This document is not an ASTM standard and is intended only to provide the user of an ASTM standard an indication of what changes have been made to the previous version. Because it may not be technically possible to adequately depict all changes accurately, ASTM recommends that users consult prior editions as appropriate. In all cases only the current version of the standard as published by ASTM is to be considered the official document.

Designation: ~~E2935 – 14~~ E2935 – 15

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 ~~are similar within~~ differ by no more than predetermined limits.

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

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

1.4 The statistical methodology for determining equivalence used is the ~~"Two~~two one-sided ~~t~~tests ~~test" (TOST).~~ (TOST) procedure. The control of risks associated with the equivalence decision is discussed.

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

1.6 *This standard does not purport to address all of the safety concerns, if any, associated with its use. It is the responsibility of the user of this standard to establish appropriate safety and health practices and determine the applicability of regulatory limitations prior to use.*

iTeh Standards
(https://standards.iteh.ai)
Document Preview

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

ASTM E2935-15
https://standards.iteh.ai/catalog/standards/sist/1f79d4e8-e28e-4773-81cf-e625e26ddd27/astm-e2935-15

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 – α, where $Pr(L \leq \theta \leq U) \geq 1 - \alpha$. **E2586**

[1] This test method is under the jurisdiction of ASTM Committee E11 on Quality and Statistics and is the direct responsibility of Subcommittee E11.20 on Test Method Evaluation and Quality Control.

Current edition approved ~~Oct. 1, 2014~~Oct. 1, 2015. Published ~~August 2013~~October 2015. Originally approved in 2013. Last previous edition approved in ~~2013~~2014 as E2935 – ~~13.~~14. DOI: ~~10.1520/E2935-14.~~10.1520/E2935-15.

[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.

3.1.3.1 *Discussion*—

The confidence level, 1 – α, reflects the proportion of cases that the confidence interval [L, U] would contain or cover the true parameter value in a series of repeated random samples under identical conditions. Once L and U are given values, the resulting

Copyright © ASTM International, 100 Barr Harbor Drive, PO Box C700, West Conshohocken, PA 19428-2959. United States

confidence interval either does or does not contain it. In this sense "confidence" applies not to the particular interval but only to the long run proportion of cases when repeating the procedure many times.

3.1.4 *confidence level, n*—the value, 1 – α, of the probability associated with a confidence interval, often expressed as a percentage. **E2586**

3.1.4.1 *Discussion*—

α is generally a small number. Confidence level is often 95 % or 99 %.

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

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

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

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

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

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

3.1.14 *repeatability conditions, n*—conditions where independent test results are obtained with the same method on identical test items in the same laboratory by the same operator using the same equipment within short intervals of time. **E177**

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

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

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

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

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

3.1.20 *test unit, n*—the total quantity of material (containing one or more test specimens) needed to obtain a test result as specified in the test method. See test result. **E2282**

3.2 *Definitions of Terms Specific to This Standard:*

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

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

3.2.2.1 *Discussion*—

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

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

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

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

3.2.5.1 *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.

iTeh Standards (https://standards.iteh.ai) Document Preview
ASTM E2935-15
https://standards.iteh.ai/catalog/standards/sist/1f79d4e8-e28e-4773-81cf-e625e26ddd27/astm-e2935-15

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

3.2.6.1 *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.

3.2.7 *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.

3.2.7.1 *Discussion—*

This design is termed a completely randomized design for a general number of populations sampled.

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

3.3 *Symbols:*

B = bias (7.1.1)
d_j = difference between a pair of test results at sampling point j (7.1.1)
$\bar{d}$ = average difference (7.1.1)
D = difference in sample means (6.1.2) (X1.1.2)
E = equivalence limit (5.2.1)
E_1 = lower equivalence limit (5.2.1.1)
E_2 = upper equivalence limit (5.2.1.1)
H_0: = null hypothesis (X1.1.1)
H_A: = alternate hypothesis (X1.1.1)
f = degrees of freedom for s (8.1.1) (X1.1.2)
f_i = degrees of freedom for s_i (6.1.1)
f_p = degrees of freedom for s_p (6.1.2)
n = sample size (number of test results) from a population (5.3) (6.1.3) (7.1.1) (8.1.1)
n_i = sample size from ith population (6.1.1)
n_1 = sample size from population 1 (6.1.2)
n_2 = sample size from population 2 (6.1.2)
s = sample standard deviation (8.1.1)
s_B = sample standard deviation for bias (8.1.2)
s_d = standard deviation of the difference between two test results (7.1.1)
s_D = sample standard deviation for mean difference (6.1.3) (X1.1.2)
s_i = sample standard deviation for ith population (6.1.1)
s_i^2 = sample variance for ith population (6.1.1)
s_1^2 = sample variance for population 1 (6.1.2)
s_2^2 = sample variance for population 2 (6.1.2)
s_p = pooled sample standard deviation (6.1.2)
s_r = repeatability sample standard deviation (6.2)
t = Student's t statistic (6.1.4) (7.1.3) (8.1.3)
$t_{1-\alpha,f}$ = (1-α)th percentile of the Student's t distribution with f degrees of freedom (X1.1.2)
X_{ij} = jth test result from the ith population (6.1)
$\bar{X}$ = test result average (8.1.1)
$\bar{X}_i$ = test result average for the ith 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-α)th percentile of the standard normal distribution (X1.5.1)
α = consumer's risk (5.2.2) (6.2) (7.2)
β = producer's risk (5.3)
Δ = true mean difference between populations (5.3)
μ = population mean (X1.4.1)
μ_i = ith population mean (X1.1.1)
ν = approximate degrees of freedom for s_D (X1.1.4)
σ = standard deviation of the test method (5.2.3)
σ_d = standard deviation of the true difference between two populations (7.2)

