Designation: ~~D6708 – 18~~ D6708 – 19

An American National Standard

Standard Practice for Statistical Assessment and Improvement of Expected Agreement Between Two Test Methods that Purport to Measure the Same Property of a Material[1]

This standard is issued under the fixed designation D6708; 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 covers statistical methodology for assessing the expected agreement between two standard test methods that purport to measure the same property of a material, and deciding if a simple linear bias correction can further improve the expected agreement. It is intended for use with results collected from an interlaboratory study meeting the requirement of Practice D6300 or equivalent (for example, ISO 4259). The interlaboratory study must be conducted on at least ten materials that span the intersecting scopes of the test methods, and results must be obtained from at least six laboratories using each method.

1.2 The statistical methodology is based on the premise that a bias correction will not be needed. In the absence of strong statistical evidence that a bias correction would result in better agreement between the two methods, a bias correction is not made. If a bias correction is required, then the *parsimony principle* is followed whereby a simple correction is to be favored over a more complex one.

NOTE 1—Failure to adhere to the parsimony principle generally results in models that are over-fitted and do not perform well in practice.

1.3 The bias corrections of this practice are limited to a constant correction, proportional correction, or a linear (proportional + constant) correction.

1.4 The bias-correction methods of this practice are method symmetric, in the sense that equivalent corrections are obtained regardless of which method is bias-corrected to match the other.

1.5 A methodology is presented for establishing the ~~95 % confidence~~ numerical limit (designated by this practice as the *between methods reproducibility*) that would be exceeded about 5 % of the time (one case in 20 in the long run) for the difference between two results where each result is obtained by a different operator using different apparatus and each applying one of the two methods *X* and *Y* on identical material, where one of the methods has been appropriately bias-corrected in accordance with this ~~practice.~~practice, in the normal and correct operation of both test methods.

NOTE 2—In earlier versions of this standard practice, the term "cross-method reproducibility" was used in place of the term "between methods reproducibility." The change was made because the "between methods reproducibility" term is more intuitive and less confusing. It is important to note that these two terms are synonymous and interchangeable with one another, especially in cases where the "cross-method reproducibility" term was subsequently referenced by name in methods where a D6708 assessment was performed, before the change in terminology in this standard practice was adopted.

NOTE 3—Users are cautioned against applying the between methods reproducibility as calculated from this practice to materials that are significantly different in composition from those actually studied, as the ability of this practice to detect and address sample-specific biases (see ~~6.8~~6.7) is dependent on the materials selected for the interlaboratory study. When sample-specific biases are present, the types and ranges of samples may need to be expanded significantly from the minimum of ten as specified in this practice in order to obtain a more comprehensive and reliable ~~95 % confidence limits for~~ between methods reproducibility that adequately cover the range of ~~sample specific~~ sample-specific biases for different types of materials.

1.6 This practice is intended for test methods which measure quantitative (numerical) properties of petroleum or petroleum products.

1.7 The statistical methodology outlined in this practice is also applicable for assessing the expected agreement between any two test methods that purport to measure the same property of a material, provided the results are obtained on the same comparison sample set, the standard error associated with each test result is known, and the sample set design meets the requirements of this practice, in particular that the statistical degree of freedom associated with all standard errors are 30 or greater.

---

[1] This practice is under the jurisdiction of ASTM Committee D02 on Petroleum Products, Liquid Fuels, and Lubricants and is the direct responsibility of Subcommittee D02.94 on Coordinating Subcommittee on Quality Assurance and Statistics.

Current edition approved ~~April 1, 2018~~May 1, 2019. Published ~~July 2018~~June 2019. Originally approved in 2001. Last previous edition approved in ~~2016~~2018 as ~~D6708 – 16b.~~D6708 – 18. DOI: ~~10.1520/D6708-18.~~10.1520/D6708-19.

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

1.8 *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]

D5580 Test Method for Determination of Benzene, Toluene, Ethylbenzene, *p/m*-Xylene, *o*-Xylene, C_9 and Heavier Aromatics, and Total Aromatics in Finished Gasoline by Gas Chromatography

D5769 Test Method for Determination of Benzene, Toluene, and Total Aromatics in Finished Gasolines by Gas Chromatography/Mass Spectrometry

D6299 Practice for Applying Statistical Quality Assurance and Control Charting Techniques to Evaluate Analytical Measurement System Performance

D6300 Practice for Determination of Precision and Bias Data for Use in Test Methods for Petroleum Products and Lubricants

D7372 Guide for Analysis and Interpretation of Proficiency Test Program Results

2.2 *ISO Standard:*[3]

ISO 4259 Petroleum Products—Determination and ~~application~~Application of ~~precision data~~Precision Data in ~~relation~~Relation to ~~methods~~Methods of ~~test.~~Test

3. Terminology

3.1 *Definitions:*

3.1.1 *between ILCP method-averages reproducibility ($R_{ILCP\_X,\ ILCP\_Y}$), n*—a quantitative expression of the random error associated with the difference between the bias-corrected ILCP average of method X versus the ILCP average of method Y from a Proficiency Testing program, when the method X has been assessed versus method Y, and an appropriate bias-correction has been applied to all method X results in accordance with this practice; it is defined as the ~~95 % confidence~~ numerical limit for the difference between two such ~~averages.~~averages that would be exceeded about 5 % of the time (one case in 20 in the long run).

