Designation: ~~D6300 – 23~~ D6300 – 23a

An American National Standard

Standard Practice for Determination of Precision and Bias Data for Use in Test Methods for Petroleum Products, Liquid Fuels, and Lubricants[1]

This standard is issued under the fixed designation D6300; 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.

INTRODUCTION

Both Research Report RR:D02-1007,[2] *Manual on Determining Precision Data for ASTM Methods on Petroleum Products and Lubricants*[2] and the ISO 4259, benefitted greatly from more than 50 years of collaboration between ASTM and the Institute of Petroleum (IP) in the UK. The more recent work was documented by the IP and has become ISO 4259.

ISO 4259 encompasses both the determination of precision and the application of such precision data. In effect, it combines the type of information in RR:D02-1007[2] regarding the determination of the precision estimates and the type of information in Practice D3244 for the utilization of test data. The following practice, intended to replace RR:D02-1007,[2] differs slightly from related portions of the ISO standard.

1. Scope*

1.1 This practice covers the necessary preparations and planning for the conduct of interlaboratory programs for the development of estimates of precision (determinability, repeatability, and reproducibility) and of bias (absolute and relative), and further presents the standard phraseology for incorporating such information into standard test methods.

1.2 This practice is generally limited to homogeneous petroleum products, liquid fuels, and lubricants with which serious sampling problems (such as heterogeneity or instability) do not normally arise.

1.3 This practice may not be suitable for products with sampling problems as described in 1.2, solid or semisolid products such as petroleum coke, industrial pitches, paraffin waxes, greases, or solid lubricants when the heterogeneous properties of the substances create sampling problems. In such instances, consult a trained statistician.

1.4 *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.*

---

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

Current edition approved ~~July 1, 2023~~Dec. 1, 2023. Published ~~August 2023~~December 2023. Originally approved in 1998. Last previous edition approved in ~~2021~~2023 as ~~D6300 – 21.~~D6300 – 23. DOI: ~~10.1520/D6300-23.~~10.1520/D6300-23A.

[2] Supporting data have been filed at ASTM International Headquarters and may be obtained by requesting Research Report RR:D02-1007. Contact ASTM Customer Service at service@astm.org.

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

2. Referenced Documents

2.1 *ASTM Standards:*[3]

D3244 Practice for Utilization of Test Data to Determine Conformance with Specifications
D3606 Test Method for Determination of Benzene and Toluene in Spark Ignition Fuels by Gas Chromatography
D6708 Practice for Statistical Assessment and Improvement of Expected Agreement Between Two Test Methods that Purport to Measure the Same Property of a Material
D7915 Practice for Application of Generalized Extreme Studentized Deviate (GESD) Technique to Simultaneously Identify Multiple Outliers in a Data Set
E29 Practice for Using Significant Digits in Test Data to Determine Conformance with Specifications
E177 Practice for Use of the Terms Precision and Bias in ASTM Test Methods
E456 Terminology Relating to Quality and Statistics
E691 Practice for Conducting an Interlaboratory Study to Determine the Precision of a Test Method

2.2 *ISO Standards:*

ISO 4259 Petroleum Products-Determination and Application of Precision Data in Relation to Methods of Test[4]

3. Terminology

3.1 *Definitions:*

3.1.1 *analysis of variance (ANOVA), n*—technique that enables the total variance of a method to be broken down into its component factors. **ISO 4259**

3.1.2 *bias, n*—the difference between the expectation of the test results and an accepted reference value.

3.1.2.1 *Discussion*—

The term "expectation" is used in the context of statistics terminology, which implies it is a "statistical expectation." **E177**

3.1.3 *between-method bias (relative bias), n*—a quantitative expression for the mathematical correction that can statistically improve the degree of agreement between the expected values of two test methods which purport to measure the same property. **D6708**

3.1.4 *degrees of freedom, n*—the divisor used in the calculation of variance, one less than the number of independent results.

3.1.4.1 *Discussion*—

This definition applies strictly only in the simplest cases. Complete definitions are beyond the scope of this practice. **ISO 4259**

3.1.5 *determinability, n*—a quantitative measure of the variability associated with the same operator in a given laboratory obtaining successive determined values using the same apparatus for a series of operations leading to a single result; it is defined as the difference between two such single determined values that would be exceeded about 5 % of the time (one case in 20 in the long run) in the normal and correct operation of the test method.

3.1.5.1 *Discussion*—

This definition implies that two determined values, obtained under determinability conditions, which differ by more than the determinability value should be considered suspect. If an operator obtains more than two determinations, then it would usually be satisfactory to check the most discordant determination against the mean of the remainder, using determinability as the critical difference (**1**).[5]

3.1.6 *mean square, n—in analysis of variance*, sum of squares divided by the degrees of freedom. **ISO 4259**

3.1.7 *normal distribution, n*—the distribution that has the probability function x, such that, if x is any real number, the probability density is

$$f(x) = (1/\sigma)(2\pi)^{-1/2}\exp[-(x-\mu)^2/2\sigma^2] \quad (1)$$

NOTE 1—μ is the true value and σ is the standard deviation of the normal distribution (σ > 0). **ISO 4259**

[3] 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.

[4] Available from American National Standards Institute (ANSI), 25 W. 43rd St., 4th Floor, New York, NY 10036, http://www.ansi.org.

[5] The bold numbers in parentheses refers to the list of references at the end of this standard.

3.1.8 *outlier, n*—a result far enough in magnitude from other results to be considered not a part of the set. **RR:D02–1007[2]**

3.1.9 *precision, n*—the degree of agreement between two or more results on the same property of identical test material. In this practice, precision statements are framed in terms of *repeatability* and *reproducibility* of the test method.

3.1.9.1 *Discussion*—

The testing conditions represented by repeatability and reproducibility should reflect the normal extremes of variability under which the test is commonly used. Repeatability conditions are those showing the least variation; reproducibility, the usual maximum degree of variability. Refer to the definitions of each of these terms for greater detail.

