**Designation: ~~E2935 – 16~~ E2935 – 17**

An American National Standard

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

## 1. Scope

1.1 This practice provides statistical methodology for conducting equivalence testing on numerical data from two sources to determine if their true means or variances differ by no more than predetermined limits.

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

1.3 The guidance in this standard applies ~~only~~ to experiments conducted on a single material at a given level of the test ~~result.~~ 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*~~ *safety, health, and* ~~*health*~~ *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.
<div align="right">E177</div>

3.1.2 *bias, n*—the difference between the expectation of the test results and an accepted reference value.
<div align="right">E177</div>

3.1.3 *confidence interval, n*—an interval estimate [L, U] with the statistics L and U as limits for the parameter $\theta$ and with confidence level $1 - \alpha$, where Pr(L $\leq \theta \leq$ U) $\geq$ ~~1~~ $\underline{1 - \alpha}$.  **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*—

$\alpha$ 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,* $\mu$, average or expected value of a characteristic in a ~~population~~ 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,* $\underline{\sigma,}$ ~~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, $\sigma^2, s^2, n$*—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 ~~populations sampled.~~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* (8.1.1) (X1.1.2)~~ |
| $f$ | = | degrees of freedom for *s* (9.1.1) (X1.1.2) |
| $F_{1-\alpha}$ | = | ~~$(1 − \alpha)th$ percentile of the *F* distribution (9.3.1)~~ |
| $F_{1-\alpha}$ | = | $(1 − \alpha)^{th}$ 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) |
| $\mathscr{F}(\bullet)$ | = | ~~the cumulative *F* distribution function (X1.6.3)~~ |
| $\mathscr{F}(\bullet)$ | = | the cumulative *F* distribution function (X1.6.3) |
| $H_0$: | = | null hypothesis (X1.1.1) |

$H_A:$ = alternate hypothesis (X1.1.1)
$H_a:$ = alternate hypothesis (X1.1.1)
$n$ = sample size (number of test results) from a population (5.4) (6.1.3) (7.1.1) (8.1.1)
$n$ = 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)
$n_i$ = 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 (5.5.3)
$R$ = 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 (8.1.1)
$s$ = sample standard deviation (9.1.1)
$s_B$ = sample standard deviation for bias (8.1.2)
$s_B$ = 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)
$s_i$ = sample standard deviation for $i^{th}$ population (6.1.1)
$s_i^2$ = sample variance 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)
$s_1^2$ = 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 (5.5.3)
$s_2^2$ = 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) (8.1.3)
$t$ = 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)
$t_{1-\alpha,f}$ = $(1-\alpha)^{th}$ 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)
$X_{ij}$ = $j^{th}$ test result from the $i^{th}$ population (6.1)
$UCL_R$ = = upper confidence limit for $\mathcal{R}$ (9.3.1)
$UCL_R$ = upper confidence limit for $\mathcal{R}$ (10.3.1)
$\bar{X}$ = test result average (8.1.1)
$\bar{X}$ = test result average (9.1.1)
$\bar{X}_i$ = test result average for the $i$th population (6.1.1)
$\bar{X}_i$ = 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)
$Z_{1-\alpha}$ = $(1-\alpha)^{th}$ percentile of the standard normal distribution (X1.6.1)
$\alpha$ = (alpha) intercept parameter (8.1.1)
$\alpha$ = consumer's risk (5.2.3) (6.2) (7.2)
$\alpha$ = consumer's risk (5.2.2) (6.2) (7.2)
$\beta$ = (beta) slope parameter (8.1.1)
$\beta$ = producer's risk (5.4.1)
$\Delta$ = true mean difference between populations (5.4.1)
$\delta$ = (delta) measurement error of $X$ (X3.1.1)
$\varepsilon$ = (epsilon) measurement error of $Y$ (X3.1.1)
$\eta$ = (eta) true mean of $Y$ (X3.1.1)
$\theta$ = (theta) angle of the straight line to the horizontal axis (8.1.4.1)
$\hat{\theta}$ = estimate of $\theta$ (8.1.4.1)
$\kappa^2$ = (kappa squared) information size (X3.3)