3.1.2 *between-method bias, n*—a quantitative expression for the mathematical correction that can statistically improve the degree of agreement between the expected values of two test methods which purport to measure the same property.

3.1.3 *between methods reproducibility (R_{XY}),n*—a quantitative expression of the random error associated with the difference between two results obtained by different operators using different apparatus and applying the two methods *X* and *Y*, respectively, each obtaining a single result on an identical test sample, when the methods have been assessed and an appropriate bias-correction has been applied in accordance with this practice; it is defined as the ~~95 % confidence~~ numerical limit for the difference between two such single and independent ~~results.~~results that would be exceeded about 5 % of the time (one case in 20 in the long run) in the normal and correct operation of both test methods.

[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 American National Standards Institute (ANSI), 25 W. 43rd St., 4th Floor, New York, NY 10036.

3.1.3.1 *Discussion*—

A statement of between methods reproducibility must include a description of any bias correction used in accordance with this practice.

3.1.3.2 *Discussion*—

Between methods reproducibility is a meaningful concept only if there are no statistically observable sample-specific relative biases between the two methods, or if such biases vary from one sample to another in such a way that they may be considered random effects. ~~(see~~(See 6.7.)

3.1.4 *centered sum of squares (CSS), n*—a statistic used to quantify the degree of agreement between the results from two test methods after bias-correction using the methodology of this practice.

3.1.5 *Interlaboratory Crosscheck Program (ILCP), n*—ASTM International Proficiency Test Program sponsored by Committee D02 on Petroleum Products, Liquid Fuels, and Lubricants; see ASTM website for current details. **D7372**

3.1.6 *total sum of squares (TSS), n*—a statistic used to quantify the information content from the inter-laboratory study in terms of total variation of sample means relative to the standard error of each sample mean.

3.2 *Symbols:*

X,Y = single X-method and Y-method results, respectively
X_{ijk}, Y_{ijk} = single results from the X-method and Y-method round robins, respectively
X_I, Y_i = means of results on the i^{th} round robin sample
S = the number of samples in the round robin
L_{Xi}, L_{Yi} = the numbers of laboratories that returned results on the i^{th} round robin sample
R_X, R_Y = the reproducibilities of the X- and Y- methods, respectively
R_{Xi}, R_{Yi} = the reproducibility of method X and Y, evaluated at the method X and Y means of the i^{th} round robin sample, respectively
$R_{ILCP\_\bar{X}, ILCP\_Y}$ = estimate of between ILCP method-averages reproducibility
s_{RXi}, s_{RYi} = the reproducibility standard deviations, evaluated at the method X and Y means of the i^{th} round robin sample
s_{rXi}, s_{rYi} = the repeatability standard deviations, evaluated at the method X and Y means of the i^{th} round robin sample
s_{Xi}, s_{Yi} = standard errors of the means i^{th} round robin sample
$\bar{X}$, $\bar{Y}$ = the weighted means of round robins (across samples)
x_i, y_i = deviations of the means of the i^{th} round robin sample results from $\bar{X}$ and $\bar{Y}$, respectively.
TSS_X, TSS_Y = total sums of squares, around $\bar{X}$ and $\bar{Y}$
F = a ratio for comparing variances; not unique—more than one use
v_X, v_Y = the degrees of freedom for reproducibility variances from the round robins
w_i = weight associated with the difference between mean results (or corrected mean results) from the i^{th} round robin sample
CSS = centered sum of squares, weighted sum of squared differences between (possibly corrected) mean results from the round robin
a,b = parameters of a linear correction: $\hat{Y} = a + bX$
t_1, t_2 = ratios for assessing reductions in sums of squares
R_{XY} = estimate of between methods reproducibility
$\hat{Y}$ = predicted Y-method value for a sample by applying the bias correction established from this practice to an actual X-method result for the same sample
$\hat{Y}_i$ = predicted i^{th} round robin sample Y-method mean, by applying the bias correction established from this practice to its corresponding X-method mean
ε_i = standardized difference between Y_i and $\hat{Y}_i$.
L_X, L_Y = harmonic mean numbers of laboratories submitting results on round robin samples, by X- and Y- methods, respectively
$R_{X\,\hat{Y}}$ = estimate of between methods reproducibility, computed from an X-method result only

4. Summary of Practice

4.1 Precisions of the two methods are quantified using inter-laboratory studies meeting the requirements of Practice D6300 or equivalent, using at least ten samples in common that span the intersecting scopes of the methods. The arithmetic means of the results for each common sample obtained by each method are calculated. Estimates of the standard errors of these means are computed.

NOTE 4—For established standard test methods, new precision studies generally will be required in order to meet the common sample requirement.

NOTE 5—Both test methods do not need to be run by the same laboratory. If they are, care should be taken to ensure the independent test result requirement of Practice D6300 is met (for example, by double-blind testing of samples in random order).

4.2 Weighted sums of squares are computed for the total variation of the mean results across all common samples for each method. These sums of squares are assessed against the standard errors of the mean results for each method to ensure that the samples are sufficiently varied before continuing with the practice.

4.3 The closeness of agreement of the mean results by each method is evaluated using appropriate weighted sums of squared differences. Such sums of squares are computed from the data first with no bias correction, then with a constant bias correction, then, when appropriate, with a proportional correction, and finally with a linear (proportional + constant) correction.

