Designation: E2935 – 16

An American National Standard

Standard Practice for Conducting Equivalence Testing in Laboratory Applications[1]

This standard is issued under the fixed designation E2935; 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 statistical methodology for conducting equivalence testing on numerical data from two sources to determine if their true means or variances differ by no more than predetermined limits.

1.2 Applications include *(1)* equivalence testing for bias against an accepted reference value, *(2)* determining means equivalence of two test methods, test apparatus, instruments, reagent sources, or operators within a laboratory or equivalence of two laboratories in a method transfer, and *(3)* determining non-inferiority of a modified test procedure versus a current test procedure with respect to a performance characteristic.

1.3 The guidance in this standard applies only to experiments conducted on a single material at a given level of the test result.

1.4 Guidance is given for determining the amount of data required for an equivalence trial. The control of risks associated with the equivalence decision is discussed.

1.5 The values stated in SI units are to be regarded as standard. No other units of measurement are included in this standard.

1.6 *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 limitations prior to use.*

2. Referenced Documents

2.1 *ASTM Standards:*[2]

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

E456 Terminology Relating to Quality and Statistics

E2282 Guide for Defining the Test Result of a Test Method

E2586 Practice for Calculating and Using Basic Statistics

2.2 *USP Standard:*[3]

USP <1223> Validation of Alternative Microbiological Methods

3. Terminology

3.1 *Definitions*—See Terminology E456 for a more extensive listing of statistical terms.

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. **E177**

3.1.2 *bias, n*—the difference between the expectation of the test results and an accepted reference value. **E177**

3.1.3 *confidence interval, n*—an interval estimate [L, U] with the statistics L and U as limits for the parameter θ and with confidence level $1 - \alpha$, where $\Pr(L \le \theta \le U) \ge 1 - \alpha$. **E2586**

3.1.3.1 *Discussion*—The confidence level, $1 - \alpha$, reflects the proportion of cases that the confidence interval [L, U] would contain or cover the true parameter value in a series of repeated random samples under identical conditions. Once L and U are given values, the resulting confidence interval either does or does not contain it. In this sense "confidence" applies not to the particular interval but only to the long run proportion of cases when repeating the procedure many times.

3.1.4 *confidence level, n*—the value, $1 - \alpha$, of the probability associated with a confidence interval, often expressed as a percentage. **E2586**

3.1.4.1 *Discussion*—α is generally a small number. Confidence level is often 95 % or 99 %.

3.1.5 *confidence limit, n*—each of the limits, L and U, of a confidence interval, or the limit of a one-sided confidence interval. **E2586**

[1] This test method 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 Nov. 15, 2016. Published January 2017. Originally approved in 2013. Last previous edition approved in 2015 as E2935 – 15. DOI: 10.1520/E2935-16.

[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] Available from U.S. Pharmacopeial Convention (USP), 12601 Twinbrook Pkwy., Rockville, MD 20852-1790, http://www.usp.org.

3.1.6 *degrees of freedom, n*—the number of independent data points minus the number of parameters that have to be estimated before calculating the variance. **E2586**

3.1.7 *equivalence, n*—condition that two population parameters differ by no more than predetermined limits.

3.1.8 *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. **E177**

3.1.9 *mean, n—of a population*, μ, average or expected value of a characteristic in a population – *of a sample*, $\bar{X}$ sum of the observed values in the sample divided by the sample size. **E2586**

3.1.10 *percentile, n*—quantile of a sample or a population, for which the fraction less than or equal to the value is expressed as a percentage. **E2586**

3.1.11 *population, n*—the totality of items or units of material under consideration. **E2586**

3.1.12 *population parameter, n*—summary measure of the values of some characteristic of a population. **E2586**

3.1.13 *precision, n*—the closeness of agreement between independent test results obtained under stipulated conditions. **E177**

3.1.14 *quantile, n*—value such that a fraction *f* of the sample or population is less than or equal to that value. **E2586**

3.1.15 *repeatability, n*—precision under repeatability conditions. **E177**

3.1.16 *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**

3.1.17 *repeatability standard deviation (s_r), n*—the standard deviation of test results obtained under repeatability conditions. **E177**

3.1.18 *sample, n*—a group of observations or test results, taken from a larger collection of observations or test results, which serves to provide information that may be used as a basis for making a decision concerning the larger collection. **E2586**

3.1.19 *sample size, n, n*—number of observed values in the sample. **E2586**

3.1.20 *sample statistic, n*—summary measure of the observed values of a sample. **E2586**

3.1.21 *standard deviation—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 from their mean divided by the sample size minus 1. **E2586**

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

3.1.23 *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**

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

3.2 *Definitions of Terms Specific to This Standard:*

3.2.1 *bias equivalence, n*—equivalence of a population mean with an accepted reference value.

3.2.2 *equivalence limit, E, n—in equivalence testing*, a limit on the difference between two population parameters.

3.2.2.1 *Discussion*—In certain applications, this may be termed *practical limit* or *practical difference*.

3.2.3 *equivalence test, n*—a statistical test conducted within predetermined risks to confirm equivalence of two population parameters.

3.2.4 *means equivalence, n*—equivalence of two population means.

3.2.5 *non-inferiority, n*—condition that the difference in means or variances of test results between a modified testing process and a current testing process with respect to a performance characteristic is no greater than a predetermined limit in the direction of inferiority of the modified process to the current process.

3.2.5.1 *Discussion*—Other terms used for *non-inferior* are "equivalent or better" or "at least equivalent as."

3.2.6 *paired samples design, n—in means equivalence testing,* single samples are taken from the two populations at a number of sampling points.

3.2.6.1 *Discussion*—This design is termed a randomized block design for a general number of populations sampled, and each group of data within a sampling point is termed a block.

3.2.7 *power, n—in equivalence testing*, the probability of accepting equivalence, given the true difference between two population means.

3.2.7.1 *Discussion*—In the case of testing for bias equivalence the power is the probability of accepting equivalence, given the true difference between a population mean and an accepted reference value.

3.2.8 *two independent samples design, n—in means equivalence testing,* replicate test results are determined independently from two populations at a single sampling time for each population.

3.2.8.1 *Discussion*—This design is termed a completely randomized design for a general number of populations sampled.

3.2.9 *two one-sided tests (TOST) procedure, n*—a statistical procedure used for testing the equivalence of the parameters from two distributions (see equivalence).

3.3 *Symbols:*

B = bias (7.1.1)
d_j = difference between a pair of test results at sampling point j (7.1.1)
$\bar{d}$ = average difference (7.1.1)
D = difference in sample means (6.1.2) (X1.1.2)

E = equivalence limit (5.2)
E_1 = lower equivalence limit (5.2.1)
E_2 = upper equivalence limit (5.2.1)
f = degrees of freedom for s (8.1.1) (X1.1.2)
$F_{1-\alpha}$ = (1 – α)th percentile of the F distribution (9.3.1)
f_i = degrees of freedom for s_i (6.1.1)
f_p = degrees of freedom for s_p (6.1.2)
$\mathcal{F}(\bullet)$ = the cumulative F distribution function (X1.6.3)
H_0: = null hypothesis (X1.1.1)
H_A: = alternate hypothesis (X1.1.1)
n = sample size (number of test results) from a population (5.4) (6.1.3) (7.1.1) (8.1.1)
n_i = sample size from ith population (6.1.1)
n_1 = sample size from population 1 (6.1.2)
n_2 = sample size from population 2 (6.1.2)
R = ratio of two sample variances (5.5.3)
$\mathcal{R}$ = ratio of two population variances (X1.6.3)
s = sample standard deviation (8.1.1)
s_B = sample standard deviation for bias (8.1.2)
s_d = standard deviation of the difference between two test results (7.1.1)
s_D = sample standard deviation for mean difference (6.1.3) (X1.1.2)
s_i = sample standard deviation for ith population (6.1.1)
s_i^2 = sample variance for ith population (6.1.1)
s_1^2 = sample variance for population 1 (6.1.2)
s_1^2 = variance of test results from the current process (5.5.3)
s_2^2 = sample variance for population 2 (6.1.2)
s_2^2 = variance of test results from the modified process (5.5.3)
s_p = pooled sample standard deviation (6.1.2)
s_r = repeatability sample standard deviation (6.2)
t = Student's t statistic (6.1.4) (7.1.3) (8.1.3)
$t_{1-\alpha,f}$ = (1-α)th percentile of the Student's t distribution with f degrees of freedom (X1.1.2)
X_{ij} = jth test result from the ith population (6.1)
UCL_R = = upper confidence limit for $\mathcal{R}$ (9.3.1)
$\overline{X}$ = test result average (8.1.1)
$\overline{X_i}$ = test result average for the ith population (6.1.1)
$\overline{X_1}$ = test result average for population 1 (6.1.3)
$\overline{X_2}$ = test result average for population 2 (6.1.3)
$Z_{1-\alpha}$ = (1-α)th percentile of the standard normal distribution (X1.6.1)
α = consumer's risk (5.2.3) (6.2) (7.2)
β = producer's risk (5.4.1)
Δ = true mean difference between populations (5.4.1)
μ = population mean (X1.4.1)
μ_i = ith population mean (X1.1.1)
ν = approximate degrees of freedom for s_D (X1.1.4)
σ = standard deviation of the test method (5.2)
σ_d = standard deviation of the true difference between two populations (7.2)
$\Phi(\bullet)$ = standard normal cumulative distribution function (X1.6.1)