RR:D02–1007[2]

3.1.10 *random error, n*—the chance variation encountered in all test work despite the closest control of variables. **RR:D02–1007[2]**

3.1.11 *repeatability (a.k.a. Repeatability Limit), n*—a quantitative expression for the random error associated with the difference between two independent results obtained under repeatability conditions that would be exceeded about 5 % of the time (one case in 20 in the long run) in the normal and correct operation of the test method.

3.1.11.1 *Discussion*—

Interpret as the limit value the absolute difference between two single test results obtained under repeatability conditions is expected to exceed with an approximate probability of 5 %.

3.1.11.2 *Discussion*—

The difference is related to the repeatability standard deviation but it is not the standard deviation or its estimate.

3.1.11.3 *Discussion*—

In 3.1.11 and 3.1.13, the term “probability” quantifies the likelihood of repeatability or reproducibility limit exceedance for the difference between a single pair of results obtained under the respective conditions. The "one case in 20 in the long run" in the parenthesis is not to be interpreted as one case in every 20, but it is over the long run. The long run concept can be illustrated using 10 cases out of 200, or 100 cases out of 2000, or 1000 cases in 20 000. The lowest numerical values of one case in 20 is used here.

3.1.11.4 *Discussion*—

The "one case in 20" is a legacy term that was carried over from RR:D02-1007 in the original development of Practice D6300.

RR:D02–1007[2]

3.1.12 *repeatability conditions, n*—conditions where independent test results are obtained with the same method on identical test items in the same laboratory by the same operator using the same equipment within short intervals of time. **E177**

3.1.13 *reproducibility (a.k.a. Reproducibility Limit), n*—a quantitative expression for the random error associated with the difference between two independent results obtained under reproducibility conditions that would be exceeded about 5 % of the time (one case in 20 in the long run) in the normal and correct operation of the test method.

3.1.13.1 *Discussion*—

Interpret as the limit value the absolute difference between two single test results obtained under reproducibility conditions is expected to exceed with an approximate a probability of 5 %.

3.1.13.2 *Discussion*—

The difference is related to the reproducibility standard deviation but is not the standard deviation or its estimate. **RR:D02–1007[2]**

3.1.13.3 *Discussion*—

In those cases where the normal use of the test method does not involve sending a sample to a testing laboratory, either because it is an in-line test method or because of serious sample instabilities or similar reasons, the precision test for obtaining reproducibility may allow for the use of apparatus from the participating laboratories at a common site (several common sites, if feasible). The statistical analysis is not affected thereby. However, the interpretation of the reproducibility value will be affected since the test data is collected under intermediate precision conditions as defined in Practice E177, and therefore, the precision statement shall, in this case, state the conditions to which the reproducibility value applies, and label this precision in a manner consistent with how the test data is obtained.

NOTE 2—The reproducibility precision outcome from 3.1.13.3 is a form of Intermediate Precision as defined in Practice E177.

3.1.14 *reproducibility conditions, n*—conditions where independent test results are obtained with the same method on identical test items in different laboratories with different operators using different equipment.

NOTE 3—Different laboratory by necessity means a different operator, different equipment, and different location and under different supervisory control.

3.1.15 *standard deviation, n*—measure of the dispersion of a series of results around their mean, equal to the square root of the variance and estimated by the positive square root of the mean square. **ISO 4259**

3.1.16 *sum of squares, n—in analysis of variance*, sum of squares of the differences between a series of results and their mean. **ISO 4259**

3.1.17 *variance, n*—a measure of the dispersion of a series of accepted results about their average. It is equal to the sum of the squares of the deviation of each result from the average, divided by the number of degrees of freedom. **RR:D02–1007[2]**

3.1.18 *variance, between-laboratory, n*—that component of the overall variance due to the difference in the mean values obtained by different laboratories. **ISO 4259**

3.1.18.1 *Discussion—*

When results obtained by more than one laboratory are compared, the scatter is usually wider than when the same number of tests are carried out by a single laboratory, and there is some variation between means obtained by different laboratories. Differences in operator technique, instrumentation, environment, and sample "as received" are among the factors that can affect the between laboratory variance. There is a corresponding definition for between-operator variance.

3.1.18.2 *Discussion—*

The term "between-laboratory" is often shortened to "laboratory" when used to qualify representative parameters of the dispersion of the population of results, for example as "laboratory variance."

3.2 *Definitions of Terms Specific to This Standard:*

3.2.1 *determination, n*—the process of carrying out a series of operations specified in the test method whereby a single value is obtained.

3.2.2 *operator, n*—a person who carries out a particular test.

3.2.3 *probability density function, n*—function which yields the probability that the random variable takes on any one of its admissible values; here, we are interested only in the normal probability.

3.2.4 *result, n*—the final value obtained by following the complete set of instructions in the test method.

3.2.4.1 *Discussion—*

It may be obtained from a single determination or from several determinations, depending on the instructions in the method. When rounding off results, the procedures described in Practice E29 shall be used.

4. Summary of Practice

4.1 A draft of the test method is prepared and a pilot program can be conducted to verify details of the procedure and to estimate roughly the precision of the test method.

4.1.1 If the responsible committee decides that an interlaboratory study for the test method is to take place at a later point in time, an interim repeatability is estimated by following the requirements in 6.2.1.

4.2 A plan is developed for the interlaboratory study using the number of participating laboratories to determine the number of samples needed to provide the necessary degrees of freedom. Samples are acquired and distributed. The interlaboratory study is then conducted on an agreed draft of the test method.

4.3 The data are summarized and analyzed. Any dependence of precision on the level of test result is removed by transformation. The resulting data are inspected for uniformity and for outliers. Any missing and rejected data are estimated. The transformation is confirmed. Finally, an analysis of variance is performed, followed by calculation of repeatability, reproducibility, and bias. When it forms a necessary part of the test procedure, the determinability is also calculated.

5. Significance and Use

5.1 ASTM test methods are frequently intended for use in the manufacture, selling, and buying of materials in accordance with

specifications and therefore should provide such precision that when the test is properly performed by a competent operator, the results will be found satisfactory for judging the compliance of the material with the specification. Statements addressing precision and bias are required in ASTM test methods. These then give the user an idea of the precision of the resulting data and its relationship to an accepted reference material or source (if available). Statements addressing determinability are sometimes required as part of the test method procedure in order to provide early warning of a significant degradation of testing quality while processing any series of samples.

