



Standard Practice for Conducting Equivalence Testing in Laboratory Applications¹

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 reappraisal. A superscript epsilon (ϵ) indicates an editorial change since the last revision or reappraisal.

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 are similar within predetermined limits.

1.2 Applications include (1) equivalence testing for bias against an accepted reference value, (2) determining equivalence of two test methods, test apparatus, instruments, reagent sources, or operators within a laboratory, and (3) equivalence of two laboratories in a method transfer.

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.

1.4 The statistical methodology for determining equivalence used is the “Two one-sided t -test” (TOST). 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

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 Aug. 1, 2013. Published August 2013. DOI: 10.1520/E2935-13.

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. 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**

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**

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* —similarity between two population parameters within 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 *population, n —the totality of items or units of material under consideration.* **E2586**

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

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

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

3.1.14 *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**

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

3.1.16 *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**

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

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

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

3.1.20 *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**

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

3.2.5.1 *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.*

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 situations it must be demonstrated that any bias or innovation will have negligible effect on test results for a characteristic of a material. This standard provides statistical methods to confirm that the mean test results from a testing process are equivalent to those from a reference standard or another testing process, where *equivalence* is defined as agreement within prescribed limits, termed *equivalence limits*.

4.1.1 The intra-laboratory applications in this practice include, but are not limited to, the following:

(1) Evaluating the bias of a test method with respect to a certified reference material,

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

(3) Qualifying new instruments, apparatus, or operators in a laboratory, and

(4) Qualifying new sources of reagents or other materials used in the test procedure.

4.1.2 This practice also supports evaluating bias in a method transfer from a developing laboratory to a receiving laboratory.

4.2 This practice currently deals only with the equivalence of population means. In this standard, a *population* refers to a hypothetical set of test results arising from a stable testing process that measures a characteristic of a single material.

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.

4.3 The data analysis for equivalence testing of population means in this practice uses a statistical methodology termed the “Two one-sided t -test” (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 Test, 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**).

4.4 This practice provides recommendations for the design of an equivalence experiment, and two basic designs are discussed. Guidance is provided for determining the amount of data required to control the risks of making the wrong decision in accepting or rejecting equivalence (see **X1.2**).

4.4.1 The consumer’s risk is the probability of accepting equivalence when the actual bias or difference in means is

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

equal to the equivalence limit. This probability is controlled to a low level so that accepting equivalence gives a high degree of assurance that differences in question are less than the equivalence limit.

4.4.2 The producer's risk is the risk of falsely rejecting equivalence. If improvements are rejected this can lead to opportunity losses to the company and its laboratories (the producers) or cause additional unnecessary effort in improving the testing process.

5. Planning the Equivalence Study

5.1 *Objectives and Design Selection*—This practice supports two equivalence study objectives: (1) determining the *bias equivalence* of a test method or (2) determining the *means equivalence* of test results from two testing processes. In both objectives two population means are compared for equivalence.

5.1.1 *Bias Equivalence*—This study requires a suitable quantity of a certified reference material 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).

5.1.2 *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.

5.2 *Design Requirements*—Inputs for carrying out the statistical test of equivalence are the equivalence limits and the consumer's risk. Additional inputs for designing the equivalence study are an estimate of the test method precision and the producer's risk profile over selected differences in the means.

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 and are determined by the subject matter expert or by industry consensus. These limits are usually symmetrical around zero and then are denoted as $-E$ and E .

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 The consumer's risk α is the probability of falsely declaring equivalence and is usually set at a value of 0.05, representing a 5% risk. Other risk levels may be selected, depending on circumstances.

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 *Sample Size Determination*—The number of test results, n , from each population controls the producer's risk β of falsely rejecting equivalence at a given true mean difference, Δ . The producer's risk may be alternatively stated in terms of the *power*, or probability $1-\beta$ of properly accepting equivalence at a given value of Δ .

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).

