Designation: D6299 – 17

An American National Standard

Standard Practice for Applying Statistical Quality Assurance and Control Charting Techniques to Evaluate Analytical Measurement System Performance[1]

This standard is issued under the fixed designation D6299; 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 practice covers information for the design and operation of a program to monitor and control ongoing stability and precision and bias performance of selected analytical measurement systems using a collection of generally accepted statistical quality control (SQC) procedures and tools.

NOTE 1—A complete list of criteria for selecting measurement systems to which this practice should be applied and for determining the frequency at which it should be applied is beyond the scope of this practice. However, some factors to be considered include (*1*) frequency of use of the analytical measurement system, (*2*) criticality of the parameter being measured, (*3*) system stability and precision performance based on historical data, (*4*) business economics, and (*5*) regulatory, contractual, or test method requirements.

1.2 This practice is applicable to stable analytical measurement systems that produce results on a continuous numerical scale.

1.3 This practice is applicable to laboratory test methods.

1.4 This practice is applicable to validated process stream analyzers.

1.5 This practice is applicable to monitoring the differences between two analytical measurement systems that purport to measure the same property provided that both systems have been assessed in accordance with the statistical methodology in Practice D6708 and the appropriate bias applied.

NOTE 2—For validation of univariate process stream analyzers, see also Practice D3764.

NOTE 3—One or both of the analytical systems in 1.5 can be laboratory test methods or validated process stream analyzers.

1.6 This practice assumes that the normal (Gaussian) model is adequate for the description and prediction of measurement system behavior when it is in a state of statistical control.

NOTE 4—For non-Gaussian processes, transformations of test results may permit proper application of these tools. Consult a statistician for further guidance and information.

2. Referenced Documents

2.1 *ASTM Standards:*[2]

D3764 Practice for Validation of the Performance of Process Stream Analyzer Systems

D5191 Test Method for Vapor Pressure of Petroleum Products (Mini Method)

D6708 Practice for Statistical Assessment and Improvement of Expected Agreement Between Two Test Methods that Purport to Measure the Same Property of a Material

D6792 Practice for Quality System in Petroleum Products and Lubricants Testing Laboratories

D7372 Guide for Analysis and Interpretation of Proficiency Test Program Results

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

E178 Practice for Dealing With Outlying Observations

E456 Terminology Relating to Quality and Statistics

E691 Practice for Conducting an Interlaboratory Study to Determine the Precision of a Test Method

3. Terminology

3.1 *Definitions:*

3.1.1 *accepted reference value, n*—a value that serves as an agreed-upon reference for comparison and that is derived as (*1*) a theoretical or established value, based on scientific principles, (*2*) an assigned value, based on experimental work of some national or international organization, such as the U.S. National Institute of Standards and Technology (NIST), or (*3*) a consensus value, based on collaborative experimental work under the auspices of a scientific or engineering group. **E177, E456**

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

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

[1] This practice 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 Jan. 1, 2017. Published January 2017. Originally approved in 1998. Last previous edition approved in 2016 as D6299 – 16. DOI: 10.1520/D6299-17.

[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**

3.1.4 *bias, n*—a 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 *lot, n*—a definite quantity of a product or material accumulated under conditions that are considered uniform for sampling purposes. **E456**

3.1.7 *precision, n*—the closeness of agreement between test results obtained under prescribed conditions. **E456**

3.1.8 *repeatability conditions, n*—conditions where mutually independent test results are obtained with the same test method in the same laboratory by the same operator with the same equipment within short intervals of time, using test specimens taken at random from a single sample of material.

3.1.9 *reproducibility conditions, n*—conditions under which test results are obtained in different laboratories with the same test method, using test specimens taken at random from the same sample of material.

3.2 *Definitions of Terms Specific to This Standard:*

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

3.2.1.1 *Discussion*—A standard test method (for example, ASTM, ISO) is an example of an *analytical measurement system*.

3.2.1.2 *Discussion*—An analytical measurement system may comprise multiple instruments being used for the same test method provided there is no statistically observable bias and precision differences between the multiple instruments.

3.2.2 *blind submission, n*—submission of a check standard or quality control (QC) sample for analysis without revealing the expected value to the person performing the analysis.

3.2.3 *check standard, n—in QC testing*, a material having an accepted reference value used to determine the accuracy of a measurement system.

3.2.3.1 *Discussion*—A check standard is preferably a material that is either a certified reference material with traceability to a nationally recognized body or a material that has an accepted reference value established through interlaboratory testing. For some measurement systems, a pure, single component material having known value or a simple gravimetric or volumetric mixture of pure components having calculable value may serve as a check standard. Users should be aware that for measurement systems that show matrix dependencies, accuracy determined from pure compounds or simple mixtures may not be representative of that achieved on actual samples.

3.2.4 *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.

3.2.5 *double blind submission, n*—submission of a check standard or QC sample for analysis without revealing the check standard or QC sample status and expected value to the person performing the analysis.

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

3.2.7 *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 exceed specified limits.

3.2.8 *proficiency testing, n*—determination of a laboratory's testing capability by participation in an interlaboratory cross-check program.

3.2.8.1 *Discussion*—ASTM Committee D02 conducts proficiency testing among hundreds of laboratories, using a wide variety of petroleum products and lubricants.

3.2.9 *quality control (QC) sample, n*—for use in quality assurance programs to determine and monitor the precision and stability of a measurement system, a stable and homogeneous material having physical or chemical properties, or both, similar to those of typical samples tested by the analytical measurement system. The material is properly stored to ensure sample integrity, and is available in sufficient quantity for repeated, long term testing.

3.2.10 *site expected value (SEV), n*—for a QC sample this is an estimate of the theoretical limiting value towards which the average of results collected from a single in-statistical-control measurement system under site precision conditions tends as the number of results approaches infinity.

