Designation: D7372 – 17

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]

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

E177 Practice for Use of the Terms Precision and Bias in ASTM Test Methods

E456 Terminology Relating to Quality and 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.

3. Terminology

3.1 *Definitions:*

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

3.1.2 *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.

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

3.1.4 *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. **E177, E456**

3.1.5 *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. **E456**

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

3.1.7 *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**

[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. Published October 2017. Originally approved in 2007. Last previous edition approved in 2012 as D7372 – 12. DOI: 10.1520/D7372-17.

[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**

3.1.8 *proficiency testing, n*—determination of a laboratory's testing capability by participation in an interlaboratory proficiency test program **D6299**

3.1.9 *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.

3.1.9.1 *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.

3.1.10 *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* ($R_{these\ data}$).

3.1.11 *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**

3.1.12 *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**

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

3.1.13 *Z'-score, n*—measure similar to the *Z-score* 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 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.

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

where:

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

3.1.13.1 *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.

3.3 *Symbols:*

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

3.3.2 *PTP or PT program*—proficiency test program.

3.3.3 *QC*—quality control.

3.3.4 *R'*—site precision.

3.3.5 $R_{these\ data}$—reproducibility determined in PT program.

3.3.6 $r_{these\ data}$—repeatability determined in PT program.

3.3.7 R_{ASTM}—published ASTM reproducibility.

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)
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 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 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, 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 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 program uses a resolution-sensitive version of the Anderson-Darling statistic referred to as ADrs. The ADrs was developed for the ASTM PT programs. The ADrs is a special case of the AD 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.

NOTE 2—Until the approach for calculating ADrs is 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 the following guidelines for interpretation of the AD and ADrs statistics. This

guide recognizes a range of AD and ADrs values where the data could be consider marginally normal.

AD, AD_{RS} <0.75	Normal	Data are likely normally distributed and the 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.
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.

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.

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

6.6.1 As mentioned in Practice D6299, subsection 7.6, it is appropriate 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 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 for more information on box and whisker plots.

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-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).

6.7.3 *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 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.

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.

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.

X2. STATISTICAL TOOLS

INTRODUCTION

The following are statistical tools available for analysis of proficiency testing program results.

X2.1 Anderson-Darling (AD) Statistic

X2.1.1 Calculate the AD statistic in accordance with Practice D6299 to determine if the data are normally distributed. If the data are distributed normally (that is, AD < 0.75) or marginally normally (AD 0.75 to 1.3), then the equations below are applicable. When the AD > 1.3, suggesting that the data are not normally distributed, then the tools described below should be used with caution.

X2.1.2 Calculate the Anderson-Darling resolution sensitive (ADrs) statistic in accordance with the report referenced in the ASTM PTP 2.0 program reports and available from the ASTM PTP Office. The same criteria for interpretation of the AD statistic above applies to the ADrs.

X2.2 Standard Error of the Mean

X2.2.1 The standard error of the mean (SE) is used to assess the confidence interval for the sample means obtained from multiple cycles of a proficiency testing for a given test parameter.

$$SE = \left(\frac{s}{\sqrt{n}}\right) \quad \text{(X2.1)}$$

where:

s = standard deviation for the PTP results (per cycle), and
n = number of valid results reported.

X2.2.2 Estimate the upper and lower 95 % confidence intervals for the mean using Eq X2.2. The "1.96 · SE" expression is also known as expanded uncertainty (see discussions in Guide E2655). Examples of the use of Eq X2.2 include the error bars shown in Appendix X3 figures.

$$95\%\ \text{confidence limits} = \bar{X} \pm (1.96 \cdot SE) \quad \text{(X2.2)}$$

X2.3 Pooled Standard Deviation

X2.3.1 Estimate the pooled standard deviation (spooled) for multiple proficiency test cycles for a given test method using Eq X2.3. This assumes a normal or near normal distribution of data and that the precision is about the same for each cycle in the pooled set (either precision does not depend on level (concentration) or the concentration varies from cycle to cycle but in a narrow range).

$$S_{pooled} = \sqrt{\frac{\sum (n_1 - 1) \cdot {s_1}^2}{(\sum n) - N}} \quad \text{(X2.3)}$$

where:

n_1 = number of labs providing data in single cycle (no outliers),
s_1 = standard deviation for single cycle, and
N = number of proficiency testing cycles in data set.

