end, if those data were not flagged already for other causes. A review of the apparent distribution of results for each test method measuring the same parameter may provide valuable insight regarding overall biases between methods. See ~~7.2~~7.3 for more information on box and whisker plots. See X3.4.

6.6.3 Another statistical approach for evaluating bias is described in Practice D6617. This guide estimates whether or not a single test result is biased compared to the consensus value from the PT program.

6.7 *Z-score, Z'-score Trends*—The Z-score or Z'-score, or both, calculated for each datum submitted by the laboratory should be reviewed with respect to the following:

6.7.1 *Sign and Magnitude of Z-score*—The sign (that is, "+" or "–") of the statistic reflects the relative bias of the individual result versus the mean of the sample group (and standardized to the standard deviation of that data set). Z-score values falling in the ranges of plus or minus 0 to 1, 1 to 2, 2 to 3, and >3 can be compared to control chart values falling in the ranges between the mean and 1-sigma, 1 to 2-sigma, 2 to 3-sigma, and >3-sigma. For normally distributed data, there is an expectation that about 68 % of the data will lie in the –1 sigma to +1 sigma range, about 95 % in the –2 sigma to +2 sigma range, and 99 % in the –3 to +3 sigma range. The further a laboratory's Z-score is from zero, the greater the relative bias and lower the probability that the data is considered within statistical control. Conduct investigations to determine the cause of any perceived bias as needed.

6.7.2 *Z-score and Run-Sum*—Collect the Z-scores or Z'-scores for each test method (parameter) for successive PT program cycles and determine the running sum for successive same sign scores. Each time the sign reverses (changes from + to – or vice versa), restart the run-sum. Use the absolute value of the run-sum (|run-sum|) to evaluate the data for potential bias relative to the PT data set as shown below. In addition, 6-in-a-row Z-scores or Z'-scores with the same sign (+ or -) signals statistical evidence of systemic bias. Plotting Z-scores (6.7.3) along with Run-Sum is useful. See X3.3 for examples.

|run-sum|	Evaluation
≤ 2.0	Generally acceptable performance
2.0 to < 4.0	Growing evidence that a bias is developing
4.0 to < 6.0	Stronger evidence suggesting that data may be biased
≥ 6.0	Statistical evidence of systemic bias

Z-Score	Evaluation
6-in-a-row on same side	Statistical evidence of systemic bias

6.7.3 *Z-scores and/or Z'-score Trends Using Data from Multiple PTP Cycles*—Collect the Z-scores or Z'-scores values for each test method (parameter) for successive PT program cycles on a control chart to show the trend over time. Plotting Z-scores or Z'-scores is more practical than plotting the signed deviations from the mean (as in 6.2.1) especially when the magnitude of means can vary considerably from PT cycle to cycle. It is recommended to use the run rules promulgated in Practice D6299 to evaluate any observed trends. Conduct investigations to determine causes as needed. According to Practice D6299, Z-score and Z'-score data for a PT program cycle and test method parameter are acceptable for trend analysis via control charts when two conditions are met: first, there are at least 16 non-outlier data for the parameter and second, the PT cycle standard deviation is not statistically greater than the reproducibility standard deviation for the test method (see F-test).

6.7.4 *Average Z-score and Average Z'-score*—Calculate the average Z-score or Z'-score for a series over a selected time period. The sign and magnitude of this result is an indication of the long-term relative bias. Conduct investigations to determine the cause of any perceived bias as needed.

6.8 *Precision Performance:*

6.8.1 *TPI (Industry)*—Assess the general capability of a test method using TPI_{IND} alone or along with other tools such as Z-score, relative standard deviation (or coefficient of variance), and the ratio of mean to standard deviation (quantitation index). Note that one can determine capability of one method versus another based using the published ASTM reproducibility, which provides the accepted or target values, and the data from a PTP, which provides results as practiced by participating laboratories. In situations when the TPI_{IND} is not calculated in a PTP report, this statistic can be calculated by the user and interpreted as indicated below.

6.8.1.1 *General TPI Implications*—Consider Table 1 for interpreting the TPI_{IND}.

