Designation: D 4855 – 97

Standard Practice for Comparing Test Methods[1]

This standard is issued under the fixed designation D 4855; 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 provides a procedure for evaluating and comparing test methods under controlled conditions using the same materials tested during the same time span. The practice describes how to obtain and compare estimates on precision, sensitivity, and bias.

1.2 This practice covers the following topics:

2. Referenced Documents

2.1 *ASTM Standards:*

D 123 Terminology Relating to Textiles[2]

D 2905 Practice for Statements on Number of Specimens for Textiles[2]

D 2906 Practice for Statements on Precision and Bias for Textiles[2]

E 456 Terminology Relating to Quality and Statistics[3]

2.2 ASTM Adjuncts:

TEX-PAC[4]

NOTE 1—Tex-Pac is a group of PC programs on floppy disks, available through ASTM Headquarters, 100 Barr Harbor Drive, Conshohocken, PA 19428, USA. The calculations for comparing the precision, sensitivity and bias of two test methods can be done using one of these programs and statements on the relative merits of the two test methods are part of the output.

3. Terminology

3.1 *Definitions:*

3.1.1 *accuracy, n—of a test method*, the degree of agreement between the true value of the property being tested (or an accepted standard value) and the average of many observations made according to the test method, preferably by many observers. (See also *bias* and *precision*.)

3.1.1.1 *Discussion*—Increased accuracy is associated with decreased bias relative to the true value; two methods with equal bias relative to the true value have equal accuracy even if one method is more precise than the other. The true value is the exact value of the property being tested for the statistical universe being sampled. When the true value is not known or cannot be determined, and an acceptable standard value is not available, accuracy cannot be established. No valid inferences on the accuracy of a method can be drawn from an individual observation.

3.1.2 *bias, n—in statistics*, a constant or systematic error in test results.

3.1.2.1 *Discussion*—Bias can exist between the accepted reference value and a test result obtained from one method, between test results obtained from two methods, or between two test results obtained from a single method, for example, between operators or between laboratories.

3.1.3 *confidence interval, n*—the interval estimate of a population parameter computed so that the statement "the population parameter lies in this interval" will be true, on the average, in a stated proportion of the times such statements are made.

3.1.4 *confidence level, n*—the stated proportion of times the confidence interval is expected to include the population parameter.

3.1.4.1 *Discussion*—Statisticians generally accept that, in the absence of special consideration, 0.95 or 95 % is a realistic confidence level. If the consequences of incorrectly estimating the confidence interval would be grave, then a higher confidence level might be considered. If the consequences of incorrectly estimating the confidence interval are of less than usual concern, then a lower confidence interval might be considered.

3.1.5 *confidence limits, n*—the two statistics that define the ends of a confidence interval.

3.1.6 *degrees of freedom, n—for a set*, the number of values that can be assigned arbitrarily and still get the same value for

[1] This practice is under the jurisdiction of ASTM Committee D-13 on Textiles and is the direct responsibility of Subcommittee D13.93 on Statistics.

Current edition approved September 10, 1997. Published August 1998. Originally published as D 4855 – 88. Last previous edition D 4855 – 91.

[2] *Annual Book of ASTM Standards*, Vol 07.01.

[3] *Annual Book of ASTM Standards*, Vol 14.02.

[4] PC programs on floppy disks are available through ASTM. For a 3½ inch disk request PCN:12-429040-18, for a 5¼ inch disk request PCN:12-429041-18.

each of one or more statistics calculated from the set of data.

3.1.6.1 *Discussion*—For example, if only an average is specified for a set of five observations, there are four degrees of freedom since the same average can be obtained with any values substituted for four of the five observations as long as the fifth value is set to give the correct total. If both the average and the standard deviation have been specified, there are only three degrees of freedom left.

3.1.7 *error of the first kind, α, n—in a statistical test,* the rejection of a statistical hypothesis when it is true. (*Syn.* Type I error.)

3.1.8 *error of the second kind, β, n—in a statistical test,* the acceptance of a statistical hypothesis when it is false. (*Syn.* Type II error.)

3.1.9 *F-test, n*—a test of statistical significance based on the use of George W. Snedecor's F-distribution and used to compare two sample variances or a sample variance and a hypothetical value.

