Designation: E177 – 14

An American National Standard

Standard Practice for Use of the Terms Precision and Bias in ASTM Test Methods[1]

This standard is issued under the fixed designation E177; 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.

This standard has been approved for use by agencies of the U.S. Department of Defense.

1. Scope

1.1 The purpose of this practice is to present concepts necessary to the understanding of the terms "precision" and "bias" as used in quantitative test methods. This practice also describes methods of expressing precision and bias and, in a final section, gives examples of how statements on precision and bias may be written for ASTM test methods.

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

2. Referenced Documents

2.1 *ASTM Standards:*[2]

E456 Terminology Relating to Quality and Statistics

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

E1488 Guide for Statistical Procedures to Use in Developing and Applying Test Methods

E2282 Guide for Defining the Test Result of a Test Method

E2586 Practice for Calculating and Using Basic Statistics

E2587 Practice for Use of Control Charts in Statistical Process Control

3. Terminology

3.1 *Definitions*—Terminology E456 provides a more extensive list of terms in E11 standards.

3.1.1 *accepted reference value, n*—a value that serves as an agreed-upon reference for comparison, and which is derived as: *(1)* a theoretical or established value, based on scientific principles, *(2)* an assigned or certified value, based on experimental work of some national or international organization, or *(3)* a consensus or certified value, based on collaborative experimental work under the auspices of a scientific or engineering group.

3.1.1.1 *Discussion*—A national or international organization, referred to in 3.1.1 (*2*), generally maintains measurement standards to which the reference values obtained are traceable.

3.1.2 *accuracy, n*—the closeness of agreement between a test result and an accepted reference value.

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

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

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

3.1.4 *characteristic, n*—a property of items in a sample or population which, when measured, counted or otherwise observed, helps to distinguish between the items. **E2282**

3.1.5 *coefficient of variation, CV, n*—for a nonnegative characteristic, the ratio of the standard deviation to the mean for a population or sample. **E2586**

3.1.6 *intermediate precision, n*—the closeness of agreement between test results obtained under specified intermediate precision conditions.

3.1.6.1 *Discussion*—The specific measure and the specific conditions must be specified for each intermediate measure of precision; thus, "standard deviation of test results among operators in a laboratory," or "day-to-day standard deviation within a laboratory for the same operator."

3.1.6.2 *Discussion*—Because the training of operators, the agreement of different pieces of equipment in the same laboratory and the variation of environmental conditions with longer time intervals all depend on the degree of within-laboratory control, the intermediate measures of precision are likely to vary appreciably from laboratory to laboratory. Thus, intermediate precisions may be more characteristic of individual laboratories than of the test method.

[1] This practice is under the jurisdiction of ASTM Committee E11 on Quality and Statistics and is the direct responsibility of Subcommittee E11.20 on Test Method Evaluation and Quality Control.

Current edition approved May 1, 2014. Published May 2014. Originally approved in 1961. Last previous edition approved in 2013 as E177 – 13. DOI: 10.1520/E0177-14.

[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.

3.1.7 *intermediate precision conditions, n*—conditions under which test results are obtained with the same test method using test units or test specimens taken at random from a single quantity of material that is as nearly homogeneous as possible, and with changing conditions such as operator, measuring equipment, location within the laboratory, and time.

3.1.8 *observation, n*—the process of obtaining information regarding the presence or absence of an attribute of a test specimen, or of making a reading on a characteristic or dimension of a test specimen. **E2282**

3.1.9 *observed value, n*—the value obtained by making an observation. **E2282**

3.1.10 *precision, n*—the closeness of agreement between independent test results obtained under stipulated conditions.

3.1.10.1 *Discussion*—Precision depends on random errors and does not relate to the accepted reference value.

3.1.10.2 *Discussion*—The measure of precision usually is expressed in terms of imprecision and computed as a standard deviation of the test results. Less precision is reflected by a larger standard deviation.

3.1.10.3 *Discussion*—"Independent test results" means results obtained in a manner not influenced by any previous result on the same or similar test object. Quantitative measures of precision depend critically on the stipulated conditions. Repeatability and reproducibility conditions are particular sets of extreme stipulated conditions.

3.1.11 *repeatability, n*—precision under repeatability conditions.

3.1.11.1 *Discussion*—Repeatability is one of the concepts or categories of the precision of a test method.

3.1.11.2 *Discussion*—Measures of repeatability defined in this compilation are *repeatability standard deviation* and *repeatability limit*.

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.

3.1.12.1 *Discussion*—See *precision*, The "same operator, same equipment" requirement means that for a particular step in the measurement process, the same combination of operator and equipment is used for every test result. Thus, one operator may prepare the test specimens, a second measure the dimensions and a third measure the mass in a test method for determining density.

3.1.12.2 *Discussion*—By "in the shortest practical period of time" is meant that the test results, at least for one material, are obtained in a time period not less than in normal testing and not so long as to permit significant change in test material, equipment or environment.

3.1.13 *repeatability limit (r), n*—the value below which the absolute difference between two individual test results obtained under repeatability conditions may be expected to occur with a probability of approximately 0.95 (95 %).

