Designation: ~~E2935 – 15~~ 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,~~ laboratory or equivalence of two laboratories in a method transfer, and *(3)* ~~equivalence of two laboratories in a method transfer.~~determining non-inferiority of a modified test procedure versus a current test procedure with respect to a performance characteristic.

1.3 The ~~current~~ guidance in this standard applies only to experiments conducted on a single ~~material. Guidance is given for determining the amount of data required for an equivalence trial.~~material at a given level of the test result.

1.4 ~~The statistical methodology~~Guidance is given for determining ~~equivalence used is the two one-sided tests (TOST) procedure.~~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 $Pr(L \leq \theta \leq U) \geq 1 - \alpha$. **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 ~~Oct. 1, 2015~~Nov. 15, 2016. Published ~~October 2015~~January 2017. Originally approved in 2013. Last previous edition approved in ~~2014~~2015 as E2935 – ~~14.~~15. DOI: ~~10.1520/E2935-15.~~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.

3.1.3.1 *Discussion—*

The confidence level, 1 – α, 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 – α, 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**

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 (5.2.1)~~
E = equivalence limit (5.2)
~~E_1~~ = ~~lower equivalence limit (5.2.1.1)~~
E_1 = lower equivalence limit (5.2.1)
~~E_2~~ = ~~upper equivalence limit (5.2.1.1)~~
E_2 = upper equivalence limit (5.2.1)
~~H_0:~~ = ~~null hypothesis (X1.1.1)~~
~~H_A:~~ = ~~alternate hypothesis (X1.1.1)~~
f = degrees of freedom for s (8.1.1) (X1.1.2)
$F_{1-\alpha}$ = $(1-\alpha)$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 (5.3) (6.1.3) (7.1.1) (8.1.1)~~
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 *i*th population (6.1.1)
s_i^2 = sample variance for *i*th 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 *i*th 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 *i*th 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 (X1.5.1)~~
$Z_{1-\alpha}$ = (1-α)th percentile of the standard normal distribution (X1.6.1)
~~α = consumer's risk (5.2.2) (6.2) (7.2)~~
α = consumer's risk (5.2.3) (6.2) (7.2)
~~β = producer's risk (5.3)~~
β = producer's risk (5.4.1)
~~Δ = true mean difference between populations (5.3)~~
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 (5.2.3)~~
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 (X1.5.1)~~
$\Phi(\bullet)$ = standard normal cumulative distribution function (X1.6.1)

3.4 *Acronyms:*

3.4.1 *ARV, n*—accepted reference value (~~5.1.2~~5.3.3) (8.1) (X1.4)

3.4.2 *CRM, n*—certified reference material (~~5.1.2~~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 (~~4.3~~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 ~~evaluate test result bias, to evaluate changes for improving the test process performance, or to validate the transfer of a test method to a new location or apparatus. In all~~make improvements in their testing processes. In these situations it must be demonstrated that any ~~bias or innovation will have negligible effect on test results for a characteristic of a material.~~ 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 ~~confirm~~demonstrate that the ~~mean~~ test results from a modified testing process are equivalent to those from ~~a reference standard or another~~ the current testing process, where *equivalence* is defined as agreement within a prescribed ~~limits,~~limit, termed an *equivalence* ~~*limits.*~~*limit.*

4.1.1 ~~The intra-laboratory applications in this practice~~ Examples of modifications to the testing process include, but are not ~~limited to,~~limited, to the following:

(1) ~~Evaluating the bias of a test method with respect to a certified reference material,~~ Changes to operating levels in the steps of the test method procedure,

~~*(2)* Evaluating bias due to a minor change in a test method procedure,~~

(2) ~~Qualifying~~ Installation of new instruments, apparatus, or ~~operators in a laboratory, and~~sources of reagents and test materials,

(3) ~~Qualifying new sources of reagents or other materials used in the test procedure.~~Evaluation of new personnel performing the testing, and

(4) Transfer of testing to a new location.

4.1.2 ~~This practice also supports evaluating systematic differences in a method transfer from a developing laboratory to a receiving laboratory.~~ 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 ~~This practice currently deals only with the equivalence of population means. In this standard, a~~Two principal types of equivalence are covered in the practice, ~~*population*~~*means equivalence* and *non-inferiority.*~~refers to a hypothetical set of test results arising from a stable testing process that measures a characteristic of a single material.~~ 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.