3.2.10.1 *Discussion*—The SEV is associated with a single measurement system; for control charts that are plotted in actual measured units, the SEV is required, since it is used as a reference value from which upper and lower control limits for the control chart specific to a batch of QC material are constructed.

3.2.11 *site precision (R'), n*—the 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 defined as 2.77 times $\sigma_{R'}$, the standard deviation of results obtained under site precision conditions.

3.2.12 *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.

3.2.12.1 *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. If multiple results are obtained within a 24-h period, then only results separated by at least 4 h should

be used in site precision calculations in order to reflect the longer term variation in the system.

3.2.13 *site precision standard deviation, n*—the standard deviation of results obtained under site precision conditions.

3.2.14 *validation audit sample, n*—a QC sample or check standard used to verify precision and bias estimated from routine quality assurance testing.

3.3 *Symbols:*

3.3.1 *ARV*—accepted reference value.

3.3.2 *EWMA*—exponentially weighted moving average.

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

3.3.4 *MR*—moving range.

3.3.5 $\overline{MR}$—average of moving range.

3.3.6 *QC*—quality control.

3.3.7 *R'*—site precision.

3.3.8 *SEV*—site expected value.

3.3.9 σ_R—site precision standard deviation.

3.3.10 *VA*—validation audit.

3.3.11 χ^2—chi squared.

3.3.12 λ—lambda.

4. Summary of Practice

4.1 QC samples and check standards are regularly analyzed by the measurement system. Control charts and other statistical techniques are presented to screen, plot, and interpret test results in accordance with industry-accepted practices to ascertain the in-statistical-control status of the measurement system.

4.2 Statistical estimates of the measurement system precision and bias are calculated and periodically updated using accrued data.

4.3 In addition, as part of a separate validation audit procedure, QC samples and check standards may be submitted blind or double-blind and randomly to the measurement system for routine testing to verify that the calculated precision and bias are representative of routine measurement system performance when there is no prior knowledge of the expected value or sample status.

5. Significance and Use

5.1 This practice can be used to continuously demonstrate the proficiency of analytical measurement systems that are used for establishing and ensuring the quality of petroleum and petroleum products.

5.2 Data accrued, using the techniques included in this practice, provide the ability to monitor analytical measurement system precision and bias.

5.3 These data are useful for updating test methods as well as for indicating areas of potential measurement system improvement.

6. Reference Materials

6.1 QC samples are used to establish and monitor the precision of the analytical measurement system.

6.1.1 Select a stable and homogeneous material having physical or chemical properties, or both, similar to those of typical samples tested by the analytical measurement system.

NOTE 5—When the QC sample is to be utilized for monitoring a process stream analyzer performance, it is often helpful to supplement the process analyzer system with a subsystem to automate the extraction, mixing, storage, and delivery functions associated with the QC sample.

6.1.2 Estimate the quantity of the material needed for each specific lot of QC sample to (*1*) accommodate the number of analytical measurement systems for which it is to be used (laboratory test apparatuses as well as process stream analyzer systems) and (*2*) provide determination of QC statistics for a useful and desirable period of time.

6.1.3 Collect the material into a single container and isolate it.

6.1.4 Thoroughly mix the material to ensure homogeneity.

6.1.5 Conduct any testing necessary to ensure that the QC sample meets the characteristics for its intended use.

6.1.6 Package or store QC samples, or both, as appropriate for the specific analytical measurement system to ensure that all analyses of samples from a given lot are performed on essentially identical material. If necessary, split the bulk material collected in 6.1.3 into separate and smaller containers to help ensure integrity over time. (**Warning—**Treat the material appropriately to ensure its stability, integrity, and homogeneity over the time period for which it is to stored and used. For samples that are volatile, such as gasoline, storage in one large container that is repeatedly opened and closed can result in loss of light ends. This problem can be avoided by chilling and splitting the bulk sample into smaller containers, each with a quantity sufficient to conduct the analysis. Similarly, samples prone to oxidation can benefit from splitting the bulk sample into smaller containers that can be blanketed with an inert gas prior to being sealed and leaving them sealed until the sample is needed.)

6.2 Check standards are used to estimate the accuracy of the analytical measurement system.

6.2.1 A check standard may be a commercial standard reference material when such material is available in appropriate quantity, quality and composition.

NOTE 6—Commercial reference material of appropriate composition may not be available for all measurement systems.

6.2.2 Alternatively, a check standard may be prepared from a material that is analyzed under reproducibility conditions by multiple measurement systems. The accepted reference value (ARV) for this check standard shall be the average after statistical examination and outlier treatment has been applied.[3]

6.2.2.1 Exchange samples circulated as part of an interlaboratory exchange program, or round robin, may be used as check standards. For an exchange sample to be usable as a check standard, the standard deviation of the interlaboratory exchange program shall not be statistically greater than the reproducibility standard deviation for the test method. An *F*-test should be applied to test acceptability.

[3] For guidance in statistical and outlier treatment of data, refer to Research Report RR:D02-1007, Practices E178 and E691, and *ASTM Standards on Precision and Bias for Various Applications*, ASTM International, 1997.

NOTE 7—The uncertainty in the ARV is inversely proportional to the square root of the number of values in the average. This practice recommends that a minimum of 16 non-outlier results be used in calculating the ARV to reduce the uncertainty of the ARV by a factor of 4 relative to the measurement system single value precision. The bias tests described in this practice assume that the uncertainty in the ARV is negligible relative to the measurement system precision. If less than 16 values are used in calculating the average, this assumption may not be valid.

NOTE 8—Examples of exchanges that may be acceptable are ASTM D02.CS92 ILCP program; ASTM D02.01 N.E.G.; ASTM D02.01.A Regional Exchanges; International Quality Assurance Exchange Program, administered by Alberta Research Council.

