Designation: D 6708 – 01$^{\epsilon 1}$

An American National Standard

Standard Practice for Statistical Assessment and Improvement of the 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 D 6708; 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.

ϵ^1 NOTE—Corrected Eq 8 editorially in November 2003.

1. Scope

1.1 This practice defines 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 D 6300 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.

NOTE 1—Examples of standard test methods are those developed by voluntary consensus standards bodies such as ASTM, IP/BSI, DIN, AFNOR, CGSB.

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 2—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 limit (designated by this practice as the *cross-method reproducibility*) 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.

NOTE 3—Users are cautioned against applying the cross-method 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) 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 cross method reproducibility that adequately cover the range of 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.

2. Referenced Documents

2.1 *ASTM Standards:*

D 5580 Test Method for Determination of Benzene, Toluene, Ethylbenzene plm-Xylene, o-Xylene, C_9 and Heavier Aromatics and Total Aromatics in Finished Gasoline by Gas Chromatography[2]

D 5769 Test Method for Determination of Benzene, Toluene, and Total Aromatics in Finished Gasoline by Gas Chromatography/Mass Spectrometry[2]

D 6299 Practice for Applying Statistical Quality Assurance Techniques to Evaluate Analytical Measurement System Performance[2]

D 6300 Practice for Determination of Precision and Bias Data for Use in Test Methods for Petroleum Products and Lubricants[2]

2.2 *ISO Standard*[3]

ISO 4259 Petroleum Products—Determination and application of precision data in relation to methods of test.

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

Current edition approved Aug. 10, 2001. Published October 2001.

[2] *Annual Book of ASTM Standards*, Vol 05.03.

[3] Available from American National Standards Institute, 11 W. 42nd St., 13th floor, New York, NY 10036.

3. Terminology

3.1 *Definitions:*

3.1.1 *closeness 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.

3.1.2 *cross-method 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 limit for the difference between two such single and independent results.

3.1.2.1 *Discussion*—A statement of cross-method reproducibility must include a description of any bias correction used in accordance with this practice.

3.1.2.2 *Discussion*—Cross-method 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 6.7.)

3.1.3 *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
s_{RXi}, s_{RYi}	=	the reproducibility standard deviations, evaluated at the means of the i^{th} round robin sample
s_{rXi}, s_{rYi}	=	the repeatability standard deviations, evaluated at the 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	=	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 cross-method reproducibility
$\hat{Y}$	=	Y-method value predicted from X-method result
$\hat{Y}_i$	=	i^{th} round robin sample Y-method mean, predicted from 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 cross-method 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 D 6300 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 D 6300 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 cross-method 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 cross-method reproducibility is estimated, and accumulated into an all-encompassing cross-method reproducibility estimate.

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 cross-method 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 95 % confidence.

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.

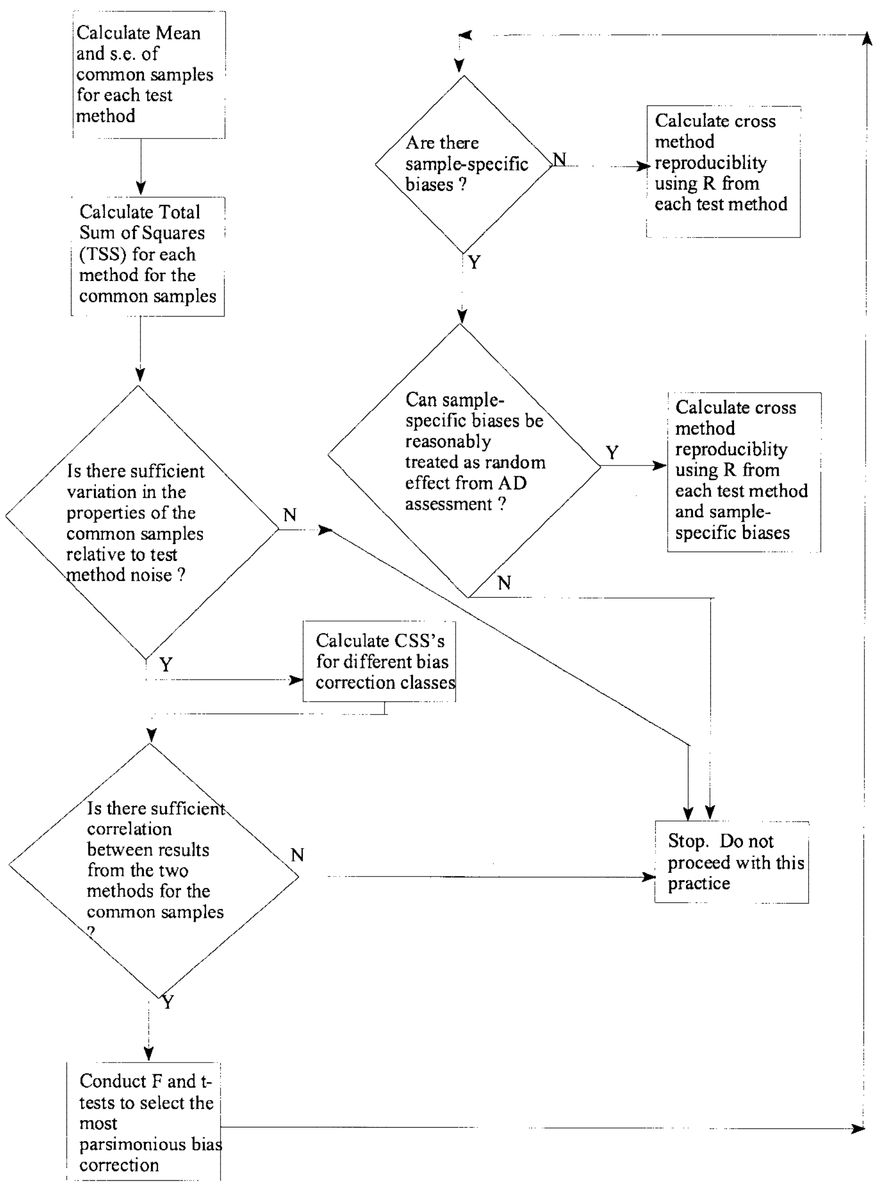

FIG. 1 Simplified Flow Diagram for this Practice

6. Procedure

NOTE 7—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 D 6300 results.