~~NOTE 1—The equivalence concept can also apply to population parameters other than means, such as precision, stated as variances, standard deviations, or relative standard deviations (coefficients of variation), linearity, sensitivity, specificity, etc.~~

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 The data analysis for equivalence testing of population means in this practice uses a statistical methodology termed the two one-sided tests (TOST) procedure which shall be described in detail in this standard (see X1.1). The TOST procedure will be adapted to the type of objective and experiment design selected.~~

~~4.3.1 Historically, this procedure originated in the pharmaceutical industry for use in bioequivalence trials (1, 2),[3] denoted as the Two One-Sided Tests Procedure, and has since been adopted for other applications, particularly in testing and measurement applications (3, 4).~~

~~4.3.2 The conventional Student's *t* test used for detecting differences is not recommended for equivalence testing as it does not properly control the consumer's and producer's risks for this application (see X1.3).~~

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 ~~probability of accepting equivalence when the actual bias or difference in means is equal to the equivalence limit. This probability is~~ 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 ~~differences in question are~~the true difference is less than the equivalence limit.

4.3.2 The producer's risk is the risk of falsely rejecting equivalence. ~~If~~ 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 ~~additional~~ unnecessary additional effort in improving the testing process.

5. Planning and Executing the Equivalence Study

5.1 ~~*Objectives and Design Selection*—~~This ~~practice supports two equivalence~~ section discusses the stages of conducting an equivalence test: (*1*) determining the information needed, (*2*~~study objectives:~~) setting up and conducting the study design, and (~~*(1)*~~*3* ~~determining~~) performing the ~~means equivalence~~statistical analysis of ~~test results from two testing processes or~~the resulting data. The study is usually ~~(2)~~conducted ~~determining the~~either in ~~bias equivalence~~a single ~~of a test method. In both objectives, two population means are compared for equivalence.~~ 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.

5.1.1 *Means Equivalence*—This study compares the average test result from the current testing process with the innovated process. A single material is selected, subdivided into test samples, and distributed for testing by each process. The material should be reasonably homogeneous, because inhomogeneity in the material will decrease the test precision.

5.1.1.1 *Design Types*—This practice provides recommendations for the design of a means equivalence experiment, and two basic designs are discussed. Section 6 discusses the two independent samples design, in which each population is sampled independently. Section 7 discusses the paired samples design in which pairs of single samples from each population are taken under different conditions of a second variable, such as time.

5.1.2 *Bias Equivalence*—This study requires a suitable quantity of a *certified reference material* (CRM) having an *accepted reference value* (ARV) for the material characteristic of interest. The ARV is considered as a known population mean with zero variability for the purpose of the equivalence study. The average of the test results conducted on the reference material is the population mean estimate to be compared with the ARV (see X1.4). Section 8 discusses the design and analysis for comparing a single sample mean with a standard value.

5.2 *Design Requirements—Inputs for*Prior information carrying out the statistical test of equivalence are the equivalence limitsrequired for the study design includes the equivalence limit *and*E, the consumer's risk. Additional inputs for designing the equivalence study are risk α, and an estimate of the test method precision and the producer's risk profile over selected differences in the means.σ.

5.2.1 The equivalence limits to be used in the TOST procedure are selected as the worst-case differences between the two population means For means equivalence tests there are two equivalence limits, $-E$ and E,are determined by the subject matter expert or by industry consensus. These limits are usually symmetrical around zero and then are denoted that are tested. Limits may be nonsymmetrical around zero, such as $-E_1$ and E.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.

5.2.1.1 In certain cases the limits may be asymmetrical and are then denoted by E_1 and E_2, where E_1 is usually a negative value. The producer's risk profile for this situation will not be treated in this practice.

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 α is the probability of falsely declaring equivalence and is usually set at a value of 0.05, representing may be determined by an industry norm or a regulatory requirement. A probability value often used is α = 0.05, which is a 5 % risk. Other risk levels may be selected, depending on circumstances.risk to the consumer that the study falsely declares equivalence.

5.2.3 The test method precision, σ, is stated as the standard deviation of the test method, or methods, used in the equivalence study. An estimate may be available from a method validation, an interlaboratory study, or other sources.

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 Determination—The number of test results,* *Sample size*n, from each population controls the producer's risk β of falsely rejectingin the design context refers to the number *equivalence*n at a given true mean difference, Δ. The producer's risk may be alternatively stated in terms of theof test results required by each testing process to manage the producer's risk. It is possible to use power,different or probability 1–β of properly accepting equivalence at a given value of Δ.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).

5.3.1 For symmetric equivalence limits, the power profile plots the probability of properly declaring equivalence versus the absolute value of Δ, due to the symmetry of the equivalence limits. This calculation can be performed using a spreadsheet computer package (see X1.5 and Appendix X2).

5.4.1 ~~An example of a set of power profiles is shown in Fig. 1. The probability scale for power on the vertical axis varies from 0 to 1. The power profile, a reversed S-shaped curve, should be close to a probability of 1 at zero~~The number of test results, symbol ~~*absoluten*~~, ~~difference and will decline to the consumer risk probability at an absolute difference of~~from each testing process controls the producer's risk β of falsely rejecting means equivalence ~~E.~~at ~~Power for absolute differences greater than~~ a given true mean difference, Δ. The producer's risk may be alternatively stated in terms of the ~~*E*~~*power,* ~~are less than the consumer risk and~~the probability 1–β of ~~decline asymptotically to~~correctly accepting equivalence ~~zero as the absolute difference increases.~~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–β 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.~~ 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 ~~A.~~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.