X2.4 F-Test—Comparison of standard Deviations from Two Test Methods

X2.4.1 Use the F-test in Eq X2.4 for comparing two standard deviations from any sources provided they are independently obtained. For purposes of this discussion we use the F-test to determine if the precisions (standard deviations) for two data sets, from two test methods (X and Y) measuring the same parameter, are statistically indistinguishable (or conversely, that the differences are not statistically significant).

$$F = \left(\frac{S_Y}{S_X}\right)^2 \quad \text{(X2.4)}$$

X2.4.2 The degrees of freedom are $n_Y - 1$ and $n_X - 1$ for the numerator and denominator, respectively. When using Excel[3] for these calculations, the probability or p-value (two-tailed) for this test is determined by:

$$p = 2 \times MIN[FDIST(F, n_Y - 1, n_X - 1), 1 - FDIST(F, n_Y - 1, n_X - 1)] \quad \text{(X2.5)}$$

X2.4.3 If $p \leq 0.05$, conclude that the precision from test method X is different from (not equal to) that of test method Y with 95 % confidence. If $p \leq 0.10$ or $p \leq 0.01$, then s_X is different from s_Y with 90 % or 99 % confidence, respectively.

X2.5 T-Test—Comparing Means from Two Test Methods – Standard Deviations Not Equal

X2.5.1 Use the t-test in Eq X2.6 to determine if the means obtained from two test methods, X and Y, are distinguishable statistically. Statistically significant differences imply a bias of one method relative to the other, hence a relative bias. It is necessary to use absolute value here for the difference in means when using Excel's TDIST function.[3]

$$t = \frac{\left|\bar{X} - \bar{Y}\right|}{\sqrt{\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}}} \quad \text{(X2.6)}$$

X2.5.2 The approximate degrees of freedom for this statistic is:

$$df = \frac{\left(\frac{s_X^2}{n_X} + \frac{s_Y^2}{n_Y}\right)^2}{\frac{s_X^4}{n_X^2(n_X - 1)} + \frac{s_Y^4}{n_Y^2(n_Y - 1)}} \quad \text{(X2.7)}$$

X2.5.3 When using Excel[3] for these calculations, the probability or p-value (two-tailed) for this test is determined by p = TDIST(t,df,2). If $p < 0.05$, conclude with 95 % confidence that the means are significantly distinguishable and there is a high probability of a bias.

[3] Trademark of Microsoft.

X2.6 T-Test, Comparing Means from Two Test Methods—Standard Deviations Equal

X2.6.1 When the standard deviations for the two test methods are equal, use the t-test in Eq X2.8 to test for bias.

$$t = \frac{\left|\bar{X} - \bar{Y}\right|}{\sqrt{\frac{(n_X - 1)s_X^2 + (n_Y - 1)s_Y^2}{n_X + n_Y - 2}\left(\frac{1}{n_X} + \frac{1}{n_Y}\right)}} \quad \text{(X2.8)}$$

where:

df = $n_X + n_Y - 2$, and
p = TDIST(t,df,2).

X2.6.2 If $p < 0.05$, conclude with 95 % confidence that the means are statistically distinguishable and there is a high probability of a bias.

X3. EXAMPLES

INTRODUCTION

Examples supporting the analyses and interpretations discussed in Sections 6 and 7.

X3.1 ASTM Standards Referenced Only in Appendix X3[2]

D1266 Test Method for Sulfur in Petroleum Products (Lamp Method)

D2622 Test Method for Sulfur in Petroleum Products by Wavelength Dispersive X-ray Fluorescence Spectrometry

D4294 Test Method for Sulfur in Petroleum and Petroleum Products by Energy Dispersive X-ray Fluorescence Spectrometry

D4951 Test Method for Determination of Additive Elements in Lubricating Oils by Inductively Coupled Plasma Atomic Emission Spectrometry

D5185 Test Method for Multielement Determination of Used and Unused Lubricating Oils and Base Oils by Inductively Coupled Plasma Atomic Emission Spectrometry (ICP-AES)

D5453 Test Method for Determination of Total Sulfur in Light Hydrocarbons, Spark Ignition Engine Fuel, Diesel Engine Fuel, and Engine Oil by Ultraviolet Fluorescence

D7039 Test Method for Sulfur in Gasoline, Diesel Fuel, Jet Fuel, Kerosine, Biodiesel, Biodiesel Blends, and Gasoline-Ethanol Blends by Monochromatic Wavelength Dispersive X-ray Fluorescence Spectrometry

X3.2 Histograms

X3.2.1 The histogram in Fig. X3.1 represents a case for 45 valid results (that is, outliers rejected) at relatively low sulfur levels (1 mg⁄kg to 4 mg/kg) for PTP #2 diesel fuel sample DF21006. An AD = 0.76, indicates normally distributed data and the TPI (Industry) = 0.82 shows fair overall performance by the participants especially at the low sulfur levels. The linearity of the normal deviate plot along with the AD statistic and visual appearance of the histogram supports the conclusion of a normal distribution of data. The slight indication of steps in Fig. X3.2 suggests that measurement resolution issues may be involved. Laboratories with flagged results or results out on the wings of the distribution should consider investigating for cause.