3.1.13.1 *Discussion*—The repeatability limit is $2.8\left(\approx 1.96\sqrt{2}\right)$ times the repeatability standard deviation. This multiplier is independent of the size of the interlaboratory study.

3.1.13.2 *Discussion*—The approximation to 0.95 is reasonably good (say 0.90 to 0.98) when many laboratories (30 or more) are involved, but is likely to be poor when fewer than eight laboratories are studied.

3.1.14 *repeatability standard deviation (s_r), n*—the standard deviation of test results obtained under repeatability conditions.

3.1.14.1 *Discussion*—It is a measure of the dispersion of the distribution of test results under repeatability conditions.

3.1.14.2 *Discussion*—Similarly, "repeatability variance" and "repeatability coefficient of variation" could be defined and used as measures of the dispersion of test results under repeatability conditions.—In an interlaboratory study, this is the pooled standard deviation of test results obtained under repeatability conditions.

3.1.14.3 *Discussion*—The repeatability standard deviation, usually considered a property of the test method, will generally be smaller than the within-laboratory standard deviation. (See *within-laboratory standard deviation*.)

3.1.15 *reproducibility, n*—precision under reproducibility conditions.

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

3.1.16.1 *Discussion—Identical material* means either the same test units or test specimens are tested by all the laboratories as for a nondestructive test or test units or test specimens are taken at random from a single quantity of material that is as nearly homogeneous as possible.

A different laboratory of necessity means a different operator, different equipment, and different location and under different supervisory control.

3.1.17 *reproducibility limit (R), n*—the value below which the absolute difference between two test results obtained under reproducibility conditions may be expected to occur with a probability of approximately 0.95 (95 %).

3.1.17.1 *Discussion*—The reproducibility limit is $2.8\left(\approx 1.96\sqrt{2}\right)$ times the reproducibility standard deviation. The multiplier is independent of the size of the interlaboratory study (that is, of the number of laboratories participating).

3.1.17.2 *Discussion*—The approximation to 0.95 is reasonably good (say 0.90 to 0.98) when many laboratories (30 or more) are involved but is likely to be poor when fewer than eight laboratories are studied.

3.1.18 *reproducibility standard deviation (s_R), n*—the standard deviation of test results obtained under reproducibility conditions.

3.1.18.1 *Discussion*—Other measures of the dispersion of test results obtained under reproducibility conditions are the "reproducibility variance" and the "reproducibility coefficient of variation."

3.1.18.2 *Discussion*—The reproducibility standard deviation includes, in addition to between-laboratory variability, the repeatability standard deviation and a contribution from the interaction of laboratory factors (that is, differences between

operators, equipment and environments) with material factors (that is, the differences between properties of the materials other than that property of interest).

3.1.19 *standard deviation, n—of a population*, σ, the square root of the average or expected value of the squared deviation of a variable from its mean; *—of a sample*, *s*, the square root of the sum of the squared deviations of the observed values in the sample divided by the sample size minus 1. **E2586**

3.1.20 *test determination, n*—the value of a characteristic or dimension of a single test specimen derived from one or more observed values. **E2282**

3.1.21 *test method, n*—a definitive procedure that produces a test result. **E2282**

3.1.22 *test result, n*—the value of a characteristic obtained by carrying out a specified test method. **E2282**

3.1.23 *test specimen, n*—the portion of a test unit needed to obtain a single test determination. **E2282**

3.1.24 *test unit, n*—the total quantity of material (containing one or more test specimens) needed to obtain a test result as specified in the test method. See *test result.* **E2282**

3.1.25 *trueness, n*—the closeness of agreement between the population mean of the measurements or test results and the accepted reference value.

3.1.25.1 *Discussion*—"Population mean" is, conceptually, the average value of an indefinitely large number of test results

3.1.26 *variance,* σ^2, s^2, *n*— square of the standard deviation of the population or sample. **E2586**

3.1.27 *within-laboratory standard deviation, n*—the standard deviation of test results obtained within a laboratory for a single material under conditions that may include such elements as different operators, equipment, and longer time intervals.

3.1.27.1 *Discussion*—Because the training of operators, the agreement of different pieces of equipment in the same laboratory and the variation of environmental conditions with longer time intervals depend on the degree of within-laboratory control, the within-laboratory standard deviation is likely to vary appreciably from laboratory to laboratory.

4. Significance and Use

4.1 Part A of the "Blue Book," *Form and Style for ASTM Standards*, requires that all test methods include statements of precision and bias. This practice discusses these two concepts and provides guidance for their use in statements about test methods.

4.2 *Precision*—A statement of precision allows potential users of a test method to assess in general terms the test method's usefulness with respect to variability in proposed applications. A statement of precision is not intended to exhibit values that can be exactly duplicated in every user's laboratory. Instead, the statement provides guidelines as to the magnitude of variability that can be expected between test results when the method is used in one, or in two or more, reasonably competent laboratories. For a discussion of precision, see 8.1.

4.3 *Bias*—A statement of bias furnishes guidelines on the relationship between a set of typical test results produced by the test method under specific test conditions and a related set of accepted reference values (see 9.1).