5.3.2 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 absolute difference and will decline to the consumer risk probability at an absolute difference of E . Power for absolute differences greater than E

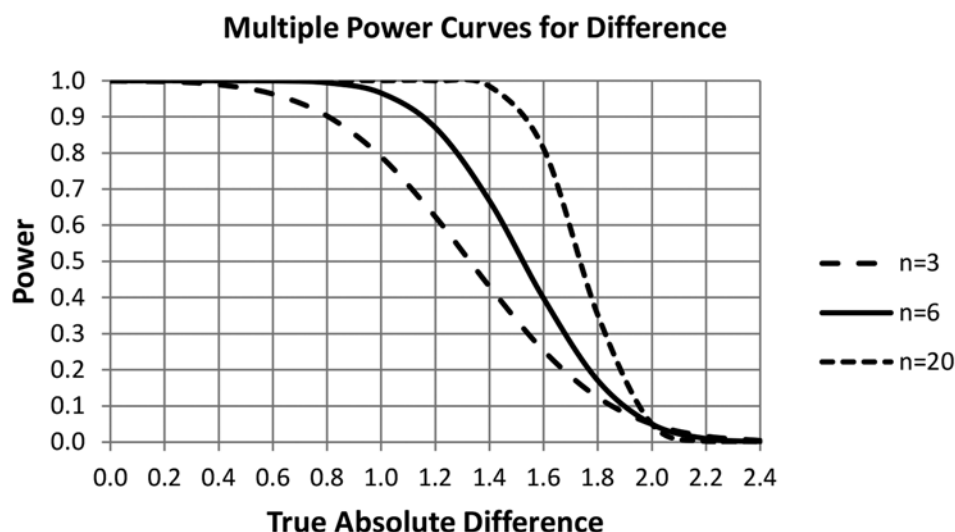


FIG. 1 Multiple Power Curves for Lab Transfer Example

are less than the consumer risk and decline asymptotically to zero as the absolute difference increases.

5.3.2.1 In Fig. 1 power profiles are shown for three different sample sizes. Increasing the sample size moves the power curve to the right, giving a greater chance of accepting equivalence for a given true difference Δ .

5.3.3 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 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).

6. The TOST Procedure for Statistical Analysis of Means Equivalence

6.1 *Statistical Analysis*—Let the data be denoted as X_{ij} = the j th test result from the i th 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 laboratory:

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

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

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

$$f_i = n_i - 1, \quad i = 1, 2 \quad (4)$$

6.1.2 Calculate the pooled standard deviation and degrees of freedom:

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

$$\text{If } n_1 = n_2 = n, \text{ then } s_p^2 = \frac{(s_1^2 + s_2^2)}{2} \\ f_p = (n_1 + n_2 - 2) \quad (6)$$

6.1.3 Calculate the difference between means and its standard error:

$$D = \bar{X}_2 - \bar{X}_1 \quad (7)$$

$$s_D = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \quad (8)$$

$$\text{If } n_1 = n_2 = n, \text{ then } s_D = s_p \sqrt{\frac{2}{n}}$$

6.1.4 *Test for Equivalence*—Compute the upper (UCL) and lower (LCL) confidence limits for the 100 (1–2 α) % two-sided confidence interval on the true difference. If the confidence interval is completely contained within the equivalence limits

(0 $\pm$ E), equivalently if $LCL > -E$ and $UCL < E$, then accept equivalence. Otherwise, reject equivalence.

$$UCL = D + ts_D \quad (9)$$

$$LCL = D - ts_D \quad (10)$$

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 = 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 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

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

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:

$$\begin{aligned} UCL &= 0.65 + (1.812)(0.310) = 1.21 \\ LCL &= 0.65 - (1.812)(0.310) = 0.09 \end{aligned}$$

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 Bias Equivalence

7.1 *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 i th 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.

7.1.1 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:

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} \quad (11)$$

$$B = \bar{X} - ARV \quad (12)$$

$$s = \sqrt{\sum_{i=1}^n (X_i - \bar{X})^2 / (n - 1)} \quad (13)$$

$$f = (n - 1) \text{ degrees of freedom (df)} \quad (14)$$

7.1.2 Calculate the standard error of the bias:

$$s_B = s / \sqrt{n} \quad (15)$$

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

$$UCL = B + ts_B \quad (16)$$

$$LCL = B - ts_B \quad (17)$$

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

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

7.2 *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%, $\alpha = 0.05$, and estimated $\sigma = 1.5$ wt% are inputs for this study.

7.2.1 *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. 2. All three curves intersect at the point (3, 0.05) as determined by the consumer's risk at the equivalence limit.

7.2.1.1 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.

7.2.1.2 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.