3.4 *Acronyms:*

3.4.1 *ARV, n*—accepted reference value (5.3.3) (8.1) (X1.4)

3.4.2 *CRM, n*—certified reference material (5.3.3) (8.1)

3.4.3 *ILS, n*—interlaboratory study (6.2)

3.4.4 *LCL, n*—lower confidence limit (6.2.5) (7.2.3)

3.4.5 *TOST, n*—two one-sided tests (5.5.1) (Section 6) (Section 7) (Section 8) (Appendix X1)

3.4.6 *UCL, n*—upper confidence limit (6.2.5) (7.2.3)

4. Significance and Use

4.1 Laboratories conducting routine testing have a continuing need to make improvements in their testing processes. In these situations it must be demonstrated that any changes will not cause an undesirable shift in the test results from the current testing process nor substantially affect a performance characteristic of the test method. This standard provides guidance on experiments and statistical methods needed to demonstrate that the test results from a modified testing process are equivalent to those from the current testing process, where *equivalence* is defined as agreement within a prescribed limit, termed an *equivalence limit*.

4.1.1 Examples of modifications to the testing process include, but are not limited, to the following:

(1) Changes to operating levels in the steps of the test method procedure,

(2) Installation of new instruments, apparatus, or sources of reagents and test materials,

(3) Evaluation of new personnel performing the testing, and

(4) Transfer of testing to a new location.

4.1.2 The equivalence limit, which represents a worst-case difference, is determined prior to the equivalence test and its value is usually set by consensus among subject-matter experts.

4.2 Two principal types of equivalence are covered in the practice, *means equivalence* and *non-inferiority*. Means equivalence implies that a sustained shift in test results between the modified and current testing processes refers to an absolute difference, meaning differences in either direction from zero. Non-inferiority is concerned with a difference only in the direction of an inferior outcome in a performance characteristic of the modified testing procedure versus the current testing procedure.

4.2.1 Equivalence testing is performed by an experiment that generates test results from the modified and current testing procedures on the same materials that are routinely tested. An exception is bias equivalence where the experiment consists of conducting multiple testing on a certified reference material (CRM) having an accepted reference value (ARV) to evaluate the test method bias.

4.2.2 Examples of performance characteristics directly applicable to the test method are bias, precision, sensitivity, specificity, linearity, and range. Additional characteristics are test cost and elapsed time to conduct the test procedure.

4.2.3 Non-inferiority may involve trade-offs in performance characteristics between the modified and current procedures. For example, the modified process may be slightly inferior to the established process with respect to assay sensitivity or precision but may have off-setting advantages such as faster delivery of results or lower testing costs.

4.3 *Risk Management*—Guidance is also provided for determining the amount of data required to control the risks of

making the wrong decision in accepting or rejecting equivalence (see Section X1.2).

4.3.1 The consumer's risk is the risk of falsely declaring equivalence. The probability associated with this risk is directly controlled to a low level so that accepting equivalence gives a high degree of assurance that the true difference is less than the equivalence limit.

4.3.2 The producer's risk is the risk of falsely rejecting equivalence. The probability associated with this risk is controlled by the amount of data generated by the experiment. If valid improvements are rejected by equivalence testing, this can lead to opportunity losses to the company and its laboratories (the producers) or cause unnecessary additional effort in improving the testing process.

5. Planning and Executing the Equivalence Study

5.1 This section discusses the stages of conducting an equivalence test: (*1*) determining the information needed, (*2*) setting up and conducting the study design, and (*3*) performing the statistical analysis of the resulting data. The study is usually conducted either in a single laboratory or, in the case of a method transfer, in both the originating and receiving laboratories. Using multiple laboratories will almost always increase the inherent variability of the data in the study, which will increase the cost of performing the study due to the need for more data.

5.2 *Prior information* required for the study design includes the equivalence limit E, the consumer's risk α, and an estimate of the test method precision σ.

5.2.1 For means equivalence tests there are two equivalence limits, $-E$ and E, that are tested. Limits may be nonsymmetrical around zero, such as $-E_1$ and E_2, but this is not usual and would require advice from a qualified statistician for a proper design setup. For non-inferiority tests only one of these limits is tested.

5.2.2 A prior estimate of the test method precision is essential for determining the number of test results required in the study design for adequate producer's risk control. This estimate can be available from method development work, from an interlaboratory study, or from other sources. The precision estimate should take into account the test conditions of the study, such as *repeatability*, *intermediate*, or *reproducibility* conditions.

5.2.3 The consumer's risk may be determined by an industry norm or a regulatory requirement. A probability value often used is $\alpha = 0.05$, which is a 5 % risk to the consumer that the study falsely declares equivalence.

5.3 The *design type* determines how the data are collected and how much data are needed to control the risk of a wrong decision. A sufficient quantity of a homogeneous material for the required number of tests is necessary. For comparing data from the modified and current testing processes, two basic designs are discussed in this practice, the Two Independent Samples Design, and the Paired Samples Design. These designs are suitable for determining either means equivalence or non-inferiority.

5.3.1 The Two Independent Samples Design for means equivalence is discussed in Section 6. In this design sets of independent test results are usually generated in a single laboratory by both testing procedures under repeatability conditions. For method transfer each laboratory generates independent test results using the same testing procedure, preferably under repeatability conditions. If this is not possible due to constraints on time or facilities, then the test results can be conducted under intermediate precision conditions, but a statistician is recommended for design and analysis of the test.

5.3.2 The Paired Samples Design for means equivalence is discussed in Section 7. In this design, multiple pairs of single test results from each testing procedure are generated under different conditions of a second variable, such as time of process sampling. This design is most useful when there are constraints on conducting the two independent samples design.

5.3.3 The design for bias equivalence is discussed in Section 8. In this design test results are generated by the current testing process on a certified reference material (CRM) having an accepted reference value (ARV) for the material characteristic of interest.

5.3.4 The statistical analysis for non-inferiority is discussed in Section 9 for evaluating two testing procedures with respect to a performance characteristic. The data can be generated by either of the designs discussed in Sections 6 and 7.

5.4 *Sample size* in the design context refers to the number n of test results required by each testing process to manage the producer's risk. It is possible to use different sample sizes for the modified and current test processes, but this can lead to poor control of the consumer's risk (see X1.1.4).

5.4.1 The number of test results, symbol n, from each testing process controls the producer's risk β of falsely rejecting means equivalence at a given true mean difference, Δ. The producer's risk may be alternatively stated in terms of the *power*, the probability $1-\beta$ of correctly accepting equivalence at a given value of Δ.

5.4.1.1 For symmetric equivalence limits in means equivalence tests the power profile plots the probability $1-\beta$ against the absolute value of Δ, due to the symmetry of the equivalence limits. This calculation can be performed using a spreadsheet computer package (see X1.6.1 and Appendix X2).

5.4.1.2 An example of a set of power profiles in means equivalence tests is shown in Fig. 1. The probability scale for power on the vertical axis varies from 0 to 1. The horizontal axis is the true absolute difference Δ. The power profile, a reversed S-shaped curve, should be close to a power probability of 1 at zero absolute difference and will decline to the consumer risk probability at an absolute difference of E. Power for absolute differences greater than E are less than the consumer risk and decline asymptotically to zero as the absolute difference increases.

5.4.1.3 In Fig. 1 power profiles are shown for three different sample sizes for testing means equivalence. Increasing the sample size moves the power curve to the right, giving a greater chance of accepting equivalence for a given true difference Δ. Equations for power profiles are shown in Section X1.5 and a spreadsheet example in Appendix X2.

5.4.2 Power curves for bias equivalence and non-inferiority are constructed by different formulas but have the same shape and interpretation as those for means equivalence.