4.3.1 An alternative term for bias is trueness, which has a positive connotation, in that greater bias is associated with less favorable trueness. Trueness is the systematic component of accuracy.

4.4 *Accuracy*—The term "accuracy," used in earlier editions of Practice E177, embraces both precision and bias (see 9.3).

5. Test Method

5.1 Section 2 of the ASTM Regulations describes a *test method* as "a definitive procedure for the identification, measurement, and evaluation of one or more qualities, characteristics, or properties of a material, product, system or service that produces a test result."

5.2 In this practice only quantitative test methods that produce numerical results are considered. Also, the word "material" is used to mean material, product, system or service; the word "property" is used herein to mean that a quantitative test result can be obtained that describes a characteristic or a quality, or some other aspect of the material; and "test method" refers to both the document and the procedure described therein for obtaining a quantitative test result for one property. For a discussion of test result, see 7.1.

5.3 A well-written test method specifies control over such factors as the test equipment, the test environment, the qualifications of the operator (explicitly or implicitly), the preparation of test specimens, and the operating procedure for using the equipment in the test environment to measure some property of the test specimens. The test method will also specify the number of test specimens required and how measurements on them are to be combined to provide a test result (7.1), and might also reference a sampling procedure appropriate for the intended use of the method.

5.4 It is necessary that the writers of the test method provide instructions or requirements for every known outside influence.

5.5 A test method conducted in a laboratory should demonstrate a long-term state of statistical control (see Refs. **(1-3)**,[3] Guide E1488, and Practice E2587).

6. Measurement Terminology

6.1 A *test result* is the value obtained by carrying out the complete protocol of the test method once, being as simple as the result of a single direct visual observation on a test specimen or the result of a complex series of automated procedures with the test result calculation performed by a computer.

6.2 The following terms are used to describe partial results of the test method: *observed value*, and *test determination*, which are more fully described in Guide E2282.

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

6.2.1 An *observed value* is interpreted as the most elemental single reading obtained in the process of making an observation. As examples, an observation may involve a zero-adjusted micrometer reading of the thickness of a test strip at one position along the strip or the weight of a subsample taken from a powder sample.

6.2.2 A *test determination* summarizes or combines one or more observed values. For example, (*1*) the measurement of the bulk density of a powder may involve the observation of the mass and the tamped volume of the sample specimen, and the calculated bulk density as the ratio mass/volume is a test determination; (*2*) the test determination of the thickness of a test specimen strip may involve averaging micrometer caliper observations taken at several points along the strip.

6.2.3 A test result summarizes or combines one or more test determinations. For example, (*1*) a test method on bulk density might require that the test determination of density for each of five subsamples of the powder sample be averaged to calculate the test result; (*2*) a test method may involve multiple automated operations, combined with a calibration procedure, with many observed values and test determinations, and the test result calculated and printed out by a computer.

6.3 Precision statements for ASTM test methods are applicable to comparisons between test results, not test determinations nor observations, unless specifically and clearly indicated otherwise.

7. Sources of Variability

7.1 *Sources of Variation of Test Results:*

7.1.1 Generation of a test result involves an *interpretation* of the written document by *an operator*, who uses a *specific unit and version of the specified test apparatus*, in the *particular environment* of this testing laboratory, to evaluate a *specified number of test specimens* of the material to be tested. Replicate test results will differ due to changes in one or more of the above emphasized experimental factors. Even when none of the experimental factors is intentionally changed, small changes usually occur. The outcome of these changes may be seen as variability among the test results.

7.1.2 Each of the above experimental factors and all others, known and unknown, that can change the test result, are potential sources of variability. Some of the more common factors are discussed in 7.2 – 7.6.

7.2 *Operator:*

7.2.1 *Clarity of Test Method*—Every effort must be made in preparing an ASTM standard test method to eliminate the possibility of serious differences in interpretation. One way to check clarity is to observe, without comment, a competent laboratory operator, not previously familiar with the method, apply the draft test method. If the operator has any difficulty, the draft most likely needs revision.

7.2.2 *Completeness of Test Method*—It is necessary that operators, who are generally familiar with the test method or similar methods, not read anything into the instructions that is not explicitly stated therein. Therefore, to ensure minimum variability due to interpretation, procedural requirements must be complete.

7.2.2.1 If requirements are not explicitly stated in the test method (see 5.4), they must be included in the instructions for the interlaboratory study (see Practice E691).

7.2.3 *Differences in Operator Technique*—Even when operators have been trained by the same instructor or supervisor to give practically identical interpretations to the various steps of the test method, different operators (or even the same operator at different times) may still differ in such things as dexterity, reaction time, color sensitivity, interpolation in scale reading, and so forth. Unavoidable operator differences are thus one source of variability between test results. The test method should be designed and described to minimize the effects of these operator sources of variability.

7.3 *Apparatus:*

7.3.1 *Tolerances*—In order to avoid prohibitive costs, only necessary and reasonable manufacturing and maintenance tolerances can be specified. The variations allowed by these reasonable specification tolerances can be one source of variability between test results from different sets of test equipment.