6.2.3 For some measurement systems, single, pure component materials with known value, or simple gravimetric or volumetric mixtures of pure components having calculable value may serve as a check standard. For example, pure solvents, such as 2,2-dimethylbutane, are used as check standards for the measurement of Reid vapor pressure by Test Method D5191. Users should be aware that for measurement systems that show matrix dependencies, accuracy determined from pure compounds or simple mixtures may not be representative of that achieved on actual samples.

6.3 Validation audit (VA) samples are QC samples and check standards, which may, at the option of the users, be submitted to the measurement system in a blind, or double blind, and random fashion to verify precision and bias estimated from routine quality assurance testing.

7. Quality Assurance (QA) Program for Individual Measurement Systems

7.1 *Overview*—A QA program (**1**)[4] can consist of five primary activities: (*1*) monitoring stability and precision through QC sample testing, (*2*) monitoring accuracy, (*3*) periodic evaluation of system performance in terms of precision or bias, or both, (*4*) proficiency testing through participation in interlaboratory exchange programs where such programs are available, and (*5*) a periodic and independent system validation using VA samples may be conducted to provide additional assurance of the system precision and bias metrics established from the primary testing activities. At minimum, the QA program must include at least item one and item two, subject to check standard availability (see 7.1.1).

7.1.1 For some measurement systems, suitable check standard materials may not exist, and there may be no reasonably available exchange programs to generate them. For such systems, there is no means of verifying the accuracy of the system, and the QA program will only involve monitoring stability and precision through QC sample testing.

NOTE 9—For guidance on the establishment and maintenance of the essentials of a quality system, see Practice D6792.

NOTE 10—For guidance on the analysis and interpretation of proficiency test (PT) program results, see Guide D7372.

7.2 *Monitoring System Stability and Precision Through QC Sample Testing*—QC test specimen samples from a specific lot are introduced and tested in the analytical measurement system on a regular basis to establish system performance history in terms of both stability and precision.

7.3 *Monitoring Accuracy:*

7.3.1 Check standards can be tested in the analytical measurement system on a regular basis to establish system performance history in terms of accuracy.

7.4 *Test Program Conditions/Frequency :*

7.4.1 Conduct both QC sample and check standard testing under site precision conditions.

NOTE 11—It is inappropriate to use test data collected under repeatability conditions to estimate the long term precision achievable by the site because the majority of the long term measurement system variance is due to common cause variations associated with the combination of time, operator, reagents, instrumentation calibration factors, and so forth, which would not be observable in data obtained under repeatability conditions.

7.4.2 Test the QC and check standard samples on a regular schedule, as appropriate. Principal factors to be considered for determining the frequency of testing are (*1*) frequency of use of the analytical measurement system, (*2*) criticality of the parameter being measured, (*3*) established system stability and precision performance based on historical data, (*4*) business economics, and (*5*) regulatory, contractual, or test method requirements.

NOTE 12—At the discretion of the laboratory, check standards may be used as QC samples. In this case, the results for the check standards may be used to monitor both stability (see 7.2) and accuracy (see 7.3) simultaneously. If check standards are expensive, or not available in sufficient quantity, then separate QC samples are employed. In this case, the accuracy (see 7.3) is monitored less frequently, and the QC sample testing (see 7.2) is used to demonstrate the stability of the measurement system between accuracy tests.

7.4.3 It is recommended that a QC sample be analyzed at the beginning of any set of measurements and immediately after a change is made to the measurement system.

7.4.4 Establish a protocol for testing so that all persons who routinely operate the system participate in generating QC test data.

7.4.5 Handle and test the QC and check standard samples in the same manner and under the same conditions as samples or materials routinely analyzed by the analytical measurement system.

7.4.6 When practical, randomize the time of check standard and additional QC sample testing over the normal hours of measurement system operation, unless otherwise prescribed in the specific test method.

NOTE 13—Avoid special treatment of QC samples designed to get a better result. Special treatment seriously undermines the integrity of precision estimates.

7.5 *Evaluation of System Performance in Terms of Precision and Bias:*

7.5.1 Pretreat and screen results accumulated from QC and check standard testing. Apply statistical techniques to the pretreated data to identify erroneous data. Plot appropriately pretreated data on control charts.

7.5.2 Periodically analyze results from control charts, excluding those data points with assignable causes, to quantify the bias and precision estimates for the measurement system.

7.6 *Proficiency Testing:*

[4] The boldface numbers in parentheses refer to the list of references at the end of this standard.

7.6.1 Participation in regularly conducted interlaboratory exchanges where typical production samples are tested by multiple measurement systems, using a specified (ASTM) test protocol, provide a cost-effective means of assessing measurement system accuracy relative to average industry performance. Such proficiency testing can be used instead of check standard testing for systems where the timeliness of the accuracy check is not critical. Proficiency testing may be used as a supplement to accuracy monitoring by way of check standard testing.

7.6.2 Participants plot their signed deviations from the consensus values (exchange averages) on control charts in the same fashion described below for check standards, to ascertain if their measurement processes are non-biased relative to industry average.

7.7 *Independent System Validation*—Periodically, at the discretion of users, VA samples may be submitted blind or double blind for analysis. Precision and bias estimates calculated using VA samples test data can be used as an independent validation of the routine QA program performance statistics.

NOTE 14—For measurement systems susceptible to human influence, the precision and bias estimates calculated from data where the analyst is aware of the sample status (QC or check standard) or expected values, or both, may underestimate the precision and bias achievable under routine operation. At the discretion of the users, and depending on the criticality of these measurement systems, the QA program may include periodic blind or double-blind testing of VA samples.

7.7.1 The specific design and approach to the VA testing program will depend on features specific to the measurement system and organizational requirements, and is beyond the intended scope of this practice. Some possible approaches are noted as follows.

