FINAL DRAFT

INTERNATIONAL STANDARD

ISO/FDIS 4259-5

ISO/TC **28**

Secretariat: **NEN**

Voting begins on:
2023-09-14

Voting terminates on:
2023-11-09

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

ISO/CEN PARALLEL PROCESSING

Reference number
ISO/FDIS 4259-5:2023(E)

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
$\Sigma_{ST,\bar{X}}$, $\Sigma_{ST,\bar{Y}}$	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 … 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 of 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, $x_1 \le x_2 \le x_3 \dots \le x_N$.

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 of 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

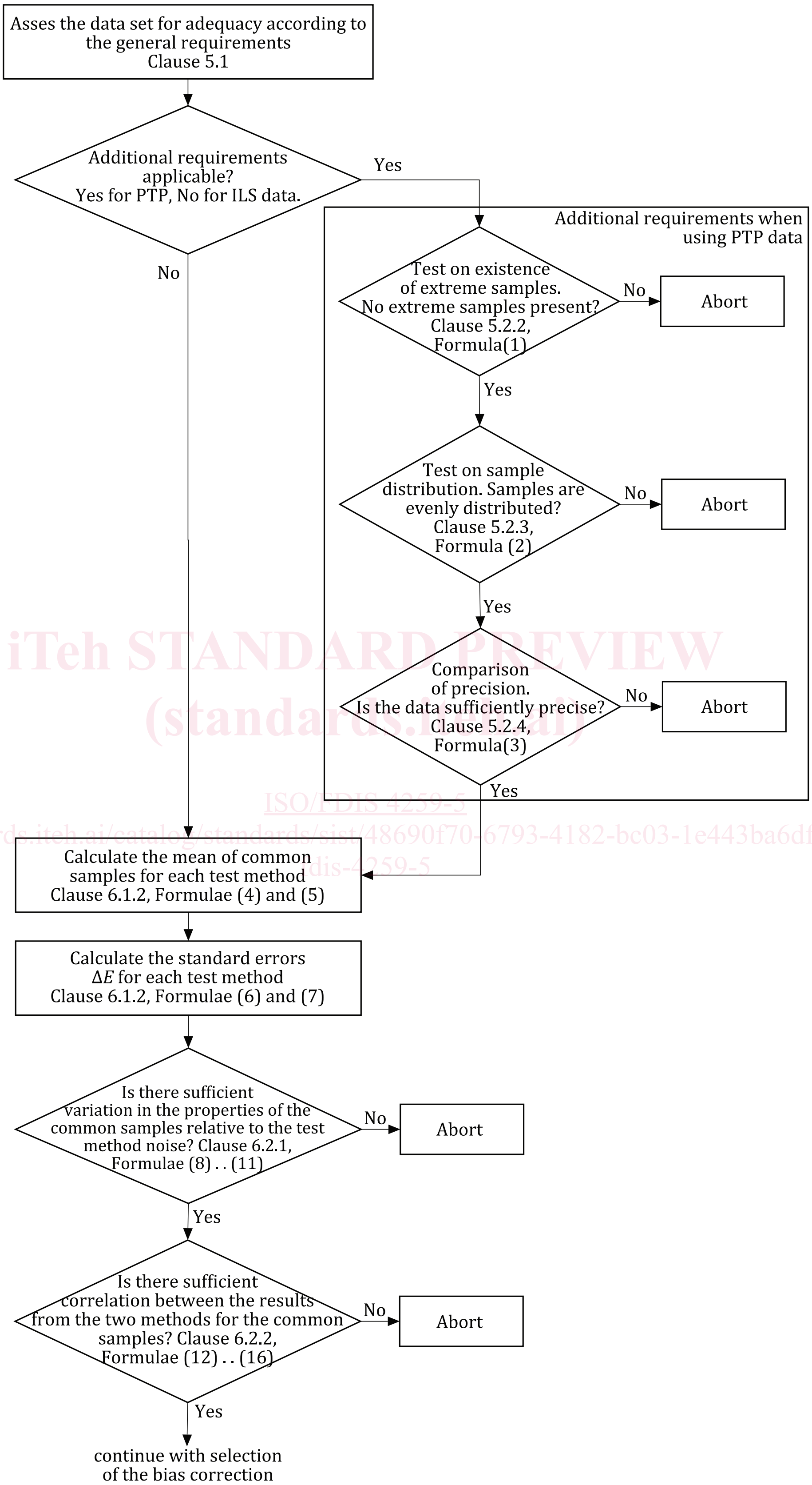

Figure 1 — Flowchart for suitability and applicability of the data