INTERNATIONAL STANDARD

ISO 4259-5

First edition
2023-12

Petroleum and related products — Precision of measurement methods and results —

Part 5: Statistical assessment of agreement between two different measurement methods that claim to measure the same property

Produits pétroliers et connexes — Fidélité des méthodes de mesure et de leurs résultats —

Partie 5: Évaluation statistique de l'accord entre deux méthodes de mesure différentes qui prétendent mesurer la même propriété

Reference number
ISO 4259-5:2023(E)

COPYRIGHT PROTECTED DOCUMENT

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva
Phone: +41 22 749 01 11
Email: copyright@iso.org
Website: www.iso.org

Published in Switzerland

Contents

Page

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular, the different approval criteria needed for the different types of ISO document should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives).

ISO draws attention to the possibility that the implementation of this document may involve the use of (a) patent(s). ISO takes no position concerning the evidence, validity or applicability of any claimed patent rights in respect thereof. As of the date of publication of this document, ISO had not received notice of (a) patent(s) which may be required to implement this document. However, implementers are cautioned that this may not represent the latest information, which may be obtained from the patent database available at www.iso.org/patents. ISO shall not be held responsible for identifying any or all such patent rights.

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation of the voluntary nature of standards, the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT), see www.iso.org/iso/foreword.html.

This document was prepared by Technical Committee ISO/TC 28, *Petroleum and related products, fuels and lubricants from natural or synthetic sources,* in collaboration with the European Committee for Standardization (CEN) Technical Committee CEN/TC 19, *Gaseous and liquid fuels, lubricants and related products of petroleum, synthetic and biological origin,* in accordance with the Agreement on technical cooperation between ISO and CEN (Vienna Agreement).

A list of all parts in the ISO 4259 series can be found on the ISO website.

Any feedback or questions on this document should be directed to the user's national standards body. A complete listing of these bodies can be found at www.iso.org/members.html.

Introduction

This document explains the statistical methodology for assessing the expected agreement between two standardized test methods that purport to measure the same property of a material. Subsequently, it is investigated whether a linear bias correction can significantly improve the expected agreement. The degree of agreement is expressed as a between-methods reproducibility after a bias correction (if necessary) has been applied.

The method uses numerical results from a set of samples that have been analysed independently using both test methods by different laboratories. The variation associated with each test method result is used for assessing the required bias correction.

Annexes A and B give worked out examples showing how the methodology is applied.

Petroleum and related products — Precision of measurement methods and results —

Part 5: Statistical assessment of agreement between two different measurement methods that claim to measure the same property

1 Scope

This document specifies statistical methodology for assessing the expected agreement between two test methods that purport to measure the same property of a material, and for deciding if a simple linear bias correction can further improve the expected agreement.

This document is applicable for analytical methods which measure quantitative properties of petroleum or petroleum products resulting from a multi-sample-multi-lab study (MSMLS). These types of studies include but are not limited to interlaboratory studies (ILS) meeting the requirements of ISO 4259-1 or equivalent, and proficiency testing programmes (PTP) meeting the requirements of ISO 4259-3 or equivalent.

The methodology specified in this document establishes the limiting value for the difference between two results where each result is obtained by a different operator using different apparatus and two methods X and Y, respectively, on identical material. One of the methods (X or Y) has been appropriately bias-corrected to agree with the other in accordance with this practice. This limit is designated as the between-methods reproducibility. This value is expected to be exceeded with a probability of 5 % under the correct and normal operation of both test methods due to random variation.

NOTE Further conditions for application of this methodology are given in 5.1 and 5.2.

2 Normative references

The following documents are referred to in the text in such a way that some or all of their content constitutes requirements of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO 4259-1, *Petroleum and related products — Precision of measurement methods and results — Part 1: Determination of precision data in relation to methods of test*

ISO 4259-3, *Petroleum and related products — Precision of measurement methods and results — Part 3: Monitoring and verification of published precision data in relation to methods of test*

ISO 4259-4, *Petroleum and related products — Precision of measurement methods and results — Part 4: Use of statistical control charts to validate 'in-statistical-control' status for the execution of a standard test method in a single laboratory*

3 Terms and definitions

For the purposes of this document, the terms and definitions in ISO 4259-1 and the following terms and definitions apply.

ISO and IEC maintain terminology databases for use in standardization at the following addresses:

— ISO Online browsing platform: available at https://www.iso.org/obp

— IEC Electropedia: available at https://www.electropedia.org/

3.1
multi-sample-multi-lab study
MSMLS
study in which one or more performance characteristics are determined on the basis of analytical results from multiple samples and multiple laboratories

Note 1 to entry: Under certain conditions, inter laboratory studies and proficiency testing schemes meet this definition of multi-sample-multi-lab study.

3.2
interlaboratory study
ILS
study specifically designed to estimate the repeatability and reproducibility of a standard test method achieved at a fixed point in time by multiple laboratories through the statistical analysis of their test results obtained on aliquots prepared from multiple materials

3.3
proficiency testing programme
PTP
programme designed for the periodic evaluation testing capability of participating laboratories of a standard test method through the statistical analysis of their test results obtained on aliquots prepared from a single batch of homogeneous material

Note 1 to entry: PTP is sometimes referred to as a proficiency testing (PT)-study or an interlaboratory cross check programme (ILCP).

3.4
between-methods bias correction
quantitative expression of the mathematical correction, when applied to the outcome of either one of two methods claiming to measure the same property, can result in a statistically significant improvement between the expected values of the two test methods claiming to measure the same property

3.5
correlation coefficient
ρ
statistical measure of the strength and direction of the relationship between two variables

Note 1 to entry: Values always range between –1 (strong negative relationship) and +1 (strong positive relationship). Values at or close to zero imply a weak or nonlinear relationship.

3.6
standard error
Δ_E
statistic estimating the standard deviation of the distribution of the average statistic obtained from the repeat random sampling of a population

3.7
sample standard deviation
s_i
estimator of the population standard deviation using the sample mean and sample size

Note 1 to entry: Sample standard deviation is also referred to as standard deviation of the sample.

3.8
between-methods reproducibility
R_{XY}
quantitative expression for the computation of the limiting value that the difference between two single results is expected to exceed with a probability of 5 % due to random variation, under the correct and normal operation of both test methods, where each result is obtained by different operators on an identical test sample using different apparatus and applying the two methods X and Y, respectively; when the methods have been assessed and an appropriate between-methods bias correction has been applied to the result from either method (X or Y) in accordance with this practice

3.9
sum of squared residuals
Σ_{SR}
statistic used to quantify the degree of agreement between the results from two test methods after *between-methods bias-correction* (3.4) using the methodology of this practice

Note 1 to entry: Σ_{SR} is used as an optimality criterion in parameter selection and bias-correction model selection.

3.10
total sum of squares
Σ_{ST}
statistic used to quantify the information content from the *interlaboratory study* (3.2) in terms of total variation of sample means relative to the *standard error* (3.6) of each sample mean

3.11
resolution
smallest difference in two results that is represented by a different value

4 Symbols

Symbol	Explanation
X, Y	reference to the X- and Y-methods, respectively
X_{ijk}, Y_{ijk}	Single k^{th} result on the i^{th} common material by the j^{th} lab using X-method and Y-method, respectively
X_i, Y_i	arithmetic mean of the i^{th} sample using X-method and Y-method, respectively
$\bar{X}, \bar{Y}$	weighted average across the samples used in the calculation of total sum of squares $\Sigma_{ST,\bar{X}}$ and $\Sigma_{ST,\bar{Y}}$ for the X-method and Y-method, respectively
$\ddot{X}, \ddot{Y}$	weighted average across the samples used in the calculation of the correlation coefficient ρ for the X-method and Y-method
Δ_{xi}, Δ_{yi}	absolute deviation of the weighted means of the i^{th} sample results from $\bar{X}$ and $\bar{Y}$, respectively
$\hat{Y}$	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} sample Y-method mean, by applying the bias correction established from this practice to its corresponding X-method mean
S	number of samples in the multi-lab-multi-sample data set
L_{Xi}, L_{Yi}	number of laboratories that returned results on the i^{th} sample using the X-method and Y-method, respectively
n_{Xij}, n_{Yij}	number of repeated results on the i^{th} sample of j^{th} lab using the X- and Y-methods, respectively
R_X, R_Y	reproducibility of the X- and Y-methods, respectively
R_{Xi}, R_{Yi}	reproducibility of the X- and Y-methods, evaluated at the method X and Y means of the i^{th} sample
R_{XY}	between-methods reproducibility
$s_{R,Xi}$, $s_{R,Yi}$	reproducibility standard deviation, evaluated at the i^{th} sample using method X and Y, respectively
$s_{r,Xi}$, $s_{r,Yi}$	repeatability standard deviation, evaluated at the i^{th} sample using method X and Y, respectively

Symbol	Explanation
ε_i	weighted residual of Y-method mean values predicted from the corresponding X-method mean values, $\hat{Y}_i$ and mean of Y-method results, Y_i on the i^{th} sample
$\Delta_{E,Xi}$, $\Delta_{E,Yi}$	standard error of the means of the i^{th} sample
$\Sigma_{SR,p}$	weighted sum of squared residuals of the mean results of Y-method and the bias-corrected mean results of the X-method for a given model p where p = 0, 1a, 1b or 2 over all samples i
$\Sigma_{ST,\bar{X}}$, $\Sigma_{ST,Y}$	total sum of squares, around the weighted averages $\bar{X}$ and $\bar{Y}$ over all samples i
F	test statistic for comparing variances, defined by the quotient of two variances
t	student t-value at a specified confidence level and specified degrees of freedom
k	class number of selected bias correction class
ν_X, ν_Y	degrees of freedom for reproducibility variances
w_i	weight associated with the difference between (corrected) mean results from the i^{th} sample
a, b	parameter of the bias correction: $\hat{Y} = a + bX$
h_i	leverage of sample i in the set of samples
Z_i	natural logarithm of the sample mean, averaged over both methods for sample i
Z	overall average of natural logarithm Z_i of all samples
t_1, t_2	ratio for assessing reductions in sums of squares
ε_i	standardized difference between Y_i and $\hat{Y}_i$, sometimes referred to as error
A, B, C	parameters of the quadratic function used for the iterative calculation of the proportional coefficient b for class 1b and class 2 correction class
D	difference statistic for confirmation of the correlation
A_i^2, A_i^{2*}	Anderson-Darling test statistic and modified test statistic, respectively
ρ	correlation coefficient

5 Procedure overview

5.1 General requirements

The procedures are intended to be executed by an analyst with sufficient working knowledge of the statistical tools and theories described in the document.

The statistical methodology is based on the premise that a bias correction is not required. In the absence of statistical evidence that a bias correction would improve the expected 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 favoured over a more complex one if the latter does not yield a statistically observable improvement over the former. Failure to adhere to this generally results in a model that is over-fitted and does not perform well in practice.

NOTE 1 The parsimony principle is that the most acceptable explanation of an occurrence, phenomenon, or event is the simplest, involving the fewest entities, assumptions.

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

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.

The methodology described in this document is applicable only if the standard error associated with each mean test result is known or can be calculated and the degrees of freedom associated with all standard errors are at least 30.

This methodology is applied to a data source derived from a MSMLS. The study shall be conducted on at least 10 independent materials that span the intersecting scopes of the test methods. The results shall be obtained from at least six (6) laboratories using each method.

The results are obtained on the same comparison set of samples and it is recommended that both test methods are not performed by the same laboratory. If this is the case, care shall be taken to ensure independence of test results, for example by double-blind testing of samples in random order.

This methodology shall not be used on the basis of interim or temporary published precision statements. Interim or temporary statements of accuracy generally lack the magnitude of the amount of data applied and, as a result, insufficient degrees of freedom are available.

Combining multiple data sources is permissible provided the quality requirements for the data set as specified in this document are met.