7.7.1.1 If all QC samples or check standards, or both, are submitted blind or double blind and the results are promptly evaluated, then additional VA sample testing may not be necessary.

7.7.1.2 QC samples or check standards, or both, may be submitted as unknown samples at a specific frequency. Such submissions should not be so regular as to compromise their blind status.

7.7.1.3 Retains of previously analyzed samples may be resubmitted as unknown samples under site precision conditions. Generally, data from this approach can only yield precision estimates as retain samples do not have ARVs. Typically, the differences between the replicate analyses are plotted on control charts to estimate the precision of the measurement system. If precision is level dependent, the differences are scaled by the standard deviation of the measurement system precision at the level of the average of the two results.

8. Procedure for Pretreatment, Assessment, and Interpretation of Test Results

8.1 *Overview*—Results accumulated from QC, check standard, and VA sample testing are pretreated and screened. Statistical techniques are applied to the pretreated data to achieve the following objectives:

8.1.1 Identify erroneous data (outliers).

8.1.2 Assess initial results to validate system stability and assumptions associated with use of control chart technique (for example, dataset normality, adequacy of variations in the dataset relative to measurement resolution).

8.1.3 Deploy, interpret, and maintain control charts.

8.1.4 Quantify long term measurement precision and bias.

NOTE 15—Refer to the annex for examples of the application of the techniques that are discussed below and described in Section 9.

8.2 *Pretreatment of Test Results*—The purpose of pretreatment is to standardize the control chart scales so as to allow for data from multiple check standards or different batches of QC materials with different property levels to be plotted on the same chart.

8.2.1 For QC sample test results, no data pretreatment is necessary if results for different QC samples are plotted in actual measurement units on different control charts.

8.2.2 For check standard sample test results that are to be plotted on the same control chart, two cases apply, depending on the measurement system precision:

8.2.2.1 *Case 1*—If either (*1*) all of the check standard test results are from one or more lots of check standard material having the same ARV(s), or (*2*) the precision of the measurement system is constant across levels, then pretreatment consists of calculating the difference between the test result and the ARV:

$$\text{Pretreated result} = \text{test result} - \text{ARV(for the sample)} \quad (1)$$

8.2.2.2 *Case 2*—Test results are for multiple lots of check standards with different ARVs, and the precision of the measurement system is known to vary with level,

$$\text{Pretreated result} = \frac{[\text{test result} - \text{check standard ARV}]}{\text{sqrt}\left[(\text{standard error of ARV})^2 + (\text{std dev of site test method at the ARV level})^2\right]} \quad (2)$$

where the standard error of the ARV is the uncertainty associated with the ARV as supplied by the check standard supplier; the standard deviation of site test method at the ARV level is the established standard deviation of the site's test method under site precision conditions at nominally the ARV level. In the event the ARV was established through round robin testing, standard deviations determined from outlier-free and normally distributed round robin test results may be used to calculate the standard error of the ARV in accordance with statistical theory. (See Note 16.)

8.2.2.3 If the ARV was not arrived at by round robin testing, a standard error of the ARV should be determined by users in a technically acceptable manner.

NOTE 16—It is recommended that the method used to determine the standard error of the ARV be developed under the guidance of a statistician.

8.2.3 Pretreatment of results for VA samples is done in the same manner as described in 8.2.1 and 8.2.2.

8.3 *Control Charts* (**1, 2**)—Individual (*I*), moving range of two (*MR*) control charts, and either Strategy 1 (additional run rules) or Strategy 2 (EWMA) are the recommended toolset (see Annex A1) for (*a*) routine recording of QC sample and check standard test results, and (*b*) immediate assessment of the "in

statistical control" (3) status of the system that generated the data. The *I* chart is intended to detect occurrence of a sudden, unique event that causes a large deviation from the expected value for the QC material. Strategy 1 (additional Run Rules) or Strategy 2 (EWMA) is intended to detect small levels of sustained shifts or drifts of the complete analytical system. MR chart is intended to detect changes in the analytical system overall variability.

NOTE 17—The control charts and statistical techniques described in this practice are chosen for their simplicity and ease of use. It is not the intent of this practice to preclude use of other statistically equivalent or more advanced techniques, or both.

8.3.1 Control charting can be viewed as a two-staged work process where:

Stage 1 comprises assessment of initial test results (for a QC material) and construction of the control chart with graphically represented assessed results and statistical values that describes the location of where future test results for this QC material from the measurement systems are expected to fall within, on the assumption that the measurement system and QC material remains unchanged.

Stage 2 comprises regular assessment of future test results (for the QC material) as they arrive in chronological order against the established expectations in Stage 1; as well as a periodic reevaluation of the expectation statistics of all accrued results to update the expectations statistics established from Stage 1, if necessary.

STAGE 1—Assessment and Chart Construction

8.4 *Assessment of Initial Results*—Assessment techniques are applied to test results collected during the initial startup phase of or after significant modifications to a measurement system (see Note 19). Perform the following assessment after at least 20 pretreated results have become available. The purpose of this assessment is to ensure that these results are suitable for deployment of control charts (described in A1.4).

NOTE 18—These techniques can also be applied as diagnostic tools to investigate out-of-control situations.

NOTE 19—During the data collection phase in Stage 1, users can deploy the procedures described in 8.7.2.3 and 8.7.3 (*Q*–procedure) to monitor measurement process performance.

8.4.1 *Screen for Suspicious Results*—Pretreated results should first be visually screened for values that are inconsistent with the remainder of the data set, such as those that could have been caused by transcription errors. Those flagged as suspicious should be investigated. Discarding data at this stage must be supported by evidence gathered from the investigation. If, after discarding suspicious pretreated results there are less than 15 values remaining, collect additional data and start over.