3.1.10 *interference, n—in testing,* an effect due to the presence of a constituent or characteristic that influences the measurement of another constituent or characteristic.

3.1.11 *least difference of practical importance, δ, n*—the smallest difference based on engineering judgment deemed to be of practical importance when considering whether a significant difference exists between two statistics or between a statistic and a hypothetical value.

3.1.12 *parameter, n—in statistics,* a variable that describes a characteristic of a population or mathematical model.

3.1.13 *precision, n*—the degree of agreement within a set of observations or test results obtained as directed in a method.

3.1.13.1 *Discussion*—The term "precision," delimited in various ways, is used to describe different aspects of precision. This usage was chosen in preference to the use of "repeatability" and "reproducibility," which have been assigned conflicting meanings by various authors and standardizing bodies.

3.1.14 *ruggedness test, n*—an experiment in which environmental or test conditions are deliberately varied in order to evaluate the effects of such variations.

3.1.15 *sensitivity, n—for a single test method,* the result of dividing (1) the derivative of measurements at different levels of a property of interest to known values of the property by (2) the standard deviation of such measurements. (Syn. *absolute sensitivity*.)

3.1.15.1 *Discussion*—The sensitivity of a single test method may be determined only with materials for which the values of the property of interest is known.

3.1.16 *sensitivity ratio, SR, n—in comparing two test methods,* the ratio of the sensitivities of the test methods with the larger sensitivity in the numerator. (Syn. *relative sensitivity*.)

3.1.16.1 *Discussion*—When the same materials are used for each test method, the sensitivity ratio may be determined using materials for which the value of the property of interest is not known.

3.1.17 *statistic, n*—a quantity that is calculated from observations on a sample and that estimates a parameter of a sample and that estimates a parameter of a population.

3.1.18 *t-test, n*—a test of statistical significance based on the use of Student's *t*-distribution and used to compare two sample averages or a sample average and a hypothetical value.

3.1.19 *Type I error*—See *error of the first kind.*

3.1.20 *Type II error*—See *error of the second kind.*

3.1.21 For definitions of textile terms used in this standard, refer to Terminology D 123. For definitions of other statistical terms used in this standard, refer to Terminology D 4392 or Terminology E 456.

4. Significance and Use

4.1 Task groups developing a test method frequently find themselves with two or more alternative procedures that must be compared. Three common situations are:

4.1.1 Two or more new test methods may have been proposed to measure a property for which there is no existing method.

4.1.2 A new test method may have been suggested to replace an existing test method.

4.1.3 Two or more existing test methods may overlap in their scopes so that one should be chosen over the other.

4.2 The selection of one test method in preference to another is not simply a statistical choice. There are many other aspects of two test methods that should be considered, which may have an influence (on the engineering judgment of the task group) equal to or greater than the statistical evidence. Some of these characteristics are discussed in Section 6.

5. Requirements for Materials

5.1 The number and type of materials to be included in a comparison study will depend on the following:

5.1.1 The range of the values of the property being measured on a given material and how the precision varies over that range,

5.1.2 The number of different materials to which the test method is applied,

5.1.3 The difficulty and expense involved in obtaining, processing, and distributing samples,

5.1.4 The difficulty of, length of time required for, and expense of performing the tests, and

5.1.5 The uncertainty of prior information on any of these points. For example, if it is already known that the precision is relatively constant or proportional to the average level over the range of values of interest, a smaller number of materials will be needed than if it is known that the precision changes erratically at different levels. A preliminary pilot or screening program may help to settle some of these questions, and may often result in the saving of considerable time and expense in the full comparison study.

5.2 In general, a minimum of three materials should be considered acceptable, and for development of broadly applicable precision statements, six or more materials should be included in the study.

5.3 Whenever feasible, the material representing any given level in a comparison study should be made as homogeneous as possible prior to its subdivision into portions or specimens that are allocated to the different methods.

5.4 For each level of material, an adequate quantity (sample) of reasonably homogeneous material should be available for subdivision for each test method. This supply of

material should include a reserve of 50 % beyond the requirements of the protocol for the comparison study for possible later use in checking results or retesting the test methods in one or more laboratories.