iTeh Standards
(https://standards.iteh.ai)
Document Preview
ASTM E2935-15
https://standards.iteh.ai/catalog/standards/sist/1f79d4e8-e28e-4773-81cf-e625e26ddd27/astm-e2935-15

$\Phi(\bullet)$ = standard normal cumulative distribution function (X1.5.1)

3.4 *Acronyms:*

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

3.4.2 *CRM, n*—certified reference material (5.1.2) (8.1)

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

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

3.4.5 *TOST, n*—two one-sided ttests ~~test~~ (4.3) (Section 6) (Section 7) (Section 8) (Appendix X1)

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

4. Significance and Use

4.1 Laboratories conducting routine testing have a continuing need to evaluate test result bias, to evaluate changes for improving the test process performance, or to validate the transfer of a test method to a new location or apparatus. In all situations it must be demonstrated that any bias or innovation will have negligible effect on test results for a characteristic of a material. This standard provides statistical methods to confirm that the mean test results from a testing process are equivalent to those from a reference standard or another testing process, where *equivalence* is defined as agreement within prescribed limits, termed *equivalence limits*.

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

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

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

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

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

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

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

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

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

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

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

4.4 *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 X1.2).

4.4.1 The consumer's risk is the probability of accepting equivalence when the actual bias or difference in means is equal to the equivalence limit. This probability is controlled to a low level so that accepting equivalence gives a high degree of assurance that differences in question are less than the equivalence limit.

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

5. Planning the Equivalence Study

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

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

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

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

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

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

5.2.1 The equivalence limits to be used in the TOST procedure are selected as the worst-case differences between the two population means and are determined by the subject matter expert or by industry consensus. These limits are usually symmetrical around zero and then are denoted as $-E$ and E.

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

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

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

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

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

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

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

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

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

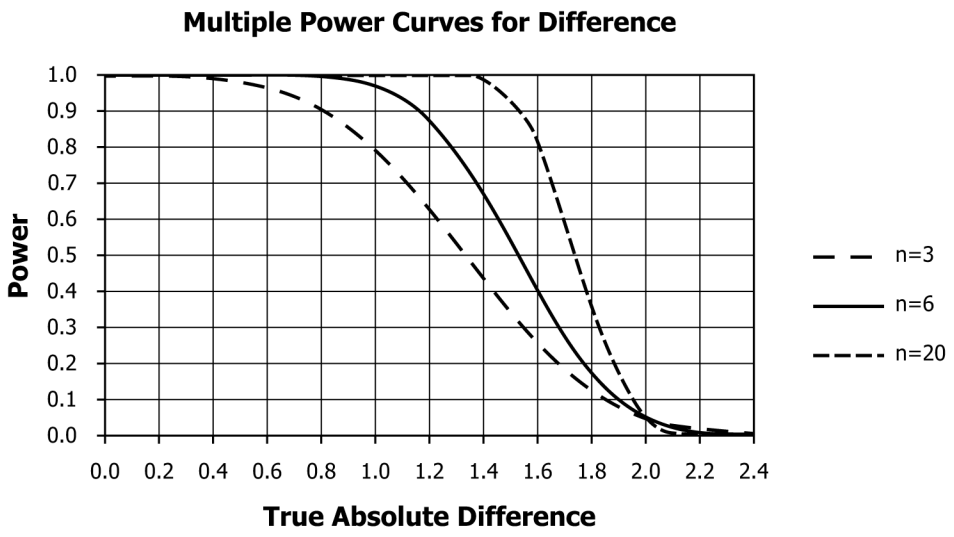

FIG. 1 Multiple Power Curves for Lab Transfer Example

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

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

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

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

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

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

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

6.1.2 Calculate the pooled standard deviation and degrees of freedom:

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

If $n_1 = n_2 = n$, then:

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

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

6.1.3 Calculate the difference between means and its standard error:

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

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

If $n_1 = n_2 = n$, then:

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

6.1.4 *Test for Equivalence*—Compute the upper (UCL) and lower (LCL) confidence limits for the 100 (1–2α) % two-sided confidence interval on the true difference. If the confidence interval is completely contained within the equivalence limits ($0 \pm E$), 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) % percentile of the Student's t distribution with $(n_1 + n_2 - 2)$ degrees of freedom.

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

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.

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

iTeh Standards (https://standards.iteh.ai) Document Preview
ASTM E2935-15