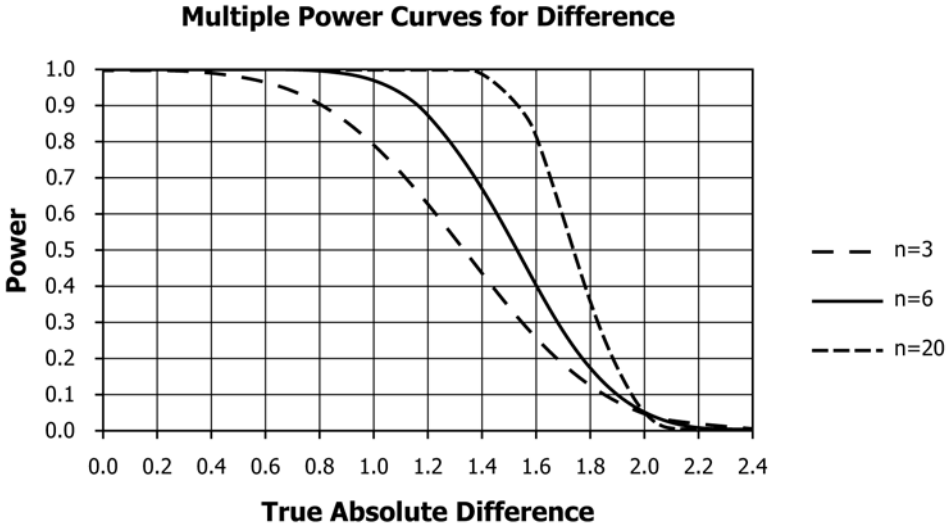

FIG. 1 Multiple Power Curves for Lab Transfer Example

5.4.2.1 For non-inferiority testing the power profile plots the probability 1–β against the true difference Δ for means (see X1.6.2) or against the true variance ratio $\mathcal{R}$ for variances (see X1.6.3).

5.4.3 Power curves are evaluated by entering different values of n and evaluating the curve shape. A practical solution is to choose n such that the power is above a 0.9 probability out to about one-half to two-thirds of the distance to E, thus giving a high probability that equivalence will be demonstrated for a range of true absolute differences that are deemed of little or no scientific import in the test result.

5.5 The *statistical analysis* for accepting or rejecting equivalence is similar for all cases and depends on the outcome of one-sided statistical hypothesis tests for means and variances. The calculations are given in detail with examples in Sections 6 – 9. The statistical theory is given in an appendix (see Section X1.1).

5.5.1 The data analysis for means equivalence testing in this practice uses a statistical methodology termed the two one-sided tests (TOST) procedure. This is based on calculating confidence limits for the true mean difference Δ as $D \pm t\, s_D$, where D is the difference between the two test result averages, s_D is the standard error of that difference, and t is a tabulated multiplier based on the number of data and a preselected confidence level. The calculation for s_D is based on the standard deviations of the two sets of data and the type of study design. Then equivalence is supported if both of the following two conditions are met:

(1) The lower confidence limit, $LCL = D - t\, s_D$, is greater than the lower equivalence limit, $-E$, and

(2) The upper confidence limit, $UCL = D + t\, s_D$, is less than the upper equivalence limit, E.

NOTE 1—Historically, this procedure originated in the pharmaceutical industry for use in bioequivalence trials (**1, 2**),[4] denoted as the Two One-Sided Tests Procedure, which has since been adopted for use in testing and measurement applications (**3, 4**).

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

5.5.1.1 The conventional Student's t test based on the null hypothesis of a zero difference is not recommended for means equivalence testing as it does not properly control the consumer's and producer's risks for this application (see Section X1.3). This test is suitable for supporting *superiority* of the modified process versus the established process instead of equivalence.

5.5.1.2 For bias equivalence the calculation for s_D is based on only a single set of data because the ARV is considered as a known mean with zero variability for the purpose of the equivalence study.

5.5.2 The data analysis for non-inferiority testing of population means uses a single one-sided test in the direction of an inferior outcome with respect to a performance characteristic determined by the test results. When the performance characteristic is defined as "higher is better", such as method sensitivity, the statistical test supports noninferiority when $LCL > -E$. Conversely, when the performance characteristic is defined as "lower is better", such as incidence of misclassifications, the statistical test supports noninferiority when $UCL < E$. Note that the means equivalence procedure comprises two one-sided statistical tests while the non-inferiority procedure performs only a single one-sided statistical test. For statistical details see Section X1.5.

5.5.3 For the equivalence testing of precision the variance is used, and "lower is better" for this parameter, so the test for non-inferiority applies. Because variances are a scale parameter, the non-inferiority test is based the ratio R of the two sample variances instead of their difference; thus $R = s_2^2/s_1^2$, where s_1^2 and s_2^2 are the calculated variances of the test results from the current and modified test processes, respectively. An upper confidence limit for the true variance ratio σ_2^2/σ_1^2, denoted UCL_R, for the given confidence level and sample sizes, can be found from the tabulated F distribution. The non-inferiority limit E is also in the form of a ratio. For example, if $E = 2$, the noninferiority limit would allow the modified process to have up to twice the variance of the

established process or up to about 1.4 times the standard deviation in the worst case. The statistical test supports noninferiority if $UCL_R < E$.

6. The TOST Procedure for Statistical Analysis of Means Equivalence — Two Independent Samples Design

6.1 *Statistical Analysis*—Let the sample data be denoted as X_{ij} = the jth test result from the ith population. The equivalence limit E, consumer's risk α, and sample sizes have been previously determined.

6.1.1 Calculate averages, variances, and standard deviations, and degrees of freedom for each sample:

$$\bar{X}_i = \frac{\sum_{j=1}^{n_i} X_{ij}}{n_i}, \quad i = 1, 2 \tag{1}$$

$$s_i^2 = \frac{\sum_{j=1}^{n_i} \left(X_{ij} - \bar{X}_i\right)^2}{(n_i - 1)}, \quad i = 1, 2 \tag{2}$$

$$s_i = \sqrt{s_i^2}, \quad i = 1, 2 \tag{3}$$

$$f_i = n_i - 1, \quad i = 1, 2 \tag{4}$$

6.1.2 Calculate the pooled standard deviation and degrees of freedom:

$$s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)}} \tag{5}$$

If $n_1 = n_2 = n$, then:

$$s_p^2 = \frac{(s_1^2 + s_2^2)}{2}$$

$$f_p = (n_1 + n_2 - 2) \tag{6}$$

6.1.3 Calculate the difference between means and its standard error:

$$D = \bar{X}_2 - \bar{X}_1 \tag{7}$$

$$s_D = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \tag{8}$$

If $n_1 = n_2 = n$, then:

$$s_D = s_p \sqrt{\frac{2}{n}}$$

6.1.4 *Test for Equivalence*—Compute the upper (UCL) and lower (LCL) confidence limits for the 100 (1–2α) % two-sided confidence interval on the true difference. If the confidence interval is completely contained within the equivalence limits (0 ± E), equivalently if $LCL > -E$ and $UCL < E$, then accept equivalence. Otherwise, reject equivalence.

$$UCL = D + ts_D \tag{9}$$

$$LCL = D - ts_D \tag{10}$$

where t is the upper 100 (1–α) % percentile of the Student's t distribution with $(n_1 + n_2 - 2)$ degrees of freedom.

6.2 *Example for Means Equivalence*—The example shown is data from a transfer of an ASTM test method from R&D Lab 1 to Plant Lab 2 (Table 1). An equivalence of limit of 2 units was proposed with a consumer risk of 5 %. An interlaboratory study (ILS) on this test method had given an estimate of s_r = 0.5 units for the repeatability standard deviation. Thus E = 2 units, α = 0.05, and estimated σ = 0.5 units are inputs for this study (the actual units are unspecified for this example).

TABLE 1 Data for Equivalence Test Between Two Laboratories

	Test Results					
Laboratory 1	96.9	97.9	98.5	97.5	97.7	97.2
Laboratory 2	97.8	97.6	98.1	98.6	98.6	98.9

6.2.1 *Sample Size Determination*—Power profiles for n = 3, 6, and 20 were generated for a set of absolute difference values ranging 0.00 (0.20) 2.40 units as shown in Fig. 1. All three curves intersect at the point (2, 0.05) as determined by the consumer's risk at the equivalence limit.