5.2 Repeatability and reproducibility are defined in the precision section of every Committee D02 test method. Determinability is defined above in Section 3. The relationship among the three measures of precision can be tabulated in terms of their different sources of variation (see Table 1).

5.2.1 When used, determinability is a mandatory part of the Procedure section. It will allow operators to check their technique for the sequence of operations specified. It also ensures that a result based on the set of determined values is not subject to excessive variability from that source.

5.3 A bias statement furnishes guidelines on the relationship between a set of test results and a related set of accepted reference values. When the bias of a test method is known, a compensating adjustment can be incorporated in the test method.

5.4 This practice is intended for use by D02 subcommittees in determining precision estimates and bias statements to be used in D02 test methods. Its procedures correspond with ISO 4259 and are the basis for the Committee D02 computer software, *Calculation of Precision Data: Petroleum Test Methods*. The use of this practice replaces that of Research Report RR:D02-1007.[2]

5.5 Standard practices for the calculation of precision have been written by many committees with emphasis on their particular product area. One developed by Committee E11 on Statistics is Practice E691. Practice E691 and this practice differ as outlined in Table 2.

6. Stages in Planning of an Interlaboratory Test Program for the Determination of the Precision of a Test Method

6.1 The stages in planning an interlaboratory test program are: preparing a draft method of test (see 6.2), planning and executing a pilot program with at least two laboratories (optional but recommended for new test methods) (see 6.3), planning the interlaboratory program (see 6.4), and executing the interlaboratory program (see 6.5). The four stages are described in turn.

6.2 *Preparing a Draft Method of Test*—This shall contain all the necessary details for carrying out the test and reporting the results. Any condition which could alter the results shall be specified. The section on precision will be included at this stage only as a heading.

6.2.1 *Interim Repeatability Study*—If the responsible committee decides that an interlaboratory study for the test method is to take place at a later point in time, using this standard, an interim repeatability standard deviation is estimated by following the steps as outlined below. This interim repeatability standard deviation can be used to meet ASTM Form and Style Requirement A21.5.1. When the committee is ready to proceed with the ILS, continue with this practice from 6.3 onwards.

6.2.1.1 *Design*—The following minimum requirements shall be met:

(1) Three (3) samples, compositionally representative of the majority of materials within the design envelope of the test method, covering the low, medium, and high regions of the intended test method range.

(2) Twelve (12) replicates per sample, obtained under repeatability conditions in a single laboratory.

6.2.1.2 *Analysis*—Carry out the following analyses in the order presented:

(1) Perform GESD Outlier Rejection as per Practice D7915 for each sample.

TABLE 1 Sources of Variation

	Method	Apparatus	Operator	Laboratory	Time
Reproducibility	Complete (Result)	Different	Different	Different	Not Specified
Repeatability	Complete (Result)	Same	Same	Same	Almost same
Determinability	Incomplete (Part result)	Same	Same	Same	Almost same

TABLE 2 Differences in Calculation of Precision in Practices D6300 and E691

Element	This Practice	Practice E691
Number of replicates	Two	Any number
Precision is written for	Test method	Each sample
Outlier tests: Within laboratories Between laboratories	Sequential Cochran test Hawkins test	Simultaneous *k*-value *h*-value
Outliers	Rejected, subject to subcommittee approval.	Rejected if many laboratories or for cause such as blunder or not following method.
	Retesting not generally permitted.	Laboratory may retest sample having rejected data.
Analysis of variance	Two-way, applied globally to all the remaining data at once.	One-way, applied to each sample separately.
Precision multiplier	$t\sqrt{2}$, where t is the two-tailed Student's t for 95 % probability.	$2.8=1.96\sqrt{2}$
	Increases with decreasing laboratories × samples particularly below 12.	Constant.
Variation of precision with level	Minimized by data transformation. Equations for repeatability and reproducibility are generated in the retransformation process.	User may assess from individual sample precisions.

(2) Calculate sample variance (v) and standard deviation (s) for each sample using non-rejected results.

(3) Perform the Hartley test for variance equality as follows:

calculate the ratio : $F_{max} = v_{max}/v_{min}$ where v_{max} and v_{min} are the largest and smallest variance obtained.

(4) If F_{max} is less than 4.85, estimate the interim repeatability standard deviation of the test method by taking the square root of the average variance calculated using individual variances from all samples as illustrated below using three samples:

Interim repeatability standard deviation = $[(v_1 + v_2 + v_3)/3]^{0.5}$, where v_1, v_2, v_3 are variances for each sample; it should be noted that if the number of non-outlying results used to calculate the variances are not the same, this equation provides an approximation only, but is suitable for the intended purpose.

(5) If F_{max} exceeds 4.85, list the averages and associated repeatability standard deviations for each sample separately.

(6) If F_{max} exceeds 4.85, and, v_{max} is associated with the sample with the lowest average, calculate the following ratio: $[10\ s_{max}]/average_{\text{sample}}$, where s_{max} is $(v_{max})^{0.5}$, and
$average_{\text{sample}}$ is the average of the sample. If this ratio is near or exceeds 1, then it is likely that this sample is at or below the limit of quantitation of the test method. If this ratio is far below 1, it is likely this is a sample-specific effect. Method developers should investigate and take appropriate steps to revise the test method scope or improve the test method precision at the low limit prior to the conduct of a full ILS.

(7) If the sample set design meets the requirement in 6.4.2, the methodology in Appendix X2 can be used to estimate an interim repeatability function by treating the repeats per sample as results from 'pseudo-laboratories' without repeats.

NOTE 4—It is highly recommended that 6.2.1.2(7) be conducted under the guidance of a statistician familiar with the methodology in Appendix X2.

6.2.1.3 *Validation of Interim Repeatability Study by Another Laboratory*—It is highly recommended that the findings from the interim repeatability study be validated by conducting a similar study at another laboratory. If the findings from the validation study do not support the functional form (constant or per Appendix X2) of the interim repeatability study obtained by the initial laboratory, or, if the ratio:

$$\left[\frac{\text{interim repeatability standard deviation from lab A}}{\text{interim repeatability standard deviation from lab B}}\right]^2$$

exceeds 2.4, where the larger of the standard deviation value is in the numerator, that is, if the repeatability standard deviation for lab A is numerically larger than B; otherwise use the repeatability standard deviation for lab B in the numerator and the repeatability standard deviation for lab A in the denominator, it can be concluded that the findings from one laboratory cannot be validated by another laboratory. The method developer is advised to consult a statistician and subject matter experts to decide on which laboratory findings are to be used.