6.8.1.2 *Specific Implications Considering TPI_{IND} and Z-score*—Consider the TPI_{IND} value calculated for the data set along with the corresponding Z-score for the laboratory's result (reference Practice D6792). A $TPI_{IND} < 0.8$ coupled with a Z-score >3 (or <–3) implies that the laboratory is likely a significant contributor to the group's poor performance. This situation warrants an investigation to look for potential causes of the apparent bias. When the $TPI_{IND} < 0.8$ and the Z-score is between 2 and 3 (or –2 and –3), then the laboratory should consider the situation a warning and consider an investigation to determine if there are any assignable causes.

6.8.2 *Precision Performance Based on F-Test*—Precision performance, an indicator introduced in the ASTM PTP 2.0 reports, is based on the outcome of the F-test. Precision performance is a quantitative estimate of the reproducibility standard deviation of the PT program versus the published ASTM reproducibility standard deviation. For the F-test, the ratio of the standard deviations squared (larger divided by smaller) is compared to the 95th percentile of Fisher's F-distribution. These two standard deviations are the published reproducibility standard deviation for the ASTM test method ($s_{ASTM\ R}$) and the standard deviation for *these data* (s_{repro}). For determining the F-distribution, the degrees of freedom for *these data* is the number of conforming data used in the calculation of the standard deviation and the degrees of freedom for the ASTM standard deviation is assumed to be 30. In the ASTM PTP 2.0 program, the risk of Type I error is held to 5 % only if the distributions are nearly normal. This statistical test evaluates whether or not the PT precision is better than, consistent with, or worse than the ASTM precision in accordance with the following table:

F-Distribution	PT Precision Performance
<0.025	Better
0.025 – 0.975	Consistent
>0.975	Worse

6.9 *PTP and Site Precision Comparison*—Compare the reproducibility standard deviation for the PT results versus the site precision value derived from the laboratory's corresponding quality control chart. The expectation is that in most cases the site precision value should be less than the PT program standard deviation. If the laboratory's site precision is greater than the PT standard deviation, then the laboratory should investigate to determine the cause. The evaluation of site precision versus the corresponding PT precision is best accomplished using the F-test and the approach described in 6.8.2.

7. Procedure—Analysis and Interpretation by Standards Development Group

7.1 This section covers the analysis and interpretation of proficiency test data by a committee, industry group, or individual interested with determining the overall implications that the published PT results have with respect to the corresponding test method or to the general users as a whole. The following cover the evaluations and analyses that any group should consider during their review in addition to the approaches covered in the previous section.

7.2 *TPI_{IND} and Precision Trends*—Compare precisions obtained over a reasonable number of rounds for a given PT program test

TABLE 1 General TPI Implications

TPI (Industry) Result	Implication
> 1.2	The performance of the group providing data is probably satisfactory relative to the corresponding ASTM published precision.
0.8 to 1.2	The performance of the group providing data may be marginal and each laboratory should consider reviewing the test method procedures to identify opportunities for improvement.
< 0.8	The performance of the test method as practiced by the group is not consistent with the ASTM published precision and laboratory method performance improvements should be investigated by all laboratories.

method (or parameter). Plotting such data series often shows the appearance of trends more clearly. The precision estimates that may be followed TPI_{IND}, standard deviations, or relative standard deviations.

7.3 *Bias via Box and Whisker Plots:*

7.3.1 Box and whisker plots provide a convenient graphical representation of the means and relative data distributions for two or more test methods that measure the same property in the PT cycle. Box and whisker plots group test data by quartiles with the center box representing the middle 50 % of test data centered on the median. The horizontal line within the box represents the median of the reported data. The whisker length is adjusted to the last data point that falls within 1.5 times the difference between the upper and lower value of the center box. Data points above or below the whisker are included in the plot unless they are off the Y-axis scale.

7.3.2 The size (length) of the box and whisker is a measure of the precision of the PT results. The position of one median relative to that in another box is a measure of the relative bias among the test methods involved. The box and whisker plots, however, do not estimate the significance of any bias observed. Further, these graphs represent the distribution of data only for one PTP cycle, so observed biases and different data distributions observed for one cycle may not be supported in subsequent cycles.