4.4 The weighted sums of squared differences for the linear correction is assessed against the total variation in the mean results for both methods to ensure that there is sufficient correlation between the two methods.

4.5 The most parsimonious bias correction is selected.

4.6 The weighted sum of squares of differences, after applying the selected bias correction, is assessed to determine whether additional unexplained sources of variation remain in the residual (that is, the individual Y_i minus bias-corrected X_i) data. Any remaining, unexplained variation is attributed to sample-specific biases (also known as method-material interactions, or matrix effects). In the absence of sample-specific biases, the between methods reproducibility is estimated.

4.7 If sample-specific biases are present, the residuals (that is, the individual Y_i minus *bias-corrected* X_i) are tested for randomness. If they are found to be consistent with a random-effects model, then their contribution to the between methods reproducibility is estimated, and accumulated into an all-encompassing between methods reproducibility estimate.

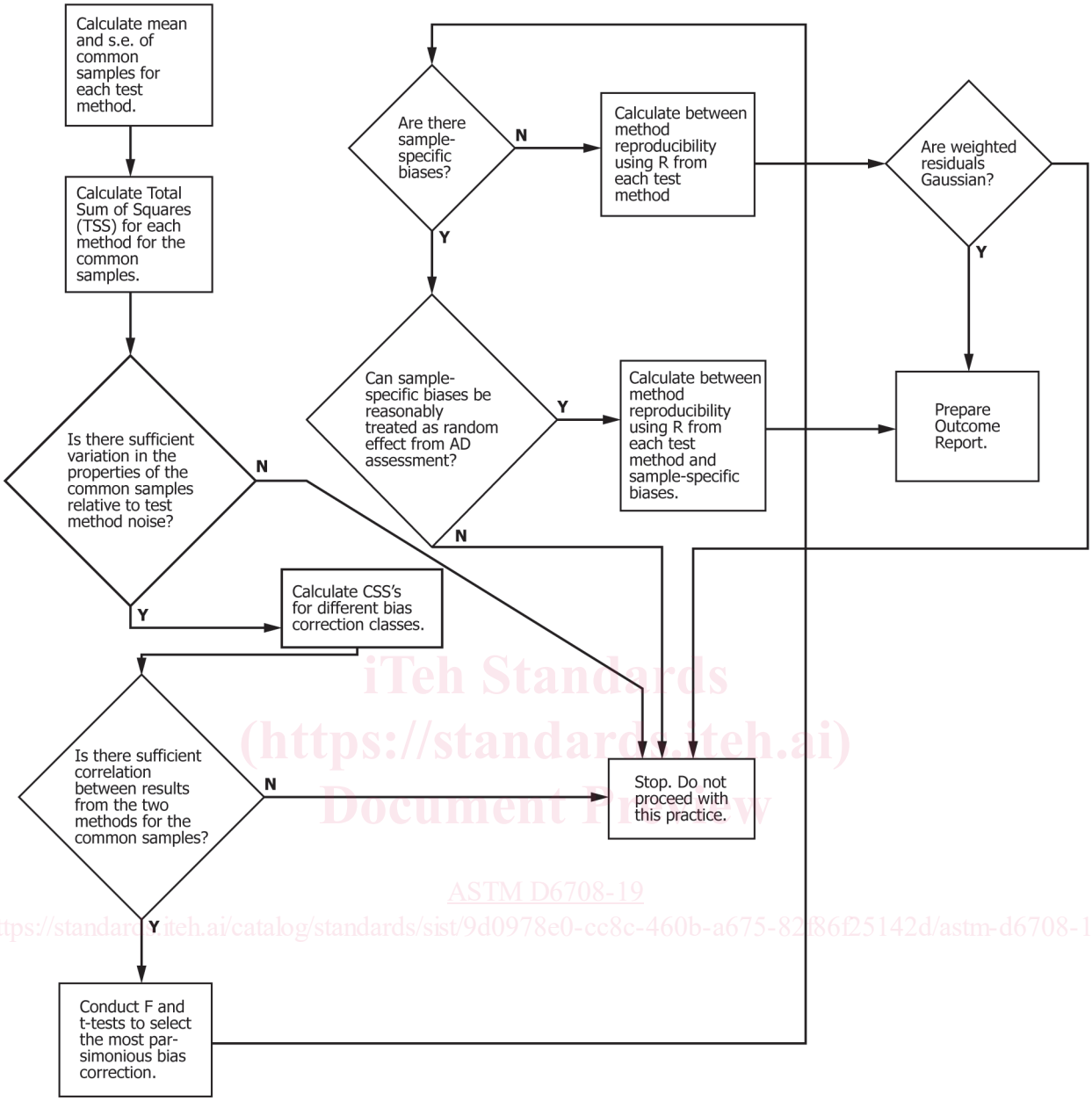

FIG. 1 Simplified Flow Diagram for this Practice

4.8 Refer to Fig. 1 for a simplified flow diagram of the process described in this practice.

5. Significance and Use

5.1 This practice can be used to determine if a constant, proportional, or linear bias correction can improve the degree of agreement between two methods that purport to measure the same property of a material.

5.2 The bias correction developed in this practice can be applied to a single result (X) obtained from one test method (method X) to obtain a *predicted* result ($\hat{Y}$) for the other test method (method Y).

NOTE 6—Users are cautioned to ensure that $\hat{Y}$ is within the scope of method Y before its use.