6.3 *Planning and Executing a Pilot Program with at Least Two Laboratories:*

6.3.1 A pilot program is recommended to be used with new test methods for the following reasons: (*1*) to verify the details in the operation of the test; (*2*) to find out how well operators can follow the instructions of the test method; (*3*) to check the precautions regarding sample handling and storage; and (*4*) to estimate roughly the precision of the test.

6.3.2 At least two samples are required, covering the range of results to which the test is intended to apply; however, include at least 12 laboratory-sample combinations. Test each sample twice by each laboratory under repeatability conditions. If any omissions or inaccuracies in the draft method are revealed, they shall now be corrected. Analyze the results for precision, bias, and determinability (if applicable) using this practice. If any are considered to be too large for the technical application, then consider alterations to the test method.

6.4 *Planning the Interlaboratory Program:*

6.4.1 There shall be at least six (6) participating laboratories, but it is recommended this number be increased to eight (8) or more in order to ensure the final precision is based on at least six (6) laboratories and to make the precision statement more representative of the qualified user population.

6.4.2 The number of samples shall be sufficient to cover the range of the property measured, and to give reliability to the precision estimates. If any variation of precision with level was observed in the results of the pilot program, then at least six samples, spanning the range of the test method in a manner than ensures the leverage (h) of each sample (see Eq 2) is less than 0.5 shall be used in the interlaboratory program. In any case, it is necessary to obtain at least 30 degrees of freedom in both repeatability and reproducibility. For repeatability, this means obtaining a total of at least 30 pairs of results in the program. In the absence of pilot test program information to permit use of Fig. 1 (see 6.4.3) to determine the number of samples, the number of samples shall be greater than five, and chosen such that the number of laboratories times the number of samples is greater than or equal to 42. Leverage calculation:

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{k=1}^{n}(x_k - \bar{x})^2} \tag{2}$$

h_{ii} = leverage of sample i,
n = total number of planned samples,
p_i = planned property level for sample i,
x_i = ln (p_i), and
$\bar{x}$ = grand average of all x_i.

6.4.3 For reproducibility, Fig. 1 gives the minimum number of samples required in terms of L, P, and Q, where L is the number of participating laboratories, and P and Q are the ratios of variance component estimates (see 8.3.1) obtained from the pilot program. Specifically, P is the ratio of the interaction component to the repeats component, and Q is the ratio of the laboratories component to the repeats component.

NOTE 5—Appendix X1 gives the derivation of the equation used. If Q is much larger than P, then 30 degrees of freedom cannot be achieved; the blank entries in Fig. 1 correspond to this situation or the approach of it (that is, when more than 20 samples are required). For these cases, there is likely to be a significant bias between laboratories. The program organizer shall be informed; further standardization of the test method may be necessary.

6.5 *Executing the Interlaboratory Program:*

6.5.1 One person shall oversee the entire program, from the distribution of the texts and samples to the final appraisal of the results. He or she shall be familiar with the test method, but should not personally take part in the actual running of the tests.

L = number of participating Laboratories
component

P = interaction variance component/repeats variance component

Q = Laboratories variance component/repeats variance

L=6

P \ Q	0	1	2	3	4	5	6	7	8	9
0	3									
1	4	11								
2	5	7								
3	5	7	14							
4	5	6	10							
5	6	6	8	15						
6	6	6	8	11						
7	6	6	7	10	15					
8	6	6	7	9	12					
9	6	6	7	8	10	15				

L=7

P \ Q	0	1	2	3	4	5	6	7	8	9
0	4									
1	5									
2	6	11								
3	6	9								
4	7	8	16							
5	7	8	12							
6	7	8	11	19						
7	7	8	10	15						
8	7	8	9	13						
9	7	8	9	11	17					

L=8

P \ Q	0	1	2	3	4	5	6	7	8	9
0	3									
1	3	5								
2	4	5	9							
3	4	5	7	14						
4	4	4	6	9	20					
5	4	4	5	7	11					
6	4	4	5	6	8	13				
7	4	4	5	6	7	10	16			
8	4	5	5	6	6	8	11	18		
9	4	5	5	5	6	7	9	13		

L=9

P \ Q	0	1	2	3	4	5	6	7	8	9
0	2									
1	3	4								
2	3	4	7							
3	3	4	5	9						
4	4	4	5	6	11					
5	4	4	5	6	7	12				
6	4	4	4	5	6	9	14			
7	4	4	4	5	6	7	10	15		
8	4	4	4	5	5	6	8	10	16	
9	4	4	4	5	5	6	7	8	11	18

L=10

P \ Q	0	1	2	3	4	5	6	7	8	9
1	2	8								
1	3	4	11							
2	3	4	5	12						
3	3	3	4	6	13					
4	3	4	4	5	7	14				
5	3	4	4	5	6	8	14			
6	3	4	4	4	5	6	9	14		
7	3	4	4	4	5	6	7	9	14	
8	3	4	4	4	5	5	6	7	10	14
9	4	4	4	4	4	5	6	6	8	10

L=11

P \ Q	0	1	2	3	4	5	6	7	8	9
0	2	4								
1	2	3	5							
2	3	3	3	7						
3	3	3	4	5	8					
4	3	3	4	4	6	8	18			
5	3	3	4	4	5	6	9	15		
6	3	3	3	4	4	5	6	9	14	
7	3	3	3	4	4	5	5	7	9	13
8	3	3	3	4	4	4	5	6	7	9
9	3	3	3	4	4	4	5	5	6	7