7.3.2 *Calibration*—One of the variables associated with the equipment is its state of calibration, including traceability to national standards. The test method must provide guidance on the frequency of verification and of partial or complete recalibration; that is, for each test determination, each test result, once a day, week, etc., or as required in specified situations. Calibration drift introduces bias into test results over time. However, frequent unnecessary calibration contributes to variability of test results.

7.4 *Environment:*

7.4.1 The properties of many materials are sensitive to temperature, humidity, atmospheric pressure, atmospheric contaminants, and other environmental factors. The test method usually specifies the standard environmental conditions for testing. However, since these factors cannot be controlled perfectly within and between laboratories, a test method must be able to cope with a reasonable amount of variability that inevitably occurs even though measurement and adjustment for the environmental variation have been used to obtain control (see 5.5). Thus, the method must be both robust to the differences between laboratories and require a sufficient number of test determinations to minimize the effect of within-laboratory variability.

7.5 *Sample (Test Specimens):*

7.5.1 A lot (or shipment) of material must be sampled. Since it is unlikely that the material is perfectly uniform, sampling variability is another source of variability among test results. In some applications, useful interpretation of test results may require the measurement of the sampling error. In interlaboratory evaluation of test methods to determine testing variability, special attention is required in the selection of the material sample (see 10.3.4 and Practice E691) in order to obtain test specimens that are as similar as possible. A small residual amount of material variability is almost always an inseparable component of any estimate of testing variability.

7.5.2 Handling and size reduction of material during sampling can affect the test result when involving exposure to heat,

light, and the atmosphere. Testing of unstable materials requires attention to all aspects of the measurement operation, not just the test method itself, to control both systematic error and variability.

7.6 *Time:*

7.6.1 Each of the above sources of variability (operator performance, equipment, environment, test specimens) may change with time; for example, during a period when two or more test results are obtained. The longer the period, the less likely changes in these sources will remain random (that is, the more likely systematic effects will enter), thereby increasing the net change and the observed differences in test results. These differences will also depend on the degree of control exercised within the laboratory over the sources of variability. In conducting an interlaboratory evaluation of a test method, the time span over which the measurements are made should be kept as short as reasonably possible (see 8.2.4).

8. Precision

8.1 *Precision:*

8.1.1 The *precision* of a measurement process, and hence the stated precision of the test method from which the process is generated, is a generic concept related to the closeness of agreement between test results obtained under prescribed conditions from the measurement process being evaluated. The greater the dispersion or scatter of the test results, the poorer the precision. (It is assumed that the resolution of the test apparatus is not so poor as to result in absolute agreement among observations and hence among test results.) Measures of dispersion, usually used in statements about precision, are, in fact, direct measures of imprecision. Although it may be stated quantitatively as the reciprocal of the standard deviation, precision is usually expressed as the standard deviation or some multiple of the standard deviation (see 10.1).

8.1.2 The precision of the measurement process will depend on what sources (7.1 – 7.6) of variability are purposely included and may also depend on the test level (see 10.3.4.1). An estimate of precision can be made and interpreted only if the experimental situation (prescribed conditions) under which the test results are obtained is carefully described. There is no such thing as *the* precision of a test method; a separate precision statement will apply to each combination of sources of variability.

8.1.3 Two conditions applicable to ASTM test methods are *repeatability conditions* and *reproducibility conditions*, and these lead to estimates of the *repeatability* and *reproducibility* of a test method. Other conditions can be defined and are known collectively as *intermediate precision conditions*. The long run variability within one laboratory is an intermediate precision condition of particular importance (see 8.4.2).

8.2 *Repeatability:*

8.2.1 Repeatability is precision determined from multiple test results conducted under repeatability conditions, where the test method is conducted by a single, well-trained operator using one set of equipment in a short period of time during which neither the equipment nor the environment is likely to change appreciably. Any variability is due to small changes in conducting the test operations and possible variation of the measured property among test samples. The latter is kept to a minimum by use of proper sampling procedures in test sample preparation. All potential sources of variability must be carefully controlled within the tolerances specified in the test method.

NOTE 1—If the test method requires a series of steps, the "single-operator-equipment" requirement means that for a particular step the same combination of operator and equipment is used for every test result and on every material. Thus one operator may prepare the test specimens, a second measure the dimensions and a third measure the breaking force. The "single-day" requirement means that the test results, at least for a particular material are obtained in the shortest practical period of time, whether this be a fraction of a day or several days.

8.2.2 The repeatability precision may be estimated in a single experiment in a laboratory as the standard deviation of the test results obtained under repeatability conditions. The number of test results should be recorded along with the standard deviation estimate.

8.2.3 An estimate of repeatability may also be provided by multiple experiments in the same laboratory over time, each experiment being conducted under repeatability conditions. The pooled standard deviation is calculated and used as the estimate of the long term repeatability for that laboratory, termed the laboratory repeatability.

8.2.4 A repeatability estimate for a test method may be provided from an interlaboratory study (ILS) where a single repeatability experiment is conducted in each of multiple laboratories. The pooled repeatability standard deviation is used as the estimate of the test method repeatability. The calculation for this estimate is given in Practice E691.