$\lambda$ = (lambda) ratio of measurement error variances of $Y$ over $X$ (8.1.1.1)
$\mu$ = population mean (X1.4.1)
$\mu_i$ = *ith* population mean (X1.1.1)
$\mu_i$ = $i^{\text{th}}$ population mean (X1.1.1)
$\nu$ = (nu) probability associated with informative confidence interval (X3.3.2)
$\nu$ = approximate degrees of freedom for $s_D$ (X1.1.4)
$\xi$ = (xi) true mean of $X$ (X3.1.1)
$\sigma$ = standard deviation of the test method (5.2)
$\sigma_d$ = standard deviation of the true difference between two populations (7.2)
$\sigma_*^2$ = measurement error variances of $Y$ (8.1.1)
$\sigma_\delta^2$ = measurement error variances of $X$ (8.1.1)
$\tau$ = (tau) perpendicular distance from line to origin (X3.1.3)
$\Phi(\bullet)$ = standard normal cumulative distribution function (X1.6.1)
$\varphi$ = (phi) half width of confidence interval for $\theta$ (8.1.4.2)
$\omega$ = (omega) width of the equivalence interval for $\theta$ (X3.2)

3.4 *Acronyms:*

3.4.1 *ARV, n*—accepted reference value (~~5.3.3~~5.5.1.1) (~~8.1~~9.1) (X1.4)

3.4.2 *CRM, n*—certified reference material (~~5.3.3~~5.5.1.1) (~~8.1~~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 ~~8~~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 ~~not~~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 ~~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.~~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 ~~Two principal types of equivalence are covered in the practice,~~Equivalence testing is performed by ~~means equivalence and non-inferiority. Means equivalence implies that a sustained shift in test results between~~a designed experiment that generates test results from 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.~~ 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* ~~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~~ 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.~~ bias of the current testing procedure. The mean test result is then compared with the ARV to estimate the occurrence of a known bias.

~~4.2.2 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.~~

4.2.4 *Non-inferiority* ~~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.~~ 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 ~~also~~ 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, ~~E,~~the consumer's ~~risk α,~~ risk, and an estimate of the test method ~~precision σ.~~precision.

5.2.1 For means equivalence tests there are two equivalence limits, $-E$ and $E$, ~~that are tested.~~ because the need to detect a potential shift in either direction. Limits may be ~~nonsymmetrical~~non-symmetrical around zero, such as $-E_1$ and $E_2$, ~~but this is not usual and would require advice from a qualified statistician for a proper design setup.~~ 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 $α = 0.05$, 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,*repeatability~~intermediate,~~ or *reproducibility*intermediate precision conditions.

5.2.4 ~~The consumer's risk may be determined by an industry norm or a regulatory requirement. A probability value often used is α =~~For slope equivalence an additional piece of required information is the ratio *0.05,λ* ~~which is a 5 % risk to the consumer that the study falsely declares equivalence.~~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. ~~A sufficient quantity of a homogeneous material for the required number of tests is necessary. For comparing~~For generating test result data from the modified and current testing processes, ~~two~~three 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.~~Design, and the Single Sample Design.

5.3.1 The Two Independent Samples Design is used for means equivalence is discussed in Section and non-inferiority testing. 6. 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 transfertransfers each laboratory generates independent test results using the same testing procedure, preferably under repeatability conditions. 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 means equivalence is discussed in Sectionslope equivalence and may also be 7. used for means equivalence. 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. 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 design for bias equivalence is discussed in Section Single Sample Design used for bias equivalence. 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. material.

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

5.4.1 The number of test results, symbol *n*, from each testing processof 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=β 1 − β 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=β 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-shapedS-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.
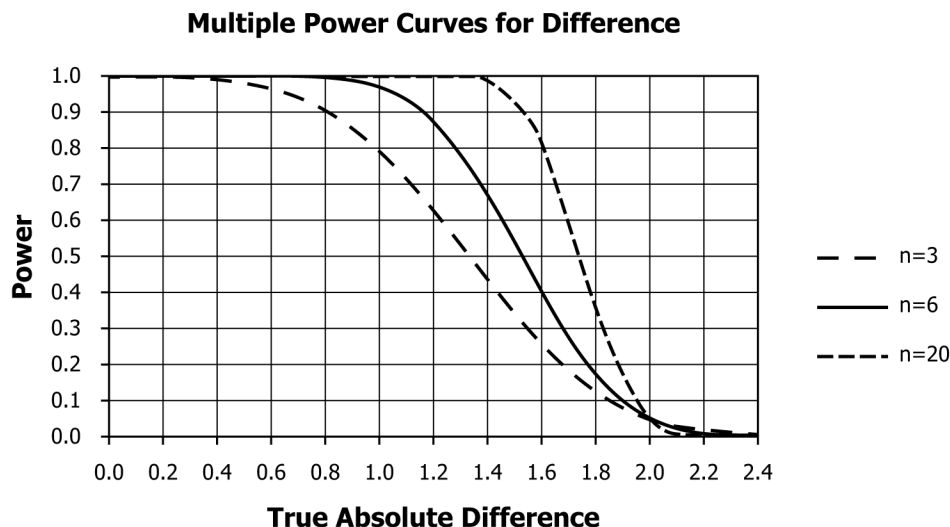