The test methods used by each laboratory shall be under statistical control, meeting the requirements in ISO 4259-4.

This methodology requires data with sufficient resolution to permit variation to be observable in a statistically meaningful manner. Statistically meaningful variation implies that the total number of unique values in a set of data, i.e. the lab results of each sample for each test method, should be sufficiently large. If, in the opinion of the analyst, the number of individual values in the data set is insufficient, the data shall be requested again from the relevant laboratories with sufficient resolution. If the data are only available with insufficient resolution, this evaluation should not be continued.

In case the data for the procedure originates from an ILS, all requirements of ISO 4259-1 shall be met and the additional requirements regarding proficiency testing programme (PTP) data do not apply.

NOTE 2 Leverage is a measure of how far away the independent variables of an observation are from those of the other observations.

NOTE 3 Cook's distance is an estimate of the influence of a data point. It is used within the context of the reference to indicate influential data points that are particularly worth checking for validity.

5.2 Additional requirements for PTP data

5.2.1 General conditions

The statistical calculations are also applicable for this evaluation, provided the results and associated statistics for the test method are obtained from a PTP, which shall meet the requirements of ISO 4259-3. A characteristic of data derived from such a PTP is that for each sample, a single result is provided by each laboratory for the test method.

The following requirements apply when using PTP data:

— the results shall be obtained from at least 10 laboratories using the test method and are equidistantly distributed over the range;

— the leverage of each sample in the data set shall not exceed the limiting value of 0,5 (see 5.2.2);

— the Anderson-Darling statistics for the tests on normal distribution of lab results per sample ≤1,12 shall be used (see 5.2.3);

— the sample standard deviations shall not significantly exceed the published reproducibility standard deviations for at least 80 % of the samples at the 0,05 significance level (see 5.2.4).

5.2.2 Test on existence of extreme samples

The leverage value h_i for each sample i in the data set is examined and may not exceed the limiting value of 0,5. If a value for h_i of a sample exceeds this limiting value, this sample is characterized as

extreme. For each of the two methods, the average of the laboratory results is calculated per sample. Subsequently, each laboratory average per sample is averaged over both test methods.

The leverage value h_i is defined by Formula (1):

$$h_i = \frac{1}{S} + \frac{(Z_i - \bar{Z})^2}{\sum_{k=1}^{S}(Z_k - \bar{Z})^2} \quad (1)$$

where

h_i is the leverage of sample i, $i = 1 \ldots S$,

S is the total number of samples,

Z_i is the natural logarithm (ln) of the sample mean, averaged over both methods,

$\bar{Z}$ is the overall average of all Z_i.

If one or more samples are characterized as extreme, they shall be removed and the procedure should be repeated. The minimum number of remaining samples shall be taken into account. If the minimum requirement for a number of samples can no longer be met, the procedure shall be discontinued.

5.2.3 Test on distribution of lab results

The distribution of the lab results for each sample are tested for normality by confirming the goodness-of-fit of the normal distribution using the Anderson-Darling statistic per sample.

NOTE 1 The Anderson-Darling test is a statistical test of whether a given sample of data are drawn from a given probability distribution. Within the context of this document, this test is used as a test on normality, with probability distribution parameters (mean and standard deviation) estimated from the sample. See Reference [7] for further details.

NOTE 2 The critical value of 1,12 is based on a significance level of approximately 1 %, taking into account the effects of rounding of the input data on the resolution.

The test statistic A_i^{2*} is calculated according to Formula (2):

$$A_i^{2*} = A_i^2\left(1 + \frac{0{,}75}{N_i} + \frac{2{,}25}{N_i^2}\right) \quad (2)$$

where

N_i is the total number of lab results in the set,

A_i^2 $-N - \frac{1}{N}\sum_{i=1}^{N}(2i-1)\{\ln[F(x_i)] + \ln[1 - F(x_{N-i+1})]\}$,

$F(x_i)$ is the cumulative normal distribution function based on sample average and standard deviation,

x_i is the data sorted in increasing order, $x_1 \leq x_2 \leq x_3 \ldots \leq x_N$.

The distribution of the results is assumed to follow a normal distribution if the corresponding A_i^{2*} value ≤1,12.

If this test shows that the distribution of one or more samples does not meet the above criterion, this sample shall be removed. The minimum number of samples for this procedure should be considered. If the minimum requirement for a number of samples can no longer be met, the procedure shall be discontinued.

Data with insufficient resolution due to rounding can overestimate the normality assessment statistics. See 5.1 for resolution provisions.

5.2.4 Comparison of precision

The sample standard deviations s_i should not significantly exceed the published reproducibility standard deviations s_{Ri} for at least 80 % of the samples at a significance level of 0,05 using a statistical F-test for the comparison of two variances s_i and s_{Ri}.

For any sample i where s_i is numerically larger than s_{Ri}, perform the following F-test specified in Formula (3):

$$F = \frac{s_i^2}{s_{Ri}^2} \tag{3}$$

where

s_i is the standard deviation of the sample i, calculated over the lab results,

s_{Ri} is the published reproducibility standard deviation evaluated at concentration level of the average results for sample i.

The number of degrees of freedom associated with s_i equals N-1, where N equals the number of result for sample i.

The number of degrees of freedom associated with s_{Ri} is preferably taken from the published precision statement of the test method or underlying research report. If s_{Ri} is not given as such, it is permitted to estimate s_{Ri} based on the published reproducibility R_i, according to $s_{Ri} = R_i/(t\sqrt{2})$, where t represents the student-t value at a confidence level of 0,05 and degrees of freedom associated with R_i.

If in this latter case the degrees of freedom for R_i is unknown, it may be estimated by the minimum value of 30, and the published reproducibility standard deviation is estimated by $s_{Ri} = (R_i/2{,}888)$.

If the above criterion is not met for one or more samples, the failing samples shall be removed. The minimum number of samples for this procedure should be considered. If the minimum requirement for a number of samples can no longer be met, the procedure shall be discontinued.

5.3 Brief sequential steps of the procedure

The following compressed overview summarizes the steps of the procedure. See Figures 1 and 2 for a flow diagram of these procedural steps.

1) Checking the adequacy of the available data

   The available data are checked against the general requirements (see 5.1). If applicable, the additional requirements when using PTP data (see 5.2, 5.2.1, 5.2.2, 5.2.3 and 5.2.4) are also checked.

2) Calculate the means and standard error of the samples

   The arithmetic means of the results for each common sample obtained by each method are calculated (see 6.1.2) and the estimates of the standard errors of these means are computed (see 6.1.3).

3) Test the suitability of the data

   Test for sufficient variation in the properties of both methods by computing the weighted sums of squared residuals 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 (see 6.2.1).

Test for sufficient correlation between both methods by assessing the weighted sums of squared residuals for the linear correction against the total variation in the mean results for both methods to ensure that there is sufficient correlation between the two methods (see 6.2.2).

4) Calculate the bias correction statistics for each bias correction class

The closeness of agreement of the mean results by each method is evaluated using appropriate weighted sums of squared residuals. 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 (see 6.3).

5) Select the appropriate bias correction class

The most parsimonious bias correction is selected based on the weighted sum of squared residuals from each bias correction and the appropriate *t*- and *F*-tests (see 6.4).

6) Test on distribution of residuals for normality

The (weighted) residuals per sample are tested for normality. The residuals are defined by the difference between each individual Y_i and bias-corrected X_i. The test for normality is performed using the Anderson-Darling test for normality. When the weighted residuals are not found to be normally distributed this practice is considered terminated (see 6.5).

7) Test for sample-specific biases

The weighted sum of squared residuals are assessed to determine whether additional unexplained sources of variation remain in the residual data (see 6.6).

Any remaining, unexplained variation is attributed to sample-specific biases, also known as method-material interactions or matrix effects. If sample-specific biases 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.

8) Compute the between-methods reproducibility

Calculate the between-methods reproducibility taking into account possible sample specific biases.

When residuals are found to be normally distributed and sample-specific biases are not found to be present, the between-methods reproducibility is defined by Formula (40).

When residuals are found to be normally distributed and sample-specific biases are present, the between-methods reproducibility is defined by Formula (41).

9) Reporting

The results of this practice are reported in the precision and bias section of the appropriate standard(s) (see Clause 7).

10) Confirmation of the correlation

The results of the assessment are periodically confirmed by users of the correlation by monitoring the difference statistics by means of control charts (see Clause 8).