7.4 *Normality Evaluations*—Plot the PT results as a QQ plot and consider the corresponding AD or ADrs statistic. Observe similar plots for the historical data sets for a given test method (parameter). Investigate situations of non-normal data. QQ plots generally are sensitive to situations where a small subset of laboratories perform the test method differently than the rest of the group. In these cases, the QQ plot shows an indication of a bimodal distribution, which can also be confirmed by a review of the corresponding histogram.

7.5 *Relative Standard Deviations:*

7.5.1 Relative standard deviation (RSD) (or the coefficient of variation, CV) expressed as a decimal or percent, is a convenient statistic to generate and interpret. Generally, the percent relative standard deviation should be low, perhaps at 10 % or lower. To establish a target, one can generate an expected percent RSD based on the published reproducibility. ~~Several examples of plots and interpretation of RSD data are provided in X3.9.~~

7.5.2 Another measure of test method capability is the quantitation index, the ratio of the mean to the standard deviation (that is, the reciprocal of the RSD). The reason for using a quantitation index relates to the use of a similar expression in evaluating limits of quantitation (that is, the point at which the ratio of mean concentration to repeatability standard deviation exceeds 10; see Practice D6259). This concept is especially important in evaluating test method performance at the lowest end of their operating ranges. ~~See the example in X3.10.~~

7.6 *Influence of Uncontrolled Variables on Robust Standard Deviations*—Use auxiliary information or data to create subsets of the PT data set and recalculate precisions and other statistics for each subset. Auxiliary information is the data/information collected by the PT program from participating laboratories to support investigations and includes topics such as instrument type or manufacturer, source of calibration standards, specific experimental conditions, etc. Contact the PT program administrator to arrange for collection of such auxiliary information. Evaluate these results with the expectation of identifying causes and potential corrective action steps.

7.7 *Contribution of Individual Laboratory Bias to Poor Reproducibility*—Identify the laboratories that are contributing to poor reproducibility (for example, those laboratories with Z-score > ±3) and evaluate the factors that may be contributing to this performance. This may involve targeting laboratories with questionnaires to gather appropriate information.

7.8 *Consultations*—Investigations are generally more successful when product experts, test method experts, and qualified statisticians are involved in the discussions.

8. Report

8.1 Laboratories and working groups should document their investigations. In the spirit of continuous improvement, laboratories and working groups are encouraged to share their findings from their investigations and analyses.

9. Keywords

9.1 precision performance; proficiency testing; quality control; test performance index; Z-score

APPENDIXES

(Nonmandatory Information)

X1. CHECKLIST FOR INVESTIGATING THE ROOT CAUSE OF UNSATISFACTORY ANALYTICAL PERFORMANCE

X1.1 For a laboratory to identify why their data may have been considered a statistical outlier or to improve the precision, or both, the following action items (not necessarily in the order of preference) are suggested. There may be additional ways to improve the performance.

X1.1.1 Check the results for typos, calculation errors, and transcription errors.

X1.1.2 Reanalyze the sample; compare the difference between this result to the original submitted result to site precision, or, if not available, test method repeatability.

X1.1.3 Review the test method, and ensure that the latest version of the ASTM test method is being used. Check the procedure step by step with the analyst.

X1.1.4 Check the instrument calibration.

X1.1.5 Check the statistical quality control chart to see if the problem developed earlier.

X1.1.6 Check the quality of the reagents and standards used and whether or not they are expired or contaminated.

X1.1.7 Check the sample for homogeneity, contamination, or that a representative sample has been analyzed.

X1.1.8 Check the equipment for proper operation against the vendor's operating manual.

X1.1.9 Perform maintenance or repairs, or both, on the equipment following guidelines established by the vendor.

X1.1.10 After the problem has been resolved, analyze a certified reference material, if one is available, or the laboratory quality control sample, to ascertain that the analytical operation is under control.

X1.1.11 Provide training to new analysts as needed, and, if necessary, refresher training to experienced analysts.

X1.1.12 Document the incident and the learnings for use in the future if a similar problem occurs.