L=12

P \ Q	0	1	2	3	4	5	6	7	8	9
0	2	4								
1	2	3	5							
2	2	3	4	6	14					
3	3	3	3	4	6	11				
4	3	3	3	4	5	6	9			
5	3	3	3	4	4	5	6	9	16	
6	3	3	3	3	4	4	5	6	9	13
7	3	3	3	3	4	4	5	5	6	8
8	3	3	3	3	4	4	4	5	5	6
9	3	3	3	3	3	4	4	4	5	6

L=13

P \ Q	0	1	2	3	4	5	6	7	8	9
0	2	3								
1	2	3	4	12						
2	2	3	3	4	8					
3	2	3	3	4	5	7	14			
4	3	3	3	3	4	5	7	10		
5	3	3	3	3	4	4	5	6	9	15
6	3	3	3	3	3	4	4	5	6	8
7	3	3	3	3	3	4	4	4	5	6
8	3	3	3	3	3	3	4	4	5	5
9	3	3	3	3	3	3	4	4	4	5

L=14

P \ Q	0	1	2	3	4	5	6	7	8	9
0	2	3								
1	2	2	3	7						
2	2	2	3	4	6	12				
3	2	2	3	3	4	5	8	18		
4	2	3	3	3	3	4	5	7	11	
5	2	3	3	3	3	4	4	5	6	8
6	3	3	3	3	3	3	4	4	5	6
7	3	3	3	3	3	3	3	4	4	5
8	3	3	3	3	3	3	3	4	4	4
9	3	3	3	3	3	3	3	3	4	4

L=15

P \ Q	0	1	2	3	4	5	6	7	8	9
0	2	2	13							
1	2	2	3	5	19					
2	2	2	3	3	4	7				
3	2	2	3	3	3	4	6	9		
4	2	2	3	3	4	4	5	7	10	
5	2	2	3	3	3	3	4	4	5	6
6	2	2	3	3	3	3	3	4	4	5
7	2	2	3	3	3	3	3	3	4	4
8	2	2	3	3	3	3	3	3	3	4
9	2	2	3	3	3	3	3	3	3	3

L=16

P \ Q	0	1	2	3	4	5	6	7	8	9
0	2	5								
1	2	2	3	4	8					
2	2	2	2	3	4	5	9			
3	2	2	2	3	3	4	4	6	9	
4	2	2	2	3	3	3	4	4	5	6
5	2	2	2	3	3	3	3	4	4	5
6	2	2	2	2	3	3	3	3	4	4
7	2	2	2	2	3	3	3	3	3	4
8	2	2	2	2	3	3	3	3	3	3
9	2	2	2	2	3	3	3	3	3	3

FIG. 1 Determination of Number of Samples Required (see 6.4.3)

6.5.2 The text of the test method shall be distributed to all the laboratories in time to raise any queries before the tests begin. If any laboratory wants to practice the test method in advance, this shall be done with samples other than those used in the program.

6.5.3 The samples shall be accumulated, subdivided, and distributed by the organizer, who shall also keep a reserve of each sample for emergencies. It is most important that the individual laboratory portions be homogeneous. Instructions to each laboratory shall include the following:

6.5.3.1 *Testing Protocol*—The protocol to be used for testing of the ILS sample set shall be provided. Factors that may affect test method outcome but are not intended to be controlled in the normal execution of the test method shall not be intentionally removed nor controlled in the testing of the ILS samples, unless explicitly permitted by the sponsoring subcommittee of the ILS for special studies where certain factors are controlled intentionally as part of the testing protocol to meet the intended ILS study objectives. To remove, control, or set limits on factors that are not intended to be controlled in the normal execution of the test method in the conduct of an ILS that is intended for the precision evaluation of the test method executed under normal operating conditions will result in overly optimistic precision. Precision statements thus generated will likely be unattainable by majority of users in the normal execution of the test method.

6.5.3.2 The agreed draft method of test;

6.5.3.3 Material Safety Data Sheets, where applicable, and the handling and storage requirements for the samples;

6.5.3.4 The order in which the samples are to be tested (a different random order for each laboratory);

6.5.3.5 The statement that two test results are to be obtained in the shortest practical period of time on each sample by the same operator with the same apparatus. For statistical reasons it is imperative that the two results are obtained independently of each other, that is, that the second result is not biased by knowledge of the first. If this is regarded as impossible to achieve with the operator concerned, then the pairs of results shall be obtained in a blind fashion, but ensuring that they are carried out in a short period of time (preferably the same day). The term *blind fashion* means that the operator does not know that the sample is a replicate of any previous run.

6.5.3.6 The period of time during which repeated results are to be obtained and the period of time during which all the samples are to be tested;

6.5.3.7 A blank form for reporting the results. For each sample, there shall be space for the date of testing, the two results, and any unusual occurrences. The unit of accuracy for reporting the results shall be specified. This should be, if possible, more digits reported than will be used in the final test method, in order to avoid having rounding unduly affect the estimated precision values.

6.5.3.8 When it is required to estimate the determinability, the report form must include space for each of the determined values as well as the test results.

6.5.3.9 A statement that the test shall be carried out under normal conditions, using operators with good experience but not exceptional knowledge; and that the duration of the test shall be the same as normal.

6.5.4 The pilot program operators may take part in the interlaboratory program. If their extra experience in testing a few more samples produces a noticeable effect, it will serve as a warning that the test method is not satisfactory. They shall be identified in the report of the results so that any such effect may be noted.

6.5.5 It can not be overemphasized that the statement of precision in the test method is to apply to test results obtained by running the agreed procedure exactly as written. Therefore, the test method must not be significantly altered after its precision statement is written.

7. Inspection of Interlaboratory Results for Uniformity and for Outliers

7.1 *Introduction:*

7.1.1 This section specifies procedures for examining the results reported in a statistically designed interlaboratory program (see Section 6) to establish:

7.1.1.1 The independence or dependence of precision and the level of results;

7.1.1.2 The uniformity of precision from laboratory to laboratory, and to detect the presence of outliers.

NOTE 6—The procedures are described in mathematical terms based on the notation of Annex A1 and illustrated with reference to the example data (calculation of bromine number) set out in Annex A2. Throughout this section (and Section 8), the procedures to be used are first specified and then illustrated by a worked example using data given in Annex A2.