6.1.1 The process of applying Practice D 6300 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 D 6300. The precision estimates from Practice D 6300 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 repeatibility 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 8—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:

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

where:

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

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-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 D 6300, 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 9—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 meaningful results since in 6.4, the quantity ($TSS_X + TSS_Y$) will be compared to a closeness sum of squares computed in the next section, to determine whether the methods are sufficiently correlated. It will be difficult to meet that correlation requirement if the samples are too similar to one another.

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

6.3.1 *Class 0*—No bias correction.

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

$$w_i = \frac{1}{S_{Yi}^2 + S_{Xi}^2} \tag{6}$$

6.3.1.2 Computes *CSS*:

$$CSS_0 = \sum_i w_i (X_i - Y_i)^2 \tag{7}$$

6.3.2 *Class 1a*—Constant bias correction.

6.3.2.1 Using the weights (w_i) from 6.3.1.1, compute the constant bias correction (a):

$$a = \sum_i \frac{w_i(Y_i - X_i)}{\sum_i w_i} = \frac{\sum w_i Y_i}{\sum w_i} - \frac{\sum w_i X_i}{\sum w_i} \tag{8}$$

6.3.2.2 Compute *CSS*:

$$CSS_{1a} = \sum_i w_i (Y_i - (X_i + a))^2 \tag{9}$$

6.3.3 *Class 1b*—Proportional bias correction.

6.3.3.1 The computations of this subsection (6.3.3) 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 $\max(Y_i) \geq 2 \min(Y_i)$.

6.3.3.2 The computations involve iterative calculation of the weights (w_i) and the proportional correction (b).

6.3.3.3 Set $b = 1$.

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

$$w_i = \frac{1}{S_{Yi}^2 + b^2 S_{Xi}^2} \tag{10}$$

6.3.3.5 Calculate b_0:

$$b_0 = \frac{\sum w_i X_i Y_i}{\sum w_i X_i^2 - \sum w_i^2 s_{Xi}^2 (Y_i - bX_i)^2} \tag{11}$$

.

6.3.3.6 If $|b - b_0| > .001\ b$, replace b with b_0 and go back to 6.3.3.4. Otherwise, the iteration can be stopped, as further iteration will not produce meaningful improvement. Replace b with b_0 and go on to 6.3.3.7.

6.3.3.7 Calculate CSS_{1b}:

$$CSS_{1b} = \sum w_i (Y_i - bX_i)^2 \tag{12}$$

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

6.3.4.1 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).

6.3.4.2 Set $b = 1$.

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

$$w_i = \frac{1}{s_{Yi}^2 + b^2 s_{Xi}^2} \tag{13}$$

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

$$\bar{X} = \frac{\sum w_i X_i}{\sum w_i} \tag{14}$$

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

6.3.4.5 Calculate the deviations from the weighted means:

$$x_i = X_i - \bar{X} \tag{15}$$

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

6.3.4.6 Calculate b_0:

$$b_0 = \frac{\sum w_i x_i y_i}{\sum w_i x_i^2 - \sum w_i^2 s_{Xi}^2 (y_i - bx_i)^2} \tag{16}$$

6.3.4.7 If $|b - b_0| > .001\ b$, replace b with b_0 and go back to 6.3.4.3, 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 6.3.4.8.

6.3.4.8 Calculate CSS_2 and a:

$$CSS_2 = \sum w_i (y_i - bx_i)^2 \tag{17}$$

$$a = \bar{Y} - b\,\bar{X} \tag{18}$$

6.4 Test whether the methods are sufficiently correlated.

6.4.1 Calculate the F-statistic:

$$F = \frac{(TSS_X + TSS_Y - CSS_2)/S}{CSS_2/(S-2)} \tag{19}$$

6.4.2 Compare F to the 95th percentile of Fisher's F distribution with S and S-2 degrees of freedom in the numerator and denominator, respectively.

6.4.2.1 If F is less than the 95th percentile value, 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.

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

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

6.5.1 The closeness sums 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 1a, 1b or 2* above) can significantly improve the expected agreement between the two methods, calculate the following ratio:

$$F = \frac{(CSS_0 - CSS_2)/2}{CSS_2/(S-2)} \tag{20}$$

6.5.2.1 Compare F to the upper 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:

$$t_1 = \sqrt{\frac{CSS_0 - CSS_1}{CSS_2/(S-2)}} \tag{21}$$

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

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

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

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 besides measurement error. Calculate the cross method reproducibility (R_{XY}) as follows:

$$R_{XY} = \sqrt{\frac{R_Y^2 + b^2 R_X^2}{2}} \tag{22}$$

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 6.3. 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 cross-method 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 cross method 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 cross-method reproducibility to be far larger than the root mean square of the reproducibilities of the methods (Eq 22).

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 effects can be treated as random effect 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 $\{\epsilon_i\}$ are given by

$$\epsilon_i = \sqrt{w_i}(Y_i - \hat{Y}_i) \tag{23}$$

where:
$\{w_i\}$ = the appropriate weights from 6.3.1-6.3.4.

6.7.2.3 Calculate the Anderson Darling (AD) statistic on the residuals $\{\epsilon_i\}$. (Refer to Practice D 6299 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 cross-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 method are not measuring the same property. Do NOT proceeed to 6.7.3.