8.4.2 *Screen for Unusual Patterns*—The next step is to examine the pretreated results for non-random patterns such as continuous trending in either direction, unusual clustering, and cycles. One way to do this is to plot the results on a run chart (see A1.3) and examine the plot. If any non-random pattern is detected, investigate for and eliminate the root cause(s). Discard the data set and start the procedure again.

8.4.3 *Test "Normality" Assumption, Independence of Test Results, and Adequacy of Measurement Resolution*—For measurement systems with no prior performance history, or as a diagnostic tool, it is useful to test that the results from the measurement system are reasonably independent, with adequate measurement resolution, and can be adequately modelled by a normal distribution. One way to do this is to use a normal probability plot and the Anderson-Darling Statistic (see A1.4). If the results show obvious deviation from normality or obvious measurement resolution inadequacy (see A1.4), follow the guidance in A1.4.2.6, Case 2.

NOTE 20—Transformations may lead to normally distributed data, but these techniques are outside the scope of this practice.

8.4.4 *Construction of Control Charts*—If no obvious unusual patterns are detected from the run charts, and no obvious deviation from normality is detected, proceed with construction of the control charts

8.4.4.1 Construct an *MR* plot and examine it for unusual patterns. If no unusual patterns are found in the *MR* plot, calculate and overlay the control limits on the *MR* plot to complete the *MR* chart.

8.4.4.2 *I Chart*—Calculate control limits and overlay them on the "run chart" to produce the *I* chart.

8.4.4.3 *EWMA Overlay*—Optionally, calculate the *EWMA* values and plot them on the *I* chart. Calculate the *EWMA* control limits and overlay them on the *I* chart.

STAGE 2—Deployment for Monitoring and Periodic Re-assessment

8.4.5 *Control Chart Deployment*—Put these control charts into operation by regularly plotting the pretreated test results on the charts and immediately interpreting the charts.

8.5 *Control Chart Interpretation :*

8.5.1 Apply control chart rules (see A1.5) to determine if the data supports the hypothesis that the measurement system is under the influence of common causes variation only (in statistical control).

8.5.2 *Investigate Out-of-Control Points in Detail*—Exclude from further data analysis those associated with assignable causes, provided the assignable causes are deemed not to be part of the normal process.

NOTE 21—All data, regardless of in-control or out-of-control status, needs to be recorded.

8.6 *Scenario 1 for Periodic Updating of Control Charts Parameters:*

8.6.1 Scenario 1 covers (*1*) control charts for a QC material where there had been no change in the system, but more data of the same level has been accrued; or (*2*) control charts for check standard pretreated results.

8.6.2 When a minimum of 20 new in-control data points becomes available, perform an *F*-test (see A1.8) of sample variances for the new data set versus the sample variance used to calculate the current control chart limits. If the outcome of the *F*-test is not significant, and, if the sample variance used to calculate the current control limits is based on less than 100 data points, statistically pool both sample variances and then update the current control limits based on this new pooled variance.

8.6.3 If the outcome of the *F*-test is not significant, and if the sample variance used to calculate the current control limits is based on more than 100 data points, the statistical pooling of both sample variances and update of the current control limits can be at the discretion of the user.

8.6.4 If the outcome of the *F*-test is significant, investigate for assignable causes. Update the current control limits based on this new sample variance if it is determined that this new variance is representative of current system performance.

8.7 *Scenario 2 for Periodic Updating of Control Charts Parameters:*

8.7.1 Scenario 2 covers control chart for QC materials where an assignable cause change in the system had occurred due to a change of QC material as the current QC material supply is exhausted. Minor or major differences in measured property level may exist between QC material batches. Since control limit calculations for the *I* chart require a center value established by the measurement system, a special transition procedure is required to ensure that the center value for a new batch of QC material is established using results produced by a measurement system that is in statistical control. This practice presents two procedures to be selected at the users' discretion.

8.7.2 *Procedure 1, Concurrent Testing:*

8.7.2.1 Collect and prepare a new batch of QC material when the current QC material supply remaining can support no more than 20 analyses.

8.7.2.2 Concurrently test and record data for the new material each time a current QC sample is tested. The result for the new material is deemed valid if the measurement process in-control status is validated by the current QC material and control chart.

8.7.2.3 Optionally, to provide an early indication of the status of the new batch of QC material, immediately start a run chart and an *MR* plot for the new material. After five valid results become available for the new material, convert the run chart into an *I* chart with trial control limits by adding a center line based on the average of the five results and control limits based on the $\overline{MR}$ from previous control charts for materials at the same nominal level. Set trial control limits for the *MR* chart based on limits from previous charts for materials at the same nominal level.

8.7.2.4 After a minimum of 20 in-control data points are collected on the new material, perform an *F*-test of sample variances for the new data set versus the historical variance demonstrated at nominal level of the new material. If the outcome of the *F*-test is not significant, and, if the historical variance demonstrated at nominal level of the new material is based on less than 100 data points, statistically pool both sample variances and then update the current control limits based on this new pooled variance.

8.7.2.5 If the outcome of the *F*-test is not significant, and, if the historical variance demonstrated at nominal level of the new material is based on more than 100 data points, the statistical pooling of both sample variances and update of the current control limits can be at the discretion of the user.

8.7.2.6 If the outcome of the *F*-test is significant, investigate for assignable causes. Update the current control limits based on this new sample variance if it is determined that this new variance is representative of current system performance.

8.7.2.7 Construct new *I* and *MR* charts (and *EWMA* overlay for strategy 2) for this new material as per Section 8, using the pooled $\overline{MR}$.

8.7.2.8 Switch over to the new *I* and *MR* charts upon depletion of current QC material.

8.7.3 *Procedure 2, Q-Procedure (see A1.9)* **(4)***:*

8.7.3.1 This procedure is designed to alleviate the need for concurrent testing of two materials. A priori knowledge of the measurement process historical standard deviation applicable at the new QC material composition and property level is required.