6.2.1.1 A sample size of n = 6 replicate assays per laboratory yielded a satisfactory power curve, in that the probability of accepting equivalence (power) was greater than a 0.9 probability (or a 90 % power) for a difference of about 1.2 units or less. Therefore, there would be less than an estimated 10 % risk to the producer that such a difference would fail to support equivalence in the actual trial.

6.2.1.2 A comparison of the three power curves indicates that the n = 3 design would be underpowered, as the power falls below 0.9 at 0.8 units. The n = 20 design gives somewhat more power than the n = 6 design but is more costly to conduct and may not be worth the extra expenditure.

6.2.2 Averages, variances, standard deviations, and degrees of freedom for the two laboratories are:

$$\bar{X}_1 = (96.9 + 97.9 + 98.5 + 97.5 + 97.7 + 97.2)/6 = 97.62 \text{ mg/g}$$

$$\bar{X}_2 = (97.8 + 97.6 + 98.1 + 98.6 + 98.6 + 98.9)/6 = 98.27 \text{ mg/g}$$

$$s_1^2 = [(96.9 - 97.62)^2 + \ldots + (97.2 - 97.62)^2]/(6 - 1) = 0.31367$$

$$s_2^2 = [(97.8 - 98.27)^2 + \ldots + (98.9 - 98.27)^2]/(6 - 1) = 0.26267$$

$$s_1 = \sqrt{0.31367} = 0.560$$

$$s_2 = \sqrt{0.26267} = 0.513$$

$$f_i = n_i - 1 = 6 - 1 = 5$$

The estimates of standard deviation are in good agreement with the ILS estimate of 0.5 mg/g.

6.2.3 The pooled standard deviation is:

$$s_p = \sqrt{\frac{(6 - 1)0.31367 + (6 - 1)0.26267}{(6 + 6 - 2)}} = \sqrt{\frac{2.8817}{10}} = 0.537 \text{ mg/g}$$

with 10 degrees of freedom.

6.2.4 The difference of means is D = 98.27 – 97.62 = 0.65 mg/g. The plant laboratory average is 0.65 mg/g higher than the development laboratory average. The standard error of the difference of means is $s_D = 0.537\sqrt{2/6} = 0.310$ mg/g with 10 degrees of freedom (same as that for s_p).

6.2.5 The 95th percentile of Student's t with 10 degrees of freedom is 1.812. Upper and lower confidence limits for the difference of means are:

$$UCL = 0.65 + (1.812)(0.310) = 1.21$$

$$LCL = 0.65 - (1.812)(0.310) = 0.09$$

The 90 % two-sided confidence interval on the true difference is 0.09 to 1.21 mg/g and is completely contained within the equivalence interval of –2 to 2 mg/g. Since 0.09 > –2 and 1.21 < 2, equivalence is accepted.

7. The TOST Procedure for Statistical Analysis of Means Equivalence — Paired Samples Design

7.1 *Statistical Analysis*—Let the sample data be denoted as X_{ij} = the test result from the *i*th population and the *j*th block, where i = 1 or 2. Each block represents a pair of single test results from each population. For example, the blocking factor may be time of sampling from a process. The equivalence limit E, consumer's risk α, and sample size (number of blocks, symbol n) have been previously determined (see Section 5).

7.1.1 Calculate the n differences, symbol d_j, between the two test results within each block, the average of the differences, symbol $\bar{d}$, and the standard deviation of the differences, symbol s_d, with its degrees of freedom, symbol f.

$$d_j = X_{1j} - X_{2j}, j = 1, ..., n \tag{11}$$

$$\bar{d} = \frac{\Sigma_{j=1}^{n} d_j}{n} = D \tag{12}$$

$$s_d = \sqrt{\frac{\Sigma_{j=1}^{n} \left(d_j - \bar{d}\right)^2}{(n - 1)}} \tag{13}$$

$$f = n - 1 \tag{14}$$

7.1.2 Calculate the standard error of the mean difference, symbol s_D.

$$s_D = \frac{s_d}{\sqrt{n}} \tag{15}$$

7.1.3 *Test for Equivalence*—Compute the upper (UCL) and lower (LCL) confidence limits for the 100(1–2α) % two-sided confidence interval on the true difference. If the confidence interval is completely contained within the equivalence limits (0 ± E), or equivalently if $LCL > -E$ and $UCL < E$, then accept equivalence. Otherwise, reject equivalence.

$$UCL = D + ts_D \tag{16}$$

$$LCL = D - ts_D \tag{17}$$

where t is the upper 100(1-α) % percentile of the Student's t distribution with (n − 1) degrees of freedom.

7.2 *Example for Means Equivalence*—Total organic carbon in purified water was measured by an on-line analyzer, wherein a water sample was taken directly into the analyzer from the pipeline through a sampling port and the test result was determined by a series of operations within the instrument. A new analyzer was to be qualified by running a TOC analysis at the same time as the current analyzer utilizing a parallel sampling port on the pipeline. The sampling time was the blocking factor, and the data from the two instruments constituted a pair of single test results measured at a particular sampling time. Sampling was to be conducted at a frequency of four hours between sampling periods.

An equivalence limit of 2 parts per billion (ppb), or 4 % of the nominal process average of 50 ppb, was proposed with a consumer risk of 5 %. A repeatability estimate of s_r = 0.7 ppb, based on previous validation work, gave an estimate for $\sigma_d = 0.7\sqrt{2}$ or approximately 1 ppb. Thus E = 2 ppb, α = 0.05, and σ_d = 1 ppb were inputs for this study.

7.2.1 *Sample Size Determination*—Because the paired samples design uses the differences of the test results within sampling periods for data analysis, the sample size equals the number of pairs for purposes of calculating the power curve. In this example, the cost of obtaining test results was not a major consideration once the new analyzer was installed in the system. Comparative power profiles for n = 10, 20, and 50 sample pairs are shown in Fig. 2. The sample size of 20 pairs yielded a satisfactory power curve, in that the probability of accepting equivalence was greater than a 0.9 (or a 90 % power) for a true difference of about 1.25 ppb. Therefore, there would be less than an estimated 10 % risk to the producer that such a difference would fail to support equivalence in the actual trial.

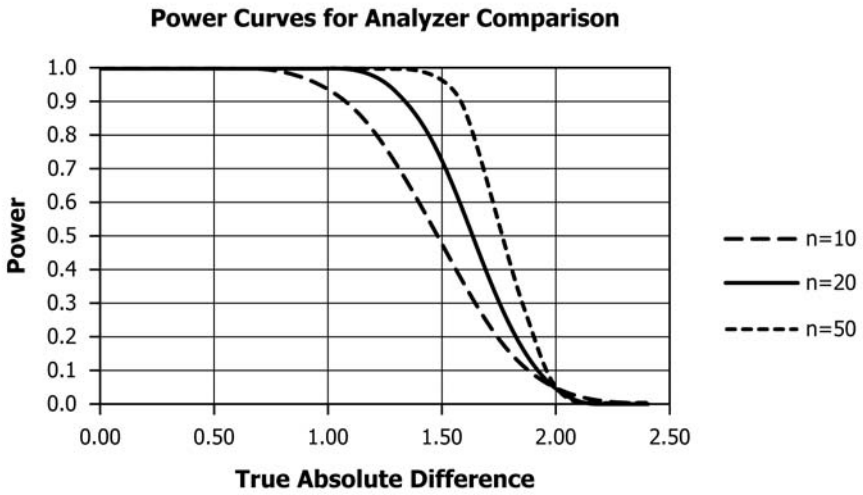

FIG. 2 Power Curves for Total Organic Carbon Analyzers Comparison

7.2.2 Test results for the two instruments at each of the 20 sampling times are listed in Table 2. The current analyzer was designated as Instrument A, and the new analyzer was designated as Instrument B. The differences d_j at each sampling time period were calculated and listed in Table 2 as differences in the test results of Instrument B minus Instrument A. The averages and standard deviations of the test results for each analyzer and their differences are also listed in Table 2.

7.2.2.1 The average difference $\bar{d}$ was 0.46 ppb and the standard deviation of the differences s_d was 1.05 ppb with f = 19 degrees of freedom. The standard error of the average difference was:

$$s_D = \frac{1.05}{\sqrt{20}} = 0.235 \text{ ppb}$$

7.2.2.2 Note that the standard deviations of test results for each analyzer over time were about 6 ppb due to process fluctuations in a range of 37–59 ppb. The source of variation due to blocks (sampling times from the process) is eliminated in the variation of the differences by pairing the test results.