5.3 The between methods reproducibility established by this practice can be used to construct an interval around $\hat{Y}$ that would contain the result of test method Y, if it were conducted, with ~~about~~approximately 95 % ~~confidence.~~probability.

5.4 This practice can be used to guide commercial agreements and product disposition decisions involving test methods that have been evaluated relative to each other in accordance with this practice.

5.5 The magnitude of a statistically detectable bias is directly related to the uncertainties of the statistics from the experimental study. These uncertainties are related to both the size of the data set and the precision of the processes being studied. A large data set, or, highly precise test method(s), or both, can reduce the uncertainties of experimental statistics to the point where the "statistically detectable" bias can become "trivially small," or be considered of no practical consequence in the intended use of the test method under study. Therefore, users of this practice are advised to determine in advance as to the magnitude of bias correction below which they would consider it to be unnecessary, or, of no practical concern for the intended application prior to execution of this practice.

NOTE 7—It should be noted that the determination of this minimum bias of no practical concern is not a statistical decision, but rather, a subjective decision that is directly dependent on the application requirements of the users.

6. Procedure

NOTE 8—For an in-depth statistical discussion of the methodology used in this section, see Appendix X1. For a worked example, see Appendix X2.

6.1 Calculate sample means and standard errors from Practice D6300 results.

6.1.1 The process of applying Practice D6300 to the data may involve elimination of some results as outliers, and it may also involve applying a transformation to the data. For this practice, compute the mean results from data that have not been transformed, but with outliers removed in accordance with Practice D6300. The precision estimates from Practice D6300 are used to estimate the standard errors of these means.

6.1.2 Compute the means as follows:

6.1.2.1 Let X_{ijk} represent the k^{th} result on the i^{th} common material by the j^{th} lab in the round robin for method X. Similarly for Y_{ijk}. (The i^{th} material is the same for both round robins, but the j^{th} lab in one round robin is not necessarily the same lab as the j^{th} lab in the other round robin.) Let n_{Xij} be the number of results on the i^{th} material from the j^{th} X-method lab, after removing outliers, that is, the number of results in *cell* (i,j). Let L_{Xi} be the number of laboratories in the X-method round robin that have at least one result on the i^{th} material remaining in the data set, after removal of outliers. Let S be the total number of materials common to both round robins.

6.1.2.2 The mean X-method result for the i^{th} material is:

$$X_i = \frac{1}{L_{xi}} \sum_j \frac{\sum_k X_{ijk}}{n_{Xij}} \tag{1}$$

where, X_i is the average of the cell averages on the i^{th} material by method X.

6.1.2.3 Similarly, the mean Y-method result for the i^{th} material is:

$$Y_i = \frac{1}{L_{Yi}} \sum_j \frac{\sum_k Y_{ijk}}{n_{Yij}} \tag{2}$$

6.1.3 The standard errors (standard deviations of the means of the results) are computed as follows:

6.1.3.1 If s_{RXi} is the estimated reproducibility standard deviation from the X-method round robin, and s_{rXi} is the estimated repeatability standard deviation, then an estimate of the standard error for X_i is given by:

$$s_{Xi} = \sqrt{\frac{1}{L_{Xi}}\left[s_{RXi}^2 - s_{rXi}^2\left(1 - \frac{1}{L_{Xi}} \sum_j \frac{1}{n_{Xij}}\right)\right]} \tag{3}$$

NOTE 9—Since repeatability and reproducibility may vary with X, even if the L_{Xi} were the same for all materials and the n_{Xij} were the same for all laboratories and all materials, the $\{s_{Xi}\}$ might still differ from one material to the next.

6.1.3.2 s_{Yi}, the estimated standard error for Y_i, is given by an analogous formula.

6.2 Calculate the total variation sum of squares for each method, and determine whether the samples can be distinguished from each other by both methods.

6.2.1 The total sums of squares (TSS) are given ~~by~~by:

$$TSS_x = \sum_i \left(\frac{X_i - \bar{X}}{s_{Xi}} \right)^2 \text{ and } TSS_y = \sum_i \left(\frac{Y_i - \bar{Y}}{s_{Yi}} \right)^2 \tag{4}$$

~~where~~where:

$$\bar{X} = \frac{\sum_i \left(\frac{X_i}{s_{Xi}^2} \right)}{\sum_i \left(\frac{1}{s_{Xi}^2} \right)} \text{ and } \bar{Y} = \frac{\sum_i \left(\frac{Y_i}{s_{Yi}^2} \right)}{\sum_i \left(\frac{1}{s_{Yi}^2} \right)} \tag{5}$$

are weighted averages of all X_i's and Y_i's respectively.

6.2.2 Compare $F = TSS_X/(S\text{-}1)$ to the 95th percentile of Fisher's F distribution with $(S\text{-}1)$ and v_x degrees of freedom for the numerator and denominator, respectively, where v_X is the degrees of freedom for the reproducibility variance (Practice D6300, paragraph 8.3.3.3) for the X-method round robin. If F does not exceed the 95th percentile, then the X-method is not sufficiently precise to distinguish among the S samples. Do not proceed with this practice, as meaningful results cannot be produced.