NOTE 10—It is possible that, by restricting the comparison to a narrower class of materials, a cross-method 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 cross-method reproducibility (R_{XY}) as follows:

$$R_{XY} = \sqrt{\frac{b^2 R_X^2}{2}\left[1+\frac{1}{L_X}\left(\frac{CSS}{S-k}-1\right)\right]+\frac{R_Y^2}{2}\left[1+\frac{1}{L_Y}\left(\frac{CSS}{S-k}-1\right)\right]} \tag{24}$$

where:

$$L_X = \frac{S}{\sum_i 1/L_{Xi}}$$

$$L_Y = \frac{S}{\sum_i 1/L_{Yi}}$$

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

NOTE 11—Eq 24 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 using a single bias-corrected result from method X, and R_{XY}.

6.8.1 Let $\hat{Y}$ be a single bias-corrected X-method result. An interval bounded by $\hat{Y} \pm R_{X\ \hat{Y}}$ can be expected to contain a single corresponding Y-method result, obtained on the identical material, with approximately 95 % confidence. Here $R_{X\ \hat{Y}}$ is computed from Eq 22 or Eq 24, as appropriate, with R_Y evaluated at $Y = \hat{Y}$.

APPENDIXES

(Nonmandatory Information)

X1. STATISTICAL BASIS

X1.1 Adequacy of Round Robin Sample Set

X1.1.1 In order to obtain a usable comparison between two methods, it is critical that the samples are sufficiently varied that they can be distinguished from one another (or at least so that some can be distinguished from some others) using the methods in question. The most straight-forward test involves the total (weighted) sum of squares, which, for the X measurement is

$$TSS_X = \sum_i \left(\frac{X_i - \bar{X}}{s_{Xi}} \right)^2 \tag{X1.1}$$

where:

$$\bar{X} = \frac{\sum_i \left(\frac{X_i}{s_{Xi}^2} \right)}{\left(\frac{1}{\sum_i s_{Xi}^2} \right)} \tag{X1.2}$$

the mean of the mean X-results weighted by the reciprocal of the squares of the standard errors $\{s_{Xi}\}$.

X1.1.2 If the S samples were all the same material, if the $\{X_i\}$ were distributed normally, and if the standard errors were known exactly, then TSS_X would have a chi-square distribution with S-1 degrees of freedom. In practice, the $\{s_{Xi}\}$ are not known exactly, but our situation approximates one in which $TSS_X/(S-1)$ would have an F distribution, with S-1 degrees of freedom in the numerator and v degrees of freedom in the denominator, where v is the degrees of freedom associated with the reproducibility estimate.

X1.1.3 If the materials were not all the same, then we would expect $TSS_X/(S-1)$ to be larger than an F-distributed variable. For round robins, hopefully samples will have been selected with a range of property values, so $TSS_X/(S-1)$ will be very much larger than the 95th percentile of F. If we come even close to failing this test, or the analogous test using the Y-method data, then the best course of action would be to start over with a more variable set of samples.

X1.2 Quantifying the Closeness of Agreement Between Two Test Methods

X1.2.1 Suppose we use a calibration function, f (X), to estimate (or *predict*) the property as measured by a reference Y-method. For the round robin samples, the mean result by the reference method, Y, can be compared to f (X) and used to quantify the closeness of agreement. In classical (weighted) regression, the weighted residual sum of squares,

$$\sum_i \frac{(Y_i - f(X_i))^2}{s_{Yi}^2} \tag{X1.3}$$

is used as a measure of the closeness of agreement. If competing calibration functions are under consideration, regression methods – classical least squares – suggest we should prefer the one with smallest sum of squares (X1.1). But this overlooks the fact that the $\{X_i\}$ are not the true values of the property as measured by the alternative method, but only estimates of those values, and they also involve random error. Let $\{h_i\}$ represent the true, unknown values of the property as measured by the reference method. The $\{h_i\}$ will be estimated from the data. Both Y_i and f (X_i) estimate h_i, which is not known. Y_i has variance s_{Yi}^2, and f (X_i) has variance approximately $f'^2(X_i)s_{Xi}^2$, where $f'(X_i)$ is the derivative of f at X_i. So an alternative measure of closeness is

$$\min_{\{h_i\}} \sum_i \left(\frac{(Y_i - h_i)^2}{S_{Yi}^2} + \frac{(f(X_i) - h_i)^2}{f'^2(X_i)s_{Xi}^2} \right) \tag{X1.4}$$

X1.2.2 This sum can be minimized term by term. The value of h_i that minimizes the ith term – and the value that is our best estimate of the true value – is

$$\hat{h}_i = \frac{f'^2(X_i)s_{Xi}^2 Y_i + s_{Yi}^2 f(X_i)}{s_{Yi}^2 + f'^2(X_i)s_{Xi}^2} \tag{X1.5}$$

and the minimized sum of squares is

$$CSS = \sum_i \frac{(Y_i - f(X_i))^2}{s_{Yi}^2 + f'^2(X_i)s_{Xi}^2} \tag{X1.6}$$

X1.2.3 Compare (Eq X1.4) to (Eq X1.1), and note that the only difference is that, in place of the variance of Y_i in the denominator of each term, (Eq X1.4) has the variance of Y_i-f (X_i).

X1.3 Properties of the Closeness Metric

X1.3.1 *Distributional Properties*:

X1.3.1.1 If the $\{X_i\}$ and $\{Y_i\}$ are independent normal, if the standard errors are known exactly, if f is linear (so that $\{f(X_i)\}$ are normal), and if E[Y_i] = E[f (X_i)] for all i, where E[Y] represents the mean or expected value of distribution of Y, then CSS has a chi-square distribution. The degrees of freedom associated with CSS is S, the number of materials (samples) common to the round robins. This may be seen by the fact that (Eq X1.2) has $2S$ terms, but S parameters $\{h_i\}$ are fitted by least-squares.

X1.3.1.2 When $E[Y_i], \neq E[f(X_i)]$, it may be because the calibration function, f, is not known exactly. If f belongs to a specific class of functions – linear functions, for example – then the unknown parameters of f (for example, a and b if $f(X) = a + b X$) may be estimated by minimizing Eq X1.4 with respect to these parameters. In this case, *CSS* would be distributed as chi-square with $S - k$ degrees of freedom.