7.2.3 The 95th percentile of Student's t with 19 degrees of freedom was 1.729. Upper and lower confidence limits for the difference of means were:

$$UCL = D + ts_D = 0.46 + (1.729)(0.235) = 0.87 \text{ ppb}$$
$$LCL = D - ts_D = 0.46 - (1.729)(0.235) = 0.05 \text{ ppb}$$

The 90 % two-sided confidence interval on the true difference is 0.05 to 0.87 ppb and is completely contained within the equivalence interval of –2 to 2 ppb. Since 0.05 > –2 and 0.87 < 2, equivalence of the two analyzers is accepted.

8. The TOST Procedure for Statistical Analysis of Bias Equivalence

8.1 *Statistical Analysis*—A number of tests are conducted on a certified reference material (CRM) in a laboratory. The average of the test results is compared with the accepted reference value (ARV) for that material. Let the data be denoted as X_i = the ith test result. The format is similar to that for the means equivalence example in Section 6, but the CRM substitutes for the first population, and its ARV is treated as a known constant. This assignment gives the correct sign for the test method bias.

TABLE 2 Data for Paired Samples Equivalence Test

Sampling Time	TOC in Water, ppb		
	Inst A	Inst B	Diff
1	46.4	48.8	2.4
2	44.2	43.5	–0.7
3	52.4	53.0	0.6
4	37.6	37.3	–0.3
5	49.3	49.1	–0.2
6	45.0	44.5	–0.5
7	51.4	51.3	–0.1
8	57.6	56.8	–0.8
9	43.4	44.9	1.5
10	45.2	44.1	–1.1
11	59.0	58.5	–0.5
12	43.1	44.1	1.0
13	39.3	40.9	1.6
14	48.2	48.4	0.2
15	48.7	49.0	0.3
16	44.4	46.1	1.7
17	52.7	53.2	0.5
18	43.3	44.6	1.3
19	54.4	56.7	2.3
20	58.4	58.4	0.0
Average	48.20	48.66	0.46
Std Dev	6.13	5.99	1.05

8.1.1 The equivalence limit E, consumer's risk α, and sample sizes have been previously determined. Calculate the average, estimated bias, standard deviation, and degrees of freedom:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \quad (18)$$

$$B = \bar{X} - ARV \quad (19)$$

$$s = \sqrt{\sum_{i=1}^{n} (X_i - \bar{X})^2 / (n - 1)} \quad (20)$$

$$f = (n - 1) \text{ degrees of freedom (d f)} \quad (21)$$

8.1.2 Calculate the standard error of the bias:

$$s_B = s / \sqrt{n} \quad (22)$$

8.1.3 *Test for Equivalence*—Calculate upper and lower confidence limits:

$$UCL = B + ts_B \quad (23)$$
$$LCL = B - ts_B \quad (24)$$

where t is the upper 100(1–α) percentile of the Student's t distribution with $(n_1 - 1)$ degrees of freedom.

If the 100(1–2α) two-sided confidence interval on the true difference is completely contained within the equivalence limits (0 ± E), equivalently if $LCL > -E$ and $UCL < E$, equivalence is accepted. Otherwise, reject equivalence.

8.2 *Example for Bias Equivalence*—The accepted reference value for the test material was given as 49.50 % by weight (wt%). An estimate of the repeatability precision from the method development validation was 1.5 wt%. An equivalence limit of 3.0 wt% was selected, based on the specification range for that material, at 5 % consumer risk. Thus E = 3 wt%, α = 0.05, and estimated σ = 1.5 wt% are inputs for this study.

8.2.1 *Sample Size Determination*—Power profiles for n = 5, 12, and 30 were generated for a set of absolute difference values ranging 0.00 (0.25) 4.00 wt% as shown in Fig. 3. All three curves intersect at the point (3, 0.05) as determined by the consumer's risk at the equivalence limit.

8.2.1.1 A sample size of 12 replicate assays yields a satisfactory power curve, in that the probability of accepting equivalence (power) was greater than a 0.9 probability (or a 90 % power) for a difference of 1.75 wt% or less. Therefore, there would be less than an estimated 10% risk to the producer that such a difference would fail to support equivalence in the actual trial.

8.2.1.2 A comparison of the three power curves indicates that the n = 5 design would be underpowered, as the power falls below 0.9 at 1.0 wt%. The n = 30 design gives somewhat more power than the n = 12 design but is more costly to conduct and may not be worth the extra expenditure.

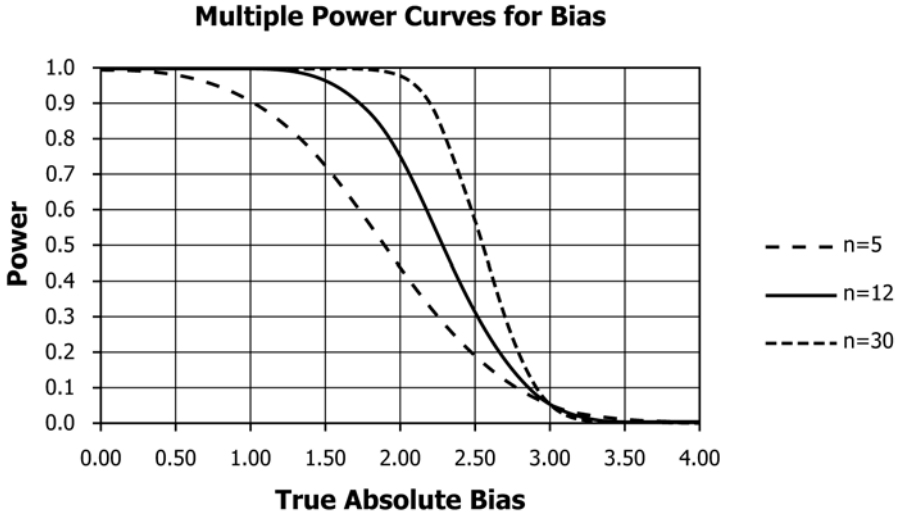

FIG. 3 Multiple Power Curves for Bias Example

8.2.2 Results for the twelve replicate assays are given in Table 3. The laboratory mean, the bias, the laboratory standard deviation, its degrees of freedom, and the standard error of the bias are:

$$\bar{X} = (48.5 + 51.0 + ... + 48.9) = 50.49 \text{ wt\%}$$

$$B = 50.49 - 49.50 = 0.99 \text{ wt\%}$$

$$s = \sqrt{[(48.5 - 50.49)^2 + ... + (48.9 - 50.49)^2]/(12 - 1)} = 1.935 \text{ wt\%}$$

$$f = 12 - 1 = 11$$

$$s_B = 1.935 / \sqrt{12} = 0.559 \text{ wt\%}$$

8.2.3 The 95th percentile of Student's t with 11 degrees of freedom is 1.796. Upper and lower confidence limits are:

UCL = 0.99 + (1.796)(0.559) = 1.99 wt%
LCL = 0.99 – (1.796)(0.559) = –0.01 wt%

8.2.4 Since –0.01 > –3 and 1.99 < 3, equivalence is accepted.

9. Procedure for Statistical Analysis of Non-Inferiority Tests Involving Means and Variances

9.1 *Statistical Analysis Involving Means*—The calculations for non-inferiority tests are essentially the same as for means equivalence with the following exceptions.

(1) The means being compared are from values of a performance characteristic, not necessarily the test result means.

(2) The scale for a performance characteristic is directional, one direction denoting inferiority of the of the modified test procedure. Thus only a single one-sided test is conducted.

TABLE 3 Data for Bias Equivalent Test

Test Results					
48.5	51.0	54.0	53.2	47.6	49.4
50.2	49.5	52.1	51.6	49.9	48.9

9.1.1 Depending on the experimental design that was used, calculate the upper (UCL) and lower (LCL) confidence limits on the difference between means. For the Two Independent Samples Design, use the calculations in 6.1. For the Paired Samples Design, use the calculations in 7.1.

9.1.2 For a performance characteristic where "higher is better", accept noninferiority for the modified test procedure with respect to the current test procedure when $LCL > -E$; otherwise denote inferiority for the modified test procedure.

9.1.3 For a performance characteristic where "lower is better", accept noninferiority for the modified test procedure with respect to the current test procedure when $UCL < E$; otherwise denote inferiority for the modified test procedure.

9.2 *Example—Non-Inferiority Test for Sensitivity of Detection*—Environmental testing for microbial contamination by the current (compendial) test method involves counting microbial colony-forming units (CFU) after plating and incubating the sample for a period of days. Newer rapid test methods give a result in shorter time and so have benefits in timeliness even though they might have slightly lower detection sensitivity than the compendial method. Therefore, the performance characteristic, sensitivity, is "higher is better" and the non-inferiority test is based on $LCL > -E$.