5.4 Flow diagram of the procedure

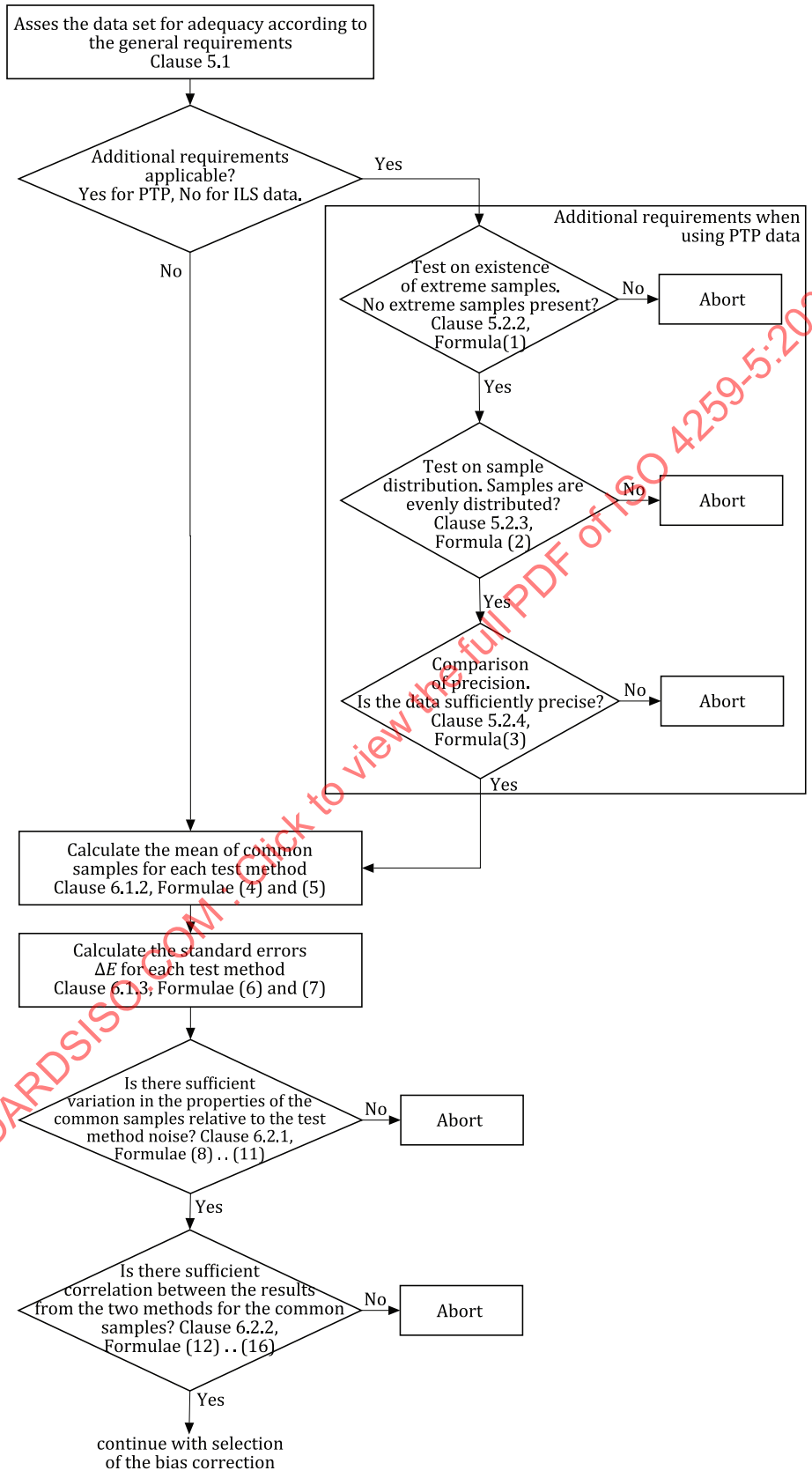

Figure 1 — Flowchart for suitability and applicability of the data

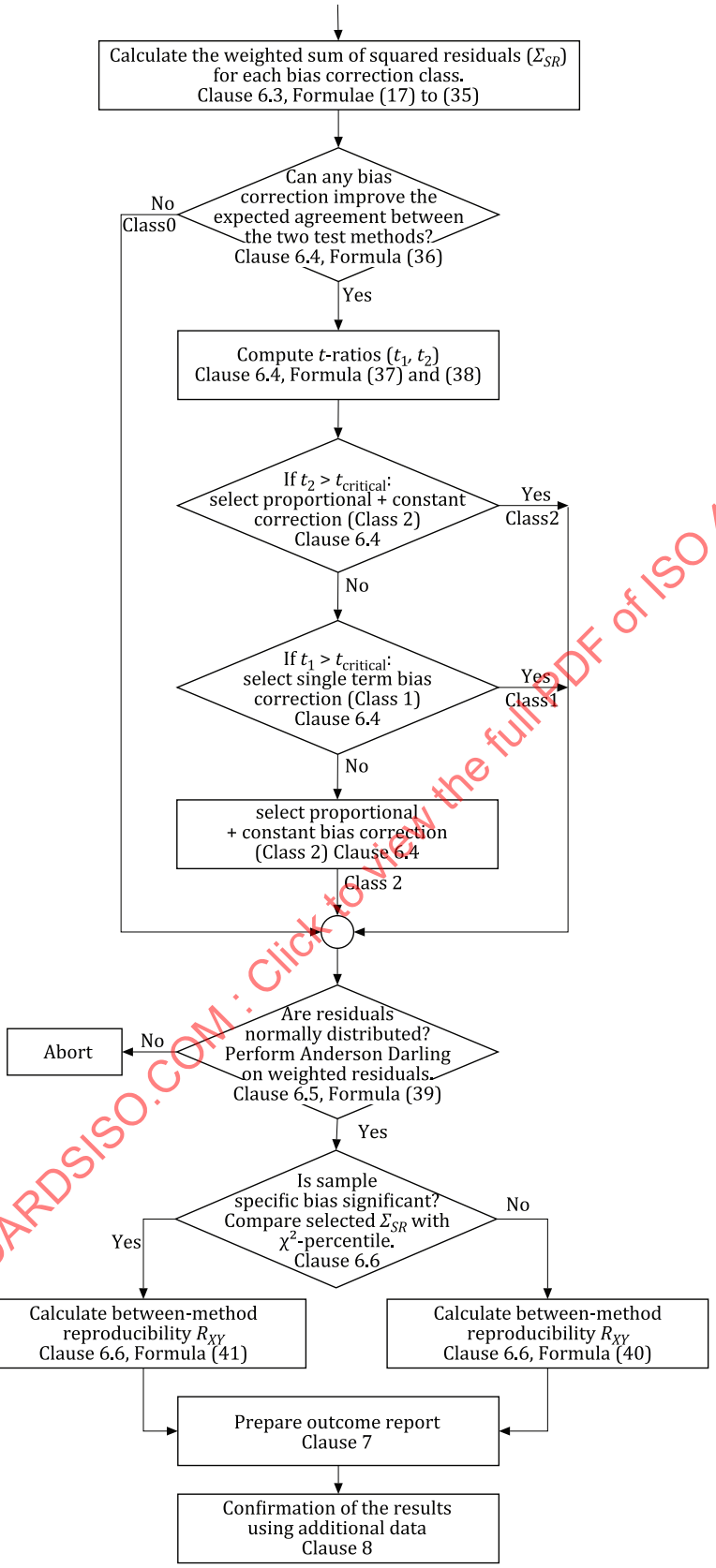

Figure 2 — Procedure for determining the bias correction

6 Procedure

6.1 Sample mean and standard error

6.1.1 General

Calculate sample means X_i and Y_i and standard errors from results from the MSMLS. Published precision estimates are used to estimate the standard errors of these means, $\Delta_{S,Xi}$ and $\Delta_{S,Yi}$.

NOTE The i^{th} material is the same for both data sets, but the j^{th} lab in one data set is not generally the same lab as the j^{th} lab in the other data set.

6.1.2 Computation of the means

The arithmetic mean X-method result for the i^{th} sample is shown in Formula (4):

$$X_i = \frac{1}{L_{Xi}} \sum_j \frac{\sum_k X_{ijk}}{n_{X_{ij}}} \tag{4}$$

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

Similarly, the mean Y-method result for the i^{th} sample is given by the analogous Formulae (5):

$$Y_i = \frac{1}{L_{Y_i}} \sum_j \frac{\sum_k Y_{ijk}}{n_{Y_{ij}}} \tag{5}$$

6.1.3 Calculation of standard errors

The standard errors are assigned to the standard deviations of the means and are calculated as follows.

If $s_{R,Xi}$ is the reproducibility standard deviation from the X-method, and $s_{r,Xi}$ is the repeatability standard deviation, then an estimate of the standard error for X_i is given by Formula (6):

$$\Delta_{E,X_i} = \sqrt{\frac{1}{L_{Xi}}\left[s^2_{R,Xi} - s^2_{r,Xi}\left(1 - \frac{1}{L_{Xi}} \sum_j \frac{1}{n_{Xij}}\right)\right]} \tag{6}$$

The estimated standard error for Y_i, is given by the analogous Formula (7):

$$\Delta_{E,Y_i} = \sqrt{\frac{1}{L_{Yi}}\left[s^2_{R,Yi} - s^2_{r,Yi}\left(1 - \frac{1}{L_{Yi}} \sum_j \frac{1}{n_{Yij}}\right)\right]} \tag{7}$$

The repeatability standard deviations and reproducibility standard deviations are calculated from published repeatability and published reproducibility by dividing these by $t\sqrt{2}$. Here, t refers to the student t-value at a confidence level of 0,05 and the number of degrees of freedom as associated with the precision figures.

In case the repeatability and reproducibility are known, but the number of degrees of freedom associated with these precision figures are unknown, a value of 30 for number of degrees of freedom is permitted.

Since repeatability and reproducibility may vary with the mean X-method results X_i, 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 $\Delta_{E,Xi}$ can still differ from one material to the next. The same is also true for method Y.

6.2 Suitability of the data

6.2.1 Test on property variation

Calculate the weighted total sum of squares for each method, and determine whether the samples can be distinguished from each other by both methods. The total sums of squares are given by Formulae (8) and (9):

$$\Sigma_{ST,\bar{X}} = \sum_i \left(\frac{X_i - \bar{X}}{\Delta_{E,Xi}} \right)^2 \tag{8}$$

$$\Sigma_{ST,\bar{Y}} = \sum_i \left(\frac{Y_i - \bar{Y}}{\Delta_{E,Yi}} \right)^2 \tag{9}$$

The weighted averages $\bar{X}$ and $\bar{Y}$ take into account their corresponding standard errors $\Delta_{S,Xi}$ and $\Delta_{S,Yi}$ and are defined by Formulae (10) and (11):

$$\bar{X} = \frac{\sum_i \left(\frac{X_i}{\Delta^2_{E,Xi}} \right)}{\sum_i \left(\frac{1}{\Delta^2_{E,Xi}} \right)} \tag{10}$$

$$\bar{Y} = \frac{\sum_i \left(\frac{Y_i}{\Delta^2_{E,Yi}} \right)}{\sum_i \left(\frac{1}{\Delta^2_{E,Yi}} \right)} \tag{11}$$

Compare $F = \Sigma_{ST,\bar{X}}/(S\text{-}1)$ to the 95th percentile of Fisher 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 for the X-method.

— 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.

— If F does exceed the 95th percentile, then the X-method is sufficiently precise to distinguish among the S samples. Proceed with the test on correlation in 6.2.2.

In a similar manner, compare $F = \Sigma_{ST,\bar{Y}}/(S\text{-}1)$ 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.

6.2.2 Correlation of the test methods

To test whether both methods are sufficiently correlated the correlation coefficient, ρ, is calculated by Formula (12):

$$\rho = \frac{\sum_i \left[w_i \left(X_i - \ddot{X} \right) \left(Y_i - \ddot{Y} \right) \right]}{\sqrt{\sum_i \left[w_i \left(X_i - \ddot{X} \right)^2 \right] \sum_i \left[w_i \left(Y_i - \ddot{Y} \right)^2 \right]}} \tag{12}$$

The weighted averages $\ddot{X}$ and $\ddot{Y}$ are calculated by Formulae (13) and (14):

$$\ddot{X} = \frac{\sum w_i X_i}{\sum w_i} \tag{13}$$

$$\ddot{Y} = \frac{\sum w_i Y_i}{\sum w_i} \tag{14}$$

Where the weights w_i are calculated by Formula (15):

$$w_i = \frac{1}{\Delta_{E,Yi}^2 + \Delta_{E,Xi}^2} \tag{15}$$

Use the correlation coefficient ρ to calculate the F-statistic according to Formula (16):

$$F = \frac{(S-2)\rho^2}{\left(1-\rho^2\right)} \tag{16}$$

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.

— If F is less than the 99th percentile value, then, this practice concludes that the test methods are too discordant to permit use of the results from one method to predict those of the other. Do not proceed with this practice.

— If F is greater than the 99th percentile value, then it can be assumed that the two test methods are sufficiently correlated to continue the procedure with the bias correction statistics in 6.3.

At this point in the procedure, it can be enlightening to graph the data with the mean X-method data (X_i) versus the mean Y-method data (Y_i) of the samples.

6.3 Bias correction selection statistics

6.3.1 General

Calculate the weighted sum of squared residuals for each of the following classes of bias-correction methodology, as specified in 6.3.2 to 6.3.5.

6.3.2 Class 0—No bias correction

Compute the weighted sum of squared residuals of the mean results for the Class 0 bias correction, $\Sigma_{SR,0}$, according to Formula (17):

$$\Sigma_{SR,0} = \sum_i w_i (X_i - Y_i)^2 \tag{17}$$

Where the weights w_i for each sample i is calculated by Formula (18):

$$w_i = \frac{1}{\Delta_{E,Yi}^2 + \Delta_{E,Xi}^2} \tag{18}$$

6.3.3 Class 1a—Constant bias correction

Using the weight w_i from Class 0 correction from Formula (18), compute the constant bias correction a according to Formula (19):

$$a = \frac{\sum_i w_i (Y_i - X_i)}{\sum_i w_i} \tag{19}$$

Compute the weighted sum of squared residuals of the means results for the Class 1a bias correction, $\Sigma_{SR,1a}$, according to Formula (20):

$$\Sigma_{SR,1a} = \sum_i w_i [Y_i - (X_i + a)]^2 \tag{20}$$

6.3.4 Class 1b — Proportional bias correction

The computations of the proportional bias correction are appropriate only if both of the following conditions apply:

a) the measured property assumes only positive values, and

b) a property value of zero has a physical significance (e.g. concentrations of specific constituents).

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

Set $b = 1$.

Compute the weight w_i for each sample i, as shown in Formula (21):

$$w_i = \frac{1}{\Delta_{E,Xi}^2 + b^2 \Delta_{E,Yi}^2} \tag{21}$$

Calculate the following three sums according to Formulae (22) to (24):

$$A = \sum_i w_i^2 X_i Y_i \Delta_{E,Xi}^2 \tag{22}$$

$$B = \sum_i w_i^2 \left(X_i^2 \Delta_{E,Xi}^2 - Y_i^2 \Delta_{E,Yi}^2 \right) \tag{23}$$

$$C = -\sum_i w_i^2 X_i Y_i \Delta_{E,Yi}^2 \tag{24}$$

Calculate the interim proportional correction b_0 according to Formula (25):

$$b_0 = \frac{-B + \sqrt{B^2 - 4AC}}{2A} \tag{25}$$

If $|b - b_0| > 0{,}001 \cdot b$, replace b with b_0 and go back to Formula (21). Otherwise, the iteration can be stopped, as further iteration will not produce meaningful improvement. Replace b with b_0 and abort the iteration.