X1.3.1.3 But if *CSS* is evaluated using an incorrect calibration equation, or by minimizing over a class of equations that does not contain the true calibration equation, or if there are sample-specific biases that cannot be accounted for by *any* calibration function, then *CSS* can be expected to be *larger* than a chi-square variable. The last of these three situations is worth special consideration. In the event that two or more different materials may have the same true value, $E[Y]$, as measured by one method, but different true values, $E[X]$, as measured by the other method, then no calibration equation can completely account for the differences between the two methods. Such sample-specific biases can be the dominant contributor to *CSS*. In fact, it almost certainly will be the dominant factor when $\{X_i\}$ and $\{Y_i\}$ are very precise, that is, when the materials are measured by sufficiently large numbers of labs. In such cases, note that an h_i of Eq X1.3 will approximate neither $E[Y_i]$ nor $E[X_i]$, but instead approximates an average of the two, an average that is weighted towards the more precise of Y_i and X_i.

X1.3.1.4 When the standard errors are not known, but approximately proportional to the same standard deviation estimate, then an F distribution may be a reasonable approximation to the distribution of CSS/S, or $CSS/(S - k)$, as appropriate.

X1.3.2 *Symmetry in X and Y*:

X1.3.2.1 Note that, if f is linear, then (Eq X1.4) is independent of which method is considered the reference method. If instead of predicting Y with $f(X)$, we wish to predict X with $f^{-1}(Y)$, then $f'(X_i) \equiv b \equiv 1/f^{-1\prime}(Y_i)$, and $Y_i - f(X_i) = b\,(f^{-1}(Y_i) - X_i)$, so b^2 cancels from the top and bottom of each term and Eq X1.4 is unchanged.

X1.3.2.2 This symmetry property is not shared by classical regression – the slope obtained from regressing Y on X is always smaller than the reciprocal of the slope from regressing X on Y. The method developed in this annex is a weighted version of what is known as regression with errors in both variables, which is discussed in many texts.[4] For non-linear f, the symmetry is lost. But for smooth f, the two equalities above are almost still true.

[4] Mandel, John, Evaluation and Control of Measurements, Marcel Dekker, 1991, Sec. 5.5.

X2. A WORKED EXAMPLE

X2.1 Example Data

X2.1.1 The data in Tables X2.1 and X2.2 are from a round robin for aromatics in gasoline conducted by seven labs. Fifteen ($S = 15$) fuels were tested by two methods. Table X2.1 are the results from Test Method D 5580, a gas chromatography (GC) method, while Table X2.2 contains the results from Test Method D 5769, gas chromatography/mass spectrometry (GC/MS). No data have been removed as outliers, but some repeat results are missing for Test Method D 5580. For purposes of this example designate Test Methods D 5580 and D 5769 as the X and Y methods, respectively.

TABLE X2.1 Aromatics by Test Method D 5580

	Fuel														
Laboratory	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	23.76	26.34	25.14	22.76	29.10	14.83	19.77	42.61	21.77	19.85	37.40	31.53	16.48	19.26	13.26
	24.22				29.16					19.81					12.99
2	24.46	25.88	25.72	22.59	29.08	15.68	19.92	41.89	21.68	19.97	37.38	31.35	16.55	19.48	13.25
	24.59	25.94	25.76	22.57	29.07	15.64	19.82	42.10	22.00	20.02	37.09	31.29	16.58	19.63	13.53
3	24.50	25.36	26.28	22.87	29.28	15.71	20.12	42.90	21.93	20.02	38.05	31.63	16.72	19.72	13.50
	24.54	25.17	26.26	22.65	29.33	15.76	20.01	42.90	21.91	20.14	38.07	31.80	16.60	19.82	13.54
4	24.74	25.23	25.72	22.82	29.31	15.51	20.35	42.52	22.24	20.32	37.03	31.77	16.50	20.03	13.63
	24.90	25.19	25.65	22.68	29.21	15.48	19.99	42.38	22.14	20.01	37.44	31.80	16.45	19.84	13.69
5	24.64	26.01	25.92	22.17	30.50	14.78	19.37	43.71	22.85	20.43	37.80	31.09	16.27	20.85	13.85
	24.70	25.87	25.87	22.20	30.69	14.88	19.66	44.00	23.50	20.30	37.84	31.31	16.55	21.01	13.85
6	24.93	26.28	26.07	22.59	30.08	15.91	20.30	43.08	22.24	20.26	38.28	32.60	16.70	19.94	13.67
	25.13	26.72	26.08	22.90	30.10	16.16	20.49	43.27	22.56	20.58	38.54	32.72	16.97	19.94	13.89
7	24.37	25.40	25.66	21.93	29.11	15.30	19.33	42.08	21.88	19.79	36.28	30.60	15.87	19.30	12.91
	24.36	25.36	25.72	21.97	29.18	15.10	19.32	41.77	21.98	19.71	37.19	30.65	15.91	19.23	12.91
Mean	24.56	25.79	25.78	22.53	29.51	15.40	19.87	42.70	22.17	20.09	37.56	31.55	16.47	19.81	13.46
Standard Error	0.177	0.181	0.181	0.170	0.193	0.140	0.159	0.234	0.168	0.160	0.219	0.201	0.145	0.159	0.131