8.2.4.1 This approach, using only one operator-day-equipment combination in each laboratory, provides a typical value for repeatability for the laboratories participating in the ILS. Therefore, it is a useful guide to the magnitude of variability that can be expected between test results in other competent laboratories. If this estimate of within- laboratory precision does not change significantly from laboratory to laboratory, then the measure of repeatability precision can be treated as a characteristic of the test method.

8.3 *Reproducibility:*

8.3.1 Reproducibility is precision determined from multiple test results conducted under reproducibility conditions, defined as conditions where the test method is conducted in several laboratories, each with its own operator, apparatus, and environmental conditions, obtaining test results on randomly-selected test samples selected from the same reasonably-uniform material. The laboratories being compared in order to obtain the reproducibility of the test method should be independent of each other, meaning that the laboratories should not be under the same supervisory control, nor should they have worked together to resolve differences. The value found for the between-laboratory variance will depend on the choice of laboratories and the selection of operators and apparatus within each laboratory.

8.3.2 The reproducibility variance (the square of the reproducibility standard deviation) is estimated as the sum of two variance components: the repeatability precision variance (8.2.4) and the between-laboratory variance. The calculation for this estimate is given in Practice E691.

8.4 *Intermediate Precision:*

8.4.1 Intermediate precision is precision determined from multiple test results conducted under intermediate precision conditions, which may be defined in various ways, depending upon which sources of variation are controlled in the experiment. Three examples of such intermediate precision are given below.

8.4.1.1 *Single-Operator-Apparatus, Multi-Day Precision*—This experiment evaluates the precision of a single operator using one set of equipment obtains replicate test results, but one on each of two or more days. Since the time interval is greater than in 8.2, there is a greater chance that the equipment (including its calibration) and the environment may change, and that the change will depend on the degree of control maintained by the laboratory over these factors. Therefore, the precision calculated in this between-day within-laboratory situation, may vary appreciably among operators and from laboratory to laboratory. While this multi-day precision has been called "repeatability" by some ASTM committees, it is better to reserve the term for the precision estimate described in 8.2.1. If information on multi-day precision is needed by a laboratory, it should be studied in that laboratory, since the estimate may vary widely from laboratory to laboratory. In the case where the test uses a comparison against the standard for the day, multi-day precision includes that variance component.

8.4.1.2 *Multi-Operator, Single-Day-Apparatus Precision*—In this experiment each of several operators in one laboratory using the same set of equipment obtains a test result. Since the operator effect may depend on the degree of training and supervision exercised in the laboratory, the precision among test results (between operators within laboratory) may vary widely from laboratory to laboratory, and therefore may not be regarded as a universal parameter of the test method. If information on multi-operator precision is needed by a laboratory, it should be studied only by that laboratory.

8.4.2 *Within-Laboratory Precision*—Single test results, or multiple test results under repeatability conditions, are periodically determined from subsamples of a bulk material involving different operators, apparatus, and other sources of variability over an extended period of time. This is also known as *laboratory precision* or *site precision* and may be used as an estimate of laboratory uncertainty.

9. Bias

9.1 *Bias:*

9.1.1 The *bias* of a measurement process is a consistent or systematic difference between an average of a set of test results and the Accepted Reference Value for the measured property. Therefore, when an accepted reference value is not available, the bias cannot be established. Test method variability includes systematic as well as random components. The systematic components can be evaluated if a certified reference material with an accepted reference value of the property being measured is available. In determining the bias, the effect of the imprecision is minimized by taking the average of a large set of test results. This average minus the accepted reference value is an estimate of the bias of the test method.

9.1.2 The magnitude of the bias may depend on what sources of variability are included, and may also vary with the test level and the nature of the material (see 10.4.8).

9.1.3 When evaluating the bias of a test method, it is usually advisable to minimize the effect of the random component of the measurement error by using at each test level the average of many (30 or more) test results, measured independently, for each of several relatively uniform materials, the reference values for which have been established by one of the alternatives in 9.2.1 (see 8.2.3).

9.1.4 If the test method has a known bias, an adjustment for the bias may be incorporated in the test method in the section on calculation or in a calibration curve and then the method would be without known bias (see 10.3.3 for a discussion of known and maximum bias).

9.1.5 The concept of bias may also be used to describe the systematic difference between two operators, two test sites (see 8.2.3), two seasons of the year, two test methods, and so forth. Such bias is not a direct property of the test method, unless one of the test sites or test methods provides the accepted reference value. The effect of such bias may be reflected in the measured reproducibility of the test method.

9.2 *Accepted Reference Value:*

9.2.1 A measurement process is generated by the application of a test method. Variability can be introduced unintentionally into the measurement process through the impact of many sources, such as heterogeneity of the material, state of maintenance and calibration of equipment, and environmental fluctuations (7.1 – 7.6). The variability may include systematic as well as random components. The systematic components may be evaluated (9.1) if an accepted reference value is available. An *accepted reference value* is a value that serves as an agreed-upon reference for comparison. It may be:

(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;

(3) A consensus value based on collaborative experimental work under the auspices of a scientific or engineering group; or

(4) For an isolated application, when no value for *(1)*, *(2)*, or *(3)* exists, an agreed upon value obtained using an accepted reference method.