Calculate the final weights w_i as in Formula (21).

Calculate the weighted sum of squared residuals of the mean results for the Class 1b bias correction, $\Sigma_{SR,1b}$, according to Formula (26):

$$\Sigma_{SR,1b} = \sum_i w_i (Y_i - bX_i)^2 \tag{26}$$

6.3.5 Class 2 — Proportional and constant bias correction

This involves iterative calculation of the weights w_i, the weighted means $\bar{X}$ and $\bar{Y}$, and the proportional term b.

Set $b = 1$.

Compute the weight w_i for each sample i as shown in Formula (27):

$$w_i = \frac{1}{\Delta_{E,Xi}^2 + b^2 \Delta_{E,Yi}^2} \tag{27}$$

Calculate the weighted averages according to Formula (28):

$$\bar{X} = \frac{\sum_i w_i X_i}{\sum_i w_i}, \quad \bar{Y} = \frac{\sum_i w_i Y_i}{\sum_i w_i} \tag{28}$$

Calculate the deviations from the weighted means as shown in Formula (29):

$$\Delta_{xi} = X_i - \bar{X}, \quad \Delta_{yi} = Y_i - \bar{Y} \tag{29}$$

Calculate the three sums A, B and C according to Formulae (30), (31) and (32):

$$A = \sum_i w_i^2 \Delta_{xi} \Delta_{yi} \Delta_{E,Xi}^2 \tag{30}$$

$$B = \sum_i w_i^2 \left(\Delta_{xi}^2 \Delta_{E,Yi}^2 - \Delta_{yi}^2 \Delta_{E,Xi}^2 \right) \tag{31}$$

$$C = -\sum_i w_i^2 \Delta_{xi} \Delta_{yi} \Delta_{E,Yi}^2 \tag{32}$$

Calculate the interim proportional correction b_0 according to Formula (33):

$$b_0 = \frac{-B + \sqrt{B^2 - 4AC}}{2A} \tag{33}$$

If $|b - b_0| > 0{,}001 \cdot b$, replace b with b_0 and go back to Formula (27), computing new values for the weights w_i, $\bar{X}$, $a = \bar{Y} - b\bar{X}$, 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 abort the iteration and calculate the final weights w_i as in Formula (27).

Calculate the weighted sum of squared residuals of the means results for the Class 2 bias correction, $\Sigma_{SR,2}$, according to Formula (34):

$$\Sigma_{SR,2} = \sum_i w_i \left(\Delta_{yi} - b \cdot \Delta_{xi} \right)^2 \tag{34}$$

Compute the constant bias correction a as shown in Formula (35)

$$a = \bar{Y} - b\bar{X} \tag{35}$$

6.4 Selection of the appropriate bias correction class

The sum of squared residuals from each class of bias correction is 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 weighted sum of squared residuals as calculated in 6.3.2 to 6.3.5 are repeated in Table 1.

Table 1 — Bias correction classes

Bias correction class	Appropriate correction	Weighted sum of squared residuals
Class 0	no correction	$\Sigma_{SR,0}$
Class 1a	constant bias correction	$\Sigma_{SR,1a}$
Class 1b	proportional bias correction	$\Sigma_{SR,1b}$
Class 2	proportional + constant bias correction	$\Sigma_{SR,2}$

To determine whether any bias correction (Class 1a, 1b or 2) can significantly improve the expected agreement between the two methods, calculate the following ratio according to Formula (36):

$$F = \frac{\left(\Sigma_{SR,0} - \Sigma_{SR,2}\right)/2}{\Sigma_{SR,2}/(S-2)} \tag{36}$$

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.

— 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 test for normal distribution of the weighted residual in 6.5.

— If the calculated F is larger, conclude that a correction can improve the expected agreement between the two methods, and compute the following t-ratios, as shown in Formulae (37) and (38):

$$t_1 = \sqrt{\frac{\Sigma_{SR,0} - \Sigma_{SR,1}}{\Sigma_{SR,2}/(S-2)}} \tag{37}$$

$$t_2 = \sqrt{\frac{\Sigma_{SR,1} - \Sigma_{SR,2}}{\Sigma_{SR,2}/(S-2)}} \tag{38}$$

where $\Sigma_{SR,1}$ is the lesser of $\Sigma_{SR,1a}$ or $\Sigma_{SR,1b}$, provided the latter is appropriate and has been calculated.

Compare t_2 to the upper 97,5th percentile of the t-distribution with S-2 degrees of freedom.

— 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 test for normal distribution of the weighted residual in 6.5.

— 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.

  — 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 $\Sigma_{SR,1b}$ is appropriate and is smaller than $\Sigma_{SR,1a}$. Proceed to test for normal distribution of residual in 6.5.

  — 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.

6.5 Confirming the normal distribution of weighted residuals

In order to make valid inferences from the selected bias-correction, the weighted residuals should follow a normal distribution. $\hat{Y}_i$ shall be the Y-method mean values predicted from the corresponding X-method mean values X_i, using the bias-correction selected in 6.4.

The weighted residuals ε_i are given by Formula (39):

$$\varepsilon_i = \sqrt{w_i}\left(Y_i - \widehat{Y}_i\right) \tag{39}$$

where w_i is the appropriate weights associated with the appropriate bias correction class.

Perform the Anderson-Darling test for normality on the residuals ε_i. See 5.2.3 for further information on the calculation and interpretation of this statistic. The distribution of the residuals is assumed to follow a normal distribution at the 5 % significance level if the corresponding $A_i^{2*} \leq 0{,}752$.

— If this test does meet the above criterion, the distribution of the residuals are considered to be normal. Proceed to 6.6 for testing for existence of sample-specific biases.

— If this test does not meet the above criterion, the distribution of the residuals is considered not to be normal. 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 methods are not measuring the same property. Do not proceed to 6.6.

6.6 Sample-specific biases

Sample-specific biases are defined by a functional dependency of (weighted) residuals and sample value. For testing the existence of sample-specific biases, a χ^2 test of independence is performed where the Σ_{SR} of the bias-correction class selected in 6.4 is compared to the 95th percentile value of a χ^2-distribution with v degrees of freedom.

where

v is S for Class 0 (no bias) correction,

v is $S - 1$ for Class 1a or Class 1b (constant or proportional) correction,

v is $S - 2$ for Class 2 (linear) correction.

— If $\Sigma_{SR} \leq \chi^2$-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.

The between-methods reproducibility is given by Formula (40):

$$R_{XY} = \sqrt{\frac{R_Y^2 + b^2 R_X^2}{2}} \tag{40}$$

where

R_Y is reproducibility of methods Y evaluated at the value of the single results from method Y,

R_X is reproducibility of methods X evaluated at the value of the single results from method X,

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

NOTE 1 For Class 0 and Class 1a bias corrections, $b = 1$.

— If $\Sigma_{SR} > \chi^2$-percentile, 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 biases are not consistent across the S common samples. The user may investigate whether the biases can be attributed to other observable properties of the samples. The user may restrict attention to a smaller class of materials to achieve an Σ_{SR} that is less than the aforementioned chi-squared percentile 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.

If the Σ_{SR} exceeds the 95th percentile value of the appropriate χ^2-distribution, 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 subclause. Even when it is appropriate to treat these sample-specific effects as random, the additional variation can cause the between-methods reproducibility to be far larger than the root mean square of the reproducibilities of the methods in Formula (40).

If the user decides to retain all samples in the study, the sample-specific bias will be treated as a random variance component and the between-methods reproducibility (R_{XY}) is calculated by Formula (41):

$$R_{XY} = \sqrt{\left(\frac{b^2 R_X^2}{2} + \frac{R_Y^2}{2}\right)\left[1 + \frac{2t^2\left(\Sigma_{SR} - S + k\right)S}{(S-k)\sum_i\left(\frac{b^2 R_{Xi}^2 + R_{Yi}^2}{b^2 \Delta_{E,Xi}^2 + \Delta_{E,Yi}^2}\right)}\right]} \tag{41}$$

where

b is the coefficient of the appropriate bias correction,

NOTE 2 For Class 0 and Class 1a, $b = 1$.

Σ_{SR} is the weighted sum of squared residuals of the appropriate bias-correction class,

S is the number of samples,

k is the Class number of selected Class,

NOTE 3 For Class 0 $k = 0$, for Class 1a or Class 1b $k = 1$ and for Class 2 $k = 2$.

$\Delta_{E,Xi}^2$, $\Delta_{E,Yi}^2$ are the squared standard errors for method X and Y of the i^{th} sample,

R_X, R_Y are the reproducibility of methods X and Y evaluated at the value of the single results X and Y,

t is the student t-value at a confidence level of 0,05 and infinite degrees of freedom.

Formulae (40) and (41) both provide an estimate of the limit which about 95 % of the differences are expected to exceed, under the correct and normal operation of both methods, when one party uses the bias-corrected X-method while another party uses the Y-method, on materials similar to the samples used for this study. Application of the methods to materials which are substantially different from the materials used for this study 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 programme.

7 Report

When reporting results for predicted method Y (represented by the symbol $\hat{Y}$) from a single bias-corrected result from method X, the calculation shall be as follows:

$$\hat{Y} = a + bX \tag{42}$$

An interval that would contain a single result from method Y (if performed on the same sample) about 95 % of the time can be constructed using a single bias-corrected result from method X, used as a predicted *Y* (represented by the symbol $\hat{Y}$), and between-methods reproducibility R_{XY}, as follows.

An interval bounded by $\hat{Y} \pm R_{XY}$ 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 Formulae (40) or (41), as appropriate, with R_Y evaluated as the bias-corrected result from method X.

The assessment findings are reported in the precision and bias clause of the appropriate standard(s) and/or test method(s).

8 Confirmation of the correlation

Users intending to use the bias-corrected test method X results as a predictor of test method Y results (or vice versa) using the correlation established by this methodology are advised to periodically check the agreement between the predicted Y result ($\hat{Y}$ or bias-corrected X) versus an actual method Y result against their application requirements. For this purpose, control chart techniques as outlined in ISO 4259-4 on the prediction error defined as ($\hat{Y}$ - Y) are used.

Confirmation should be performed using the following difference statistic D as shown in Formula (42), or other statistically equivalent techniques. For a single value D, the assessment findings are considered validated if the absolute value is less than or equal to 3. For a control chart, the D values are expected to randomly vary on either side of zero.

$$D = \frac{\left(\bar{Y} - \hat{Y}\right)}{\sqrt{\Delta^2_{E,\bar{Y}} + \Delta^2_{E,\hat{Y}}}} \tag{43}$$

where

$\hat{Y}$ $\quad \bar{Y}\,\bar{X}\,\hat{Y} = b\bar{X} + a$

$\bar{Y}$ is the average of Y method of the same material,

$\bar{X}$ is the average of X method of the same material,

a, b are the bias correction coefficients from the bias assessment, if applicable,

$\Delta^2_{E,\bar{Y}}$ is the standard error of $\bar{Y}$ given by $\Delta_{E,\bar{Y}} = s_{R,Y}/\sqrt{L_Y}$, where $s_{R,Y}$ refers to the published repeatability standard deviation of method Y according to $s_{R,Y} = R_Y/(t\sqrt{2})$,

$\Delta^2_{E,\hat{Y}}$ $\quad \hat{Y}\bar{X}\ \Delta^2_{E,\hat{Y}} = \dfrac{bs_{R,X}}{\sqrt{L_X}}\, s_{R,X} s_{R,X}\, R_X t$

R_X, R_Y are the published reproducibilities for methods X and Y,

L_X, L_Y are the number of non-rejected results used to calculate the average for methods X and Y, where the protocol is for a single test result to be reported by each participant,

t represent the student-t value with a confidence level of 0,05 and degrees of freedom from the published precision statement of the method Y.