6.2.3 In a similar manner, compare $F = TSS_Y/(S\text{-}1)$ to the 95th percentile of Fisher's F distribution, using the degrees of freedom of the reproducibility variance of the Y-method, v_Y, in place of v_X. Similarly, do not proceed with this practice if F does not exceed the 95th percentile.

NOTE 10—If one or both of the conditions of 6.2.2 and 6.2.3 are satisfied only marginally, it is unlikely that this practice will produce a meaningful outcome. The test in the next subsection will almost certainly fail.

6.3 Test whether the methods are sufficiently correlated.

6.3.1 Using the weights $\{w_i\}$ as computed in 6.4.1.1, Eq 6, calculate the weighted *correlation coefficient* ~~*r*~~*r:*

$$r = \frac{\sum w_i\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum w_i\left(X_i - \bar{X}\right)^2 \sum w_i\left(Y_i - \bar{Y}\right)^2}} \tag{6}$$

where X^- and Y^- are $\sum w_i X_i / \sum w_i$ and $\sum w_i Y_i / \sum w_i$, respectively.

6.3.2 Use r to calculate the F-statistic:

$$F = \frac{(S - 2)r^2}{1 - r^2} \tag{7}$$

6.3.3 Compare F to the 99th percentile of Fisher's F distribution with 1 and S-2 degrees of freedom in the numerator and denominator, respectively.

6.3.3.1 If F is less than the 99th percentile value, ~~then,~~then this practice concludes that the methods are too discordant to permit use of the results from one method to predict those of the other.

6.3.3.2 If F is greater than the tabled value, proceed to 6.5.

6.4 Calculate the centered sum of squares *(CSS)* statistic for each of the following classes of bias-correction methodology.

NOTE 11—The revised algorithms presented in this version of D6708 were developed in order to correct very rare cases in which the algorithms of previous versions do not converge to the optimal linear models. The rare cases generally involved data sets with poor correlations between the two methods. In the vast majority of data sets, including worked example of this practice, the old and the new algorithms converge to exactly the same optimal models. Continuing to use the old algorithms is a reasonable option provided the user verifies that the computed value of CSS1b is never larger than CSS0, and that the computed value of CSS2 is never larger than either CSS1a or CSS1b. If the aforementioned situation is detected using the old algorithms, then the outcome from this version is deemed to be the correct outcome.

6.4.1 *Class 0*—No bias correction.

6.4.1.1 Compute the weights (w_i) for each sample i:

$$w_i = \frac{1}{s_{Yi}^2 + s_{Xi}^2} \tag{8}$$

6.4.1.2 Compute *CSS*:

$$CSS_0 = \sum_i w_i (X_i - Y_i)^2 \tag{9}$$

6.4.2 *Class 1a*—Constant bias correction.

6.4.2.1 Using the weights (w_i) from 6.4.1.1, compute the constant bias correction (a):

$$a = \sum_i \frac{w_i (Y_i - X_i)}{\sum_i w_i} = \frac{\sum w_i Y_i}{\sum w_i} - \frac{\sum w_i X_i}{\sum w_i} \tag{10}$$

6.4.2.2 Compute *CSS*:

$$CSS_{1a} = \sum_i w_i \left(Y_i - (X_i + a)\right)^2 \tag{11}$$

6.4.3 *Class 1b*—Proportional bias correction.

6.4.3.1 The computations of this subsection (6.4.3) are appropriate only if both of the following conditions apply: (*1*) the measured property assumes only non-negative values, and (*2*) a property value of *zero* has a physical significance (for example, concentrations of specific constituents). In addition, it is not mandatory but highly recommended that $\max(Y_i) \geq 2 \min(Y_i)$.

6.4.3.2 The computations involve iterative calculation of the weights $\{w_i\}$ and the proportional correction b.

6.4.3.3 Set $b = 1$.

6.4.3.4 Compute the weight w_i for each sample i:

$$w_i = \frac{1}{s_{Yi}^2 + b^2 s_{Xi}^2} \tag{12}$$

6.4.3.5 Calculate the following three sums:

$$A = \sum w_i^2 X_i Y_i s_{Xi}^2 \tag{13}$$

$$B = \sum w_i^2 (X_i^2 s_{Yi}^2 - Y_i^2 s_{Xi}^2) \tag{14}$$

$$C = -\sum w_i^2 X_i Y_i s_{Yi}^2 \tag{15}$$

6.4.3.6 Calculate b_0:

$$b_0 = \frac{-B + \sqrt{B^2 - 4AC}}{2A} \tag{16}$$

6.4.3.7 If $|b - b_0| > .001\ b$, replace b with b_0 and go back to 6.4.3.4. Otherwise, the iteration can be stopped, as further iteration will not produce meaningful improvement. Replace b with b_0 and go on to 6.4.3.8.

6.4.3.8 Calculate the final weights $\{w_i\}$ as in 6.4.3.4.

6.4.3.9 Calculate CSS_{1b}:

$$CSS_{1b} = \sum w_i (Y_i - bX_i)^2 \tag{17}$$

6.4.4 *Class 2*—Linear (proportional + constant) bias correction.

6.4.4.1 This involves iterative calculation of the weights $\{w_i\}$, the weighted means of X_i's and Y_i's, and the proportional term b.