NOTE 2—When the accepted reference value is a theoretical value, it is sometimes referred to as the "true" value, but this usage is not recommended.

9.2.2 Many test methods derive the test determination from the ratio of a sample response to that of a reference material of known (or accepted) value, or using a standard curve. The reference material response provides a calibration specific to the occasion of the test, which reduces effects of equipment and environmental factors.

9.2.2.1 Tests carried out in a group of analyses carried out using the same preparation of reference material share observed data from the reference. For this reason, variability among tests within the group will be smaller than from completely independent replications of the test method.

9.3 *Accuracy:*

9.3.1 Accuracy is a concept of exactness related to the closeness of agreement between a single test result and an accepted reference value (**4-6**). It depends on both the imprecision and the bias of the test method.

9.3.1.1 Although accuracy may be defined as the mean square error, the square root of the sum of squared bias and precision variance, this usage is not recommended.

9.3.1.2 In some fields accuracy is a synonym for bias, but this usage is not recommended.

9.3.2 In order to avoid confusion resulting from use of accuracy, only the terms precision and bias should be used as descriptors of ASTM test methods.

10. Statements of Precision and Bias for Test Methods

10.1 *Indexes of Precision:*

10.1.1 *Standard Deviation*—The preferred index of precision is the sample estimate of the standard deviation (symbol *s*) of test results for a given set of conditions (for example, repeatability conditions). The number of test results should be sufficiently large (at least 30 is recommended) so that the sample standard deviation is a good approximation to the standard deviation of the population of all test results (symbol σ) that could be obtained for that type of precision.

10.1.1.1 A purpose of an ASTM ILS is to determine the repeatability standard deviation (symbol s_r) and the reproducibility standard deviations (symbol s_R) of a test method (see Practice E691) and these shall be stated in the precision statement of the test method.

10.1.2 *Precision Limits*—The repeatability limit (symbol *r*) and the reproducibility limit (symbol *R*) are useful for comparing test results within laboratories and between laboratories, respectively. These limits are calculated as: $r = 2.8\ s_r$, and $R = 2.8\ s_R$. These limits may also be stated in the precision statement of the test method, and are recommended as being useful for future evaluation of test results.

10.1.2.1 The factor 2.8 is rationalized as follows (**7**). Approximately 95 % of all pairs of test results conducted under repeatability conditions from laboratories similar to those in the ILS can be expected to differ in absolute value by less than $1.960\ \sqrt{2}\ s_r = 2.77\ s_r$ (or about $2.8\ s_r$). The multiplier is 1.960 because 95 % of a normal population is within 1.960 standard deviations of the mean. The index *r* is also known as the 95 % limit on the difference between two test results run under repeatability conditions. Similarly the index *R* is the 95 % limit on the difference between two test results run under reproducibility conditions and is useful for comparing single test results from two laboratories.

10.1.3 *Precision Relative to the Test Result Level*—In some situations there may be some advantage in expressing the precision index as a fraction or percentage of the average test result; that is, in terms of the *coefficient of variation* (symbol *CV*) or *relative standard deviation* (symbol *RSD*). The *CV* is the ratio of the sample standard deviation to the sample average, usually expressed as a percentage: $CV = 100\ s/\bar{X}$.

10.2 *Preferred Statements of Bias for ASTM Test Methods:*

10.2.1 If a certified reference material is used in an ILS an estimate of bias shall be stated in the test method and how this correction is accommodated therein. An adjustment for what is known about the bias may be incorporated in the calculations or calibration curves as part of the test method.

10.2.2 If the bias of a test method, or the uncorrected balance of the bias, is not known because there is no accepted reference value, but upper and lower bounds can be estimated by a theoretical analysis of potential systematic errors, credible bounds for this uncorrectable balance of the bias should be given in the bias statement (**6**).

NOTE 3—No formula for combining the precision and the bias of a test method into a single numerical value of accuracy is likely to be useful. Instead separate statements of precision and bias should be presented.

10.3 *Elements of a Statement of Precision and Bias:*

10.3.1 The precision and bias section of a test method should include, as a minimum, the elements specified in this section.

10.3.1.1 A brief description of the interlaboratory test program on which the statement is based, including: *(1)* what materials were tested, *(2)* number of laboratories, *(3)* number of test results per laboratory per material, and the *(4)* interlaboratory practice (usually Practice E691) followed in the design of the study and analysis of the data. This section should give the ASTM Research Report number for the interlaboratory data and analysis.

10.3.1.2 A description of any deviation from complete adherence to the test method for each test result, such as preparation in one laboratory of the cured test sheets and distribution thereof to the participating laboratories, when curing is a specified part of the test method.

10.3.1.3 The number of test determinations and their combination to form a test result, if not clearly defined in the body of the test method.

