Designation: E2489 – 21

An American National Standard

Standard Practice for Statistical Analysis of One-Sample and Two-Sample Interlaboratory Proficiency Testing Programs[1]

This standard is issued under the fixed designation E2489; 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 describes methods for the statistical analysis of laboratory results obtained from interlaboratory proficiency testing programs. As in accordance with Practice E1301, proficiency testing is the use of interlaboratory comparisons for the determination of laboratory testing or measurement performance. Conversely, collaborative study (or collaborative trial) is the use of interlaboratory comparisons for the determination of the precision of a test method, as covered by Practice E691.

1.1.1 Method A covers testing programs using single test results obtained by testing a single sample (each laboratory submits a single test result).

1.1.2 Method B covers testing programs using paired test results obtained by testing two samples (each laboratory submits one test result for each of the two samples). The two samples should be of the same material or two materials similar enough to have approximately the same degree of variation in test results.

1.2 Methods A and B are applicable to proficiency testing programs containing a minimum of 10 participating laboratories.

1.3 The methods provide direction for assessing and categorizing the performance of individual laboratories based on the relative likelihood of occurrence of their test results, and for determining estimates of testing variation associated with repeatability and reproducibility. Assumptions are that a majority of the participating laboratories execute the test method properly and that samples are of sufficient homogeneity that the testing results represent results obtained from each laboratory testing essentially the same material. Each laboratory receives the same instructions or protocol.

1.4 *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.*

1.5 *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

E178 Practice for Dealing With Outlying Observations

E456 Terminology Relating to Quality and Statistics

E691 Practice for Conducting an Interlaboratory Study to Determine the Precision of a Test Method

E1301 Guide for Proficiency Testing by Interlaboratory Comparisons (Withdrawn 2012)[3]

E2586 Practice for Calculating and Using Basic Statistics

3. Terminology

3.1 *Definitions*—Unless otherwise noted in this standard, all terms relating to quality and statistics are defined in Terminology E456.

3.1.1 *collaborative study, n*—interlaboratory study in which each laboratory uses the defined method of analysis to analyze identical portions of homogeneous materials to assess the performance characteristics obtained for that method of analysis. **Horwitz**[4]

3.1.2 *collaborative trial, n*—see *collaborative study*.

3.1.3 *interlaboratory comparison, n*—organization, performance, and evaluation of tests on the same or similar test items by two or more laboratories in accordance with predetermined conditions.

---

[1] This practice 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 Dec. 1, 2021. Published December 2021. Originally approved in 2006. Last previous edition approved in 2016 as E2489 – 16. DOI: 10.1520/E2489-21.

[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] The last approved version of this historical standard is referenced on www.astm.org.

[4] Horwitz, W., "Protocol for the Design, Conduct and Interpretation of Collaborative Studies," *Pure and Applied Chemistry*, Vol 60, No. 6, 1988, pp. 855–864.

***A Summary of Changes section appears at the end of this standard**

3.1.4 *median, $\tilde{X}$, n*—the 50th percentile in a population or sample. **E2586**

3.1.4.1 *Discussion*—The sample median is the $[(n + 1)/2]$ order statistic if the sample size n is odd and is the average of the $[n/2]$ and $[n/2 + 1]$ order statistics if n is even.

3.1.5 *outlier, n*—see *outlying observation*. **E178**

3.1.6 *outlying observation, n*—observation that appears to deviate markedly in value from other members of the sample in which it appears. **E178**

3.1.7 *proficiency testing, n*—determination of laboratory testing performance by means of interlaboratory comparisons.

3.1.8 *repeatability standard deviation (S_r), n*—standard deviation of test results obtained under repeatability conditions. **E177**

3.1.9 *reproducibility standard deviation (S_R), n*—standard deviation of test results obtained under reproducibility conditions. **E177**

3.2 *Definitions of Terms Specific to This Standard:*

3.2.1 *hinge (upper or lower), n*—median of the upper or lower half of a set of data when the data is arranged in order of size.

3.2.1.1 *Discussion*—When there is an odd number of items in the data set, the middle value is included in both the upper and lower halves. The upper hinge is an estimate of the 75th percentile; the lower hinge is an estimate of the 25th percentile.

3.2.2 *inner fence (upper or lower), n*—value equal to the upper or lower hinge of a data set plus (upper) or minus (lower) 1.5 times the difference between upper and lower hinges.

3.2.3 *interquartile range, n*—distance between the upper and lower hinges of a data set.

3.2.4 *outer fence (upper or lower), n*—value equal to the upper or lower hinge of a data set plus (upper) or minus (lower) three times the difference between upper and lower hinges.

4. Summary of Practice

4.1 This practice describes methods of displaying interlaboratory data that visually show individual laboratory results.

4.2 The methods described in this practice can be applied to large and small sample populations from any distribution expected to have a general mound shape. It is recommended that in cases in which it is suspected that the data may be highly unsymmetrical or very unusual in some other manner a statistician should be consulted regarding the applicability of the analysis method.

4.2.1 The median is used as the "consensus" value of the measured test property.

4.2.2 The interquartile range (IQR) is used as the basis for estimating the spread in the data. Because the median and the interquartile range are not affected by the magnitude of extreme values of a data set, the analysis approach presented in this practice effectively eliminates the need to identify outlying observations (outliers).

4.3 Laboratory results are categorized according to how far the results lie outside of the interquartile range.

4.4 The upper and lower ends of the interquartile range are referred to as the hinges. The limits for categorizing laboratory results lying outside of the interquartile range are determined by multiplying the extent of the interquartile range by the fixed factors of 1.5 and 3.0. The upper and lower limits lying a distance of 1.5 times the range of the IQR beyond the hinges are referred to as the inner fences. The upper and lower limits for results lying at 3.0 times the range of the IQR beyond the hinges are referred to as the outer fences.

4.5 Guidance is provided for proficiency testing programs wishing to establish additional limits (or fences). The user is directed to Guide E1301 for additional guidance.

4.6 When using the methods in this practice, the number of participating laboratories should be at least ten. Since the degree of confidence is lower for analyses performed on small sample populations, caution should be used in applying statistics obtained from small sample populations.

4.7 When possible, it is generally desirable to have 30 or more participants when estimating the precision of test methods.

4.8 Estimates of the repeatability standard deviation and the reproducibility standard deviation are determined by dividing the interquartile ranges of appropriate data sets by a factor of 1.35.

4.8.1 The number 1.35 used in determining the repeatability and reproducibility standard deviations is based on an assumption of similarity to a normal distribution. Therefore, the estimate of the standard deviation using the methods described in this practice may not supply the desired accuracy if the distribution differs too much from the general shape of a normal curve. It is beyond the scope of this practice to describe procedures for determining when the analysis methods described in this practice are not applicable.

5. Significance and Use

5.1 This practice is specifically designed to describe simple robust statistical methods for use in proficiency testing programs.

5.2 Proficiency testing programs can use the methods in this practice for the purpose of comparing testing results obtained from a group of participating laboratories. The practice describes evaluation of individual laboratory results using the interquartile range and Tukey inner and outer fences.

5.3 In addition, the data obtained in proficiency testing programs may contain information regarding repeatability (within-lab) and reproducibility (between-lab) testing variation. Repeatability information is possible only if the program uses more than one sample. See Method B. Proficiency testing programs often have a greater number of participants than might be available for conducting an interlaboratory study to determine the precision of a test method (such as described in Practice E691). Precision estimates obtained for the larger number of participants in a proficiency testing program, along with the corresponding wider variation of test conditions, can provide useful information to standards developers regarding