NOTE 22—It is recommended that this standard deviation estimate be based on at least 50 data points.

8.7.3.2 When the *Q*-procedure is operational (minimum of two data points), it can be used in conjunction with a *MR* chart constructed using the observations to provide QA of the measurement process.

8.7.3.3 Because the *Q*-procedure is technically equivalent to the I chart procedure, after 20 data points have been accrued (by the *Q*-procedure), the user can either follow the steps described in 8.7.2 on Concurrent Testing after 20 data points have been accrued to construct a new *I/MR* control chart for the new batch of QC material, or continue to operate the *Q*–chart and MR chart for measurement process stability and precision monitoring, respectively, using the new batch of QC material.

8.7.3.4 It is necessary to start a new *Q*-chart with each new batch of QC material if the plotted results are not pre-treated, or, if the new batch of material has a different historical standard deviation and the plotted results are not pre-treated.

8.7.3.5 A common *Q*-chart and MR chart can be used for pre-treated results as per Case I and Case II in 8.2. For Case I, the standard deviation shall be the applicable standard deviation for the QC material; for Case II, the standard deviation is the value in the denominator of Eq 2.

8.8 *Short Run Scenario*—The *Q*-procedure (described in 8.7.3) can also be used to address short run situations where a single batch of QC material can provide only a limited number (less than 20) of QC test results and replacement of exactly the same material is not feasible or possible. For these short run QC batches, since there is insufficient data to properly characterize the mean of batch, the *Q*-procedure, in conjunction with the MR chart, can be used to monitor stability and precision of the measurement process, respectively.

8.9 *Instrument Replacement or Post Overhaul Scenario*—The *Q*-procedure (described in 8.7.3) may be used to address situations where an instrument is taken out of service and is replaced by another qualified instrument, or, when the primary instrument is returned to service after a major overhaul such as replacement of critical parts or factory re-calibration. For these situations, the existing system precision parameters can be used with the Q-procedure, in conjunction with the MR chart, to monitor stability and precision of the replacement or overhauled measurement process, respectively, based on the assumption that the existing system precision parameter is still valid. After sufficient data is accrued, a statistical assessment

shall be performed to confirm this assumption, or update the system precision parameters accordingly. Use of the existing precision will enable the system to be immediately put into service, while providing a safeguard against the situation where the new system performance with replacement or overhauled instrument is statistically worse than the previous system performance. Use of the Q-procedure is in addition to any steps such as calibration and running check standards needed to qualify replacement instruments.

9. Evaluation of System Performance in Terms of Precision and Bias

9.1 *Site Precision Estimated from Testing of QC Samples:*

9.1.1 Estimate the site precision of the measurement system at the level corresponding to a specific lot of QC sample using the root-mean-square (rms) formula for standard deviation ($\sigma_{R'}$).

$$\sigma_{R'} = \sqrt{\frac{\sum_{i=1}^{n}\left(I_i - \bar{I}\right)^2}{n-1}} \quad (3)$$

$$R' = 2.77 \times \sigma_{R'} \quad (4)$$

9.1.1.1 Alternatively, in the absence of auto-correlation in the data (see A1.4), R' may be estimated as 2.46 times the average of the moving range $(\overline{MR})$ from the MR chart for that specific lot.

$$R' = 2.46 \times \overline{MR} \quad (5)$$

NOTE 23—The site precision standard deviation ($\sigma_{R'}$) is estimated from the MR chart as $R'/2.77 = (\overline{MR})/1.128$.

9.1.1.2 For estimate of site precision standard deviation ($\sigma_{R'}$) using retain results, first obtain the standard deviation of differences by applying the root-mean-square formula below to the differences between the original and retest results for samples with same nominal property level. If measurement process precision is known to be level independent, retest results from samples with different property levels can be used. Otherwise, sample pairs with nominally similar property level (general rule is within 2R) should be used to estimate the site precision at the nominal property level. Divide the standard deviation of differences by 1.414 to obtain the estimate for site precision standard deviation. ($\sigma_{R'}$).

$$\text{standard deviation of differences} = \sqrt{\frac{\sum (\text{individual difference} - \text{average difference})^2}{\text{total number of differences}}} \quad (6)$$

$$\sigma_{R'} = (\text{standard deviation of differences}) \div 1.414 \quad (7)$$

9.1.2 Compare R' to published reproducibility of the test method at the same level, if available. R' is expected to be less than or equal to the published value. Use the χ^2 test described in A1.7.

9.2 *Measurement System Bias Estimated from Multiple Measurements of a Single Check Standard*—If a minimum of 15 test results is obtained on a single check standard material under site precision conditions, then calculate the average of all the in-control individual differences plotted on the *I* chart. Perform a *t*-test (see A1.6) to determine if the average is statistically different from zero.

9.2.1 If the outcome of the *t*-test is that the average is not statistically different from zero, then the bias in the measurement process is negligible.

9.2.2 If the outcome of the *t*-test is that the average is statistically different from zero, then the best estimate of the measurement process bias at the level of the check standard is the average. If bias is deemed to be of practical significance by the user, investigate for root causes, and take corrective measures.

9.3 *Measurement System Bias Estimated from Measurements of Multiple Check Standards*—When using multiple check standards, determine if there is a relationship between the bias and the measurement level.

9.3.1 Plot the pretreated results as per Section 8 versus their corresponding ARVs. Examine the plot for patterns indicative of level-dependent bias.

9.3.2 If there is no discernible pattern, perform the *t*-test as described in 9.2 to determine if the average of all the pretreated differences plotted on the *I* chart is statistically different from zero.

9.3.2.1 If the outcome of the *t*-test is that the average is not statistically different from zero, then the bias in the measurement process is negligible.

9.3.2.2 If the outcome of the *t*-test is that the average is statistically different from zero, then there is evidence that the measurement system is biased. The bias may be level dependent. However, the statistical methodology for estimating the bias/level relationship is beyond the scope of this practice.