In case the degree of freedom is unknown, the minimum value of 30 may be used.

Sustained values of D on either the positive or negative side of zero should trigger activities for a reassessment.

Annex A
(informative)

Worked example using ILS data

A.1 General

The purpose of this annex is to illustrate the methodology using an ILS data set containing results for the determination of cetane numbers (CN) of diesel fuel oil in compression ignition engines using test methods ISO 5165 and EN 16906. This data set is a subset of a larger data set that has been used to establish a precision statements for both test methods.

Within this evaluation, the test method from ISO 5165 is indicated as test method X and the test method from EN 16906 is indicated as test method Y.

A.2 Relevant property selection

A.2.1 General

Consider and summarize the relevant properties of both test methods, X and Y.

A.2.2 Test method X

ISO 5165:2017[1)] establishes the rating of diesel fuel oil in terms of an arbitrary scale of cetane numbers (CNs) using a standard single cylinder, four-stroke cycle, variable compression ratio, indirect injected diesel engine. The application and precision ranges are listed in Table A.1.

Table A.1 — Scope of test method X

Typical application range, CN	30 to 65
Precision range	52,4 to 73,8
Published reproducibility, *R*	0,125·CN–2,2
Published repeatability, *r*	0,01·CN+0,42

The expression of repeatability and reproducibility functions are constructed by regression of fixed precision values at specified CN levels as shown in Table A.2.

Table A.2 — Published precision for test method X

Cetane number CN	Repeatability *r*	Reproducibility *R*
40	0,8	2,8
44	0,9	3,3
48	0,9	3,8
52	0,9	4,3
56	1,0	4,8

In addition, the precision statement in ISO 5165 expresses that the average standard deviation for each CN level has been multiplied by 2,772 to obtain the respective repeatability and reproducibility values. This means that for the calculation of the reproducibility standard deviation and repeatability standard deviation from the reproducibility and repeatability, a factor of 2,772 should be used, not of 2,888.

1) Withdrawn.

A.2.3 Test method Y

EN 16906 specifies a test method for the determination of numbers (CNs) in diesel fuel in the range from CN 0 to CN 100 (CN 40 to CN 75 typical), using a standard single cylinder, four-stroke cycle, indirect injection engine. The cetane number provides a measure of the ignition characteristics of diesel fuels in compression ignition engines. The relevant scope details are listed in Table A.3.

Table A.3 — Scope of test method Y

Typical application range, CN	0 to 100 (40 to 75 typically)
Precision range	44 to 66
Published reproducibility, *R*	1,5
Published repeatability, *r*	0,64

A.3 Laboratory data

The laboratory data of all samples are collected and discard all outliers. Since the entire data set has already been assessed with a procedure from ISO 4259-1 or an equivalent standard to generate a precision statement, it is fair to assume that no outliers are present in the data set. The cetane number data for method X is given in Table A.4 and for method Y in Table A.5, where the laboratories are indicated in vertical columns as 'L1', 'L2', etc. and the samples are indicated in horizontal rows as 'S1', 'S2', etc. The vertical column labelled 'Repeat no.' refers to the first or second measurement result of a lab for a sample.

Table A.4 — Cetane number data for test method X

		Sample no. S_i														
Laboratory no. L_i	**Repeat no.**	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15
L1	1	52,8	52,1	67,6	61,3	54,4	49,1	43,8	53,9	51,9	52,1	53,8	53,5	56,5	51,4	52,5
	2	52,6	52,8	68,6	62,2	54,9	49,4	43,5	53,3	52,1	51,7	53,2	54,0	56,5	51,2	52,1
L2	1	52,0	52,3	65,9	60,8	53,7	47,9	43,4	56,5	53,3	52,2	54,8	51,6	55,4	50,7	52,9
	2	52,3	52,1	66,5	60,1	53,1	48,0	43,6	56,8	52,9	52,3	53,8	51,2	54,9	51,2	53,4
L3	1	51,8	53,3	66,4	60,0	53,8	49,4	44,1	56,3	52,4	51,9	52,7	51,7	55,0	51,0	53,6
	2	52,7	52,4	65,7	60,6	54,2	49,7	44,2	55,9	51,9	51,3	52,7	52,0	54,9	50,9	53,0
L4	1	52,3	50,5	67,0	59,7	55,0	50,6	42,7	57,9	53,0	52,7	52,2	51,3	56,9	51,3	54,4
	2	52,4	51,0	66,4	60,0	55,1	50,5	41,7	56,7	53,6	53,2	52,3	51,5	57,3	51,3	54,6
L5	1	51,9	51,7	65,7	60,5	53,7	49,7	44,4	53,8	53,2	53,3	54,0	52,0	55,4	50,2	52,4
	2	51,6	51,9	64,7	61,1	54,1	49,6	44,9	53,9	52,8	53,4	53,7	52,3	55,0	49,4	52,3
L6	1	53,0	52,6	66,9	59,5	54,1	48,6	42,5	54,2	52,8	53,5	52,3	52,4	55,5	50,7	52,8
	2	53,7	52,6	66,4	59,9	54,3	49,2	43,5	54,5	52,0	53,6	53,0	53,3	55,7	51,6	53,0
L7	1	52,5	50,1	63,3	60,8	53,9	46,4	42,6	56,5	51,8	53,7	53,7	51,8	54,2	51,7	54,0
	2	52,3	51,3	63,4	61,3	54,6	48,3	42,5	56,3	51,8	53,5	54,1	52,1	53,8	51,2	53,5
L8	1	51,7	51,6	66,6	60,9	53,8	48,2	43,9	55,2	52,5	52,9	53,8	51,8	55,1	52,6	52,4
	2	51,1	51,7	65,8	60,2	53,5	47,4	44,4	55,1	52,9	52,1	54,0	52,5	54,6	52,1	52,4
L9	1	52,0	52,6	66,9	58,9	52,4	48,9	42,4	53,8	51,2	53,4	55,0	52,5	55,0	51,0	53,3
	2	51,9	52,1	67,7	58,0	51,2	47,8	42,9	54,9	51,5	53,8	54,5	53,2	55,9	51,6	53,8

Table A.5 — Cetane number for test method Y

		Sample no. S_i														
Laboratory no. L_i	**Repeat no.**	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	S15
L1	1	51,6	50,8	65,3	60,0	53,7	48,0	44,6	53,8	52,7	52,6	53,8	51,6	55,2	51,0	53,0
	2	51,7	51,6	66,1	59,9	53,8	48,2	44,1	54,1	52,5	52,5	53,7	51,6	54,9	50,9	52,8
L2	1	52,3	52,2	66,1	59,2	52,8	47,9	43,4	55,0	52,1	52,6	53,4	52,1	54,8	50,7	53,0
	2	52,2	51,9	66,3	59,0	53,3	47,4	43,8	55,2	52,2	52,9	53,3	52,4	54,8	50,8	52,8
L3	1	52,1	51,5	66,9	59,9	52,8	47,7	43,3	54,5	52,1	52,8	53,7	51,2	55,4	51,8	53,0
	2	52,2	52,0	67,4	60,1	52,8	47,8	43,3	54,7	52,3	52,6	53,0	51,2	55,8	52,0	53,9
L4	1	52,2	52,1	65,8	59,3	54,4	48,7	43,9	53,9	51,8	52,3	52,8	52,8	54,8	50,0	52,4
	2	52,0	52,0	65,3	59,6	54,8	48,8	44,0	54,2	51,3	53,0	52,8	52,6	55,0	49,7	52,3
L5	1	51,8	52,4	65,5	60,0	53,2	48,0	43,6	55,0	51,1	52,7	52,9	52,6	55,1	50,8	52,2
	2	51,9	52,1	65,4	60,1	53,0	48,4	44,1	54,4	51,5	52,7	52,8	52,2	55,0	50,9	52,2
L6	1	52,6	51,0	66,1	61,3	53,4	49,2	43,5	55,6	52,8	53,3	53,2	52,3	56,9	52,0	52,2
	2	52,4	50,6	65,8	61,1	53,2	48,3	43,4	55,5	52,9	53,4	53,0	52,3	56,6	52,2	52,3
L7	1	51,9	52,9	63,6	59,6	53,8	47,3	44,1	56,2	52,2	52,4	51,9	51,9	56,1	51,4	52,3
	2	52,0	52,5	64,6	59,5	53,5	48,0	43,3	55,9	52,2	52,7	52,2	52,1	56,4	51,1	52,4
L8	1	51,7	51,4	65,9	59,7	54,2	48,1	42,9	55,1	51,5	51,6	53,4	51,6	55,2	51,7	53,4
	2	52,0	52,0	66,4	60,2	54,7	48,0	42,2	54,9	51,9	51,9	53,4	52,2	55,4	51,8	53,6
L9	1	51,6	50,5	65,6	59,3	53,6	49,1	42,8	54,2	52,5	52,1	54,3	51,6	54,6	50,5	52,9
	2	52,3	51,1	65,4	59,8	53,7	49,5	42,3	54,9	53,2	51,6	53,7	52,0	55,3	50,3	52,5

For testing the amount of variation or resolution of the data, the number of unique values within the data set is counted and evaluated. It is considered to be sufficient to have 117 (43 %) unique values out of a total of 270 values for the test method ISO 5165, and 114 (42 %) unique values out of a total 270 the variation in the data set.

A.4 Calculate the averages over the repeats and rearrange

Construct new tables with cells presenting the average over the repeats of all lab/sample combinations with one additional decimal. This information is listed in Table A.6 for method X and in Table A.7 for method Y. For convenient data processing, the table is transposed to display samples horizontally and laboratories vertically.

Table A.6 — Cetane Number Data for test method X

	Laboratory no. L_i												
Sample no. S_i	L1	L2	L3	L4	L5	L6	L7	L8	L9	**Average** X_i	**Standard deviation** s_{Xi}	**No. of laboratories** L_{Yi}	**Relative standard deviation** s_{Xi} (%)
S1	52,70	52,15	52,25	52,35	51,75	53,35	52,40	51,40	51,95	52,26	0,561	9	1,1
S2	52,45	52,20	52,85	50,75	51,80	52,60	50,70	51,65	52,35	51,93	0,776	9	1,5
S3	68,10	66,20	66,05	66,70	65,20	66,65	63,35	66,20	67,30	66,19	1,342	9	2,0
S4	61,75	60,45	60,30	59,85	60,80	59,70	61,05	60,55	58,45	60,32	0,934	9	1,5
S5	54,65	53,40	54,00	55,05	53,90	54,20	54,25	53,65	51,80	53,88	0,924	9	1,7
S6	49,25	47,95	49,55	50,55	49,65	48,90	47,35	47,80	48,35	48,82	1,036	9	2,1
S7	43,65	43,50	44,15	42,20	44,65	43,00	42,55	44,15	42,65	43,39	0,840	9	1,9
S8	53,60	56,65	56,10	57,30	53,85	54,35	56,40	55,15	54,35	55,31	1,346	9	2,4
S9	52,00	53,10	52,15	53,30	53,00	52,40	51,80	52,70	51,35	52,42	0,655	9	1,2