10.3.1.4 Report the repeatability and reproducibility standard deviations (or percent coefficients of variation) among test results. A statement of the precision between test results expressed in terms of the 95 % repeatability limit and the 95 % reproducibility limit may also be included. The variation of these statistics with test level or material should also be stated. Finally, state that repeatability and reproducibility are used as directed in this practice.

10.3.2 If precision under additional conditions (for example, operator-to-operator or day-to-day) has been determined, report the number of operators or days per laboratory. Include a careful description of the additional conditions, and the precision values obtained, using such terminology as 95 % limit (operator-to-operator within laboratory).

10.3.3 A bias statement gives what is known about bias, including how the method has been modified to adjust for what is known about bias and that it is now without known bias. If the value of the property being measured can be defined only in terms of the test method, state this and whether the method is generally accepted as a reference method. If an estimate of the maximum bias of the method can be made on theoretical grounds (for example, by examining the maxi4mum probable contributions of various steps in the procedure to the total bias), then describe these grounds in this section. Give the ASTM Research Report number on the theoretical or experimental study of bias.

10.3.4 *Variation of Precision and Bias with Material:*

10.3.4.1 A test method is intended to cover a class of materials. Any one material within the class differs from any other in the following two basic ways: the level of the property that is being measured; and the matrix of the material. The matrix is the totality of all properties, other than the level of the property to be measured, that can have an effect on the measured value. Precision and bias of the test method can be functions of both property level and matrix.

10.4 *Examples of Statements of Precision and Bias:*

10.4.1 The illustrative example Ex.1 is an example in which only two materials have been used but with the required minimum number (six) of participating laboratories:

Ex.1 Precision

Ex.1.1 *Interlaboratory Test Program*—An interlaboratory study was run in which randomly drawn test specimens of two materials (kraft envelope paper and wove envelope paper) were tested for tearing strength in each of six laboratories, with each laboratory testing two sets of five specimens of each material. Except for the use of only two materials, Practice E691 was followed for the design and analysis of the data, the details are given in ASTM Research Report No. XXXY.

Ex.1.2 *Test Result*—The precision information given below in the units of measurement (grams force) is for the comparison of two test results, each of which is the average of five test determinations:

Ex.1.3 *Precision:*

	Material A	Material B
Average test value	45 gf	100 gf
Repeatability standard deviation	1.07 gf	2.5 gf
Reproducibility standard deviation	2.14 gf	4.3 gf
95 % repeatability limit (within laboratory)	3 gf	7 gf
95 % reproducibility limit (between laboratories)	6 gf	12 gf

The above terms are used as specified in Practice E177.

10.4.2 If a sufficient number of different materials to cover the test range are included in the interlaboratory study (at least three or more in accordance with Practice E691), then the approximate variation in precision with test level may be determined. Since two distinctly separate classes of material are tested by the method shown in illustrative example Ex.2, two separate interlaboratory studies were made. In the first study, the repeatability was found to be essentially proportional to the test value (with minor variation from material to material as shown), whereas the reproducibility had a more complex linear relationship (that is, a constant as well as a proportional term). In the second study, the repeatability and the reproducibility were each found to be proportional to the test value.

Ex.2 Precision

Coarse-fiber materials

Test range	30 to 150 g
95 % repeatability limit (within laboratory)	7 % (6 to 8.5 %) of the test result
95 % reproducibility limit (between laboratories)	2 g + 10 % (8 to 12 %) of the test result

Well beaten (fine-fiber) materials

Test range	20 to 75 g
95 % repeatability limit (within laboratory)	4 % (3.5 to 5 %) of the test result
95 % reproducibility limit (between laboratories)	7 % (5 to 8 %) of the test result

Ex.2.1 The values shown above for the limits are the average and range) in each case as found in separate interlaboratory studies for the coarse and fine-fiber materials. The terms repeatability limit and reproducibility limit are used as specified in Practice E177. The respective standard deviations among test results may be obtained by dividing the above limit values by 2.8.

10.4.3 Precision information can often be obtained from studies made for other purposes. Example below illustrates this approach and also illustrates another way of showing variation from material to material.

Ex.3 Precision

Ex.3.1 *Interlaboratory Test Program*—The information given below is based on data obtained in the TAPPI Collaborative **(8)** Reference Program for self-evaluation of laboratories, Reports 25 through 51 (Aug. 1973 through Jan. 1978). Each report covers two materials with each of approximately 16 laboratories testing 5 specimens of each material.

Ex.3.2 *Test Result*—The precision information given below has been calculated for the comparison of two test results, each of which is the average of 10 test determinations.

Ex.3.3 *95 % Repeatability Limit (within laboratory)*—The repeatability is 5.3 % of the test result. For the different materials the repeatability ranged from 3.7 to 9.6 %. The range of the central 90 percent of the repeatability values was 3.9 to 8.7 %.

Ex.3.4 *95 % Reproducibility Limit (between laboratories)*—The reproducibility is 16.2 % of the test result. For the different materials the range of all of the calculations of reproducibility was 6.4 to 45.4 %. The range of the central 90 percent of the calculations was 9.2 to 25.5 %.

Ex.3.5 *Definitions and Standard Deviations*—The above terms repeatability limit and reproducibility limit are used as specified in Practice E177. The respective percent coefficients of variation among test results may be obtained by dividing the above numbers by 2.8.

