Designation: E2935 – 17

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 to experiments conducted on a single material at a given level of the test result or on multiple materials covering a range of selected test results.

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, health, and environmental practices and determine the applicability of regulatory limitations prior to use.*

1.7 *This international standard was developed in accordance with internationally recognized principles on standardization established in the Decision on Principles for the Development of International Standards, Guides and Recommendations issued by the World Trade Organization Technical Barriers to Trade (TBT) Committee.*

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

E3080 Practice for Regression Analysis

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**

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.

[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, 2017. Published November 2017. Originally approved in 2013. Last previous edition approved in 2016 as E2935 – 16. DOI: 10.1520/E2935-17.

[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.

3.1.4 *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*—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 *range equivalence, n*—equivalence of two population means over a range of test result values.

3.2.9 *slope equivalence, n*—equivalence of the slope of a linear statistical relationship with the value one (1).

3.2.10 *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.

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

3.2.11 *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:*

a = intercept estimate (8.1.3)
B = bias (7.1.1)
b = slope estimate (8.1.3)
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)
e_i = residual estimate (8.3.3)
f = degrees of freedom for s (9.1.1) (X1.1.2)
$F_{1-\alpha}$ = $(1-\alpha)^{th}$ percentile of the F distribution (10.3.1)
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) (9.1.1)
n_i = sample size from i^{th} 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 (5.5.2.1)
r = sample correlation coefficient (8.3.2)
$\mathcal{R}$ = ratio of two population variances (X1.6.3)
S_{XX} = sum of squared deviations of X from their mean (8.1.3.2)
S_{XY} = sum of products of deviations of X and Y from their means (8.1.3.2)
S_{YY} = sum of squared deviations of Y from their mean (8.1.3.2)
s = sample standard deviation (9.1.1)
s_B = sample standard deviation for bias (9.1.2)
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 (10.3.1)
s_2^2 = sample variance for population 2 (6.1.2)
s_2^2 = variance of test results from the modified process (10.3.1)
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) (9.1.3)
$t_{1-\alpha,f}$ = $(1-\alpha)^{th}$ percentile of the Student's t distribution with f degrees of freedom (X1.1.2)
X_{ij} = j^{th} test result from the i^{th} population (6.1)
UCL_R = upper confidence limit for $\mathcal{R}$ (10.3.1)
$\bar{X}$ = test result average (9.1.1)
$\bar{X}_i$ = test result average for the i^{th} population (6.1.1)
$\bar{X}_1$ = test result average for population 1 (6.1.3)
$\bar{X}_2$ = test result average for population 2 (6.1.3)
$Z_{1-\alpha}$ = $(1-\alpha)^{th}$ percentile of the standard normal distribution (X1.6.1)
α = (alpha) intercept parameter (8.1.1)
α = consumer's risk (5.2.2) (6.2) (7.2)
β = (beta) slope parameter (8.1.1)
β = producer's risk (5.4.1)
Δ = true mean difference between populations (5.4.1)
δ = (delta) measurement error of X (X3.1.1)
ε = (epsilon) measurement error of Y (X3.1.1)
η = (eta) true mean of Y (X3.1.1)
θ = (theta) angle of the straight line to the horizontal axis (8.1.4.1)
$\hat{\theta}$ = estimate of θ (8.1.4.1)
κ^2 = (kappa squared) information size (X3.3)
λ = (lambda) ratio of measurement error variances of Y over X (8.1.1.1)
μ = population mean (X1.4.1)
μ_i = i^{th} population mean (X1.1.1)
ν = (nu) probability associated with informative confidence interval (X3.3.2)
ν = approximate degrees of freedom for s_D (X1.1.4)
ξ = (xi) true mean of X (X3.1.1)
σ = standard deviation of the test method (5.2)
σ_d = standard deviation of the true difference between two populations (7.2)
σ_ε^2 = measurement error variances of Y (8.1.1)
σ_δ^2 = measurement error variances of X (8.1.1)
τ = (tau) perpendicular distance from line to origin (X3.1.3)
$\Phi(\bullet)$ = standard normal cumulative distribution function (X1.6.1)
φ = (phi) half width of confidence interval for θ (8.1.4.2)
ω = (omega) width of the equivalence interval for θ (X3.2)

3.4 *Acronyms:*

3.4.1 *ARV, n*—accepted reference value (5.5.1.1) (9.1) (X1.4)

3.4.2 *CRM, n*—certified reference material (5.5.1.1) (9.1)

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 9) (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 neither 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 The equivalence limit, which represents a worst-case difference or ratio, is determined prior to the equivalence test and its value is usually set by consensus among subject-matter experts.