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 ~~half~~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.

~~5.3.4 For comparing two populations, equal numbers of test samples from each population are recommended. Equal numbers of replicate test results will assure nearly constant consumer risk even when there is a difference in the variability for the two data samples (see X1.1.4).~~

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

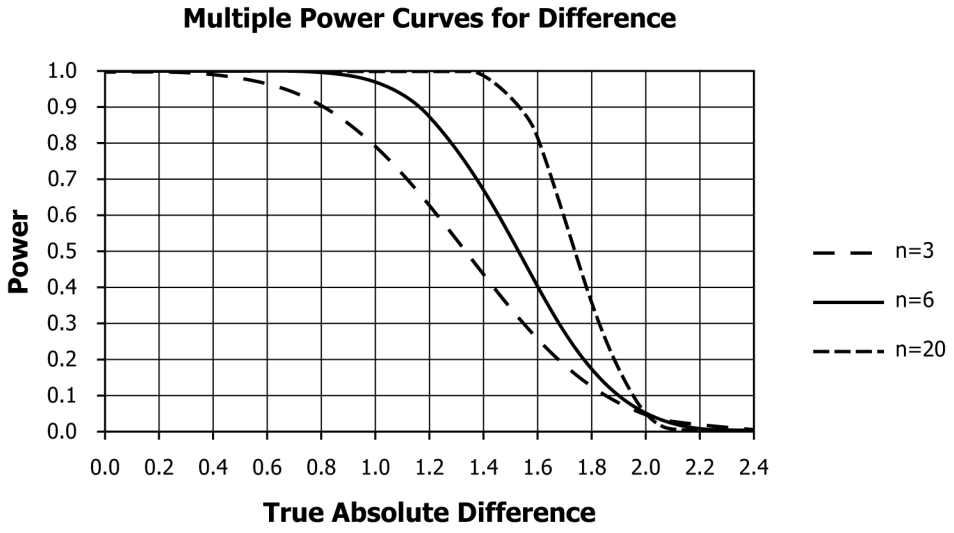

FIG. 1 Multiple Power Curves for Lab Transfer Example

(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**).

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:

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

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

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:

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

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

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).

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:

$$\bar{X}_1 = (96.9 + 97.9 + 98.5 + 97.5 + 97.7 + 97.2)/6 = 97.62 \text{ mg/g}$$
$$\bar{X}_2 = (97.8 + 97.6 + 98.1 + 98.6 + 98.6 + 98.9)/6 = 98.27 \text{ mg/g}$$

$$s_1^2 = [(96.9 - 97.62)^2 + \dots + (97.2 - 97.62)^2]/(6-1) = 0.31367$$
$$s_2^2 = [(97.8 - 98.27)^2 + \dots + (98.9 - 98.27)^2]/(6-1) = 0.26267$$

$$s_1 = \sqrt{0.31367} = 0.560$$
$$s_2 = \sqrt{0.26267} = 0.513$$

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

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,

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

symbol $\bar{d}$, and the standard deviation of the differences, symbol s_d, with its degrees of freedom, symbol f.

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

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

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

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

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.

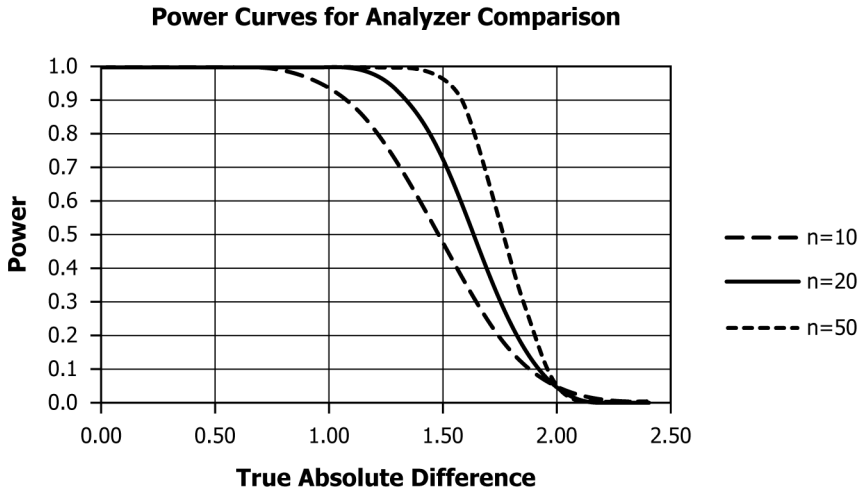

FIG. 2 Power Curves for Total Organic Carbon Analyzers Comparison