X3.2.2 Fig. X3.3 shows the histogram for 4 mg⁄kg to 8 mg⁄kg sulfur by Test Method D5453 for a diesel sample (DF20906). In this case, there are 154 valid results, the TPI (Industry) = 1.32, and the AD = 0.93. The appearance of the histogram, the AD statistic and the approximate linearity of the normal deviate plot in Fig. X3.4 suggest that the data are normally distributed. The appearance of minor steps in the normal deviate plot indicates that there may be some measurement resolution issues. This may be related to the application of Test Method D5453 at the low sulfur levels, which represents the lower operation range for this test method.

X3.2.3 The histogram in Fig. X3.5 reflects the distribution for 70 valid results with a TPI (Industry) = 0.99 and AD = 1.07. The TPI (Industry) indicates good overall performance by the participants. The histogram shows a small node or bump near the first percentile limit. The slightly high value for the AD and the indications of non-linearity at the lower end of the normal deviate plot in Fig. X3.6 suggests non-normal behavior. This bimodality should be of concern to the participants especially those with results in the lower node on the left side of the histogram. Laboratories with results located in the area of the lower node should investigate looking for root causes.

X3.2.4 The histogram in Fig. X3.7 for sample DF20906 is a case for 49 valid results, TPI (Industry) = 0.51 and AD =2.04 for sulfur values down in the 3 mg⁄kg to 12 mg/kg range, the lower operating range for the test method. We found that the normal deviate plot in Fig. X3.8 shows a distinct step function indicative of measurement resolution problems for the method. The skewed shape of the histogram, the AD statistic, and the normal deviate plot generally support a conclusion that the data are not normally distributed. Based on these statistics, laboratories should investigate rejected data and Z-score > 3 outliers, but other less critical flagged data may not require significant effort to find a root cause.

X3.2.5 A laboratory could choose to develop histograms in support of their investigations. For example, use a histogram to view the distribution of data derived from several test methods measuring the same parameter for a single crosscheck sample. The histogram in Fig. X3.9 was generated in Excel[3] using PTP data from the DF20910 cycle for sulfur in #2 diesel fuel. In this case, where the mean sulfur level is in the 7 mg/kg range, the distribution of results for Test Methods D2622, D5453, and D7039 are similar and as expected. The distribution for Test Method D4294, however, is very different. This observation along with evaluations presented later in this chapter suggests

that Test Method D4294, as practiced by the laboratories in this PTP, appears to be less capable in this low sulfur range. In addition to Excel,[3] there are a number of other software tools available to the investigator for creating histograms.

X3.3 Bias (Deviation from Mean)

X3.3.1 Fig. X3.10 is a control chart for a single laboratory's deviations for determination of sulfur in ULSD by Test Method D5453. The use of a control chart in this case is acceptable in that both the corresponding AD and normal deviate plot (not shown) indicate a normal distribution of data. This control chart shows reasonably good behavior for the laboratory in that data are within the designated control limits and the moving average (EWMA) generally tracks about the mean showing no indications of bias. In cases where the precision varies with analyte level or the data cover a large range, then it may be more useful to plot the corresponding Z-scores rather than the raw deviations.

X3.4 Z-Scores

X3.4.1 Fig. X3.11 represents Z-score results simulated for two laboratories collected for nearly four years of monthly samples for sulfur in ULSD using Test Method D7039. In this case, a moving average (from Excel[3]) that is similar to the EWMA shows how the averages move about the centerline. This is similar to plotting Z-scores on a quality control chart. A review of the dispersion of Z-scores over time is different for the two labs. For the entire data set, an F-test shows that the standard deviations for all Z-scores for these two labs are statistically distinguishable. An examination of the chart for the most recent cycles would suggest that the standard deviations should be similar. This is substantiated using the F-test and t-test and the results show that the standard deviations and mean for the twelve most recent data are not statistically distinguishable. This means Lab 2 was able to improve performance over time eventually matching that for Lab 1.