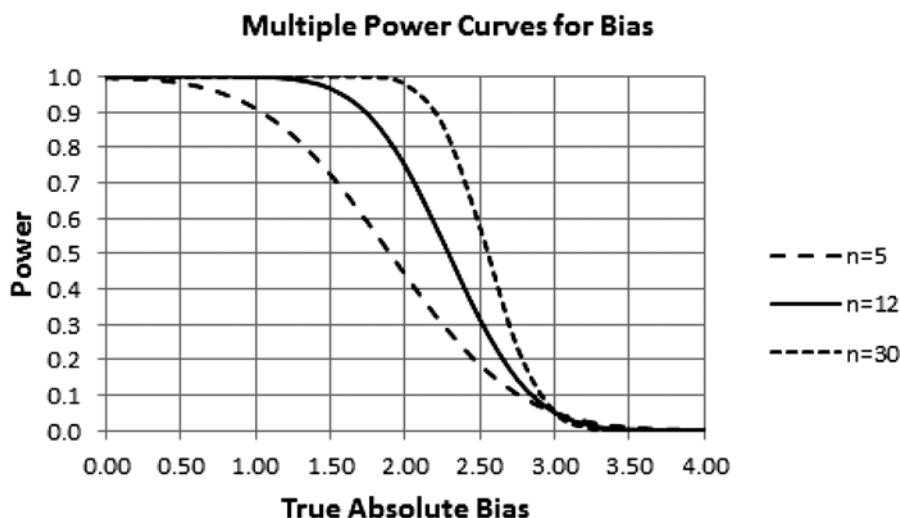


FIG. 2 Multiple Power Curves for Bias Example

7.2.2 Results for the twelve replicate assays are given in Table 2. The laboratory mean, the bias, the laboratory standard deviation, its degrees of freedom, and the standard error of the bias are:

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

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

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

$$f = 12 - 1 = 11$$

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

TABLE 2 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

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

$$UCL = 0.99 + (1.796)(0.559) = 1.99 \text{ wt\%}$$

$$LCL = 0.99 - (1.796)(0.559) = -0.01 \text{ wt\%}$$

7.2.4 Since $-0.01 > -3$ and $1.99 < 3$, equivalence is accepted.

8. Keywords

8.1 bias equivalence; confidence interval; equivalence; equivalence limit; means equivalence; two one-sided t test (TOST)

APPENDIX

(Nonmandatory Information)

X1. STATISTICAL HYPOTHESIS TESTS FOR EQUIVALENCE

X1.1 Two One-Sided t -Test (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.

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

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 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:

$$s_D = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \quad (\text{X1.1})$$

With approximate degrees of freedom (5):

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

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 $\text{MIN}(n_1 - 1, n_2 - 1)$ 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.

TOST declares that:	Populations are truly:	
	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, $\mu - \text{ARV}$, 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 Power Profiles

X1.5.1 The power function of the equivalence test was been examined (6, 7) where the emphasis is on finding a sample size n for a given value of the true difference. 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 good 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 is:

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

where:

- $\Phi(\bullet)$ = the standard normal cumulative distribution function,
- Δ = $\mu_1 - \mu_2$, the true difference parameter,
- σ_D = the standard error of the statistic D , based on σ and n , and
- z_α = the $(1-\alpha)$ 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 zero chance of accepting equivalence.

X1.6 Alternative Designs

X1.6.1 In addition to the "two independent samples design", an alternate is a "paired samples design", where, for example, two instruments measuring air particulates are placed in the same environment over multiple time periods. This constitutes a randomized blocks design where the blocks are "Times". The design and analysis of such experiments are outside the current scope of this practice.

X1.6.2 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 (8, 9).

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) 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.
- (7) 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.
- (8) 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.
- (9) Schwenke, J., and O’Connor, D., “Design and Analysis of Analytical Method Transfer Studies,” *J. Pharma. Biosciences*, Vol. 18, No. 5, 2008, pp. 1013-1033.

ASTM International takes no position respecting the validity of any patent rights asserted in connection with any item mentioned in this standard. Users of this standard are expressly advised that determination of the validity of any such patent rights, and the risk of infringement of such rights, are entirely their own responsibility.

This standard is subject to revision at any time by the responsible technical committee and must be reviewed every five years and if not revised, either reapproved or withdrawn. Your comments are invited either for revision of this standard or for additional standards and should be addressed to ASTM International Headquarters. Your comments will receive careful consideration at a meeting of the responsible technical committee, which you may attend. If you feel that your comments have not received a fair hearing you should make your views known to the ASTM Committee on Standards, at the address shown below.

This standard is copyrighted by ASTM International, 100 Barr Harbor Drive, PO Box C700, West Conshohocken, PA 19428-2959, United States. Individual reprints (single or multiple copies) of this standard may be obtained by contacting ASTM at the above address or at 610-832-9585 (phone), 610-832-9555 (fax), or service@astm.org (e-mail); or through the ASTM website (www.astm.org). Permission rights to photocopy the standard may also be secured from the ASTM website (www.astm.org/COPYRIGHT/).