10.4.4 Precision is often constant for low test values and proportional for higher test values, as shown in the following example:

Ex.4 Precision

Test range	0.010 to 1200 mm
95 % repeatability limit (within laboratory)	0.002 mm or 2.5 % of the average, whichever is larger
95 % reproducibility limit (between laboratories)	0.005 mm or 4.2 % of the average, whichever is larger

The above terms repeatability limit and reproducibility limit are used as specified in Practice E177. The respective standard deviations and percent coefficients of variation among test results may be obtained by dividing the above limit values by 2.8.

10.4.5 A table may be used especially if the precision indexes vary irregularly from material to material. Note in the following example that the materials have been arranged in increasing order of test value:

Ex.5 Precision

Material	Glucose in Serum, Average	Repeatability Standard Deviation	Reproducibility Standard Deviation	Repeatability Limit	Reproducibility Limit
A	41.518	1.063	1.063	2.98	2.98
B	79.680	1.495	1.580	4.19	4.42
C	134.726	1.543	2.148	4.33	6.02
D	194.717	2.625	3.366	7.35	9.42
E	294.492	3.935	4.192	11.02	11.74

Ex.5.1 *Interlaboratory Test Program*—An interlaboratory study of glucose in serum was conducted in accordance with Practice E691 in eight laboratories with five materials, with each laboratory obtaining three test results for each material. See ASTM Research Report No. XXXX.

Ex.5.2 The terms repeatability limit and reproducibility limit in Ex.5 are used as specified in Practice E177.

10.4.6 An example of a bias statement when bias has been removed through comparison with a reference method is given in below.

Ex.6 Bias

Ex.6.1. *Bias*—The original draft of this abbreviated method was experimentally compared in one laboratory with the appropriate reference method (ASTM DXXXX) and was found to give results approximately 10 % high, as theoretical considerations would suggest (see ASTM Research Report No. XXXW). An adjustment for this bias is now made in Section XX on calculations, so that the final result is now without known bias.

10.4.7 A similar statement would apply for any accepted reference value, for example, from an accepted reference material. If bias depends on other properties of the material, a statement such as the following might be used:

Ex.7 Bias

Ex.7.1. *Bias*—A ruggedness study (ASTM Research Report No. XXXZ) showed that test results are temperature dependent, with the dependence varying with the type of material. Therefore, if the test temperature cannot be maintained within the specified limits, determine the temperature dependence for the specific material being tested and correct test results accordingly.

10.4.8 A maximum value for the bias of a test method may be estimated by an analysis of the effect of apparatus and procedural tolerances on the test results, as illustrated below:

Ex.8 Bias

Ex.8.1. *Bias*—Error analysis shows that the absolute value of the maximum systematic error that could result from instrument and other tolerances specified in the test method is 3.2 % of the test result.

10.4.9 Even when a quantitative statement on bias is not possible, it is helpful to the user of the method to know that the developers of the method have considered the possibility of bias. In such cases, a statement on bias based on one of the following examples may be used:

Ex.9 Bias

Ex.9.1 *Bias*—This method has no bias because (insert the name of the property) is defined only in terms of this test method.
Ex.9.2 *Bias*—Since there is no accepted reference material, method, or laboratory suitable for determining the bias for the procedure in this test method for measuring (insert the name of the property), no statement on bias is being made.
Ex.9.3 *Bias*—No justifiable statement can be made on the bias of the procedure in this test method for measuring (insert the name of the property) because (insert the reason).

11. Keywords

11.1 accepted reference value; accuracy; bias; interlaboratory study; precision; precision conditions; repeatability; reproducibility; standard deviation

REFERENCES

(1) Shewhart, W. A., *Statistical Method from the Viewpoint of Quality Control*, The Graduate School of the Department of Agricultural, Washington, DC, 1939.

(2) Mandel, J., *The Statistical Analysis of Experimental Data*, Interscience-Wiley Publishers, New York, NY, 1964 (out of print); corrected and reprinted by Dover Publishers, New York, NY, 1984, p. 105.

(3) Manual on Presentation of Data and Control Chart Analysis, *MNL7A*, ASTM 2002.

(4) Murphy, R. B., "On the Meaning of Precision and Accuracy," *Materials Research and Standards*, ASTM, April 1961, pp. 264–267.

(5) Eisenhart, C., "The Reliability of Measured Values—Part I: Fundamental Concepts," *Photogrammetric Engineering*, June 1952, pp. 542–561.

(6) Eisenhart, C., "Realistic Evaluation of the Precision and Accuracy of Instrument Calibration Systems," *Journal of Research of the National Bureau of Standards*, *67C*, 1963, pp. 161–187.

(7) Mandel, J., and Lashof, T. W., "The Nature of Repeatability and Reproducibility," *Journal of Quality Technology*, Vol 19, No. 1, January 1987, pp. 29–36.

(8) TAPPI Collaborative Reference Program, Reports 25 through 51, August 1973 through January 1978, Technical Association of the Pulp and Paper Industry.