4.1.2 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.

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

4.2 Equivalence testing is performed by a designed experiment that generates test results from the modified and current testing procedures on the same types of materials that are routinely tested. The design of the experiment depends on the type of equivalence needed as discussed below. Experiment design and execution for various objectives is discussed in Section 5.

4.2.1 *Means equivalence* is concerned with a potential shift in the mean test result in either direction due to a modification in the testing process. Test results are generated under repeatability conditions by the modified and current testing processes on the same material, and the difference in their mean test results is evaluated.

4.2.1.1 In situations where testing cannot be conducted under repeatability conditions, such as using in-line instrumentation, test results may be generated in pairs of test results from the modified and current testing processes, and the mean differences among paired test results are evaluated.

4.2.2 *Range equivalence* evaluates the differences in means over a selected wider range of test results and the experiment uses materials that cover that range. The slope of the linear statistical relationship between the test results from the two testing procedures is calculated. If the slope is equivalent to the value one (1), then the two testing processes meet slope equivalence. The combination of slope equivalence and means equivalence defines range equivalence.

4.2.3 *Bias equivalence* is a special case of means equivalence applied to a performance characteristic. A single set of test results is generated on a certified reference material (CRM) having an accepted reference value (ARV) to evaluate the test method bias of the current testing procedure. The mean test result is then compared with the ARV to estimate the occurrence of a known bias.

4.2.4 *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. Non-inferiority may involve the comparisons of means, standard deviations, or other statistical parameters.

4.2.4.1 Non-inferiority testing 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 test results or lower testing costs.

4.3 *Risk Management*—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 5.4 and 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, the consumer's risk, and an estimate of the test method precision.

5.2.1 For means equivalence tests there are two equivalence limits, $-E$ and E, because the need to detect a potential shift in either direction. Limits may be non-symmetrical around zero, such as $-E_1$ and E_2, and this will usually be the case for slope equivalence. For non-inferiority tests only one of these limits is tested.

5.2.2 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 user of the test results that the study falsely declares equivalence due to the modification of the testing process.

5.2.3 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* or *intermediate precision* conditions.

5.2.4 For slope equivalence an additional piece of required information is the ratio λ of the measurement variability of the modified and current test methods, expressed as variances. These estimates are usually available from experience or from method development work, but see 5.3.2.1.

5.3 The *design type* determines how the data are collected and how much data are needed to control the producer's risk, or the risk of a wrong decision. For generating test result data from the modified and current testing processes, three basic designs are discussed in this practice, the Two Independent Samples Design, the Paired Samples Design, and the Single Sample Design.

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

5.3.2 The Paired Samples Design is used for slope equivalence and may also be used for means equivalence. In this design, pairs of single test results from each testing procedure are generated on the same material over different time periods, or on various materials that are sampled either from a manufacturing process over time or from a set of materials that cover a predetermined range.

5.3.2.1 If information on measurement error is not available for slope equivalence testing, the experiment design can be modified to run duplicate test results by each testing process on each of the n materials to provide these precision estimates needed for estimation of their ratio.

5.3.3 The Single Sample Design used for bias equivalence. In this design, test results are generated by the current testing process on a certified reference material.

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 of the two testing processes 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*, defined as 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.