9.3.3 If there is a discernible pattern in the plot in 9.3.1, then the measurement system may exhibit a level dependent bias. The statistical methodology for estimating the bias/level relationship is beyond the scope of this practice.

9.3.4 If a bias is detected in 9.3.2.2, or if the plot in 9.3.3 exhibits discernible patterns, investigate for root cause(s).

9.3.4.1 If there is evidence of a bias versus level relationship, or, if users wish to perform a more rigorous examination of the bias versus level relationship with multiple check standards, it is recommended that the principles of Practice D6708 be employed under the guidance of qualified statistical expertise.

10. Validation of System Performance Estimates Using VA Samples

10.1 If the users decide to include VA sample testing as part of their QA program, then they should periodically evaluate the results obtained on the VA samples. The purpose of the evaluation is to establish whether the system performance estimates described in Section 9 are reasonably applicable to routinely tested samples.

10.2 VA sample test results should be evaluated independently through an internal or external audit system, or both. It is recommended that the internal audit team not be limited to the operators of the measurement system and their immediate supervisors.

10.3 Insofar as possible, analyze the results obtained on the VA samples separately and in the same manner as those from the routine QC and check standard testing program.

10.4 Using *F*- or *t*- tests, or both (see A1.8 and A1.6), statistically compare the system performance estimates obtained from the VA sample testing program to the measurement system accuracy and precision estimates from the QC sample testing program.

10.5 If the comparison reveals that the two estimates of the measurement system performance are not statistically equivalent, there is cause for concern that the actual performance of the measurement system may be significantly worse than estimated. Investigate thoroughly for the assignable cause(s) of this inconsistency, and eliminate it. Until the causes are identified and eliminated, the lab precision estimates of Section 9 should be considered suspect.

ANNEX

(Mandatory Information)

A1. STATISTICAL QUALITY CONTROL TOOLS

A1.1 Purpose of this Annex

A1.1.1 The purpose of this annex is to provide guidance to practitioners, including worked examples, for the proper execution of the statistical procedures described in this practice. See Tables A1.1-A1.13 and Figs. 1–15.

NOTE A1.1—For some examples in this annex, 15 data points are used to illustrate calculation and plotting methodologies; it is not the intention of this annex to override the mandatory requirement of 20 minimum data points (see 8.4). Work is underway to revise the annex examples to use 20 data points for all examples.

A1.2 Pretreatment of Test Results (8.1 to 8.2.3)

A1.2.1 Throughout this annex, $\{Y_i: i=1. . .n\}$ denotes a sequence of as measured test results. $\{I_i: i=1. . .n\}$ will signify a sequence of test results after pretreatment, if necessary.

A1.2.2 If $\{Y_i: i=1. . .n\}$ is a sequence of results from a single QC sample, then

$$I_i = Y_i \quad \text{(A1.1)}$$

with no pretreatment being required.

A1.2.2.1 An example of a sequence of results, Y_i, from a single QC sample is given in Columns 2 and 4 of Table A1.3.

A1.2.3 If $\{Y_i: i=1. . .n\}$ is a sequence of results from a single check standard, from multiple check standards having nominally the same ARV, or from multiple check standards having different ARVs where the precision of the measurement system does not vary with level, and if $\{X_i: i=1. . .n\}$ is the sequence of corresponding ARVs, then

$$I_i = Y_i - X_i \quad \text{(A1.2)}$$

The site precision (R') of the measurement process must be essentially the same for all values $\{X_i\}$.

A1.2.3.1 An example of a sequence of results from a single check standard is given in Table A1.4. The preprocessed result, I_i, is given in Column 4 of Table A1.4.

A1.2.4 If $\{Y_i\}$ is a sequence of results from different check standards, and if the reproducibility varies with the level of the accepted reference values, $\{X_i\}$, then

$$I_i = (Y_i - X_i)/\sigma_i \quad \text{(A1.3)}$$

where σ_i are estimates of the standard deviation under site precision conditions of the measurement process at levels $\{X_i\}$.

A1.2.4.1 Table A1.5 shows an example of results for multiple check standards where the precision of the measurement system is level dependent.

A1.2.4.2 *Discussion*—Site precision (R') estimates at ARV values that are significantly different from those in the site's historical database can also be estimated proportionally using the published R at the ARV level. Calculate the fraction of R' and R at the ARV level with known R' and multiply this fraction by R at the new ARV level with unknown R' to arrive at the estimated R' at the new ARV level. This approach is based on the assumption that the fraction of R' and R is constant among different ARV levels. Users are cautioned that this assumption may not be valid if the published precision has different functional forms between r and R. Note that this fraction is the inverse of TPI (Test Performance Index) as defined in Practice D6792.

Example:

R' of site (calculated from actual QC data) at sulfur level 10 ppm = 2 ppm (published R at sulfur level of 10 ppm = 3 ppm).

Fraction of R'/R at 10 ppm = 2/3

Estimated R' of site at sulfur level at 15 ppm is estimated as: (2/3)* (published R at sulfur level of 15 ppm).

A1.3 The Run Chart

A1.3.1 A run chart is a plot of results in chronological order that can be used to screen data for unusual patterns. Preferably, pretreated results are plotted. Use a run chart to screen data for unusual patterns such as continuous trending in either direction, unusual clustering, and cycles. Several non-random patterns are described in control chart literature. When control parameters have been added to a run chart, it becomes a control chart of individual values (*I* chart).

A1.3.2 Plot results on the chart. Plot the first result at the left, and plot each subsequent point one increment to the right of its predecessor. The points may be connected in sequence to facilitate interpretation of the run chart.