TABLE X2.2 Aromatics by Test Method D 5769

Laboratory	Fuel 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	21.33	21.37	22.21	20.90	26.19	10.88	15.88	38.58	18.66	16.81	33.14	27.87	14.74	17.72	11.78
	22.01	21.12	21.99	20.98	25.88	10.93	16.07	38.39	18.41	17.21	33.76	28.39	14.77	17.68	12.12
2	21.70	21.32	22.20	20.79	26.85	11.60	16.26	40.33	19.29	17.41	34.32	29.28	14.99	18.10	12.31
	21.79	21.15	22.60	20.69	26.57	11.84	16.25	38.86	18.79	17.28	33.99	28.48	14.86	18.13	12.24
3	24.09	23.36	24.71	22.40	27.99	12.45	17.31	41.40	20.65	19.83	35.18	29.96	16.24	19.81	12.94
	24.32	23.57	24.93	22.26	28.08	12.31	17.26	41.36	20.88	18.94	36.35	29.82	16.43	19.42	12.81
4	23.43	22.59	24.15	21.55	27.58	12.23	17.09	41.04	20.14	18.53	35.80	30.28	15.39	18.23	12.52
	23.08	22.54	23.99	21.61	27.50	12.36	17.15	41.11	20.37	18.46	35.98	30.12	15.43	18.23	12.59
5	23.63	22.65	24.54	21.26	28.10	12.52	17.49	41.79	20.47	18.73	35.67	30.01	15.74	18.99	12.31
	24.33	22.69	24.88	22.36	28.24	12.48	17.26	40.71	20.29	18.31	35.84	30.03	16.03	18.73	12.30
6	22.38	20.43	22.70	20.13	26.34	11.27	15.72	38.89	18.74	17.13	34.29	27.73	14.97	18.56	12.17
	22.53	20.40	22.86	20.39	26.44	11.24	15.54	39.13	18.71	17.26	34.74	27.85	15.01	18.59	12.05
7	22.84	21.79	22.90	20.85	27.10	11.33	16.36	40.88	19.50	17.76	34.93	28.80	15.05	17.82	12.01
	22.72	21.76	23.32	20.25	26.47	11.33	16.79	40.27	19.42	17.50	34.71	29.11	14.87	17.56	11.99
Mean	22.87	21.91	23.43	21.17	27.10	11.77	16.60	40.20	19.59	17.94	34.91	29.12	15.32	18.40	12.30
Standard Error	0.345	0.330	0.353	0.319	0.408	0.177	0.250	0.606	0.295	0.270	0.526	0.439	0.231	0.277	0.185

NOTE X2.1—Note: All equations referenced are from this standard except as noted.

X2.1.2 The repeatabilities and reproducibilities were estimated from the round robins in accordance with Practice D 6300. These are shown in Table X2.3. The degrees of freedom are also from the precision analysis. The standard deviations associated with repeatability and reproducibility are obtained by dividing the precision estimates by $t_{.975}\sqrt{2}$, where $t_{.975}$ is the 97.5$^{\text{th}}$ percentile of the t-distribution with the applicable number of degrees of freedom.

X2.2 Calculation of the Mean Results and Standard Errors

X2.2.1 Both round robins included seven participants, and all participants measured every sample, so $L_{Xi}= L_{Yi}= 7$ for all i. As an example, for the second sample from method X, X_2 is calculated using (Eq 1) as follows:

$$X_2 = \frac{1}{7}\left(\frac{26.34}{1} + \frac{25.88 + 25.94}{2} + \frac{25.36 + 25.17}{2} + ... + \frac{25.4 + 25.36}{2}\right) \quad \text{(X2.1)}$$

$$= \frac{1}{7}(26.34+25.91+25.265+25.21+26.94+26.5+25.38)=25.79$$

X2.2.2 Note that this is not the same as the average of the thirteen X-method results on this sample. The remaining X_i and Y_i are computed in a similar fashion.

X2.2.3 The standard error of each mean is calculated using Eq 3. Again for the second sample X-method results, the n_{i2} are all equal to 2, except $n_{1,2}= 1$, so

$$\frac{1}{L_{Xi}}\sum_j \frac{1}{n_{Xij}} = \frac{4}{7} \text{ and } s_{Xi} = \sqrt{\frac{1}{7}\left[.0964^2 - 0.0296^2\left(\frac{3}{7}\right)\right]}\sqrt{25.79} = 0.181. \quad \text{(X2.2)}$$

X2.2.4 The means and standard errors for each fuel by both methods are found at the bottoms of their respective tables (Tables X2.1 and X2.2).

X2.3 Calculate the Total Variation Sum of Squares

X2.3.1 Table X2.4 demonstrates the application of Eq 4 and 5 to obtain the total sum of squares for the Y-method means. The weighted mean, $\bar{Y}$, is found to be 3333.81/186.8 = 17.85. TSS_Y= 6564.8. We compare 6564.8/14 = 469 to the 95$^{\text{th}}$ percentile of the F distribution with 14 and 9 degrees of freedom for the numerator and denominator, respectively. The F percentile is 3.03. Hence, we conclude TSS_Y is highly statistically significant. Similarly, a high degree of significance is also found for TSS_X.

X2.4 Calculate the Closeness Sums of Squares (*CSS*)

X2.4.1 *Class 0*—No correction. The first three columns of Table X2.5 display the computations from Eq 6 and Eq 7. As shown in the next-to-last line in the table, CSS_0 turns out to be 812.46.

X2.4.2 *Class 1a*—Constant correction. Table X2.5 contains these computations, also. Note that $\bar{Y}_i$ is smaller than $\bar{X}_i$ for all samples, so it is not surprising that CSS_{1a} is quite a bit smaller than CSS_0. $a = \bar{Y} - \bar{X} = 18.36 - 20.62 = -2.26$.

X2.4.3 *Class 1b*—Proportional correction.