Table A.6 *(continued)*

	Laboratory no. L_i												
Sample no. S_i	L1	L2	L3	L4	L5	L6	L7	L8	L9	**Average** X_i	**Standard deviation** s_{Xi}	**No. of laboratories** L_{Yi}	**Relative standard deviation** s_{Xi} (%)
S10	51,90	52,25	51,60	52,95	53,35	53,55	53,60	52,50	53,60	52,81	0,775	9	1,5
S11	53,50	54,30	52,70	52,25	53,85	52,65	53,90	53,90	54,75	53,53	0,834	9	1,6
S12	53,75	51,40	51,85	51,40	52,15	52,85	51,95	52,15	52,85	52,26	0,765	9	1,5
S13	56,50	55,15	54,95	57,10	55,20	55,60	54,00	54,85	55,45	55,42	0,916	9	1,7
S14	51,30	50,95	50,95	51,30	49,80	51,15	51,45	52,35	51,30	51,17	0,661	9	1,3
S15	52,30	53,15	53,30	54,50	52,35	52,90	53,75	52,40	53,55	53,13	0,737	9	1,4

Table A.7 — Cetane Number Data for test method Y

	Laboratory no. L_i												
Sample no. S_i	L1	L2	L3	L4	L5	L6	L7	L8	L9	**Average** Y_i	**Standard deviation** s_{Yi}	**No. of laboratories** L_{Yi}	**Relative standard deviation** s_{Yi} %
S1	51,65	52,25	52,15	52,10	51,85	52,50	51,95	51,85	51,95	52,028	0,253	9	0,5
S2	51,20	52,05	51,75	52,05	52,25	50,80	52,70	51,70	50,80	51,700	0,654	9	1,3
S3	65,70	66,20	67,15	65,55	65,45	65,95	64,10	66,15	65,50	65,750	0,812	9	1,2
S4	59,95	59,10	60,00	59,45	60,05	61,20	59,55	59,95	59,55	59,867	0,592	9	1,0
S5	53,75	53,05	52,80	54,60	53,10	53,30	53,65	54,45	53,65	53,594	0,616	9	1,1
S6	48,10	47,65	47,75	48,75	48,20	48,75	47,65	48,05	49,30	48,244	0,574	9	1,2
S7	44,35	43,60	43,30	43,95	43,85	43,45	43,70	42,55	42,55	43,478	0,606	9	1,4
S8	53,95	55,10	54,60	54,05	54,70	55,55	56,05	55,00	54,55	54,839	0,674	9	1,2
S9	52,60	52,15	52,20	51,55	51,30	52,85	52,20	51,70	52,85	52,156	0,556	9	1,1
S10	52,55	52,75	52,70	52,65	52,70	53,35	52,55	51,75	51,85	52,539	0,483	9	0,9
S11	53,75	53,35	53,35	52,80	52,85	53,10	52,05	53,40	54,00	53,183	0,574	9	1,1
S12	51,60	52,25	51,20	52,70	52,40	52,30	52,00	51,90	51,80	52,017	0,454	9	0,9
S13	55,05	54,80	55,60	54,90	55,05	56,75	56,25	55,30	54,95	55,406	0,676	9	1,2
S14	50,95	50,75	51,90	49,85	50,85	52,10	51,25	51,75	50,40	51,089	0,737	9	1,4
S15	52,90	52,90	53,45	52,35	52,20	52,25	52,35	53,50	52,70	52,733	0,496	9	0,9

From a quality point of view, the data integrity can be assessed somewhat by considering the relative standard deviations on this point for both test methods. Outliers should be considered further.

A.5 Performing the basic checks on the data set

A.5.1 The data set is made up of 15 samples and nine laboratories presenting averages over the repeated results. This satisfies the requirements for ILS data that the results originate from at least 10 samples and six laboratories.

A.5.2 The samples in the data set span the CN range from 43,6 to 66,2 for ISO 5165 and from 43,5 to 65,8 for EN 16906. As shown in Table A.8, these ranges are compared with the published application and precision ranges of the test methods.

Table A.8 — Application and precision ranges of the test methods compared with data set

Reference for the test method	Test method X	Test method Y
Typical published application range	30 to 65	45 to 63
Published precision range	52,4 to 73,8	47 to 61
Range of the data set	43,6 to 66,2	43,5 to 65,8

Comparing these ranges shows that the results of the evaluation are limited to a CN range of 52,4 to 61.

A.6 The test samples for leverage

Table A.9 shows the leverages of the samples, calculated based on the averages of the methods. In this case, a leverage is not performed for each sample for both test methods, but the average over the test methods is used to determine the leverage per sample.

Table A.9 — Leverage of each sample, averaged over both methods

Sample no. S_i	Sample mean $(X_i + Y_i)/2$	Leverage h_i
S1	52,14	0,07
S2	51,81	0,07
S3	65,97	0,46
S4	60,09	0,19
S5	53,74	0,07
S6	48,53	0,14
S7	43,43	0,42
S8	55,07	0,08
S9	52,29	0,07
S10	52,68	0,07
S11	53,36	0,07
S12	52,14	0,07
S13	55,41	0,08
S14	51,13	0,08
S15	52,93	0,07

For each sample the calculated leverage value, h_i, remains below the critical value of 0,5, so the test has passed.

The sample distribution is examined for normality using the Anderson Darling (AD)-test and part of the overview table.

This test is not required for data originating from an ILS study, which have already gone through an entire procedure of ISO 4259-1 or an equivalent standard. However, it is recommended to add it for consistency with the PTP-data set evaluation in Annex B.

A.7 Constructing overview tables with summary statistics per sample

To check the data against the requirements, the data are examined for the distribution of the laboratory results per sample. For each sample, at least six laboratories returned a result and the precision of the data are compared to the published precision figures. Only the test for the minimum number of laboratories returning a result is required for data originating from an ILS study. The results are given in Table A.10 for test method X and in Table A.11 for test method Y.

Table A.10 — Overview table for test method X

Sample no. S_i [a]	Average X_i [b]	Standard deviation s_{Xi} [c]	No. of laboratories L_{Xi} [d]	Reproducibility R_{Xi} [e]	Repeatability r_{Xi} [f]	Reproducibility standard deviation $s_{R,Xi}$ [g]	Repeatability standard deviation $s_{r,Xi}$ [h]	Standard error $\Delta_{E,Xi}$ [i]	Leverage h_i [j]	Distribution value A_i^{2*} [k]	Result of test $L \geq 10$ [l]	Result of F-test $s_{Xi} \leq s_{R,Xi}$ [m]
S1	52,256	0,561	9	= 0,125·X-2,2	= 0,01·X+0,42	1,563	0,340	0,515	0,070	0,229	yes	yes
S2	51,928	0,776	9			1,548	0,339	0,510	0,072	0,466	yes	yes
S3	66,194	1,342	9			2,191	0,390	0,726	0,464	0,439	yes	yes
S4	60,322	0,934	9			1,927	0,369	0,637	0,194	0,268	yes	yes
S5	53,878	0,924	9			1,636	0,346	0,539	0,068	0,512	yes	yes
S6	48,817	1,036	9			1,408	0,328	0,462	0,138	0,199	yes	yes
S7	43,389	0,840	9			1,163	0,308	0,379	0,416	0,253	yes	yes
S8	55,306	1,346	9			1,700	0,351	0,561	0,077	0,370	yes	yes
S9	52,422	0,655	9			1,570	0,341	0,517	0,069	0,194	yes	yes
S10	52,811	0,775	9			1,588	0,342	0,523	0,067	0,458	yes	yes
S11	53,533	0,834	9			1,620	0,345	0,534	0,067	0,378	yes	yes
S12	52,261	0,765	9			1,563	0,340	0,515	0,070	0,396	yes	yes
S13	55,422	0,916	9			1,706	0,351	0,563	0,081	0,380	yes	yes
S14	51,172	0,661	9			1,514	0,336	0,498	0,080	0,753	yes	yes
S15	53,133	0,737	9			1,602	0,343	0,528	0,067	0,290	yes	yes
									pass	pass	pass	100 % pass

[a] The sample ID, S_i.

[b] The average of the sample over the laboratories, X_i, Y_i.

[c] The standard deviation of the sample over the laboratories, s_{Xi}, s_{Yi}.

[d] The number of laboratories presenting at least one result per sample, L_{Xi}, L_{Yi}.

[e] The published reproducibility given by the published test method, R_{Xi}, R_{Yi}.

[f] The published repeatability given by the published test method, r_{Xi}, r_{Yi}.

[g] Reproducibility standard deviation, $s_{R,Xi}$, $s_{R,Yi}$.

[h] Reproducibility standard deviation, $s(r_{Xi})$, $s(r_{Yi})$.

[i] The standard error, $\Delta_{E,Xi}$, $\Delta_{E,Yi}$, which is calculated using Formula (6) where $s_R = R/2{,}772$ and $s_r = r/2{,}772$ for method X (see ISO 5165). See A.2.1 for an explanation regarding why 2,772 is used instead of 2,888 to calculate s_R and s_r from R and r for this method. For method Y, (see EN 16906) $s_R = R/2{,}888$ and $s_r = r/2{,}888$.

[j] The leverage h_i is calculated according to A.5 with the overall test result in the final row indicating whether all the leverage tests for samples meet the critical value of 0,5.

[k] The Anderson-Darling test parameter for a normal distribution with the overall test result in the final row indicating whether the AD-tests for all samples meet the critical value of 1,12.

[l] Test results for $L \geq 6$ (ILS requirement) with the overall test result in the final row, indicating whether all the test parameters meet the critical value of at least 6.

[m] The result of the F-test where the sample standard deviation is compared with the published reproducibility standard deviation. For at least 80 % of the samples, the sample standard deviation is not allowed to significantly exceed the reproducibility standard deviation.

Table A.11 — Overview table for test method Y

Sample no. S_i [a]	Average Y_i [b]	Standard deviation s_{Yi} [c]	No. of laboratories L_{Yi} [d]	Reproducibility R_{Xi} [e]	Repeatability r_{Xi} [f]	Reproducibility standard deviation $s_{R,Yi}$ [g]	Repeatability standard deviation $s_{r,Yi}$ [h]	Standard error $\Delta_{E,Yi}$ [i]	Leverage h_i [j]	Distribution value A_i^{2*} [k]	Result of test $L \geq 10$ [l]	Result of F-test $s_{Yi} \leq s_{R,Yi}$ [m]
S1	52,028	0,253	9	= 1,5	= 0,64	0,519	0,222	0,165	0,070	0,236	yes	yes
S2	51,700	0,654	9			0,519	0,222	0,165	0,072	0,316	yes	yes
S3	65,750	0,812	9			0,519	0,222	0,165	0,464	0,489	yes	yes
S4	59,867	0,592	9			0,519	0,222	0,165	0,194	0,646	yes	yes
S5	53,594	0,616	9			0,519	0,222	0,165	0,068	0,364	yes	yes
S6	48,244	0,574	9			0,519	0,222	0,165	0,138	0,445	yes	yes
S7	43,478	0,606	9			0,519	0,222	0,165	0,416	0,391	yes	yes
S8	54,839	0,674	9			0,519	0,222	0,165	0,077	0,225	yes	yes
S9	52,156	0,556	9			0,519	0,222	0,165	0,069	0,317	yes	yes
S10	52,539	0,483	9			0,519	0,222	0,165	0,067	0,736	yes	yes
S11	53,183	0,574	9			0,519	0,222	0,165	0,067	0,283	yes	yes
S12	52,017	0,454	9			0,519	0,222	0,165	0,070	0,161	yes	yes
S13	55,406	0,676	9			0,519	0,222	0,165	0,081	0,785	yes	yes
S14	51,089	0,737	9			0,519	0,222	0,165	0,080	0,214	yes	yes
S15	52,733	0,496	9			0,519	0,222	0,165	0,067	0,510	yes	yes
									pass	pass	pass	100 % pass

[a] The sample ID, S_i.

[b] The average of the sample over the laboratories, X_i, Y_i.

[c] The standard deviation of the sample over the laboratories, s_{Xi}, s_{Yi}.

[d] The number of laboratories presenting at least one result per sample, L_{Xi}, L_{Yi}.

[e] The published reproducibility given by the published test method, R_{Xi}, R_{Yi}.

[f] The published repeatability given by the published test method, r_{Xi}, r_{Yi}.