FIG. 1 Multiple Power Curves for Lab Transfer Example

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−β~~ $1 − \beta$ against the true mean difference $\Delta$ ~~for means~~ (see X1.6.2) or against the true variance ratio $\mathscr{R}$ 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 for means and variances. The~~ tests. These calculations are given in detail with examples in Sections ~~6 − 9~~6, 7, 9, and 10, with statistical theory given in Appendix X1. The statistical ~~theory~~ analysis for range equivalence is given in ~~an appendix (see~~ Section X1.18). , with statistical theory given in Appendix X3.

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±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~~The initial hypothesis ~~s_D is based on the standard deviations of the~~ that the average difference between two sets of data ~~and the type of study design. Then equivalence is supported if both of the following two conditions are met:~~ 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.

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

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

NOTE 1—Historically, this procedure originated in the pharmaceutical industry for use in bioequivalence trials (**1, 2**),[4] 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 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.~~

5.5.1.1 For bias equivalence, the ~~calculation for test~~ $s_D$ is based on only a single set of data ~~because the ARV is considered as~~ 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",~~better," such as method sensitivity, the statistical test supports ~~noninferiority~~non-inferiority when $LCL > −E$. Conversely, when the performance characteristic is defined as "lower is ~~better",~~better," such as incidence of misclassifications, the statistical test supports ~~noninferiority~~non-inferiority 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.~~

5.5.2.1 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.

~~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=s₂²/s₁², where s₁² and s₂² 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 σ₂²/σ₁², 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 $j$~~th~~th test result from the $i$~~th~~th population. The equivalence limit *E*, consumer's risk $\alpha$, and sample sizes have been previously determined.

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

---

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

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

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

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

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

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

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

6.1.2 Calculate the pooled standard deviation and degrees of freedom:

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

If $n_1 = n_2 = n$, then:

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

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

6.1.3 Calculate the difference between means and its standard error:

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

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

If $n_1 = n_2 = n$, then:

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

6.1.4 *Test for Equivalence*—Compute the upper (UCL) and lower (LCL) confidence limits for the 100 ~~(1−2α)~~ $(1 - 2\alpha)$ % 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−α)~~ $(1 - \alpha)$ % 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~~ 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, $\alpha = 0.05$, and estimated $\sigma = 0.5$ 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.

**TABLE 1 Data for Equivalence Test Between Two Laboratories**

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

6.2.1.1 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:

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

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

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

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

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:

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

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 $s_{\bar{p}p}$).

6.2.5 The 95th~~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~~th~~ population and the $j$th~~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 $\alpha$, 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 $f$.

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

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

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

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

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)~~ $100(1 - 2\alpha)$ % two-sided confidence interval on the true difference. If the confidence interval is completely contained within the equivalence limits $(0 \pm E)$, 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)~~ $100(1 - α)$ % 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 $σ_d = 0.7\sqrt{2}$ 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.

7.2.3 The ~~95th~~ $95^{th}$ 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.
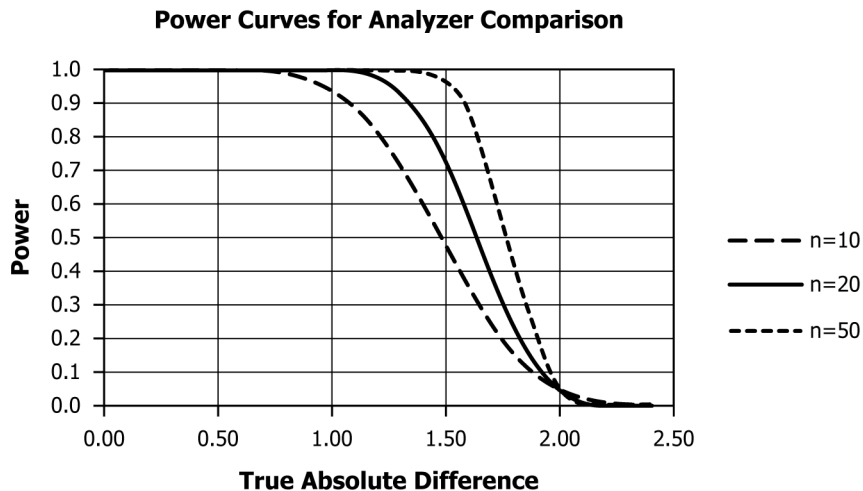


**FIG. 2 Power Curves for Total Organic Carbon Analyzers Comparison**