NOTE 7—It is assumed throughout this section that all the deviations are either from a single normal distribution or capable of being transformed into such a distribution (see 7.2). Other cases (which are rare) would require different treatment that is beyond the scope of this practice. Also, see **(2)** for a statistical test of normality.

7.2 *Transformation of Data:*

7.2.1 In many test methods the precision depends on the level of the test result, and thus the variability of the reported results is different from sample to sample. The method of analysis outlined in this practice requires that this shall not be so and the position is rectified, if necessary, by a transformation.

7.2.1.1 Prior to commencement of analysis to determine if transformation is necessary, it is a good practice to examine information gathered from ILS participants to determine compliance with agreed upon ILS protocol and method of test. As part of this examination, the raw data as reported should be inspected for existence of extreme or outlandish values that are visually obvious. Exclusion of extreme or outlandish results from transformation analysis is recommended if assignable causes can be found in order to help ensure test data dependability, transformation reliability, and subsequent computation efficiency. If assignable causes cannot be found, exclusion of extreme or outlandish results from transformation analysis should be confirmed for each sample using a formal statistical test such as the General Extreme Studentized Deviation (GESD) multi-outlier technique (see Practice D7915) or other technically equivalent techniques at the 99 % confidence level on the difference and average (or sum) of the two replicate results as submitted by each ILS participant for each sample as follows:

(1) Compute the difference of the two replicates submitted by each participant for the sample;

(2) Perform GESD on the differences from all participants for the sample;

(3) For each difference that is identified as outlier, reject the result that is farthest from the median of all results for that sample;

(4) Compute the average (or sum) of the two replicates for each participant for the sample; for participants who submitted only a single result, or, if one of the submitted replicates is rejected in (*3*), use the remaining result as the average (or 2 × the remaining result as sum) for the participant;

(5) Perform GESD on the averages (or sums) from all participants for the sample;

(6) Reject all results identified as outliers in (*5*); and

(7) Continue execution of the remainder of this practice using the retained results.

It is recommended that such statistical tests be conducted under the guidance of a statistician.

7.2.2 The laboratories' standard deviations D_j, and the repeats standard deviations d_j (see Annex A1) are calculated and plotted separately against the sample means m_j . If the points so plotted may be considered as lying about a pair of lines parallel to the m-axis, then no transformation is necessary. If, however, the plotted points describe non-horizontal straight lines or curves of the form $D = f_1(m)$ and $d = f_2(m)$, then a transformation will be necessary.

7.2.3 The relationships $D = f_1(m)$ and $d = f_2(m)$ will not in general be identical. It is frequently the case, however, that the ratios $u_j = \frac{d_j}{D_j}$ are approximately the same for all m_j, in which case f_1 is approximately proportional to f_2 and a single transformation will be adequate for both repeatability and reproducibility. The statistical procedures of this practice are greatly facilitated when a single transformation can be used. For this reason, unless the u_j clearly vary with property level, the two relationships are combined into a single dependency relationship $D = f(m)$ (where D now includes d) by including a dummy variable T. This will take account of the difference between the relationships, if one exists, and will provide a means of testing for this difference (see A4.1).

7.2.4 In the event that the rations u_j do vary with level (mean, m_j), as confirmed with a regression of u_j on m_j, or $\log(u_j)$ on $\log(m_j)$, follow the instructions in Annex A5. Otherwise, continue with 7.2.5.

7.2.5 The single relationship $D = f(m)$ is best estimated by weighted linear regression analysis. Strictly speaking, an iteratively weighted regression should be used, but in most cases even an unweighted regression will give a satisfactory approximation. The derivation of weights is described in A4.2, and the computational procedure for the regression analysis is described in A4.3. Typical forms of dependence $D = f(m)$ are given in A3.1. These are all expressed in terms of at most two (2) transformation parameters, B and B_0.

7.2.6 The typical forms of dependence, the transformations they give rise to, and the regressions to be performed in order to estimate the transformation parameters B, are all summarized in A3.2. This includes statistical tests for the significance of the regression (that is, is the relationship $D = f(m)$ parallel to the m-axis), and for the difference between the repeatability and reproducibility relationships, based at the 5 % significance level. If such a difference is found to exist, follow the procedures in Annex A5.

7.2.7 If it has been shown at the 5 % significance level that there is a significant regression of the form $D = f(m)$, then the appropriate transformation $y = F(x)$, where x is the reported result, is given by the equation

$$F(x) = K\int \frac{dx}{f(x)} \tag{3}$$

where K = a constant. In that event, all results shall be transformed accordingly and the remainder of the analysis carried out in terms of the transformed results. Typical transformations are given in A3.1.

7.2.8 The choice of transformation is difficult to make the subject of formalized rules. Qualified statistical assistance may be required in particular cases. The presence of outliers may affect judgement as to the type of transformation required, if any (see 7.7).

7.2.9 *Worked Example:*

7.2.9.1 Table 3 lists the values of *m, D*, and *d* for the eight samples in the example given in Annex A2, correct to three significant digits. Corresponding degrees of freedom are in parentheses. Inspection of the values in Table 3 shows that both *D* and *d* increase with *m*, the rate of increase diminishing as *m* increases. A plot of these figures on log-log paper (that is, a graph of log *D* and log *d* against log *m*) shows that the points may reasonably be considered as lying about two straight lines (see Fig. A4.1 in Annex A4). From the example calculations given in A4.4, the gradients of these lines are shown to be the same, with an estimated value of 0.638. Bearing in mind the errors in this estimated value, the gradient may for convenience be taken as 2/3.

$$\int x^{-\frac{2}{3}}\,dx = 3x^{\frac{1}{3}} \tag{4}$$

7.2.9.2 Hence, the same transformation is appropriate both for repeatability and reproducibility, and is given by the equation. Since the constant multiplier may be ignored, the transformation thus reduces to that of taking the cube roots of the reported bromine numbers. This yields the transformed data shown in Table A1.3, in which the cube roots are quoted correct to three decimal places.

7.3 *Tests for Outliers:*

7.3.1 The reported data or, if it has been decided that a transformation is necessary, the transformed results shall be inspected for outliers. These are the values which are so different from the remainder that it can only be concluded that they have arisen from some fault in the application of the test method or from testing a wrong sample. Many possible tests may be used and the associated significance levels varied, but those that are specified in the following subsections have been found to be appropriate in this practice. These outlier tests all assume a normal distribution of errors.