6.4.4.2 Set $b = 1$.

6.4.4.3 Compute the weight w_i for each sample i:

$$w_i = \frac{1}{s_{yi}^2 + b^2 s_{xi}^2} \tag{18}$$

6.4.4.4 Calculate the weighted means of $\{X_i\}$ and $\{Y_i\}$ respectively:

$$\bar{Y} = \frac{\sum w_i Y_i}{\sum w_i} \tag{19}$$

$$\bar{X} = \frac{\sum w_i X_i}{\sum w_i}$$

6.4.4.5 Calculate the deviations from the weighted means:

$$x_i = X_i - \bar{X} \tag{20}$$

$$y_i = Y_i - \bar{Y}$$

6.4.4.6 Calculate the three sums:

$$A = \sum w_i^2 x_i y_i s_{Xi}^2 \tag{21}$$

$$B = \sum w_i^2 (x_i^2 s_{Yi}^2 - y_i^2 s_{Xi}^2) \tag{22}$$

$$C = -\sum w_i^2 x_i y_i s_{Yi}^2 \tag{23}$$

6.4.4.7 Calculate b_0:

$$b_0 = \frac{-B + \sqrt{B^2 - 4AC}}{2A} \tag{24}$$

6.4.4.8 If $|b - b_0| > .001\ b$, replace b with b_0 and go back to 6.4.4.3, computing new values for the weights $\{w_i\}$, $\bar{X}$,$\bar{Y}$, $\{x_i\}$, $\{y_i\}$, and b_0. Otherwise, the iteration can be stopped, as further iteration will not produce meaningful improvement. Replace b with b_0 and go to 6.4.4.9.

6.4.4.9 Calculate the final weights $\{w_i\}$ as in 6.4.4.3.

6.4.4.10 Calculate CSS_2 and a:

$$CSS_2 = \sum w_i (y_i - bx_i)^2 \tag{25}$$

$$a = \bar{Y} - b\,\bar{X} \tag{26}$$

6.5 Conduct tests to select the most parsimonious bias correction class needed.

6.5.1 The centered sum of squares for differences from each class of bias correction are used to select the most parsimonious bias correction class that can improve the expected degree of agreement between the $\hat{Y}$ (the predicted Y-method result using X-method result) and the actual Y-method result on the same material. The classes of bias correction and the associated *CSS* as calculated earlier are repeated in the following table.

Bias Correction Class	*CSS*
Class 0–no correction	CSS_0
Class 1a–constant bias correction	CSS_{1a}
Class 1b–proportional bias correction (when appropriate)	CSS_{1b}
Class 2–linear (proportional + constant bias correction)	CSS_2

6.5.2 To determine whether *any* bias correction (~~*Classes*~~*Class 1a, 1b, or 2* above) can significantly improve the expected agreement between the two methods, calculate the following ratio:

$$F = \frac{(CSS_0 - CSS_2)/2}{CSS_2/(S-2)} \tag{27}$$

6.5.2.1 Compare F to the upper ~~95th~~95th percentile of the F distribution with 2 and S-2 degrees of freedom for the numerator and denominator, respectively.

6.5.2.2 If the calculated F is smaller, conclude that a bias correction of *Class 1a, 1b, or 2* does not sufficiently improve the expected agreement between the two methods, relative to *Class 0* (no bias correction). Proceed to 6.6.

6.5.2.3 If the calculated F is larger, conclude that a correction can improve the expected agreement between the two methods, and continue in 6.5.3.

6.5.3 If the F-value calculated in 6.5.2 is larger than the 95th percentile of F, compute the following t-ratios:

$$t_1 = \sqrt{\frac{CSS_0 - CSS_1}{CSS_2/(S-2)}} \tag{28}$$

$$t_2 = \sqrt{\frac{CSS_1 - CSS_2}{CSS_2/(S-2)}}$$

where, CSS_1 is the lesser of CSS_{1a} or CSS_{1b}, provided the latter is appropriate and has been calculated.

6.5.3.1 Compare t_2 to the upper 97.5th percentile of the t distribution with S-2 degrees of freedom.

6.5.3.2 If t_2 is larger, conclude that a bias correction of *Class 2* (proportional + constant correction) can improve the expected agreement over that of a single term (constant or proportional) correction alone (*Class 1*). Proceed to 6.6.

6.5.3.3 If t_2 is smaller than the t-percentile, compare t_1 to the same upper 97.5th percentile of the t distribution with (S-2) degrees of freedom.

6.5.3.4 If t_1 is larger, conclude that a single term bias correction of *Class 1* is preferred to a bias correction of *Class 2*. Use the constant correction unless CSS_{1b} is appropriate and is smaller than CSS_{1a}. Proceed to 6.6.

6.5.3.5 If t_1 is smaller, then neither t_1 nor t_2 is statistically significant. A bias correction of *Class 2* is preferred over single-term (constant or proportional) correction of *Class 1*.

6.6 Test for existence of sample-specific biases.

6.6.1 Compare the *CSS* of the bias-correction class selected in 6.5 to the 95th percentile value of a chi-square distribution with v degrees of ~~freedom~~freedom.

where:

~~v = S for *Class 0* (no bias) correction,~~
~~v = $S - 1$ for *Class 1a* or *Class 1b* (constant or proportional) correction~~
~~v = $S - 2$ for *Class 2* (linear) correction~~

v = S for *Class 0* (no bias) correction,
v = $S - 1$ for *Class 1a* or *Class 1b* (constant or proportional) correction, and
v = $S - 2$ for *Class 2* (linear) correction.