X3.4.2 Fig. X3.12 shows Z-scores for sulfur in ULSD by Test Method D5453. A visual inspection of the chart indicates that the variability of the Z-scores seems to get worse over time as highlighted by the two overlaid ovals. Analysis of this data shows that the standard deviation during the earlier period (s = 0.71) was noticeably better than the precision (s = 1.06) in the more recent period. Further, an F-test of this data revealed that the respective standard deviations are statistically distinguishable. Therefore, the real question for the laboratory would be, what happened to cause the Z-score precision to increase and how can they return to the previous precision level. It is possible, however, that the Z-scores might be getting worse because the standard deviations representing the performance of all participating labs might be getting better over time.

X3.4.3 Fig. X3.13 shows sequential historic Z-scores plotted for two laboratories for determination of calcium in Lube Oil by Test Method D4951. This chart also shows the linear trend lines (from the Excel[3] spreadsheet). For these cases, we suggest using the graphics along with the corresponding performance indicators (PI) to enhance the analysis. The story for Lab A is a good one in that the trend for Z-score over time shows improvement from scores in the –1 to –2 range to the 0 to –1 range. Since PI > 0.8 there is no indication that the lab precision needs improvement. Lab B has a PI < 0.8, so lab precision may be in need of improvement, as is obvious from the scatter of data in the graph. There does not appear to be any decrease in variability over the more recent cycles, although the corresponding trend line is in the right direction. The differences in precision between the Labs A and B in Fig. X3.10 is obvious even without referring to the PI scores.

X3.5 TPI (Industry)

X3.5.1 Fig. X3.14 is a composite of sections of an ILCP report showing how a low TPI (Industry) score and a high Z-score along with the scatter plot and histogram lead to similar conclusions. This data indicates that Lab 26 should investigate the cause of their contribution to poor precision and bias. Even though the AD > 1.0 in this case, the other evidence is strong enough to support taking action to improve performance. See the next section for further discussions regarding analyses of TPI (Industry) data by the responsible technical groups.

X3.6 Box and Whisker Plots

X3.6.1 Fig. X3.15 shows the box and whisker plots for sulfur determinations in an ILCP jet fuel sample. The following discussion demonstrates how a laboratory or responsible work group might proceed with an investigation or analysis using these box and whisker graphs. The plots in Fig. X3.12 for a 2009 cycle have mean sulfur levels in the 1240 mg/kg (0.1240 % by mass) range. The data distributions (precision) and the means (relative biases) among the three test methods (Test Methods D2622, D4294, and D5453) appear to be similar. The data displayed for Test Method D1266 is largely ignored because there are only two data reported and the significance of the reported mean is uncertain. For the results displayed in Fig. X3.15, it is difficult just from a review of the displayed plots to determine if the observed differences in means and precisions among the three test methods are significant. Analyses using F-tests and t-tests show that the means and precisions among the pairs are statistically significantly distinguishable, with the exception of the Test Methods D2622 and D5453 pair where the means are not statistically distinguishable. The individual laboratory should determine where their results fall within the graph and evaluate any implications.

X3.7 Mean ($\bar{X}$) Graphs

X3.7.1 Perhaps the simplest graphical approach for evaluating long-term bias is to plot the means obtained for each test method on the same chart for multiple cycle results. This is more meaningful when results are available for numerous PT cycles. In these cases, one might be able to observe whether there appears to be any significant relative biases among the methods. Fig. X3.16 shows the relative biases for four test methods analyzing for sulfur in RFG. The ovals on the chart are used only to highlight the four results for a specific cycle/sample. Based on this chart, one would readily conclude that there is a reasonable chance that Test Method D4294 results are biased high relative to the means reported by the

other three methods. For the scale used in this plot the results for Test Method D2622, D5453, and D7039 are packed too closely to make any immediate conclusion regarding relative biases. Even with the scales exploded, it is still difficult to discern any bias. Although this is an easy plot to make, without also knowing the corresponding standard deviations it is difficult to know for sure if any of the observed differences are significant.

X3.7.2 An effective graphical approach is to plot the means for one or more test methods versus another test method. This shows relative biases more directly, especially if error bars ($\bar{X} \pm 1.96 \cdot SE$) are also used. Refer to Fig. X3.17 and Fig. X3.18 for the same data set discussed above. The overall bias for Test Method D4294 relative to Test Method D2622 is obvious in Fig. X3.17, especially considering the distance that most the error bars are from the parity line (vertical error bars are for Test Method D4294 and horizontal bars are for D2622). We determined using the t-test that this relative bias is statistically significant at the 95 % confidence level and using the F-test that the precisions are statistically distinguishable with Test Method D4294 having the larger precision.