A1.3.3 Allow sufficient space in the *x*-axis direction to accommodate as many results as should be obtained from a

consistent batch of material. Allow enough space in the *y*-axis direction to accommodate the expected minimum and maximum of the data.

A1.3.4 *Example of a Run Chart for QC Results*—The first 15 results from Column 2 of Table A1.3 are plotted in sequence as they are collected as shown in Fig. A1.1. The data would be examined for unusual patterns.

A1.3.5 *Example of a Run Chart for Multiple Results from a Single Check Standard*—The first 15 preprocessed results (differences) from Column 4 of Table A1.4 are plotted in sequence as they are collected as shown in Fig. A1.2. The data would be examined for unusual patterns.

A1.3.6 *Example of a Run Chart for Results from Multiple Check Standards*—The first 15 preprocessed results (differences scaled by σ_i) from Table A1.5 are plotted in sequence as they are collected as shown in Fig. A1.3. The data would be examined for unusual patterns.

A1.4 Normality, Data Independence, and Resolution Adequacy Checks

A1.4.1 A normal probability plot (a special case of a *q-q* plot) is used to visually assess the validity of the assumption that the observations are normally distributed. Since the control chart and limits prescribed in this practice are based on the assumption that the data behavior is adequately modeled by the normal distribution, it is recommended that a test of this normality assumption be conducted.

A1.4.1.1 To construct a normal probability plot:

(1) Create a column of the observations sorted in ascending order.

(2) Select the appropriate column from Fig. A1.4, based on the number of observations (*n*).

(3) Plot each observation in the sorted column (*y*-value) against its corresponding value from Fig. A1.4 (*z*-value).

A1.4.1.2 Visually inspect the plot for an approximately linear relationship. If the results are normally distributed, the plot should be approximately linear. Major deviations from linearity are an indication of nonnormal distributions of the differences.

NOTE A1.2—The assessment methodology of the normal probability plot advocated in this practice is strictly visual due to its simplicity. For statistically more rigorous assessment techniques, users are advised to use the Anderson-Darling technique described below, and consult a statistician.

A1.4.2 *Anderson-Darling Statistic* —The Anderson-Darling (A-D) statistic is used to objectively test for normality, data independence, and adequacy of measurement resolution relative to the overall variation in the dataset. Two A-D statistics ($A\text{-}D_{rms}$, $A\text{-}D_{MR}$) are calculated using the identical procedure outlined as follows, where $A\text{-}D_{rms}$, $A\text{-}D_{MR}$ are the A-D statistic calculated using numerical estimates of the sample standard deviation(*s*) as per the *rms* (root-mean-square) and the *MR*- (moving range of 2) techniques, respectively. The calculation steps are as follows:

A1.4.2.1 Order the non-outlying results such that $x_1 \le x_2 \le \ldots x_n$

A1.4.2.2 Obtain standardized variate from the x_i values as follows:

$$w_i = (x_i - \bar{x})/s \quad \text{(A1.4)}$$

for (i= 1 . . . n), where s is sample standard deviation of the results using either the *rms* or *MR* technique, and $\bar{x}$ is the average of the results.

NOTE A1.3—One standard deviation estimate ~ 0.89 × [average MR] of the dataset.

A1.4.2.3 Convert the w_i values to standard normal cumulative probabilities p_i values using the cumulative probability table for the standardized normal variate z (see Fig. A1.5):

$$p_i = \text{Probability}\ (z < w_i) \quad \text{(A1.5)}$$

A1.4.2.4 Compute A^2 as:

$$A^2 = -\frac{\sum_{i=1}^{n}(2i-1)\left[\ln(p_i)+\ln(1-p_{n+1-i})\right]}{n} - n \quad \text{(A1.6)}$$

A1.4.2.5 Compute the quantity A^{2*} as:

$$A^{2*} = A^2\left(1+\frac{0.75}{n}+\frac{2.25}{n^2}\right) \quad \text{(A1.7)}$$

The quantity A^{2*} is referred to as the A-D statistic (A-D).

A1.4.2.6 *Guidance on Interpretation of the Two A-D Statistics (A-D_{rms} and A-D_{MR}):* CASE 1—Both $A\text{-}D_{rms}$ and $A\text{-}D_{MR}$ are << 1.0. This is to be interpreted as, "no compelling evidence to reject the hypotheses that the data is normal, independent, with adequate measurement resolution." Proceed to construct control chart with either the rms-based or the MR-based standard deviation estimate.

CASE 2—Both $A\text{-}D_{rms}$ and $A\text{-}D_{MR}$ are >>> 1.0, and the q-q plot shows a few distinct "staircases," which really means the majority of the data is clustered into a few unique values. This is strong evidence that there is inadequate variation in the dataset due to inadequate numerical resolution. Under these circumstances, if the total number of unique values in the data set is less than six, increase data resolution (carry an additional decimal) and reevaluate both A-D statistics for the purpose of control charting. Note that because results are used for internal QA purposes, this should not be considered as a deviation from test method reporting requirements. If additional data resolution is not possible, or, if the total number of unique values in the data set is six or greater, or, if after increase in data resolution, both A-D statistics are still >>1.0, users can still use regular plotting of chronological QC data to monitor for occurrence of an abnormal event. For the purpose of the latter, it is recommended that the run-chart be used with a lower and upper percentile-based action limits, provided that there is no visual indication of process trending in the data set used to determine the action limits. The suggested percentiles are 1st and 99th, based on a data set of at least 75 results, collected under site precision conditions. It is not the intent of this practice to exclude use of other percentiles, or, use of other user-defined action limits, provided the limits meet the application requirements. Users are advised to seek qualified statistical guidance on how to determine the appropriate action limits and associated implications.

CASE 3—$A\text{-}D_{rms}$ is << 1.0, but $A\text{-}D_{MR}$ > 1.0. This is indicative that the test results are serially correlated, or not independent. A direct consequence of this non-independence is