[g] Reproducibility standard deviation, $s_{R,Xi}$, $s_{R,Yi}$.

[h] Reproducibility standard deviation, $s(r_{Xi})$, $s(r_{Yi})$.

[i] The standard error, $\Delta_{E,Xi}$, $\Delta_{E,Yi}$, which is calculated using Formula (6) where $s_R = R/2{,}772$ and $s_r = r/2{,}772$ for method X (see ISO 5165). See A.2.1 for an explanation regarding why 2,772 is used instead of 2,888 to calculate s_R and s_r from R and r for this method. For method Y, (see EN 16906) $s_R = R/2{,}888$ and $s_r = r/2{,}888$.

[j] The leverage h_i is calculated according to A.5 with the overall test result in the final row indicating whether all the leverage tests for samples meet the critical value of 0,5.

[k] The Anderson-Darling test parameter for a normal distribution with the overall test result in the final row indicating whether the AD-tests for all samples meet the critical value of 1,12.

[l] Test results for $L \geq 6$ (ILS requirement) with the overall test result in the final row, indicating whether all the test parameters meet the critical value of at least 6.

[m] The result of the F-test where the sample standard deviation is compared with the published reproducibility standard deviation. For at least 80 % of the samples, the sample standard deviation is not allowed to significantly exceed the reproducibility standard deviation.

For a few samples of method Y, the standard deviation s_{Yi} exceeds the published reproducibility standard deviation of the method $s_{R,Yi.}$ However, the F-test indicates that it cannot be proved that s_i is significantly larger than R_{Yi}.

The tests on leverage, distribution, minimum number of laboratories and the sample standard deviation are not required for data originating from an ILS study, which have already gone through an entire ISO 4259-1 or equivalent procedure. However, for uniformity with the PTP-data set evaluation in Annex B it is recommended to add those test results anyway.

A.8 Testing on sufficient variation

A test on sample variation is performed for each test method to determine whether the samples can be distinguished from each other by comparing the F-test results with their critical F-test values. The summary statistics for this test on variation are given in Table A.12.

Table A.12 — Summary statistics for test on variation

	Method X	Method Y
Weighted average, $\bar{X}$, $\bar{Y}$	52,23	53,24
Total sum of squares, $\Sigma_{ST,X}$, $\Sigma_{ST,Y}$	1 215,8	12 476,6
Number of samples, S	15	15
Confidence level, α	0,05	0,05
Calculated F-test value	86,8	891,2
Critical F-test value	2,04	2,04
Result of the F-test	pass	pass

These statistics show that the test has passed for both methods, meaning that the test methods are sufficiently precise to distinguish between noise and sample variation.

A.9 Correlation test

Perform a test to determine whether the methods are sufficiently correlated by calculating the correlation coefficient and perform the t-test. The summary statistics for this test on correlation are given in Table A.13.

Table A.13 — Summary statistics for correlation test

Weighted average, $\ddot{X}$	52,36
Weighted average, $\ddot{Y}$	52,10
Correlation coefficient, ρ	0,999 4
Number of samples, S	15
F-statistic for significance test, F	10 553,88
Confidence level, α	0,01
Critical value for F	9,07
Result of the F-test	pass

The calculated F-value is significantly larger than the critical F-value with which this test passes. This means that the data from both methods are sufficiently correlated.

A.10 Calculation of bias correction classes

The optimal bias correction shall first be determined. The residuals of this model are then examined for normal distribution and subsequently tested for the presence of sample specific biases. The residuals of each of the bias correction model are examined for distribution and each model is tested for the presence of a sample specific bias. The results are given in Tables A.14 to A.17.

Table A.14 — No bias correction, Class 0

Calculation of Class 0 – No bias correction						Test for normal distribution of residuals (AD-test)			
$\Sigma_{SR,0}$	$= \sum[w_i \cdot (X_i - Y_i)^2]$			5,1		number of samples, S			15
						confidence level, α			0,05
						A_i^{2*} calculate			0,47
						A_i^{2*} critical			0,71
						p-value			0,24
						Are residuals normally distributed?			Yes
						Test for existence of sample-specific bias			
R_X	From published test method			0,125·X-2,2		Degree of freedom, ν			14
R_Y	From published test method			1,5		confidence level, α			0,05
	Residuals normally distributed			yes		Σ_{SR}			5,1
	Sample specific bias significant			no		χ^2-critical			23,7
R_{XY}	$= \sqrt{[(R^2_X + R^2_Y)/2]}$					p-value			0,98
						Is the sample specific bias significant?			No

Table A.15 — Constant correction, Class 1a

Calculation of Class 1a – Constant bias correction						Test for normal distribution of residuals (AD-test)			
$\Sigma_{SR,1a}$				1,8		number of samples, S			15
a				−0,258		confidence level, α			0,05
						A_i^{2*} calculate			0,45
						A_i^{2*} critical			0,71
						p-value			0,28
						Are residuals normally distributed?			Yes
						Test for existence of sample-specific bias			
R_X	From published test method			0,125·X-2,2		Degree of freedom, ν			14
R_Y	From published test method			1,5		confidence level, α			0,05
	Residuals normally distributed			yes		Σ_{SR}			1,8
	Sample specific bias significant			no		χ^2-critical			23,7
R_{XY}	$= \sqrt{[(R^2_X + R^2_Y)/2]}$					p-value			1,00
						Is the sample specific bias significant?			No

Table A.16 — Proportional correction, Class 1b

Calculation of Class 1b – Proportional bias correction						Test for normal distribution of residuals (AD-test)			
$\Sigma_{SR,1b}$				1,6		number of samples, S			15
b				0,995		confidence level, α			0,05
A				127 557,9		AD-calculate			0,46
B				−113 789,0		AD-critical			0,71
C				−13 058,8		p-value			0,26
						Are residuals normally distributed?			**Yes**
						Test for existence of sample-specific bias			

Table A.16 *(continued)*

R_X	From published test method	0,125·X-2,2		Degree of freedom, ν		14
R_Y	From published test method	1,5		confidence level, α		0,05
	Residuals normally distributed	yes		Σ_{SR}		1,6
	Sample specific bias significant	no		χ^2-critical		23,7
R_{XY}	$= \sqrt{[(R^2_X + R^2_Y)/2]}$			p-value		1,00
				Is the sample specific bias significant?		No

Table A.17 — Linear correction, Class 2

Calculation of Class 2 – Linear bias correction						**Test for normal distribution of residuals (AD-test)**		
$\Sigma_{SR,2}$				1,3		number of samples, S		15
$\bar{X}$				52,364		confidence level, α		0,05
$\bar{Y}$				52,106		AD-calculate		0,62
a				0,801		AD-critical		0,71
b				0,980		p-value		0,11
A				1 026,2		Are residuals normally distributed?		Yes
B				−884,9				
C				−118,2				
						Test for existence of sample-specific bias (χ^2-test)		
R_X	From published test method			0,125·X-2,2		Degree of freedom, ν		13
R_Y	From published test method			1,5		confidence level, α		0,05
	Residuals normally distributed			yes		Σ_{SR}		1,3
	Sample specific bias significant			no		χ^2-critical		22,4
R_{XY}	$= \sqrt{[(R^2_X + R^2_Y)/2]}$					p-value		1,00
						Is the sample specific bias significant?		No

A.11 Selecting the most appropriate class bias correction

A.11.1 For the selection of the appropriate class of correction, the results of each class are summarized, as shown in Table A.18.

Table A.18 — Summary of the results of the correction classes

Bias correction class	Constant bias correction coefficient a	Proportional correction coefficient b	Weighted sum of squared residuals Σ_{SR}	Degrees of freedom ν	Critical value for distribution parameter χ^2	Is the sample specific bias significant?	Do residuals follow a normal distribution?
Class 0	-	-	5,1	15	25,0	no	yes
Class 1a	−0,258	-	1,8	14	23,7	no	yes
Class 1b	-	0,995	1,6	14	23,7	no	yes
Class 2	0,801	0,980	1,3	13	22,4	no	yes
number of samples, S			15				

A.11.2 The data for the optimal bias correction class is calculated and presented in Table A.19.

Table A.19 — Summarized statistics for the selection of the optimal bias correction

A. Check whether any bias correction can significantly improve the agreement (test Class 2 against Class 0)	
$F = [(\Sigma_{SR,0}\text{-}\Sigma_{SR,2})/2] \,/\, [\Sigma_{SR,2}/(S\text{-}2)]$	18,50
confidence level	0,05
F-critical	3,81
F-test	A bias correction (1a, 1b or 2) can sufficiently improve the agreement.
Next step to follow:	Testing on Class 1 (a or b) and Class 2. (Continue with B.)
B. Testing on Class 1 (a or b) and Class 2	
B.0. Which Class 1 (a or b) is preferred over the other?	Class 1b
B.1. Check whether a Class 2 can improve the expected agreement over a Class 1 (a or b).	
$t_2 = \sqrt{\{(\Sigma_{SR,1}\text{-}\Sigma_{SR,2})/[\Sigma_{SR,2}/(S\text{-}2)]\}}$	1,58
confidence level	0,025
t_2-critical	2,53
t_2-test	A Class 1 (a or b) is preferred over a Class 2 to improve the expected agreement.
Next step to follow	Test if Class 2 can improve the expected agreement over a Class 1 (a or b). (To B.2.)
B.2. Check whether a Class 2 can improve the expected agreement over a Class 1 (a or b).	
$t_1 = \sqrt{\{(\Sigma_{SR,0}\text{-}\Sigma_{SR,1})/[\Sigma_{SR,2}/(S\text{-}2)]\}}$	5,87
confidence level	0,025
t_1-critical	2,53
t_1-test	A Class 1 (a or b) is preferred over a Class 2 to improve the expected agreement.
Next step to follow	Select Class 1 (Class 1b in this case) and continue with test on sample specific biases. (C)
C. Check on sample specific biases	
Class selected	Class 1b
a	-
b	0,995
Do residuals follow a normal distribution?	yes
Is the sample specific bias significant for Class 1b ?	no
R_{XY}	$\sqrt{[(R^2_Y + b^2 \cdot R^2_X)/2]}$

A.12 Between-methods reproducibility

The result is that a bias correction Class 1b is optimal, with the residuals having a normal distribution and a bias sample specific bias absent. The between methods reproducibility is thus given as shown in Formula (A.1):

$$R_{XY} = \sqrt{\frac{R_Y^2 + b \cdot R_X^2}{2}} \tag{A.1}$$

where

b is 0,995;

R_X is 0,125·CN – 2,2;

R_Y is 1,5.

With a valid CN range of 52,4 up to 61 (see A.5.2).

A.13 Confirmation

For this example, no additional data are available for confirmation of the results.

Annex B
(informative)

Worked example using PTP data

B.1 General

The purpose of this annex is to illustrate the procedure using a PTP data set containing results for the determination of benzene in spark ignition engine fuels by gas chromatography using the test methods in ASTM 6839[5] and ASTM D5580.[6] This data set is a subset of a larger data set that has been used to establish a precision statements for both test methods.

Within this evaluation, the test method from ASTM D6839:2017 is indicated as test method X and the test method from ASTM D5580:2020 indicated as test method Y.

B.2 Relevant property selection

B.2.1 General

Consider and summarize the relevant properties of both test methods.

B.2.2 Test method X

The test method ASTM D6839:2017 covers the quantitative determination of saturates, olefins, aromatics, and oxygenates in spark ignition engine fuels by multidimensional gas chromatography. Only the content benzene is the subject of this example. The relevant application and precisions ranges are listed in Table B.1.