6. Evaluating Test Methods

6.1 *Each Proposed New Test Method*—When evaluating one or more test methods, take into account the following features that are desirable in a proposed test method:

6.1.1 The relationship between the test results and the property of interest is clearly understood.

6.1.2 There is a small or non-existent bias over a wide range of test results.

6.1.3 The test method is precise enough to satisfy the requirements of the application.

6.1.4 The test method has acceptable ruggedness and sensitivity.

6.1.5 Any potential interferences are known and small enough to tolerate.

6.1.6 There is a low cost for making an observation with short times for learning to run the test, getting ready to run the test and cleaning up after running the test.

6.1.7 The test method may have other special attributes that encourage its selection as a preferred method.

6.1.8 Data are available from the advocates of the test method to support the above claims.

6.2 *Two or More New Test Methods*—When two or more new test methods are being evaluated, the task group should also consider the possibility that:

6.2.1 One test method may be more suitable for one range of values and another for a second range of values.

6.2.2 One method may be better suited as a referee method while the other is better for routine testing.

6.3 *New Versus Existing Test Method*—When looking for a new test method the task group wants improved precision, improved sensitivity, a shorter elapsed time to get test results, or a reduced cost without unduly disturbing any other characteristics of the test method.

7. Sensitivity Criterion

7.1 Sometimes a test method that is more precise than another test method has less discriminating power from the standpoint of detecting changes in the level in the property of interest. The sensitivity criterion is a quantitative measure of the relative merit of two test methods which:

7.1.1 Combines the precision of each method with the ability of the test method to measure differences in the property of interest.

7.1.2 Permits the comparison of test methods for which test results are reported in different units of measure. For this reason, comparisons of the sensitivity of two methods may be more meaningful than comparisons of their precisions.

7.2 When comparing test methods on the basis of data collected, it is important that the task group has formulated and evaluated a plan for analysis of the data so as to arrive at a correct decision, before conducting any tests. Statistical tests of significance are recommended as a means of helping make the decisions for these reasons: they are objective, they require a clear statement of the problem, they make more efficient use of the observed data than subjective techniques, and they allow control of the probability of concluding two test methods are different when they are really alike, as well as the probability of concluding two test methods are alike when they are really different.

8. Basic Statistical Design

8.1 Decide whether the precision, the sensitivity, the accuracy, or the bias of the two test methods is to be compared.

8.2 Specify the values of probability of Type I error, α, probability of Type II error, β, and the least difference of practical importance, δ, to be used in determining the number of observations required for each level and method (see Fig. 1).

8.3 It is common practice to arbitrarily set $\alpha = 0.05$ and $\beta = 0.10$. The use of an α error of 0.05 is a compromise between the increased cost of experimenting when α is smaller and the greater risk of falsely stating that two equivalent methods are different that exists when α is larger. The β error of 0.10 takes into account the fact that the risk of failing to detect a true difference between two methods becomes rapidly smaller when the actual difference exceeds δ. If the experimenter believes that risks should be revised because the consequences of error are unusually grave and because the values of $\alpha = 0.05$ or $\beta = 0.10$ lead to high cost of evaluation, qualified statistical assistance is recommended.

		Our Decision	
		Methods are Equivalent	Methods are Different
True Situation	Methods are Equivalent	Decision is Correct	Alpha
	Methods are Different by Delta Units	Beta	Decision is Correct

FIG. 1 Schematic of Decision Procedure

TABLE 1 Comparing Methods for Precision—Two-Sided Test[A]

NOTE 1—See Appendix X1 for the basis for this table.

Difference, %[B]	$S_{Larger}/S_{Smaller}$	Observations per Cell, *r*			
		1 Level of Material	2 Levels of Material	3 Levels of Material	4 Levels of Material
30	1.30	155	78	53	40
40	1.40	95	48	33	25
60	1.60	50	26	18	14
80	1.80	33	17	12	9
100	2.00	24	13	9	7
120	2.20	19	10	7	6
140	2.40	16	9	6	5
160	2.60	14	8	6	5
180	2.80	12	7	5	4
200	3.00	11	6	5	4
225	3.25	10	6	4	4
250	3.50	9	5	4	3
275	3.75	8	5	4	3
300	4.00	8	5	4	3

[A] $\alpha = 0.05$; $\beta = 0.10$.
[B] The minimum experiment should include at least the number of observations shown for a 100 % difference. Differences of 120 % or more require so few observations that internal estimates of precision will be too variable. Observations per cell for differences of 120 % or more are shown only to illustrate the large differences that may be overlooked with smaller than recommended experiments. For example an experiment that will probably detect a difference of 100 % in the size of the population variances when comparing four levels of materials requires seven observations per material at the specified values of α and β.