X3.7.3 In some cases, the graphical approach does not lead one to a clear conclusion regarding biases. For example, the plot in Fig. X3.18 seems to suggest that Test Method D5453 results may not be biased relative to Test Method D2622. In this case, we observe that the error bars are generally close to, or overlapping, the parity line. In this case, one should resort to using a t-test for clearer guidance. Such analyses show that the means for Test Method D5453 are not statistically significantly distinguishable from D2622.

X3.7.4 From Fig. X3.19 it would appear that for the ILCP cycles observed, the means from Test Method D4294 are statistically significantly distinguishable from those using Test Method D2622. Statistical analyses to determine pooled standard deviations along with the F-test and t-test statistics show that the precision for Test Method D4294 is statistically significantly distinguishable compared to Test Method D2622. Further, the means for Test Method D4294 versus D2622 are not distinguishable at the 95 % confidence level; however, the difference appears to be statistically significant at the 90 % confidence level. These observations suggest further investigation is needed. Note that the approach used here is not quantitative because the standard deviations may vary with concentrations across the cycles.

X3.8 TPI (Industry)

X3.8.1 Fig. X3.20 and Fig. X3.21 are the graphs for TPI (Industry) versus either the test cycle or the mean sulfur content, respectively, for sulfur in RFG. For the TPI (Industry) versus test cycle chart, the linear trend line for Test Method D5453 shows an increasing trend, which means that overall, laboratory precision is improving over time and that interlaboratory variation is decreasing. The more recent data shows that the average TPI (Industry) values are approaching 1.0, which indicates improving capability. The TPI (Industry) trend for Test Method D4294 is decreasing indicating deteriorating performance over time. Although not apparent from the chart, the number of laboratories using this method is also decreasing. Test Method D4294 consistently posts the lowest TPI (Industry) values thus showing the poorest capability, as practiced by industry laboratories, for RFG and other products especially at low sulfur levels.

X3.8.2 Fig. X3.21 shows the TPI (Industry) data as a function of its corresponding reported mean sulfur level for a different data set representing sulfur determinations in ULSD. Although there is a fair scatter of TPI (Industry) results across the range of mean sulfur levels shown, the general tendency for TPI (Industry) for Test Method D5453 is in the 0.6 to 1.0 range. Test Method D4294, on the other hand, has relatively low values across this sulfur range. One could conclude that Test Method D4294, as practiced by the participating laboratories, is not very capable for sulfur measurements in RFG in the <100 mg⁄kg range.

X3.9 Relative Standard Deviation (RSD)

X3.9.1 Fig. X3.22 shows the relationship between RSD and the mean sulfur content of the sample across a number of products using Test Method D2622. The plots also show two expected RSD lines based on precision statements specifically for gasoline range and diesel range samples. This chart demonstrates that the capability of the test method (as practiced by industry laboratories) is relatively independent of product type across the concentration range studied for #2 diesel, jet, motor gasoline (Mogas) and RFG. Test Method D2622 performance is slightly better (that is, a lower RSD) for the ULSD product in the 5 mg⁄kg to 15 mg/kg sulfur range. Further investigation would be needed to understand the reasons for such observations.

X3.9.2 An analysis of Test Method D4294 capability from Fig. X3.23 shows that this method appears to be much less capable especially at sulfur levels below 100 mg/kg. Here, capability seems to be independent of the product types included. More investigation is warranted, but reaching such conclusions is not the purpose of this chapter.

X3.9.3 Fig. X3.24 shows good capability for Test Method D5453 across a range of ILCP products. Similar to the observations for Test Method D2622 (Fig. X3.22), the capability for Test Method D5453ppears to be slightly better for the ULSD product. These differences are just about as pronounced as it was for Test Method D2622 in the lowest sulfur range. Understanding the rationale supporting these differences would require further investigation by the responsible technical group.

X3.10 Ratio of Mean to Standard Deviation – Quantitation Index

X3.10.1 Fig. X3.25 examines the performance of Test Method D5453 for sulfur determinations in the 0 mg⁄kg to 15 mg⁄kg range, the lower end of its operating range. For these determinations, the capability of Test Method D5453 appears to be much better for ULSD product than for the others. The quantitation index for the ULSD product is in the 10 to 20 range compared to the 0 to 10 range for the other products. The base data set for Fig. X3.25 is the same as used in Fig. X3.24. The next step would be to understand why this occurs, a task