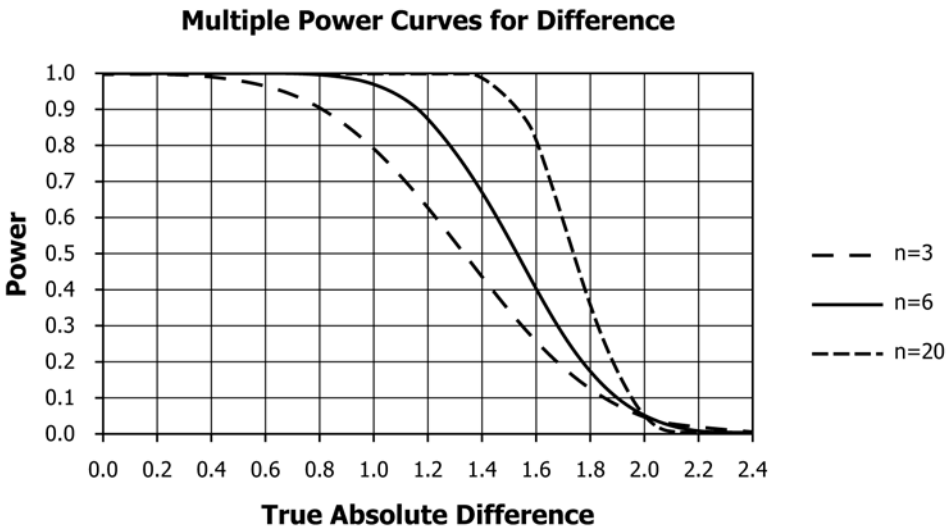

FIG. 1 Multiple Power Curves for Lab Transfer Example

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 $1 - \beta$ against the true mean difference Δ (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 from zero 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.

5.4.4 Appendix X3 provides criteria for determining the number of samples required to meet power requirements for slope equivalence.

5.5 The statistical analysis for accepting or rejecting equivalence of means and variances for a single material is similar for all cases and depends on the outcome of one-sided statistical hypothesis tests. These calculations are given in detail with examples in Sections 6, 7, 9, and 10, with statistical theory given in Appendix X1. The statistical analysis for range equivalence is given in Section 8, with statistical theory given in Appendix X3.

5.5.1 The data analysis for means equivalence uses a statistical methodology termed the two one-sided tests (TOST) procedure. The initial hypothesis is that the average difference between two sets of data exceeds an equivalence limit in one of the directions from zero, and this is tested in both directions. If the hypothesis is rejected in both directions then the alternate hypothesis that the mean difference is less than the equivalence limit is accepted and the two sources of data are deemed means equivalent.

NOTE 1—Historically, this procedure originated in the pharmaceutical industry for use in bioequivalence trials (**1, 2**),[4] and was 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 For bias equivalence, the test is based on only a single set of data conducted on a certified reference material (CRM) because its accepted reference value (ARV) is considered to be 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 non-inferiority when $LCL > -E$. Conversely, when the performance characteristic is defined as "lower is better," such as incidence of misclassifications, the statistical test supports non-inferiority when $UCL < E$.

5.5.2.1 For the non-inferiority testing of precision, the variances of the two data sets are used, and "lower is better" for this parameter, so the test for non-inferiority applies. Because variances are a scale parameter, the single non-inferiority test is based the ratio R of the two sample variances, and the non-inferiority limit E is also in the form of a ratio.

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 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 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:

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

6.1.4 *Test for Equivalence*—Compute the upper (UCL) and lower (LCL) confidence limits for the 100 $(1 - 2\alpha)$ % 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.

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

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

where t is the upper 100 $(1 - \alpha)$ % 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

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

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

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:

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

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

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

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

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

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

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

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.

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

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

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

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

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.

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.

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

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. Procedure for Equivalence of Test Results Over a Range of Values

8.1 *Range equivalence* is defined as the condition that means equivalence holds over a predetermined range of test