X2.4.3.1 Aromatics concentration having a true zero, and as $\max(Y_i) = 40.2 > 23.54 = 2\min(Y_i)$, it is appropriate to also consider a proportional correction. Table X2.6 shows the computations for the first two iterations. Starting with $b = 1$, the first iteration proceeds using w_i's from Table X2.5. Computing b_0:

TABLE X2.3 Precision Estimates and Associated Standard Deviations[A]

Precision Estimates	Degrees of Freedom	t (.975)	Standard Deviations
r_X=0.0831 $\sqrt{X}$	94	1.986	s_{rX}=0.0290 $\sqrt{X}$
R_X=0.2792 $\sqrt{X}$	28	2.048	s_{RX}=0.0964 $\sqrt{X}$
r_Y=0.0292 Y	105	1.983	s_{rY}=0.0104 Y
R_Y=0.1292 Y	9	2.262	s_{RX}=0.0404 Y

[A] This inter-laboratory study did not meet the minimum degrees of freedom requirement (30) as recommended in Practice D 6300. The low degrees of freedom for R_X and R_Y suggest the need for further inter-laboratory standardization, and the latter could be a contributing factor towards the sample-specific biases observed.

TABLE X2.4 Total Variation Sum of Squares for Y-Method

i	Y_i	s_{Yi}	$1/s_{Yi}^2$	$Y_i\ /s_{Yi}^2$	$(Y_i - \bar{Y})^2/s_{Yi}^2$
1	22.87	0.345	8.42	192.57	212.48
2	21.91	0.330	9.17	201.01	151.48
3	23.43	0.353	8.02	187.99	249.90
4	21.17	0.319	9.82	208.01	108.70
5	27.10	0.408	6.00	162.54	513.12
6	11.77	0.177	31.80	374.21	1174.31
7	16.60	0.250	15.98	265.27	24.75
8	40.20	0.606	2.73	109.57	1361.51
9	19.59	0.295	11.47	224.77	35.04
10	17.94	0.270	13.68	245.49	0.12
11	34.91	0.526	3.61	126.17	1052.00
12	29.12	0.439	5.19	151.22	660.32
13	15.32	0.231	18.76	287.42	119.47
14	18.40	0.277	13.01	239.38	3.95
15	12.30	0.185	29.13	358.18	897.59
Sum			186.80	3333.81	6564.75
Wt Avg				17.85	

TABLE X2.5 CSS_0 and CSS_{1a}

i	$Y_i - X_i$	w_i	$w_i(Y_i - X_i)^2$	w_iX_i	w_iY_i	$w_i(Y_i - X_i - \bar{Y} + \bar{X})^2$
1	−1.69	6.67	19.1	163.8	152.5	2.16
2	−3.88	7.05	106.2	181.7	154.4	18.52
3	−2.36	6.35	35.3	163.7	148.7	0.06
4	−1.36	7.66	14.2	172.6	162.2	6.21
5	−2.42	4.90	28.7	144.6	132.7	0.12
6	−3.63	19.56	257.4	301.2	230.3	36.57
7	−3.27	11.37	121.7	225.9	188.7	11.63
8	−2.51	2.37	14.9	101.3	95.4	0.14
9	−2.58	8.66	57.6	192.0	169.7	0.88
10	−2.15	10.15	46.8	203.8	182.0	0.13
11	−2.65	3.08	21.7	115.7	107.5	0.47
12	−2.42	4.29	25.2	135.5	125.1	0.12
13	−1.15	13.45	17.8	221.6	206.1	16.54
14	−1.41	9.79	19.4	193.9	180.1	7.08
15	−1.17	19.45	26.5	261.9	239.2	23.20
Sum		134.80	CSS_0=812.46	2779.2	2474.5	CSS_{1a} = 123.86
Wt Avg				20.62	18.36	

TABLE X2.6 Iterating *Class 1b*

		First Iteration				Second Iteration			Final Step
i	w_i	$w_iX_iY_i$	$w_iX_i^2$	$w_i^2s_{Xi}^2(Y_i - bX_i)^2$	w_i	$w_iX_iY_i$	$w_iX_i^2$	$w_i^2s_{Xi}^2(Y_i - bX_i)^2$	$w_i\ (Y_i - bX_i)^2$
1	6.67	3746.7	4023.7	3.962	6.94	3900.2	4188.5	0.861	4.83
2	7.05	3981.3	4686.7	24.633	7.37	4164.3	4902.1	3.077	11.19
3	6.35	3834.2	4220.1	7.374	6.61	3992.2	4394.0	0.065	0.56
4	7.66	3654.0	3888.7	3.120	7.99	3813.2	4058.0	1.442	7.31
5	4.90	3917.4	4267.2	5.259	5.07	4058.4	4420.8	0.263	1.91
6	19.56	3545.2	4637.9	99.028	21.10	3823.8	5002.3	38.358	88.48
7	11.37	3751.0	4490.2	35.110	12.03	3967.8	4749.7	6.120	18.21
8	2.37	4073.2	4327.1	1.927	2.43	4175.5	4435.8	0.988	8.63
9	8.66	3761.9	4257.1	14.122	9.08	3945.1	4464.4	0.319	0.82
10	10.15	3656.4	4094.1	12.101	10.67	3844.6	4304.9	0.062	0.07
11	3.08	4038.5	4345.3	3.199	3.17	4154.8	4470.3	0.574	4.63
12	4.29	3945.4	4273.9	4.367	4.44	4079.1	4418.6	0.411	2.97
13	13.45	3395.2	3650.3	5.043	14.21	3587.4	3856.9	1.022	4.18
14	9.79	3567.4	3840.7	4.816	10.27	3743.0	4029.7	0.850	4.03
15	19.45	3220.2	3526.1	8.817	20.76	3436.4	3762.9	0.222	0.97
Sum		56088.3	62529.0	232.88		58685.8	65459.0	54.63	CSS_{1b} = 158.79

$$b_0 = \frac{\sum w_iX_iY_i}{\sum w_iX_i^2 - \sum w_i^2s_{Xi}^2(Y_i - bX_i)^2} = \frac{56088.3}{62529 - 232.88} = 0.9003 \quad (X2.3)$$

X2.4.3.2 As $|b - b_0| = 0.0997 > .001\ b$, we must iterate as shown.