TABLE 2 Comparing Methods for Average Level—Two-Sided Tests[A]

NOTE 1—See Appendix X2 for the basis for this table.

E[B]	Observations per Cell, *r*
0.5	86
0.6	60
0.7	44
0.8	34
0.9	27
1.0	23
1.1	19
1.2	16
1.3	14
1.4	12
1.5	11
1.6	10
1.7	9
1.8	8
1.9	7
2.0	7

[A] $\alpha = 0.05$; $\beta = 0.10$.
[B] *E* is calculated using Eq 1.

8.4 Choose the appropriate test statistic. This will be a *t*-test or an *F*-test. If there is doubt as to the correct test statistic, get qualified statistical help.

8.5 Utilize the preselected levels of α, β, and δ, as inputs to Tables 1 and 2, to estimate the required size of the experiment as directed in 9.3.

8.6 Plan and conduct an experiment which compares the methods across the range of conditions which are of interest.

8.7 Analyze the data and calculate the test statistic in 8.4. Compare the calculated test statistic with a critical value found in an appropriate table of *t*-values or *F*-values.[5,6] Based on this comparison, decide whether the methods differ significantly.

9. Experimental Procedure

9.1 This basic experimental procedure is designed so that it has enough flexibility that it can be utilized to compare methods on the basis of precision, sensitivity, accuracy, and bias.

9.2 The layout of the basic procedure, as shown in Fig. 2, requires *r* test observations be obtained by each method on two levels of material.

9.2.1 This experimental procedure requires a series of specimens being tested for the low level of the property and a series of specimens for the high level of the property, with the full range of interest for the property being covered, when possible. See Practice D 2905 for determination for number of specimens.

9.2.2 Test the specimens over a period of three to four weeks, or until *r* test observations have been obtained for each level.

9.3 Determine the size of the basic procedure by:

9.3.1 Choosing the smallest difference in variability that is of practical importance to detect.

9.3.2 Expressing the difference as a percent increase in the measure of variability of the more variable method as compared to the less variable method. For example, selecting a minimum practical difference of 60 % means that we are only interested in detecting a measure of variability in one method that is larger than the comparable measure of variability of the other method by 60 % or more.

9.3.3 Choosing the smallest difference in average of the property being tested for which the detection is of practical importance.

9.3.3.1 Expressing this difference by using Eq 1:

$$E = \delta/s_p \quad (1)$$

where:

δ = the smallest difference of practical importance expressed in units of measure,
E = the smallest difference of practical importance as a multiple of the standard deviation, and
s_p = the best available estimate of the average standard deviation for individual observations for the two test methods.

9.3.4 Estimating *r*, the required number of observations for each combination of methods and levels, using both Tables 1 and 2.

9.3.5 Using the larger of the estimates of *r* obtained in 9.3.4 as the number of observations for each combination of methods and levels to be tested.

10. Procedure for Comparing Precision

10.1 When comparing the precision of the two test methods, plan the experimental procedure as directed in Section 9. See Practice D 2906.

10.2 Calculate the average, $\bar{X}_{ij}$ and standard deviation, s_{ij} for each level and each method tested by using Eq 2 and 3:

$$\bar{X}_{ij} = \frac{\Sigma X_{ij}}{r_{ij}} \quad (2)$$

[5] Davies, O. L., *The Design and Analysis of Industrial Experiments*, Hafner Publishing Company, 1954, Table H, p. 614 and pp. 609–610.
[6] Dixon, W. J., and Masey, F. J., Jr., *Introduction to Statistical Analysis*, 4th Ed., McGraw-Hill Book Company, 1983.