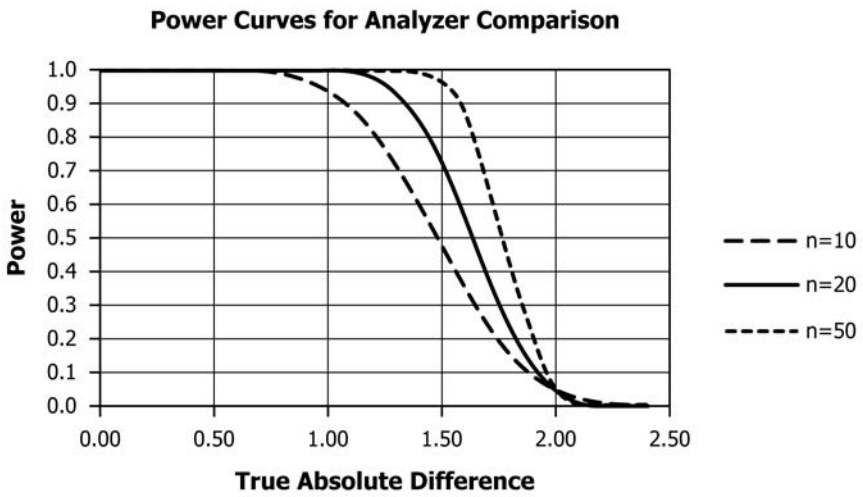

FIG. 2 Power Curves for Total Organic Carbon Analyzers Comparison

method values. Two tests for range equivalence are required, means equivalence (covered in previous sections) and *slope equivalence*. The statistical procedure for slope equivalence involves the estimation and evaluation of a straight line representing a linear statistical relationship between paired test results from the modified and current testing processes on materials the covering the test result range. Slope equivalence is accepted if the slope is equivalent to the value 1, representing a 45 degree line relationship. Range equivalence is accepted when both the slope equivalence and the mean equivalence at the averages of the test results are met.

8.1.1 The linear statistical function is $Y=\alpha+\beta X$, describing the straight line relationship between pairs of test results from the modified testing process Y_i, and from the current testing process X_i. The *function parameters* are the *intercept* α and the *slope* β. The intercept is the value of Y when $X = 0$, and the slope is the amount of change in Y units for a unit change in X. Unlike the similar simple linear regression model (see Practice E3080), both X and Y are subject to measurement errors, and their variances are denoted as σ_δ^2, σ_ε^2, respectively.

8.1.1.1 To calculate the slope estimate it is necessary to have an estimate of the precision ratio of Y with respect to X, denoted as $\lambda=\sigma_\varepsilon^2/\sigma_\delta^2$. The measurement error variances can be estimated from experience with current method use and method development data for the modified method. Alternatively, the comparison experiment can conduct duplicate test results from both methods for estimating these variances, as noted in the referenced article (**5**).

8.1.1.2 For simplicity, this section will assume that the two test methods have similar measurement error, thus dealing only with the case that $\lambda = 1$. Appendix X3 discusses methodology for the situation where $\lambda \neq 1$.

8.1.2 The experimental design consists of single tests run by each method on n samples, and the resulting data set of test results is designated as the pairs (X_i, Y_i), with the sample index i ranging from 1 through n. Information for determining sample size for the experiment design is given in Appendix X3. If replicate tests are conducted on all samples, the averages of the replicate test results are used for calculations in this section.

8.1.3 The model parameters α and β are estimated by a procedure known as Orthogonal Least Squares, which finds their corresponding estimates a and b that minimize the sum of the squares of the *perpendicular* distances between the data points and the estimated line. The calculations for estimation follow.

8.1.3.1 Calculate the averages of X and Y:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \tag{18}$$

$$\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n} \tag{19}$$

8.1.3.2 Calculate the sums of squared deviations of X and Y from their averages, respectively S_{XX} and S_{YY}, and the sum of cross products of the X and Y deviations from their means, S_{XY}:

$$S_{XX} = \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2 \tag{20}$$

$$S_{YY} = \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2 \tag{21}$$

$$S_{XY} = \sum_{i=1}^{n} \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right) \tag{22}$$

8.1.3.3 Calculate b, the estimate of the slope β:

$$b = \frac{S_{YY} - S_{XX} + \sqrt{(S_{YY} - S_{XX})^2 + 4S_{XY}^2}}{2S_{XY}} \tag{23}$$

8.1.3.4 Calculate a, the estimate of the intercept α:

$$a = \bar{Y} - b\bar{X} \tag{24}$$

8.1.4 For slope equivalence, a confidence interval (LCL, UCL) is constructed for the slope and the acceptance criteria is $LCL > E_1$ and $UCL < E_2$, where E_1 and E_2 are the respective lower and upper equivalence limits. Because the intercept may fall well outside the range of the data, it is recommended that the TOST procedure for means equivalence (that is, for $\bar{X}=\bar{Y}$) applied to the complete set of data be used as a surrogate test for $\alpha = 0$. To obtain the slope confidence interval it is necessary to transform the linear relationship to angular scales for calculation purposes and back transform to the original scale of the data.