7.3.1.1 The total percentage of outliers rejected, as defined by 100× (no. of rejected results/no. of reported results), shall be reported explicitly to the ILS Program Manager for approval by the sponsoring subcommittee and main committee.

7.3.2 *Uniformity of Repeatability*—The first outlier test is concerned with detecting a discordant result in a pair of repeat results. This test **(3)** involves calculating the $e_{ij}^{\ 2}$ over all the laboratory/sample combinations. Cochran's criterion at the 1 % significance level is then used to test the ratio of the largest of these values over their sum (see A1.5). If its value exceeds the value given in Table A2.2, corresponding to one degree of freedom, *n* being the number of pairs available for comparison, then the member of the pair farthest from the sample mean shall be rejected and the process repeated, reducing *n* by 1, until no more rejections are called for. In certain cases, specifically when the number of digits used in reporting results leads to a large number of repeat ties, this test can lead to large proportion of rejections. If this is so, consideration should be given to cease this rejection test and retain some or all of the rejected results. A decision based on judgement in consultation with a statistician will be necessary in this case.

7.3.3 *Worked Example*—In the case of the example given in Annex A2, the absolute differences (ranges) between transformed repeat results, that is, of the pairs of numbers in Table A1.3, in units of the third decimal place, are shown in Table 4. The largest range is 0.078 for Laboratory G on Sample 3. The sum of squares of all the ranges is

TABLE 3 Computed from Bromine Example Showing Dependence of Precision on Level

Sample Number	3	8	1	4	5	6	2	7
m	0.756	1.22	2.15	3.64	10.9	48.2	65.4	114
D	0.0669 (14)	0.159 (9)	0.729 (8)	0.211 (11)	0.291 (9)	1.50 (9)	2.22 (9)	2.93 (9)
d	0.0500 (9)	0.0572 (9)	0.127 (9)	0.116 (9)	0.0943 (9)	0.527 (9)	0.818 (9)	0.935 (9)

TABLE 4 Absolute Differences Between Transformed Repeat Results: Bromine Example

Laboratory	Sample							
	1	2	3	4	5	6	7	8
A	42	21	7	13	7	10	8	0
B	23	12	12	0	7	9	3	0
C	0	6	0	0	7	8	4	0
D	14	6	0	13	0	8	9	32
E	65	4	0	0	14	5	7	28
F	23	20	34	29	20	30	43	0
G	62	4	78	0	0	16	18	56
H	44	20	29	44	0	27	4	32
J	0	59	0	40	0	30	26	0

$$0.042^2 + 0.021^2 + \ldots + 0.026^2 + 0^2 = 0.0439.$$

Thus, the ratio to be compared with Cochran's criterion is

$$\frac{0.078^2}{0.0439} = 0.138 \tag{5}$$

where 0.138 is the result obtained by electronic calculation of unrounded factors in the expression. There are 72 ranges and as, from Table A2.2, the criterion for 80 ranges is 0.1709, this ratio is not significant.

7.3.4 *Uniformity of Reproducibility:*

7.3.4.1 The following outlier tests are concerned with establishing uniformity in the reproducibility estimate, and are designed to detect either a discordant pair of results from a laboratory on a particular sample or a discordant set of results from a laboratory on all samples. For both purposes, the Hawkins' test **(4)** is appropriate.

7.3.4.2 This involves forming for each sample, and finally for the overall laboratory averages (see 7.6), the ratio of the largest absolute deviation of laboratory mean from sample (or overall) mean to the square root of certain sums of squares (A1.6).

7.3.4.3 The ratio corresponding to the largest absolute deviation shall be compared with the critical 1 % values given in Table A1.5, where n is the number of laboratory/sample cells in the sample (or the number of overall laboratory means) concerned and where v is the degrees of freedom for the sum of squares which is additional to that corresponding to the sample in question. In the test for laboratory/sample cells v will refer to other samples, but will be zero in the test for overall laboratory averages.

7.3.4.4 If a significant value is encountered for individual samples the corresponding extreme values shall be omitted and the process repeated. If any extreme values are found in the laboratory totals, then all the results from that laboratory shall be rejected.

7.3.4.5 If the test leads to large proportion of rejections, consideration should be given to cease this rejection test and retain some or all of the rejected results. A decision based on judgement in consultation with a statistician will be necessary in this case.

7.3.5 *Worked Example:*

7.3.5.1 The application of Hawkins' test to cell means within samples is shown below.

7.3.5.2 The first step is to calculate the deviations of cell means from respective sample means over the whole array. These are shown in Table 5, in units of the third decimal place. The sum of squares of the deviations are then calculated for each sample. These are also shown in Table 5 in units of the third decimal place.

7.3.5.3 The cell to be tested is the one with the most extreme deviation. This was obtained by Laboratory D from Sample 1. The appropriate Hawkins' test ratio is therefore:

$$B^* = \frac{0.314}{\sqrt{0.117+0.015+\ldots+0.017}} = 0.7281 \tag{6}$$

7.3.5.4 The critical value, corresponding to n = 9 cells in sample 1 and v = 56 extra degrees of freedom from the other samples is interpolated from Table A1.5 as 0.3729. The test value is greater than the critical value, and so the results from Laboratory D on Sample 1 are rejected.

TABLE 5 Deviations of Cell Means from Respective Sample Means: Transformed Bromine Example

	Sample							
Laboratory	1	2	3	4	5	6	7	8
A	20	8	14	15	10	48	6	3
B	75	7	20	9	10	47	6	3
C	64	35	3	20	30	4	22	25
D	314	33	18	42	7	39	80	50
E	32	32	30	9	7	18	18	39
F	75	97	31	20	30	8	74	53
G	10	34	32	20	20	61	9	62
H	42	13	4	42	13	21	8	50
J	1	28	22	29	14	8	10	53
Sum of Squares	117	15	4	6	3	11	13	17

7.3.5.5 As there has been a rejection, the mean value, deviations, and sum of squares are recalculated for Sample 1, and the procedure is repeated. The next cell to be tested will be that obtained by Laboratory F from Sample 2. The Hawkins' test ratio for this cell is:

$$B^* = \frac{0.097}{\sqrt{0.006+0.015+\ldots+0.017}} = 0.3542 \tag{7}$$

7.3.5.6 The critical value corresponding to $n = 9$ cells in Sample 2 and $v = 55$ extra degrees of freedom is interpolated from Table A1.5 as 0.3756. As the test ratio is less than the critical value there will be no further rejections.