9.2.1 In this example the acceptance criterion is based on a ratio rather than on a difference. The industry standard USP <1223> stipulates that "The alternate method should provide an estimate of viable microorganisms not less than 70 % of the estimate provided by the traditional method …", thus the noninferiority limit for the ratio of the CFU counts (rapid/compendial) would be $-E = 0.7$. For this situation, a logarithmic transformation gives a natural scale for this acceptance criterion in terms of a mean difference. Let $\bar{X}_1 =$ the average count by the rapid method and let $\bar{X}_2 =$ the average count by the compendial method. In the log metric, the log of the ratio is equal to the difference in the log means thus $\log_{10}(\bar{X}_1 / \bar{X}_2) = \log_{10}(\bar{X}_1) - \log_{10}(\bar{X}_2) = D$. Therefore, the equivalence limit $-E$ is equal to $\log_{10}(0.7) = -0.1549$ in the log metric.

9.2.2 Eighteen independent bioassays were conducted at the same time, nine each by the compendial and rapid test methods, sampling from a single microorganism suspension having approximately 50 CFU. This was a Two Independent Samples Design. The 6.1 calculations were made on the log-transformed count data using an equivalence limit of $-E = -0.1549$, and these calculations are summarized in Table 4. The average recovery by the rapid method was lower than the compendial method (50.4 CFU versus 54.3 CFU) with a ratio of 0.928, or a 7.2 % reduction. The lower confidence limit (*LCL*) on the log difference *D* was –0.0828, which was higher than the equivalence limit –0.1549, and thus non-inferiority was supported.

9.2.3 Note that the use of a normal distribution for log counts was justified in this situation because the count range is small. This was confirmed by a normality test on each source of nine log-transformed counts (not shown here).

9.2.4 Fig. 4 shows a post-facto power curve based on $n = 9$, $\alpha = 0.05$, and $\sigma = 0.06$ log CFU. The curve intersects the point (–0.1549, 0.05) confirming that the power is 5 % at the given equivalence limit. Power is above 90 % at a log CFU Ratio down to near –0.1 (about a 20 % reduction in sensitivity) for this design. This supports the sample size that was used for this non-inferiority test.

9.3 *Statistical Analysis Involving Variances*—The test for non-inferiority of precision is conducted using variances as the test statistic. Non-inferiority tests are used for variances because precision is a performance characteristic in which "smaller is better". The statistical procedure is based on a one-sided F test. The proper design is the Two Independent Samples Design, so use the calculations in 6.1, Eq 2-4, for the variances.

9.3.1 Calculate the ratio *R* of the variances (modified/current) and its upper confidence limit, where s_1^2 is the variance estimate of the current procedure with f_1 degrees of freedom and s_2^2 is the variance estimate of the modified procedure with f_2 degrees of freedom:

$$R = s_2^2/s_1^2 \tag{25}$$

The upper confidence limit for the true variance ratio $\mathcal{R} = \sigma_2^2/\sigma_1^2$ is:

$$UCL_R = R\ F_{1-\alpha} \tag{26}$$

where $F_{1-\alpha}$ is the upper 100(1-α)th percentile of the F distribution with f_1 and f_2 degrees of freedom (see X1.5.3).

9.3.2 *Test for Non-Inferiority of Population 2 Precision*—Because precision stated inversely as variance is a performance characteristic where "lower is better", accept noninferiority for the modified test procedure with respect to the current test procedure when $UCL_R < E$; otherwise denote inferiority for the modified test procedure.

9.3.3 The needed sample sizes for variance tests will be much larger than those for means. It will usually be difficult, if not impossible, to generate 30 or more test results at the same time by each test method under repeatability conditions. This means that the tests will be conducted under intermediate precision conditions using a set of control samples that are homogeneous and stable. Fig. 5 shows power curves for equal sample sizes $n = 31$, 51, and 101 (degrees of freedom, $f = 30$, 50, and 100) with $\alpha = 0.05$ and $E = 4$. A power of 0.8 can be attained at $\mathcal{R} = 1.6$, 2.0, and 2.4, respectively. Note that the three power profile curves intersect at the point (4, 0.05).

9.3.4 If a control sample is unavailable, an alternate design would be to run duplicate tests by each test method on a series of routine samples. Each duplicate will provide a one degree of freedom estimate of test variance under repeatability conditions. These variances can then be pooled to obtain a repeatability estimate for each test method.

TABLE 4 Data and Calculations for Non-inferiority Test on Microbial Recovery

	Counts, CFU		Log Counts		Equation Number
	Comp	Rapid	Comp	Rapid	
Data	53	59	1.7243	1.7709	
	62	53	1.7924	1.7243	
	61	41	1.7853	1.6128	
	43	60	1.6335	1.7782	
	54	58	1.7324	1.7634	
	47	47	1.6721	1.6721	
	66	46	1.8195	1.6628	
	54	43	1.7324	1.6335	
	49	47	1.6902	1.6721	
n	9	9	9	9	
Average	54.3	50.4	1.7313	1.6989	Eq 1
Std Dev	7.52	7.21	0.0605	0.0620	Eq 3
Degrees of Freedom, *f*			8	8	Eq 4
Pooled Standard Deviation				0.0612	Eq 5
Degrees of Freedom				16	Eq 6
Difference (Rapid-Comp)				–0.0325	Eq 7
Standard Error of Difference				0.0289	Eq 8
95 % Confidence Limit:					
Student's *t*, *f* = 16, 95th Percentile				1.746	
Lower Confidence Limit				–0.0828	Eq 10
Equivalence Limit				–0.1549	
Non-Inferiority Test				Pass	

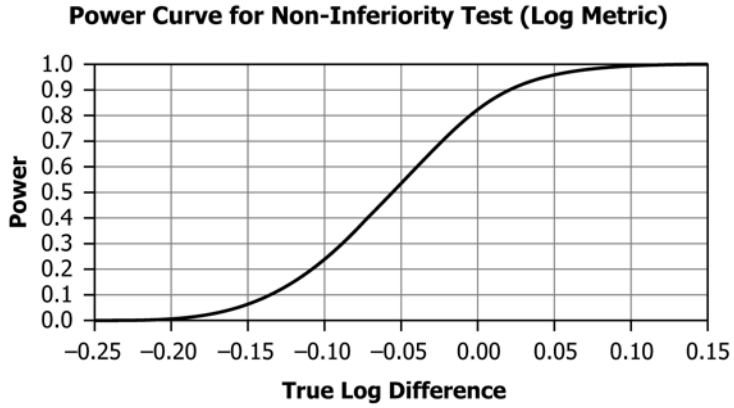

FIG. 4 Power Curve for Microbial Detection Example

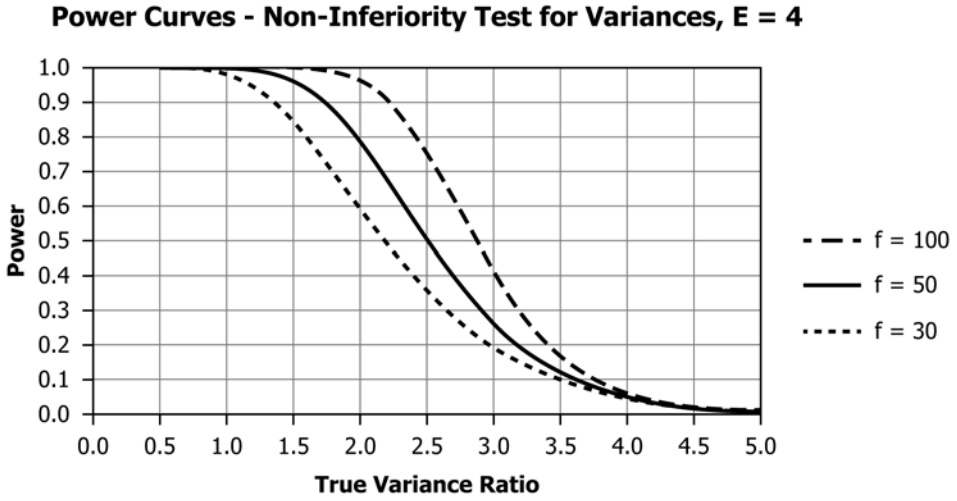

FIG. 5 Power Curves for Variance Non-Inferiority Tests where $E = 4$

10. Keywords

10.1 bias equivalence; confidence interval; equivalence; equivalence limit; means equivalence; non-inferiority; two one-sided tests (TOST) procedure

APPENDIXES

(Nonmandatory Information)

X1. STATISTICAL HYPOTHESIS TESTS FOR EQUIVALENCE

X1.1 Two One-Sided Tests (TOST) Procedure (1)

X1.1.1 Data from two populations (sources) are assumed to arise independently from normally distributed populations having distinct means, denoted as μ_1, μ_2, and a common standard deviation, denoted as σ. The TOST procedure sets up two null hypotheses (H_0) and corresponding alternate hypotheses (H_a) on the difference between the two population means as follows:

	Hypothesis 1	Hypothesis 2
Null hypothesis	H_{01}: $\mu_2-\mu_1 \geq E$	H_{02}: $\mu_2-\mu_1 \leq -E$
Alternative hypothesis	H_{a1}: $\mu_2-\mu_1 < E$	H_{a2}: $\mu_2-\mu_1 > -E$

The value E is termed the equivalence limit, representing the worst case difference between the two means.