Table B.1 — Scope for benzene for test method X

Typical application range, volume fraction %	0 to 2
Precision range, volume fraction %	0,5 to 1,6
Published reproducibility, volume fraction %	$0{,}053 \cdot x^{1,6}$
Published repeatability, volume fraction %	$0{,}019 \cdot x^{1,6}$
Key x measured value	

B.2.3 Test method Y

The test method specified in ASTM D5580 covers the determination of benzene, toluene, ethylbenzene, the xylenes, C9 and heavier aromatics, and total aromatics in finished motor gasoline by gas chromatography. Only the content benzene is the subject of this example. The relevant application and precisions ranges are listed in Table B.2.

Table B.2 — Scope for benzene for test method Y

Typical application range, volume fraction %		0,1 to 5
Precision range, volume fraction %		0,11 to 1,5
Published reproducibility, volume fraction %		$0{,}108\ 7 \cdot x^{0{,}64}$
Published repeatability, volume fraction %		$0{,}025\ 9 \cdot x^{0{,}64}$
Key x measured value		

B.3 Laboratory data

The laboratory data of all samples are collected and all outliers are discarded. According to the reporting methods, the data are presented with 2 decimals. The benzene content data for method X is given in Table B.3 and for method Y in Table B.4 where the laboratories are indicated in vertical columns as 'L1', 'L2', etc. and the samples are indicated in horizontal rows as 'S1', 'S2', etc.

Table B.3 — Benzene content for test method X

	Sample no. S_i											
Laboratory no. L_i	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12
L1	0,47		0,24	1,36	0,57	0,64	0,42	0,99		0,5	1,61	0,57
L2	0,49	0,86	0,24	1,42	0,56	0,63	0,41	0,98		0,48	1,59	0,58
L3	0,48	0,88	0,24	1,4	0,56	0,64	0,41		0,92	0,5	1,55	0,57
L4		0,89	0,24	1,36	0,55	0,64	0,41	0,99	0,92		1,58	0,58
L5	0,48	0,86		1,42	0,58	0,64	0,42	0,98	0,9	0,48	1,55	0,58
L6	0,48	0,88	0,24	1,43	0,56	0,64	0,41	1,01	0,92		1,55	0,57
L7	0,48	0,88	0,24	1,39		0,64	0,42	1,02	0,91	0,49	1,59	0,57
L8	0,48	0,86		1,4	0,55		0,42	0,99	0,9	0,48	1,59	0,57
L9	0,48	0,86	0,24	1,37	0,57		0,42	1,02	0,92	0,48	1,59	0,57
L10	0,49		0,24	1,43	0,55	0,64	0,42	0,99	0,91	0,48		0,57
L11		0,86	0,24	1,35	0,55		0,41	1,01	0,92	0,51	1,55	0,57
L12	0,48	0,87		1,41	0,57	0,63	0,42	1,03	0,92	0,49	1,56	0,58
L13		0,85	0,24	1,39	0,56	0,65	0,41	1,01	0,92		1,57	0,57
L14	0,47	0,85	0,24	1,43		0,63	0,42	1,03	0,88	0,51	1,56	0,57
L15	0,47	0,86	0,24	1,41	0,55	0,65	0,42	1,03	0,93	0,49	1,58	0,58

Table B.4 — Benzene content by test method Y

	Sample no. S_i											
Laboratory no. L_i	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12
L1	0,44	0,87	0,24		0,54	0,62	0,43	1,04	0,89		1,5	0,52
L2	0,47	0,82	0,25	1,36	0,54		0,44	1,02	0,81	0,44		
L3	0,46	0,88	0,23	1,34	0,54		0,41	0,96	0,85	0,51		0,54
L4	0,45	0,86		1,38	0,54	0,62	0,44	1,01		0,47	1,55	0,51
L5	0,45	0,81	0,25	1,35		0,63	0,44	1	0,87		1,52	0,56
L6	0,46		0,24	1,33	0,55	0,64	0,42	1	0,84	0,5	1,46	
L7	0,47	0,87	0,25	1,36	0,55		0,44	1,03	0,84	0,41		0,55
L8	0,47	0,82	0,25		0,53	0,61	0,41	1,01		0,44	1,6	0,55
L9	0,46	0,83		1,33	0,56	0,62	0,41	1,02	0,89		1,49	
L10	0,46	0,84	0,24	1,39	0,55	0,63	0,41	1,03		0,5	1,58	0,56
L11	0,46	0,88	0,24	1,41		0,63	0,41	1,01	0,82		1,48	0,53
L12		0,84	0,25	1,43	0,58		0,44	1	0,79	0,46	1,53	0,51
L13	0,47	0,84	0,24	1,35	0,55	0,63		0,97	0,83	0,46	1,55	0,56
L14		0,82	0,23	1,37	0,53	0,61	0,41	0,97		0,46	1,48	0,55
L15	0,45	0,89	0,25	1,41		0,6	0,43	0,98	0,9	0,43		0,54

For testing the amount of variation or resolution of the data, the number of unique values within the data set is counted and evaluated. It is considered to be sufficient to have 119 (74 %) unique values out of a total of 180 values with 20 outliers or missing values for the test method in ASTM D6839:2017, and 130 (87 %) unique values out of a total 180 values with 30 outliers or missing data.

B.4 Rearrange the data set

A summary is presented with statistics per sample, including the average, standard deviation and laboratory number for each sample. To gain insight into the extent of the spread over the laboratories, the relative standard deviation can also be calculated. This information is listed in Table B.5 for method X and in Table B.6 for method Y.

Table B.5 — Rearranged data for benzene for test method X

	Laboratory no. L_i																		
Sample no. S_i	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	L11	L12	L13	L14	L15	**Average** Xi	**Standard deviation** s_{Xi}	**No. of laboratories** L_{Xi}	**Relative standard deviation** s_{Xi} %
S1	0,47	0,49	0,48		0,48	0,48	0,48	0,48	0,48	0,49		0,48		0,47	0,47	0,479	0,007	12	1,4
S2		0,86	0,88	0,89	0,86	0,88	0,88	0,86	0,86		0,86	0,87	0,85	0,85	0,86	0,866	0,013	13	1,5
S3	0,24	0,24	0,24	0,24		0,24	0,24		0,24	0,24	0,24		0,24	0,24	0,24	0,240	0,000	12	0,0
S4	1,36	1,42	1,40	1,36	1,42	1,43	1,39	1,40	1,37	1,43	1,35	1,41	1,39	1,43	1,41	1,398	0,027	15	2,0
S5	0,57	0,56	0,56	0,55	0,58	0,56		0,55	0,57	0,55	0,55	0,57	0,56		0,55	0,560	0,010	13	1,8
S6	0,64	0,63	0,64	0,64	0,64	0,64	0,64			0,64		0,63	0,65	0,63	0,65	0,639	0,007	12	1,0
S7	0,42	0,41	0,41	0,41	0,42	0,41	0,42	0,42	0,42	0,42	0,41	0,42	0,41	0,42	0,42	0,416	0,005	15	1,2
S8	0,99	0,98		0,99	0,98	1,01	1,02	0,99	1,02	0,99	1,01	1,03	1,01	1,03	1,03	1,006	0,019	14	1,9
S9			0,92	0,92	0,90	0,92	0,91	0,90	0,92	0,91	0,92	0,92	0,92	0,88	0,93	0,913	0,013	13	1,4
S10	0,50	0,48	0,50		0,48		0,49	0,48	0,48	0,48	0,51	0,49		0,51	0,49	0,491	0,012	12	2,4
S11	1,61	1,59	1,55	1,58	1,55	1,55	1,59	1,59	1,59		1,55	1,56	1,57	1,56	1,58	1,573	0,020	14	1,3
S12	0,57	0,58	0,57	0,58	0,58	0,57	0,57	0,57	0,57	0,57	0,57	0,58	0,57	0,57	0,58	0,573	0,005	15	0,9

Table B.6 — Rearranged data for benzene for test method Y

Sample no. S_i	Laboratory no. L_i															Average Y_i	Standard deviation s_{Yi}	No. of laboratories L_{Yi}	Relative standard deviation s_{Yi} %
	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	L11	L12	L13	L14	L15				
S1	0,44	0,47	0,46	0,45	0,45	0,46	0,47	0,47	0,46	0,46	0,46		0,47		0,45	0,459	0,010	13	2,1
S2	0,87	0,82	0,88	0,86	0,81		0,87	0,82	0,83	0,84	0,88	0,84	0,84	0,82	0,89	0,848	0,027	14	3,1
S3	0,24	0,25	0,23		0,25	0,24	0,25	0,25		0,24	0,24	0,25	0,24	0,23	0,25	0,243	0,008	13	3,1
S4		1,36	1,34	1,38	1,35	1,33	1,36		1,33	1,39	1,41	1,43	1,35	1,37	1,41	1,370	0,032	13	2,3
S5	0,54	0,54	0,54	0,54		0,55	0,55	0,53	0,56	0,55		0,58	0,55	0,53		0,547	0,014	12	2,5
S6	0,62			0,62	0,63	0,64		0,61	0,62	0,63	0,63		0,63	0,61	0,60	0,622	0,012	11	1,9
S7	0,43	0,44	0,41	0,44	0,44	0,42	0,44	0,41	0,41	0,41	0,41	0,44		0,41	0,43	0,424	0,014	14	3,3
S8	1,04	1,02	0,96	1,01	1,00	1,00	1,03	1,01	1,02	1,03	1,01	1,00	0,97	0,97	0,98	1,003	0,024	15	2,4
S9	0,89	0,81	0,85		0,87	0,84	0,84		0,89		0,82	0,79	0,83		0,90	0,848	0,036	11	4,2
S10		0,44	0,51	0,47		0,50	0,41	0,44		0,50		0,46	0,46	0,46	0,43	0,462	0,032	11	6,8
S11	1,50			1,55	1,52	1,46		1,60	1,49	1,58	1,48	1,53	1,55	1,48		1,522	0,045	11	2,9
S12	0,52		0,54	0,51	0,56		0,55	0,55		0,56	0,53	0,51	0,56	0,55	0,54	0,540	0,019	12	3,4

From a quality point of view, the data integrity can be assessed somewhat by considering the relative standard deviations on this point for both test methods. Outliers should be considered further.

B.5 Performing the basic checks on the data set

B.5.1 The data set is made up of 12 samples and 15 laboratories presenting single results. This satisfies the requirements for PTP data that the results originate from at least 10 samples and 10 laboratories.

B.5.2 The samples in the data set span the benzene content range from 0,19 to 1,57 for test method X and from 0,21 to 1,52 for test method Y. As shown in Table B.7, these ranges are compared with the published application and precision ranges of the test methods (see Table B.3).

Table B.7 — Application and precision ranges for both test methods

Reference for the test method	Test method X	Test method Y
Typical published application range, volume fraction %	0 to 2	0,1 to 5
Published precision range, volume fraction %	0,5 to 1,6	0,11 to 1,5
Range of the data set, volume fraction %	0,24 to 1,57	0,4 to 1,52

This comparison shows that the results of the evaluation are limited to a benzene range of a volume fraction of 0,5 % to 1,5 %.

B.6 Testing the samples for leverage

As part of the additional requirements for PTP data, the leverages of the samples are calculated based on the averages of the methods (see Table B.4). In this case, a leverage is not performed for each sample for both test methods, but the average over the test methods is used to determine the leverage per sample. Details are given in Table B.8.

Table B.8 — Leverage of each sample, averaged over both methods

Sample no. S_i	Sample mean $(X_i + Y_i)/2$	Leverage h_i
S1	0,469	0,12
S2	0,857	0,10
S3	0,242	0,41
S4	1,384	0,26
S5	0,553	0,09
S6	0,630	0,08
S7	0,420	0,15
S8	1,005	0,14
S9	0,881	0,11
S10	0,476	0,12
S11	1,547	0,32
S12	0,557	0,09

For each sample the calculated leverage value, h_i, remains below the critical value of 0,5, so the test has passed.

B.7 Constructing overview tables with summary statistics per sample

To check the data against the additional requirements, the data are examined for the distribution of the lab results per sample. For each sample at least 10 laboratories returned a result and the precision