7.4 *Rejection of Complete Data from a Sample:*

7.4.1 The laboratories standard deviation and repeats standard deviation shall be examined for any outlying samples. If a transformation has been carried out or any rejection made, new standard deviations shall be calculated.

7.4.2 If the standard deviation for any sample is excessively large, it shall be examined with a view to rejecting the results from that sample.

7.4.3 Cochran's criterion at the 1 % level can be used when the standard deviations are based on the same number of degrees of freedom. This involves calculating the ratio of the largest of the corresponding sums of squares (laboratories or repeats, as appropriate) to their total (see A1.5). If the ratio exceeds the critical value given in Table A2.2, with n as the number of samples and v the degrees of freedom, then all the results from the sample in question shall be rejected. In such an event, care should be taken that the extreme standard deviation is not due to the application of an inappropriate transformation (see 7.1), or undetected outliers.

7.4.4 There is no optimal test when standard deviations are based on different degrees of freedom. However, the ratio of the largest variance to that pooled from the remaining samples follows an F-distribution with v_1 and v_2 degrees of freedom (see A1.7). Here v_1 is the degrees of freedom of the variance in question and v_2 is the degrees of freedom from the remaining samples. If the ratio is greater than the critical value given in A2.6, corresponding to a significance level of $0.01/S$ where S is the number of samples, then results from the sample in question shall be rejected.

7.4.5 *Worked Example:*

7.4.5.1 The standard deviations of the transformed results, after the rejection of the pair of results by Laboratory D on Sample 1, are given in Table 6 in ascending order of sample mean, correct to three significant digits. Corresponding degrees of freedom are in parentheses.

TABLE 6 Standard Deviations of Transformed Results: Bromine Example

Sample number	3	8	1	4	5	6	2	7
m	0.9100	1.066	1.240	1.538	2.217	3.639	4.028	4.851
D	0.0278	0.0473	0.0354	0.0297	0.0197	0.0378	0.0450	0.0416
	(14)	(9)	(13)	(11)	(9)	(9)	(9)	(9)
d	0.0214	0.0182	0.028	0.0164	0.0063	0.0132	0.0166	0.0130
	(9)	(9)	(8)	(9)	(9)	(9)	(9)	(9)

7.4.5.2 Inspection shows that there is no outlying sample among these. It will be noted that the standard deviations are now independent of the sample means, which was the purpose of transforming the results.

7.4.5.3 The values in Table 7, taken from a test program on bromine numbers over 100, will illustrate the case of a sample rejection.

7.4.5.4 It is clear, by inspection, that the laboratories standard deviation of Sample 93 at 15.76 is far greater than the others. It is noted that the repeats standard deviation in this sample is correspondingly large.

7.4.5.5 Since laboratory degrees of freedom are not the same over all samples, the variance ratio test is used. The variance pooled from all samples, excluding Sample 93, is the sum of the sums of squares divided by the total degrees of freedom, that is

$$\frac{(8\times5.10^2+9\times4.20^2+\ldots+8\times3.85^2)}{(8+9+\ldots+8)}=19.96 \tag{8}$$

7.4.5.6 The variance ratio is then calculated as

$$\frac{15.26^2}{19.96}=11.66 \tag{9}$$

where 11.66 is the result obtained by electronic calculation without rounding the factors in the expression.

7.4.5.7 From Table A1.8 the critical value corresponding to a significance level of 0.01/8 = 0.00125, on 8 and 63 degrees of freedom, is approximately 4. The test ratio greatly exceeds this and results from Sample 93 shall therefore be rejected.

7.4.5.8 Turning to repeats standard deviations, it is noted that degrees of freedom are identical for each sample and that Cochran's test can therefore be applied. Cochran's criterion will be the ratio of the largest sum of squares (Sample 93) to the sum of all the sums of squares, that is

$$2.97^2/(1.13^2+0.99^2+\ldots+1.36^2)=0.510 \tag{10}$$

This is greater than the critical value of 0.352 corresponding to n = 8 and v = 8 (see Table A2.2), and confirms that results from Sample 93 shall be rejected.

7.5 *Estimating Missing or Rejected Values:*

7.5.1 *One of the Two Repeat Values Missing or Rejected*—If one of a pair of repeats (Y_{ij1} or Y_{ij2}) is missing or rejected, this shall be considered to have the same value as the other repeat in accordance with the least squares method.

7.5.2 *Both Repeat Values Missing or Rejected:*

7.5.2.1 If both the repeat values are missing, estimates of a_{ij} (= Y_{ij1} + Y_{ij2}) shall be made by forming the laboratories × samples interaction sum of squares (see Eq 18), including the missing values of the totals of the laboratories/samples pairs of results as unknown variables. Any laboratory or sample from which all the results were rejected shall be ignored and new values of L and S used. The estimates of the missing or rejected values shall be those that minimize the interaction sum of squares.

7.5.2.2 If the value of single pair sum a_{ij} has to be estimated, the estimate is given by the equation:

$$a_{ij}=\frac{1}{(L-1)(S'-1)}(LL_1+S'S_1-T_1) \tag{11}$$

where:

L_1 = total of remaining pairs in the ith laboratory,

TABLE 7 Example Statistics Indicating Need to Reject an Entire Sample

Sample number	90	89	93	92	91	94	95	96
m	96.1	99.8	119.3	125.4	126.0	139.9	139.4	159.5
D	5.10	4.20	15.26	4.40	4.09	4.87	4.74	3.85
	(8)	(9)	(8)	(11)	(10)	(8)	(9)	(8)
d	1.13	0.99	2.97	0.91	0.73	1.32	1.12	1.36
	(8)	(8)	(8)	(8)	(8)	(8)	(8)	(8)