X2.4.3.3 *From the Second Iteration*:

$$b_0 = \frac{\sum w_i X_i Y_i}{\sum w_i X_i^2 - \sum w_i^2 s_{Xi}^2 (Y_i - bX_i)^2} = \frac{58685.8}{65459.0 - 54.63} = 0.8973 \quad (X2.4)$$

X2.4.3.4 Again, $|b - b_0| = 0.0030 > .001\ b$, so a third iteration (not shown) is required. From the third iteration, $b_0 = 0.8972$, $|b - b_0| = 0.0001 < .001\ b$, and iteration may stop. The final step, computation of $CSS_{1b} = 158.79$, is shown in the last column of Table X2.6.

X2.4.4 *Class 2*—Linear correction.

X2.4.4.1 Tables X2.7 and X2.8 demonstrate two iterations of the algorithm for fitting the linear model. Starting with $b = 1$, the first iteration proceeds as in *Class 1*, shown in Tables X2.5-X2.7. Computing b_0:

$$b_0 = \frac{\sum w_i x_i y_i}{\sum w_i x_i^2 - \sum w_i^2 s_{Xi}^2 (y_i - bx_i)^2} = \frac{5069.01}{5228.26 - 38.08} = 0.97665 \quad (X2.5)$$

X2.4.4.2 As $|b - b_0| = 0.02335 > .001\ b$, we must iterate as shown in Table X2.8.

X2.4.4.3 From the Second Iteration:

$$b_0 = \frac{5121.63}{5282.30 - 38.44} = 0.97669 \quad (X2.6)$$

X2.4.4.4 Now $|b - b_0| = 0.00004 < .001\ b$, and iteration may stop. The final step, computation of $CSS_2 = 121.03$, is shown in the last column of Table X2.8. Using equation (Eq 18), $a = 18.34 - 0.9767 \times 20.60 = -1.78$.

X2.5 Test Whether the Methods are Sufficiently Correlated

X2.5.1 From Eq 19 compute:

$$F = \frac{(TSS_X + TSS_Y - CSS_2)/S}{CSS_2/(S-2)} \quad (X2.7)$$

$$= \frac{(26182.3 + 6564.7 - 121.03)/15}{121.03/13} = 233.6$$

X2.5.2 The 95th percentile of the F distribution, with 15 and 13 degrees of freedom, is 2.53. As the computed F is (very much) larger than 2.53, the methods are sufficiently correlated.

X2.6 Conduct Tests to Select the Most Parsimonious Bias Correction Class Needed.

X2.6.1 From Eq 20 compute:

$$F = \frac{(CSS_0 - CSS_2)/2}{CSS_2/(S-2)} = \frac{(812.46 - 121.03)/2}{121.03/13} = 37.13 \quad (X2.8)$$

X2.6.2 The 95th percentile of the F distribution, with 2 and 13 degrees of freedom, is 3.81. As the computed F is larger than 3.81, we conclude that a bias correction (of class yet to be determined) will significantly improve the expected agreement between the two methods.

X2.6.3 As CSS_{1a} is smaller than CSS_{1b}, the t-ratios of equation Eq 21 are:

$$t_1\sqrt{\frac{CSS_0 - CSS_{1a}}{CSS_2/(S-2)}} = \sqrt{\frac{812.46 - 123.86}{121.03/13}} = 8.60 \quad (X2.9)$$

and

$$t_2\sqrt{\frac{CSS_{1a} - CSS_2}{CSS_2/(S-2)}} = \sqrt{\frac{123.86 - 121.03}{121.03/13}} = 0.55. \quad (X2.10)$$

X2.6.4 The 97.5th percentile of Student's t distribution, with 13 degrees of freedom, is 2.16. As t_2 is smaller than 2.16, we compare t_1 to the same percentile, as discussed in 6.5.3.3. t_1 exceeds 2.16, so we conclude that a constant bias correction is preferred to a linear (proportional + constant) bias correction. The preferred bias correction is to subtract (since a has a negative sign) 2.26 volume % aromatics from any Test Method D 5580 result, in order to predict a Test Method D 5769 result on the same material. Note that the predicted Test Method D 5769 result should be within the scope of D 5769 in order for it to be meaningful.

X2.7 Test for Existence of Sample-Specific Biases

X2.7.1 The CSS of the selected bias correction is 123.86, with S-1 = 14 degrees of freedom. The 95th percentile value of the chi-square distribution is 23.68. As the CSS is larger, we conclude that there are likely sample-specific biases between the methods.

TABLE X2.7 First Iteration of *Class 2* Model Fitting

i	w_i	w_iX_i	w_iY_i	x_i	y_i	$w_ix_iy_i$	$w_ix_i^2$	$w_i^2s_{Xi}^2(y_i - bx_i)^2$
1	6.67	163.8	152.5	3.94	4.51	118.68	103.70	0.45
2	7.05	181.7	154.4	5.17	3.55	129.50	188.61	4.30
3	6.35	163.7	148.7	5.17	5.07	166.28	169.47	0.01
4	7.66	172.6	162.2	1.91	2.82	41.29	28.08	1.37
5	4.90	144.6	132.7	8.90	8.74	380.80	387.73	0.02
6	19.56	301.2	230.3	−5.22	−6.59	672.94	533.28	14.07
7	11.37	225.9	188.7	−0.74	−1.76	14.84	6.29	3.36
8	2.37	101.3	95.4	22.08	21.84	1144.46	1157.30	0.02
9	8.66	192.0	169.7	1.56	1.24	16.66	20.96	0.22
10	10.15	203.8	182.0	−0.53	−0.42	2.24	2.85	0.03
11	3.08	115.7	107.5	16.94	16.55	863.63	884.04	0.07
12	4.29	135.5	125.1	10.93	10.77	505.33	513.04	0.02
13	13.45	221.6	206.1	−4.14	−3.03	169.13	230.94	4.68
14	9.79	193.9	180.1	−0.81	0.04	−0.32	6.43	1.75
15	19.45	261.9	239.2	−7.15	−6.06	843.55	995.54	7.71
Sum	134.80	2779.21	2474.55			5069.01	5228.26	38.08
Avg		20.62	18.36					