X1.1.2 The TOST procedure is carried out using the data sampled from the two populations, as illustrated in 6.1 with an example. A one-sided t test at the α significance level tests each of the two null hypotheses.

Let $D=\bar{X}_2-\bar{X}_1$ and $s_D=s_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$

where $s_p=\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{(n_1+n_2-2)}}$, with $f=(n_1+n_2-2)$ degrees of freedom.

The t statistics are $t_1=(E-D)/s_D$ and $t_2=(E+D)/s_D$ for hypotheses 1 and 2, respectively. Both null hypotheses are rejected when $t_1>t_{1-\alpha,f}$ and $t_2>t_{1-\alpha,f}$ where $t_{1-\alpha,f}$ is the upper $(1-\alpha)$th quantile of the Student's t distribution with f degrees of freedom. If both hypotheses are rejected, then it is asserted that $-E<\mu_1-\mu_2<E$ and the two sources are said to be equivalent; otherwise, the two data sources are deemed non-equivalent.

X1.1.3 The TOST procedure is operationally identical to constructing a two-sided $100(1-2\alpha)$ % confidence interval on the difference between two means (2). If the confidence interval is completely contained within the interval $(-E, E)$ then equivalence is accepted. The interval $(-E, E)$ is termed the equivalence interval.

X1.1.4 It is strongly recommended (5) that the sample sizes from each population be equal to minimize the effect of a departure from equal population variances. If the variances differ greatly the standard error of the difference may be calculated as:

$$s_D=\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}} \tag{X1.1}$$

With approximate degrees of freedom:

$$v=\frac{\left(s_1^2/n_1+s_2^2/n_2\right)^2}{\left[\frac{\left(s_1^2/n_1\right)^2}{(n_1-1)}+\frac{\left(s_2^2/n_2\right)^2}{(n_2-1)}\right]} \tag{X1.2}$$

In many statistical software packages this calculation is used in the option "assume unequal variances" for a t test. The resulting degrees of freedom are bounded between $MIN(n_1-1, n_2-1)$ and n_1+n_2-2.

X1.2 Decision Errors and Risks

X1.2.1 In any statistical hypothesis testing situation a decision is made to either accept or reject the null hypothesis based on outcome of the procedure. Since the data are subject to variation, this will create uncertainty in the final decision. There are two kinds of errors associated with the final decision:

(1) Rejecting the null hypothesis when it is true (Type I error), and

(2) Not rejecting the null hypothesis when it is false (Type II error).

X1.2.2 For the equivalence application of a hypothesis test, the null hypothesis is that the two populations are *not* equivalent, so the Type I error is declaring equivalence when the two populations are truly not equivalent. The Type I error is considered a consumer's risk, since acceptance of a non-equivalent testing process will affect customers (patients, regulators, etc.) by creating erroneous test results in release of product and other quality management activities. This risk is set by choosing the significance level of the two hypothesis tests in the TOST procedure, so that the consumer's risk is directly controlled.

X1.2.3 The Type II error is failing to declare equivalence when the two populations are truly equivalent. The Type II error is considered a producer's risk, since this will create additional investigational work to make a desired improvement. This risk is controlled by choosing an adequate sample size to be taken from each population by consideration of power profiles from various sample sizes.

X1.2.4 The table below summarizes the four situations that may occur for a given TOST procedure.

	Populations are truly:	
TOST declares that:	Equivalent	Not Equivalent
Populations are equivalent	Decision is correct	Type I Error
Populations are not equivalent	Type II Error	Decision is correct

X1.3 Criticism of the Use of the Conventional *t* Test for Equivalence Testing

X1.3.1 In the conventional two sample t test a single hypothesis test is set up as follows:

Null hypothesis	H_0: $\mu_1-\mu_2=0$
Alternative hypothesis	H_a: $\mu_1-\mu_2\neq 0$

The null hypothesis is rejected if the two-sided confidence interval on the difference between the population means excludes zero and is not rejected if the confidence interval includes zero. If used for equivalence testing, equivalence would be rejected if the null hypothesis was rejected. This is

operationally the same as rejecting the null hypothesis if the two-sided confidence interval on the mean does not include zero.

X1.3.2 The Type I Error for the *t* test is the error of falsely declaring a non-zero difference, or the error of falsely declaring non-equivalence, which is the producer's risk. As hypothesis tests are set up to directly control the Type I error (often at the 0.05 significance level) the conventional *t* test is not directly protecting the customer in the equivalence application. The consumer's risk is indirectly controlled by the samples sizes selected.

X1.3.3 If the variances of the population means are small, either reflecting a precise test method, large sample sizes, or both, the confidence interval on the difference may not include zero, thus rejecting equivalence, even for small differences that are not of scientific importance. On the other hand, if the variances of the population means are large, the confidence interval on the difference may include zero, but may be extremely wide, thus masking critical differences. For these reasons, the conventional *t* test is not recommended for equivalence testing.

X1.4 Equivalence Testing for Bias

X1.4.1 The TOST procedure may also be used for bias equivalence testing. In this situation population mean μ_1 is the accepted reference value (*ARV*) with zero variance. The experiment consists of comparing μ_2 with the *ARV*. The population mean is re-designated as μ and the sample mean and variance calculated for the single data set is used for estimating the bias, $\mu - ARV$, and its confidence limits for testing against the equivalent limit, or worst-case bias. The only change from the two population case is the calculation of the standard error and its degrees of freedom.

X1.5 Equivalence Testing for Non-Inferiority

X1.5.1 Non-inferiority in this practice compares a modified testing process to the current process with respect to a performance characteristic, where the acceptance criterion is stated in terms of a difference in means or a ratio of variances. The statistical procedure for non-inferiority testing uses a single one-sided hypothesis test where the null hypothesis states that the modified testing process is inferior to the current process. If the null hypothesis is rejected, the modified process is declared non-inferior to the current process for that performance characteristic.

X1.5.2 For performance characteristics comparing means, the hypothesis sets in X1.5.1 are used with μ_1 defined as the mean of the current process and μ_2 defined as the mean of the modified process. For an acceptance criterion where "lower is better" use Hypothesis 1, and for an acceptance criterion where "higher is better" use Hypothesis 2. The TOST procedure will supply the necessary one-sided hypothesis test calculations.

X1.5.3 For performance characteristics comparing variances, with σ_1^2 as the variance of the current process and σ_2^2 as the variance of the modified process, the hypothesis set is $H_0: \sigma_2^2 \geq E\,\sigma_1^2$ and $H_A: \sigma_2^2 \leq E\,\sigma_1^2$. The equivalence limit E is set in the form of the ratio $\mathcal{R} = \sigma_2^2/\sigma_1^2$ that represents the worst case increase of variance.

X1.5.3.1 The statistical test involves the ratio $R = s_2^2/s_1^2$, using s_1^2 as the variance estimate of the current procedure with f_1 degrees of freedom and s_2^2 as the variance estimate of the modified procedure with f_2 degrees of freedom, is the test statistic for the one-sided F test. The acceptance criterion for non-inferiority is $R\,F_{1-\alpha} < E$, where $F_{1-\alpha}$ is the upper 100(1-α)th percentile of the F distribution with f_1 and f_2 degrees of freedom.

X1.5.4 A reference for non-inferiority procedures is M. Rothmann, et. al. (**6**). Although their context is directed to clinical trials for pharmaceuticals, many numerical examples are included, and these are easily translatable to test method evaluation.