8.1.4.1 Calculate the estimate of the angle θ that the line makes with the horizontal (X) axis:

$$\hat{\theta} = \arctan(b) \tag{25}$$

8.1.4.2 Calculate the half width φ of the two-sided 90 % confidence interval for θ:

$$\varphi = 0.5 \arcsin\left[t_{n-2,0.95} \frac{2}{\sqrt{n-2}} \sqrt{\frac{S_{YY}S_{XX} - S_{XY}^2}{(S_{YY} - S_{XX})^2 + 4S_{XY}^2}} \right] \tag{26}$$

where $t_{n-2,0.95}$ is the upper 95th quantile of the Student's t distribution with $n - 2$ degrees of freedom.

8.1.4.3 The lower and upper two-sided 90 % confidence limits on θ are:

$$LCL(\theta) = \hat{\theta} - \varphi \tag{27}$$

$$UCL(\theta) = \hat{\theta} + \varphi \tag{28}$$

8.1.4.4 The lower and upper two-sided 90 % confidence limits on β are the tangents of the confidence limits for θ:

$$LCL(\beta) = \tan\left(\hat{\theta} - \varphi\right) \tag{29}$$

$$UCL(\beta) = \tan\left(\hat{\theta} + \varphi\right) \tag{30}$$

The confidence limits for β will not be symmetrical around the slope estimate b.

8.2 *Example*—A new in-line instrument (Instrument B) for measuring total organic carbon (TOC) in purified water is compared with an existing instrument (Instrument A) in 7.2 for range equivalence. This is a paired comparison design and the TOC test results by each instrument for 20 water samples are listed in Table 2. The data covered a TOC range of 35 to 60

parts per billion (ppb) and the data supported means equivalence (see 7.2) with an equivalence limit of $E = 2$ ppm. This example will now be used for the linear statistical relationship model approach to establish slope equivalence for the new instrument as compared with the current instrument.

8.2.1 Measurement error was assumed to be approximately the same for both instruments, thus for the slope equivalence test $\lambda = 1$, and the use of the estimation of the linear statistical relationship by orthogonal least squares. The equivalence limits on the slope were set at $E_1 = 0.8$ and $E_2 = 1.25$. (Slope units were ppb/ppb, thus dimensionless.) For a discussion on equivalence limits for slopes, see X3.2.

8.2.2 The TOC data from Table 2 are again listed for each instrument in Table 3 for 20 water sampling times. The current Instrument A was designated as the X variable and the new Instrument B was designated as the Y variable. The averages (Eq 18, Eq 19) were $\bar{X} = 48.20$ ppb and $\bar{Y} = 48.66$ ppb, and these values represent the centroid of the X, Y data set. The sums of squares and cross products (Eq 20-22) gave $S_{XX} = 714.62$, $S_{YY} = 681.37$, and $S_{XY} = 687.53$. The slope and intercept estimates (Eq 23, Eq 24) were $b = 0.9761$ and $a = 1.61$, respectively.

8.2.3 To obtain the 90 % two-sided confidence interval for the slope, the conversion to the angular scale was made, resulting in $\hat{\theta} = \arctan(b) = \arctan(0.9761) = 0.7733$ radians (Eq 25), which was close to the 45 degree line having $\arctan(1.0) = \pi/4$ or 0.7854 radians. The confidence half-interval for θ was calculated (Eq 26) as:

$$\varphi = 0.5 \arcsin\left[t \frac{2}{\sqrt{n-2}} \sqrt{\frac{S_{YY}S_{XX} - S_{XY}^2}{(S_{YY} - S_{XX})^2 + 4S_{XY}^2}} \right]$$