TABLE X2.8 Second Iteration of *Class 2* Model Fitting

i	w_i	w_iX_i	w_iY_i	x_i	y_i	w_ix_iyi	$w_ix_i^2$	$w_i^2s_{XI}^2(y_i - bx_i)^2$	$w_i\,(y_i - bx_i)^2$
1	6.73	165.41	154.03	3.96	4.53	120.82	105.61	0.62	2.96
2	7.12	183.68	156.04	5.19	3.57	131.99	191.91	3.76	16.02
3	6.41	165.26	150.15	5.18	5.09	169.03	172.25	0.00	0.00
4	7.74	174.35	163.83	3.120	1.93	42.35	28.88	1.54	6.93
5	4.94	145.82	133.86	8.91	8.76	385.56	392.54	0.00	1.01
6	19.92	306.67	234.42	−5.20	−6.57	681.05	539.39	17.28	44.10
7	11.52	228.98	191.29	−0.73	−1.74	14.55	6.08	3.56	12.18
8	2.39	101.94	95.96	22.10	21.86	1153.15	1166.05	0.02	0.17
9	8.76	194.19	171.60	1.57	1.25	17.28	21.67	0.17	0.70
10	10.27	206.27	184.22	−0.51	−0.40	2.11	2.70	0.03	0.10
11	3.10	116.49	108.27	16.96	16.57	871.36	891.90	0.00	0.00
12	4.33	136.56	126.07	10.95	10.78	511.02	518.79	0.01	0.04
13	13.63	224.51	208.82	−4.13	−3.02	169.67	232.07	4.01	14.00
14	9.90	196.15	182.19	−0.79	0.06	−0.46	6.23	1.72	6.87
15	19.76	265.98	242.90	−7.14	−6.04	852.16	1006.23	5.72	16.95
Sum	136.51	2812.26	2503.64			5121.63	5282.30	38.44	CSS_2 = 121.03
Avg		20.60	18.34						

X2.8 Examine Residuals to Assess Reasonableness of Random Effect Assumption

X2.8.1 The (standardized) residuals, $\epsilon_i = \sqrt{w_i}(\,Y_i - \hat{Y}_i)$, are shown in Table X2.9. For example, the residual for the first sample (first in Tables X2.1-X2.8) is $\sqrt{6.67}$ (22.87 − (24.56 − 2.26)) = 1.47, which is found in the eleventh row. (The table has been sorted in order of increasing ϵ_i.) $\{w_i\}$ are taken from Table X2.5, which is appropriate for the selected bias correction.

X2.8.2 *Anderson-Darling Statistic*:

X2.8.2.1 From Eq X1.4 of Practice D 6299, the residuals, $\{\epsilon_i\}$, are again normalized. To avoid a conflict in notation, what are called w_i in that practice are called $v_i = (\epsilon_i - \bar{\epsilon})/s_\epsilon$ here and in Table X2.9, where $\bar{\epsilon}$ = -.06 is the mean of the $\{\epsilon_i\}$, and s_ϵ=2.97 is the standard deviation. The $\{p_i\}$ are from tables of the standard normal distribution. From Eq. A1.6 and A1.7 of Practice D 6299,

$$A^2 = \frac{\sum(2i-1)[1n(p_i) + 1n(1-p_{n+1-i})]}{n} - n = 0.361 \quad \text{(X2.11)}$$

$$A^{2*} = A^2\left(1 + \frac{0.75}{n} + \frac{2.25}{n^2}\right) = 0.382 \quad \text{(X2.12)}$$

X2.8.2.2 As A^{2*}(0.382) is less than the .05 level critical value (0.752) for the Anderson Darling statistic, the distribution of the residuals cannot be distinguished from the normal distribution.

X2.8.3 *Cross-Method Reproducibility*:

X2.8.3.1 Estimate the cross-method reproducibility (R_{XY}) as follows:

$$R_{XY} = \sqrt{\frac{R_X^2}{2}\left[1 + \frac{1}{L_X}\left(\frac{CSS}{S-k} - 1\right)\right] + \frac{R_Y^2}{2}\left[1 + \frac{1}{L_Y}\left(\frac{CSS}{S-k} - 1\right)\right]} \quad \text{(X2.13)}$$

$$= \sqrt{\frac{0.2792^2\,X}{2}\left[1 + \frac{1}{7}\left(\frac{123.86}{14} - 1\right)\right] + \frac{0.1292^2\,Y^2}{2}\left[1 + \frac{1}{7}\left(\frac{123.86}{14} - 1\right)\right]}$$

$$= \sqrt{0.0865X + .01851Y^2}$$

X2.8.3.2 Because of the sample-specific biases (which could be due to the need for further standardization in one of the methods as noted earlier), this is almost 50 % larger than the root mean squares of the individual reproducibilities.

TABLE X2.9 Residuals

Rank	Original Sequence No.	Sorted Residual	v_i	p_i	i^{th} Term in Eq X2.1
1	6	−6.05	−2.01	0.022	−0.45
2	2	−4.30	−1.43	0.077	−1.01
3	7	−3.41	−1.13	0.130	−1.25
4	9	−0.94	−0.30	0.383	−1.21
5	11	−0.69	−0.21	0.416	−1.24
6	8	−0.38	−0.11	0.457	−1.17
7	5	−0.35	−0.10	0.460	−1.23
8	12	−0.34	−0.10	0.462	−1.39
9	3	−0.25	−0.06	0.475	−1054
10	10	0.36	0.14	0.555	−1.52
11	1	1.47	0.51	0.696	−1.26
12	4	2.49	0.86	0.804	−1.08
13	14	2.66	0.91	0.820	−0.56
14	13	4.07	1.39	0.917	−0.30
15	15	4.82	1.64	0.949	−0.14