X1.6 Power Profiles

X1.6.1 The power function of the means equivalence test has been examined (**7, 8**) where the emphasis is on finding a sample size *n* for a given value of the true difference in means. Power functions involving the non-central and central Student's *t* distributions were considered, along with incorporating an upper confidence limit on the variance estimate with a normal distribution power function. The normal distribution approximation should be adequate when a strong estimate of σ is used (or the use of an upper confidence limit on σ if a more conservative estimate is desired.) The normal approximation of power, given a true difference Δ, for equal sample sizes *n* is:

$$Power = \Phi\left(\frac{E-\Delta}{\sigma_D} - z_{1-\alpha}\right) - \Phi\left(\frac{-E-\Delta}{\sigma_D} + z_{1-\alpha}\right) \quad \text{(X1.3)}$$

where:

$\Phi(\bullet)$ = the standard normal cumulative distribution function,
Δ = $\mu_1 - \mu_2$, the true difference parameter,
σ_D = $\sigma\sqrt{2/n}$, the standard error of the test statistic *D*, and
$z_{1-\alpha}$ = the (1–α)th percentile of the standard normal distribution.

If the sample sizes are too small, the upper confidence limit on $-E$ may exceed the lower confidence limit on E, and there will be a zero chance of accepting equivalence.

X1.6.2 The power function for the non-inferiority test for means depends on the direction of inferiority and uses the appropriate part of equation (Section X1.7).

For a performance characteristic where "higher is better use:

$$Power = 1 - \Phi\left(\frac{-E-\Delta}{\sigma_D} + z_{\alpha}\right) \quad \text{(X1.4)}$$

For a performance characteristic where "lower is better use:

$$Power = \Phi\left(\frac{E-\Delta}{\sigma_D} - z_{\alpha}\right) \quad \text{(X1.5)}$$

X1.6.3 The power function, plotted against values of $\mathcal{R} = \sigma_2^2/\sigma_1^2$, of the non-inferiority test for variances uses the F distribution:

$$Power = 1 - \mathcal{F}(\mathcal{R}\,F_{1-\alpha}/E) \quad \text{(X1.6)}$$

where:

$\mathcal{F}(\bullet)$ = the cumulative F distribution function with f_1 and f_2 degrees of freedom,

E = the equivalence limit expressed as the hypothesized ratio σ_2^2/σ_1^2, and

$F_{1-\alpha}$ = the upper 100(1-α)th percentile of the F distribution with f_1 and f_2 degrees of freedom.

X1.7 Alternative Designs

X1.7.1 Designs conducted using intermediate precision conditions may involve other sources of variation, thus making the analysis more complicated and possibly raising side issues, such as differences among operators or instruments within laboratories (**9**, **10**).

X2. SPREADSHEET FOR POWER PROFILE CURVES

X2.1 Power Profile for Means or Bias Equivalence Using a Single Sample or Two Independent Samples of Equal Sample Size

X2.1.1 *Data Entry*—A spreadsheet example for generating power profiles is shown in Fig. X2.1. See Section X1.6 for background information. Five input variables are entered into cells B3–B7 as follows:

- In B3, enter the estimate of the standard deviation of the test results, σ
- In B4, enter the consumer risk, α
- In B5, enter 1 for a single sample design or 2 for a two independent samples design
- In B6, enter the equivalence limit, *E*
- In B7, enter the sample size, *n*
- In A10, downward enter a range of true differences starting with zero and exceeding the equivalence limit, *E*, and adjust the horizontal axis of the graph accordingly.

X2.1.2 *Calculation*—Cells E3 and E4 list results for intermediate calculations of Z_α and σ_D. The power for a given true difference is calculated from *E*, Δ, Z_α, and σ_D, and the function equation for this appears in Row 24. The calculated power curve values will appear in B10 downward

X2.1.3 *Graph*—The graph plots the power on the vertical axis versus the absolute true difference on the horizontal axis. The curve is anchored at the point (*E*, α). For different ranges for the true difference the axes may have to be altered by the user.

X2.2 *Disclaimer*—This spreadsheet example is not supported by ASTM, and the user of this standard is responsible for its use. For questions pertaining to use of this spreadsheet example please contact Subcommittee E11.20.

	A	B	C	D	E	F	G	H	I
1	**Power curve for one sample or two independent samples case with equal n**								
2	**Input Variables**			***Intermediate Calculations***					
3	**Standard Deviation**	**0.5**		***Z alpha***	1.645	=NORMSINV(1-B4)			
4	**Consumer Risk**	**0.05**		***Std Error***	0.289	=B3*SQRT(B5/B7)			
5	**Design (# samples)**	**2**							
6	**Equivalence Limit**	**2**							
7	**Sample size, n**	**6**							
8									
9	**True difference**	**Power**							
10	0.00	1.0000							
11	0.20	1.0000							
12	0.40	1.0000							
13	0.60	0.9993							
14	0.80	0.9940							
15	1.00	0.9656							
16	1.20	0.8700							
17	1.40	0.6677							
18	1.60	0.3977							
19	1.80	0.1705							
20	2.00	0.0500							
21	2.20	0.0097							
22	2.40	0.0012							
23									
24	Power: B10=NORMSDIST(((B$6-A10)/E$4)-E$3)-NORMSDIST(((-B$6-A10)/E$4)+E$3)								
25									

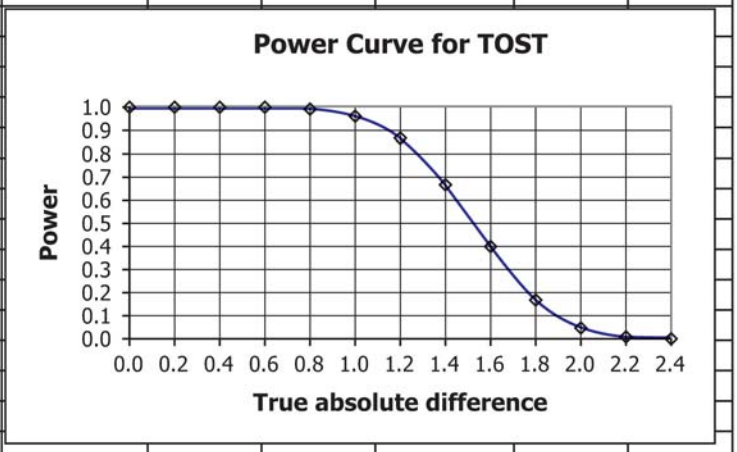

FIG. X2.1 Power Profile Spreadsheet Example for Means

REFERENCES

(1) Schuirmann, D. J.,"A Comparison of the Two One-sided Tests Procedure and the Power Approach for Assessing the Equivalence of Average Bioavailability," *Journal of Pharmacokinetics and Biopharmaceutics*, Vol 15, 1987, pp. 657–680.

(2) Westlake, W. J.,"Response to T. B. L. Kirkwood: Bioequivalence Testing – A Need to Rethink," *Biometrics*, Vol 37, 1981, pp. 589–594.

(3) Limentani, G. B., Ringo, M. C., Ye, F., Bergquist, M. L., and McSorley, E. O., "Beyond the *t*-Test: Statistical Equivalence Testing," *Analytical Chemistry*, June 1, 2005, pp. 221A–226A.

(4) Chambers, D., Kelly, G., Limentani, G., Lister, A., Lung, K. R., and Warner, E., "Analytical Method Equivalency – An Acceptable Analytical Practice," *Pharmaceutical Technology*, September 2005, pp. 64–80.

(5) Welch, B. L., "The Significance of the Difference Between Two Means When the Population Variances are Unequal," *Biometrika*, Vol 29, 1938, pp. 350–362.

(6) Rothmann, M. D., Wiens, B. L., and Chan, S. F. I., *Design and Analysis of Non-Inferiority Trials*, Chapman & Hall/CRC, Taylor & Francis Group, Boca Raton, FL, 2012.

(7) Bristol, D. R., "Probabilities and Sample Sizes for the Two One-sided Tests Procedure," *Communications in Statistics – Theory Methods*, Vol 22, 1993, pp. 1953–1961.

(8) Stein, J., and Doganaksoy, N., "Sample Size Considerations for Assessing the Equivalence of Two Process Means," *Quality Engineering*, Vol 12, No. 1, 1999, pp. 105–110.

(9) Kringle, R., Khan-Malek, R., Snikeris, F., Munden, P., Agut, C., and Bauer, M., "A Unified Approach for Design and Analysis of Transfer Studies for Analytical Methods," *Drug Information Journal*, Vol 35, 2001, pp. 1271–1288.

(10) Schwenke, J., and O'Connor, D., "Design and Analysis of Analytical Method Transfer Studies," *Journal of Pharmaceutical and BioSciences*, Vol 18, No. 5, 2008, pp. 1013–1033.