$$= 0.5 \arcsin\left[1.7341 \frac{2}{\sqrt{20-2}} \sqrt{\frac{(681.37)(714.62) - 687.53^2}{(681.37 - 714.62)^2 + 4(687.53)^2}} \right]$$

$$= 0.5 \arcsin(0.0709) = 0.0355 \text{ radians}$$

using $t = 1.7341$ based on 18 degrees of freedom (df). Then the 90 % confidence limits for θ (Eq 27, Eq 28) were (0.7378, 0.8088) and the 90 % confidence limits for β (Eq 29, Eq 30) were (0.9091, 1.0479). The confidence interval for β fell within the equivalence range of (0.8, 1.25) thus accepting slope equivalence. The estimated intercept $a = 1.61$ was close to zero but no statistical test was conducted on $\alpha = 0$. Because of the acceptance of slope equivalence and means equivalence (see 7.2), the acceptance of individual equivalence was also supported.

8.3 *Evaluation of the Relationship*—Three important diagnostics for evaluation of the statistical relationship are the scatter plot, the correlation coefficient, and examination of the residuals, which are the perpendicular distances from the data points to the fitted line.

8.3.1 The scatter plot in Fig. 3 shows that the data points appear to fall along a straight line around the fitted statistical relationship line, which should be close to a 45 degree line (slope of 1).

8.3.2 The sample correlation coefficient is a dimensionless statistic intended to measure the strength of a linear relationship between two variables. The estimated correlation coefficient, r, from a set of paired data (X_i, Y_i) is calculated from the three statistics, S_{XX}, S_{YY}, and S_{XY}:

$$r = \frac{S_{XY}}{S_{XX}S_{YY}} \quad (31)$$

The value of the correlation coefficient ranges between –1 and +1, but for this application r should be close to 1 in order to meet slope equivalence.

8.3.2.1 *For the TOC example,* $r = 0.9853$. The two regression lines Y on X (X the regressor variable) and X on Y (Y the regressor variable) are virtually superimposed on the relationship line, and all three lines travel through the centroid of the

TABLE 3 Data and Calculations for Straight Line Regression on TOC Example

Sample	InstA	InstB	Deviations from Avg		Residual		
Point, i	X_i	Y_i	$X_i - \bar{X}$	$Y_i - \bar{Y}$	e_i	Statistic	Results
1	46.4	48.8	–1.80	0.14	1.36	S_{XX}	714.62
2	44.2	43.5	–4.00	–5.16	–0.90	S_{YY}	618.37
3	52.4	53.0	4.20	4.34	0.17	S_{XY}	687.53
4	37.6	37.3	–10.60	–11.36	–0.73	Slope, b_1	0.9761
5	49.3	49.1	1.10	0.44	–0.45	Intercept, b_0	1.61
6	45.0	44.5	–3.20	–4.16	–0.74	Angle, θ	0.7733
7	51.4	51.3	3.20	2.64	–0.35	Half interval, φ	0.0355
8	57.6	56.8	9.40	8.14	–0.74	90 % Conf. Int. for θ:	
9	43.4	44.9	–4.80	–3.76	0.66	LCL, θ	0.7378
10	45.2	44.1	–3.00	–4.56	–1.17	UCL, θ	0.8088
11	59.0	58.5	10.80	9.84	–0.50	90 % Conf. Int. for β:	
12	43.1	44.1	–5.10	–4.56	0.30	LCL, β	0.9091
13	39.3	40.9	–8.90	–7.76	0.66	UCL, β	1.0479
14	48.2	48.4	0.00	–0.26	–0.19	Corr. Coef., r	0.9853
15	48.7	49.0	0.50	0.34	–0.11	Equiv. Limit for β:	
16	44.4	46.1	–3.80	–2.56	0.82	Lower, E_1	0.80
17	52.7	53.2	4.50	4.54	0.11	Upper, E_2	1.25
18	43.3	44.6	–4.90	–4.06	0.52		
19	54.4	56.7	6.20	8.04	1.42		
20	58.4	58.4	10.20	9.74	–0.15		
Average	48.20	48.66	0.00	0.00	0.00		
Variance	37.6116	35.8615			0.5402		