6.6.2 If the *CSS* is smaller than the chi-square percentile, it is reasonable to conclude that there are no sample-specific biases, that is, that there are no other sources of variation that are statistically observable above the measurement error. Perform the Anderson-Darling (A-D) assessment on the residuals as per 6.7.2.2 and 6.7.2.3. If the outcome is not significant at the 5 % level, calculate the between methods reproducibility (R_{XY}) as per Eq 29 below. If the A-D assessment is significant, application of the practice is considered terminated with failure at this point, as the statistical evidence suggests that a single between-method reproducibility (R_{XY}) cannot be found that is applicable to all materials covered by the intersecting scope of both test methods. It is reasonable to conclude that, at least for some materials, the test methods are not measuring the same property.

$$R_{XY} = \sqrt{\frac{R_Y^2 + b^2 R_X^2}{2}} \tag{29}$$

where:

b = the coefficient of the appropriate bias correction. (For *Class 0* and *Class 1a* bias corrections, b=1.)

6.6.3 If the *CSS* is larger than the chi-square percentile (see 6.6.1), there is strong evidence that biases between the methods have not been adequately corrected by the bias-corrections of 6.4. In other words, the relative biases are not consistent across the S common samples of the round robins. The user may wish to investigate whether the biases can be attributed to other observable properties of the samples. Or he or she may wish to restrict attention to a smaller class of materials for the purpose of establishing a between methods reproducibility. Such investigations are beyond the scope of this practice, as the issues typically are not statistical in nature. This practice does recommend investigating whether it is reasonable to treat the sample-specific biases as random effects, as described in 6.7.

6.7 *Treatment of Sample-Specific Relative Bias as a Variance Component:*

6.7.1 If the *CSS* exceeds the 95th percentile value of the appropriate chi-square distribution (see 6.6.1), there is strong evidence that sources other than measurement error are contributing towards the variation of the expected agreement between the two methods. In this practice, these sources are attributed to sample-specific effects (also known as matrix effects or method-material interactions). In some cases these sample-specific effects can be treated as *random* effects, and hence can be incorporated as an additional source of variation into a between methods reproducibility as described in this section. Note that, even when it is appropriate to treat these sample-specific effects as random, the additional variation may cause the between methods reproducibility to be far larger than the root mean square of the reproducibilities of the methods (Eq 29).

6.7.2 Examine residuals to assess reasonableness of *random effect* assumption.

6.7.2.1 Assess the reasonableness of the assumption that the sample-specific biases can be treated as random effects by examination of the distribution of the residuals. While there are numerous statistical tools available to perform this assessment, this practice recommends use of the Anderson-Darling normality test, based on its simplicity and ease of use. It is not the intent of this practice to exclude other tools for this purpose.

6.7.2.2 Let $\{\hat{Y}_i\}$ be the Y-method values predicted from the corresponding X-method mean values $\{X_i\}$, using the bias-correction selected in 6.5. The (standardized) residuals $\{\varepsilon_i\}$ are given by:

$$\varepsilon_i = \sqrt{w_i}\left(Y_i - \hat{Y}_i\right) \tag{30}$$

where:

$\{w_i\}$ = the appropriate weights from 6.4.1 – 6.4.4.

6.7.2.3 Calculate the Anderson Darling (AD) statistic on the residuals $\{\varepsilon_i\}$. (Refer to Practice D6299 for guidance on calculation and interpretation of this statistic.)

6.7.2.4 If the AD statistic is not significant at the 5 % significance level, conclude that the sample-specific relative bias may be treated as a variance component. Proceed to 6.7.3.

6.7.2.5 If the AD statistic is significant, there is strong evidence that the sample-specific effects cannot be treated as random effects. Application of this practice is considered terminated at this point, as the statistical evidence suggests that a single between methods reproducibility (R_{XY}) cannot be found that is applicable to all materials covered by the intersecting scope of both test methods. It is reasonable to conclude that, at least for some materials, the test ~~method~~methods are not measuring the same property. Do NOT proceeed to 6.7.3.

NOTE 12—It is possible that, by restricting the comparison to a narrower class of materials, a between methods reproducibility can be obtained (for that narrower class) that does not have sample-specific biases, or, has sample-specific biases that can be treated as a random effect. However, individual outlier materials should not be excluded from this study, after-the-fact, based on the statistics only, without other evidence that they clearly belong to a separate and identifiable class.

6.7.3 Calculate the between methods reproducibility (R_{XY}) as follows:

$$R_{XY} = \sqrt{\left(\frac{b^2R_X^2}{2} + \frac{R_Y^2}{2}\right)\left(1 + \frac{2(1.96)^2\,(CSS - S + k)S}{(S-k)\sum\frac{b^2R_{Xi}^2 + R_{Yi}^2}{b^2s_{Xi}^2 + s_{Yi}^2}}\right)} \tag{31}$$

where b and *CSS* are appropriate to the selected bias-correction, and k is 0 if the bias-correction is *Class 0*; k is 1 if the bias correction is *Class 1a* or *Class 1b*; or k is 2 if the bias-correction is *Class 2*.

NOTE 13—Eq 31 provides an estimate of the magnitude below which about 95 % of the differences are expected to fall, when one party uses the bias-corrected X-method while another party uses the Y-method, on materials similar to the round robin samples. Application of the methods to materials which are substantially different from these round robin materials may affect both the average bias and the variance of the random component. *Laboratories which engage in routine substitution of one method for another are advised to periodically monitor the deviations between methods, as a regular part of their quality assurance program.*

6.8 Construction of ~~a 95 % confidence interval for a single result from method Y~~ an interval using a single bias-corrected result from method X, and R_{XY}, that may contain, about 95 % of the time, a single result from method Y, if the latter is conducted on the same sample.

6.8.1 Let $\hat{Y}$ be a single bias-corrected X-method result. An interval bounded by $\hat{Y} \pm R_{XY\hat{}}$ can be expected to contain a single corresponding Y-method result, obtained on the identical ~~material, with approximately 95 % confidence.~~ material about 95 % of the time. Here $R_{XY\hat{}}$ is computed from Eq 29 or Eq 31, as appropriate, with R_Y evaluated at $Y = \hat{Y}$.

7. Report

7.1 Upon completion of the calculations, it is recommended that the assessment findings be reported in the Precision and Bias section of the appropriate test method(s). In the event that one of the test methods assessed is cited as a referee test method, with the other test method being an alternative, this practice recommends the following naming convention, indicating the publication year for method D YYYY by the addition of suffix "-yy", and the publication year for method XXXX by the addition of the suffix "-xx":

Referee Test Method designation: Test Method D YYYY-yy
Alternative Test Method designation: Test Method D XXXX-xx

TABLE 1 Summary of Findings[A]

A	B	C	D1	D2	D3	Assessment Outcome
Is there adequate variation in the property level of the sample set relative to Test Method XXXX and Test Method YYYY reproducibilities?	Is there adequate correlation between the test results from Test Method XXXX and Test Method YYYY?	Will a scaling/bias correction significantly improve the agreement between the results from Test Method XXXX and Test Method YYYY over and above their combined reproducibilities?	Are there sample-specific biases?	If yes to (D1), can these biases be treated as a random effect?	If no to (D1), are the residuals randomly scattered?	
Yes	Yes	No	No	N/A	Yes	Pass (A1)
Yes	Yes	No	No	N/A	**No**	**Fail (B4)**
Yes	Yes	No	Yes	Yes	N/A	Pass (A2)
Yes	Yes	No	Yes	**No**	N/A	**Fail (B3)**
Yes	Yes	Yes	No	N/A	Yes	Pass (A3)
Yes	Yes	Yes	No	N/A	**No**	**Fail (B4)**
Yes	Yes	Yes	Yes	Yes	N/A	Pass (A4)
Yes	Yes	Yes	Yes	**No**	N/A	**Fail (B3)**
Yes	**No**	N/A	N/A	N/A	N/A	**Fail (B2)**
No	N/A	N/A	N/A	N/A	N/A	**Fail (B1)**

[A] Boldfaced type indicates reason for failure.

7.2 Report assessment findings in the Precision and Bias section of the appropriate test method, under a subsection titled "Between-Method Bias," as follows:

~~*Degree of Agreement between results by Test Method D XXXX and Test Method D YYYY-yy*—Results on the same materials produced by Test Method D XXXX and Test Method D YYYY-yy have been assessed in accordance with procedures outlined in Practice D6708. The findings are: *(report the findings here)*~~

Degree of Agreement between results by Test Method D XXXX and Test Method D YYYY-yy—Results on the same materials produced by Test Method D XXXX and Test Method D YYYY-yy have been assessed in accordance with procedures outlined in Practice D6708. The findings are: *(report the findings here).*

7.2.1 To choose the appropriate findings, see Table 1. (A) represents passing, and (B) represents failure. Choose one of the following findings (A1, A2, A3, A4, B1, B2, B3, or B4).

7.2.1.1 If the finding is **A1**, and R_X, estimated with at least 30 degrees of freedom, is less than or equal to 1.2 published R_Y, report the following for property range where R_X satisfies the aforementioned requirement.

No bias-correction considered in Practice D6708 can further improve the agreement between results from Test Method D XXXX and Test Method D YYYY-yy for the materials studied (reference Research Report ZZZZ). For applications where Test Method X is used as an alternative to Test Method Y, results from Test Method D XXXX and Test Method D YYYY-yy may be considered to be statistically indistinguishable, for sample types and property ranges listed below. No sample-specific bias, as defined in Practice D6708, was observed for the materials studied.

~~Sample types and property range where results from method D XXXX and DYYYY-yy may be considered to be statistically indistinguishable are: *(list applicable sample types and property ranges here)*~~
Sample types and property range where results from method D XXXX and DYYYY-yy may be considered to be statistically indistinguishable are: *(list applicable sample types and property ranges here).*

7.2.1.2 If the finding is **A1**, for property range where R_X does not meet the requirement listed above, report the following:

~~No bias-correction considered in Practice D6708 can further improve the agreement between results from Test Method D XXXX and Test Method D YYYY-yy for the materials studied (reference Research Report ZZZZ). No sample-specific bias, as defined in Practice D6708, was observed for the materials and property range listed below. *(list sample types and property ranges for above findings here)*~~
No bias-correction considered in Practice D6708 can further improve the agreement between results from Test Method D XXXX and Test Method D YYYY-yy for the materials studied (reference Research Report ZZZZ). No sample-specific bias, as defined in Practice D6708, was observed for the materials and property range listed below. *(List sample types and property ranges for above findings here.)*