Institute for
Interlaboratory Studies

iis Interlaboratory Studies: Protocol for the Organization, Statistics and Evaluation

Institute for Interlaboratory Studies
Spijkenisse, The Netherlands

Authors: Mr. R.J. Starink BSc., Mr. M. Audier BSc., & Mrs. C.M. Nijssen-Wester BSc.
Report: iis-protocol (version 4.0, October 2024)

SUMMARY OF CHANGES

This protocol replaces the previous version 3.5 of June 2018.

The whole document was changed to the new reporting format with regards to lay out, font type and spelling and the company details were added on the last page.
Furthermore the list of literature articles and methods have been updated and the numbering changed (and also the reference numbers to the literature in the texts)

The following parts of the protocol have been revised:

- Paragraph 1.2, 1.5 and 7.1: update numbers mentioned
- Paragraph 1.4: updated for new version ISO/IEC 17043 and RVA membership ILAC
- Paragraph 2.1: explanation purpose of PTs
- Paragraph 3.2: added text about treatment of PT samples, accreditation requirements homogeneity testing laboratory and minimum amount of participants
- Paragraph 3.3: added text about recipient assistance in transport of samples
- Paragraph 3.4: added link to iis website for data entry portal access
- Paragraph 3.5: Added text about choice of methods
- Paragraph 4.3, 4.4 and 4.5 and 8: updated versions of ISO/IEC 17043 and/or ISO 13528 and/or ISO 5725
- Paragraph 6: added text to reflect current report contents

Removed: references to method evaluation studies, on request iis can provide more information about this separately.
Removed: Appendices – no longer relevant for this document

All revisions are made in blue text.

CONTENTS

1 INTRODUCTION

1.1 THE INSTITUTE FOR INTERLABORATORY STUDIES (iis)

The independent Institute for Interlaboratory Studies (iis) organizes global interlaboratory studies on petroleum products, liquid fuels, petrochemicals and consumer products since 1994.
Studies are usually performed on commercially relevant products and involve testing on full specifications. Besides its annual program, iis organizes tailor made studies on request.

This report provides a comprehensive description of the organization, statistics and evaluation used in iis interlaboratory studies. This includes studies for proficiency testing, for the preparation of reference materials and for method evaluation.

For the most recent information about iis and its activities is referred to the Institute's website www.iisnl.com.

1.2 WORLD-WIDE PROGRAM

iis acts world-wide and participants in its interlaboratory studies can be found all over the world. For the iis proficiency tests for example, more than 1900 laboratories from over 120 countries were actively participating in the last three years.

1.3 CONFIDENTIALITY

iis handles all information supplied by the participating laboratories with great care and strictly confidential. No information is passed to third parties unless prior permission is received. The identity of individual participants is always maintained confidential and is only accessible to authorized iis-personnel.

The Institute is aware of the fact that participants of an interlaboratory study do not (always) wish to enclose their performance to third parties. Therefore, in the iis reports the results, methods and all other information provided by a laboratory is only presented under a lab code number.

1.4 QUALITY

The Institute for Interlaboratory Studies in Spijkenisse, the Netherlands, is accredited (R007) in agreement with ISO/IEC 17043:2010 [1] since January 2000, by the Dutch Accreditation Council (Raad voor Accreditatie). See www.rva.nl for the actual accreditation scope. A new version of ISO/IEC 17043 [2] has been published to which iis will comply before the implementation date of May 2026. The performance of a laboratory that participates in an iis proficiency test will be accepted with confidence by a National Accreditation Body, for our accreditation body (RVA) is a member of the International Laboratory Accreditation Cooperation (ILAC).

The employees are highly qualified and experienced in the design, implementation and reporting of interlaboratory studies. Specialists of iis play leading roles in the field of

proficiency testing, such as in Eurachem committees. All of our staff members are fully qualified and their qualifications are documented in records.

1.5 UNIQUE SET-UP

The proficiency tests program of iis is unique in many aspects:

- Its world-wide set-up: more than 1900 laboratories from over 120 laboratories have been registered and are actively participating.
- Its short turn-around time: normally, the complete time span from sample dispatch up to and including the publication of the final report does not exceed three months.
- Its wide scope: iis aims to use natural matrix materials, which are investigated on complete profiles (analysis of full specification).
- Its advanced parametric and evaluation statistics: the parametric statistics use: normality checks of data, outlier detection routines and calculation of the usual statistical precision parameters like mean, standard deviation and reproducibility.
- Target z-scores for evaluation of performance ‘over time’: z-scores are calculated with the use of a fixed standard deviation taken from the corresponding, internationally accepted test method (e.g. ISO, ASTM, EN, IP, IEC or another accepted standard in the industry).

Based on the analytical results in a proficiency test, each participant receives an indication of its performance. The z-score is used by iis as performance indicator, which gives an indication of the laboratories competence. The performance is evaluated per test, per laboratory and - if requested or desired - per group. Performances are measured with reference to internationally accepted analytical standard test methods (ISO, ASTM, EN, IP, IEC or other accepted industrial standards). Graphical tools are used to facilitate the interpretation of all data per test.

1.6 ANNUAL PT-PROGRAM

iis works with an annual schedule. The contents of its PT-program is discussed and decided upon during the advisory board meetings.
The criteria for priority selection of products and tests for each year’s program are chosen on the basis of an evaluation of commercial risks (claims, near-misses and complaints), findings in previous programs, requests from participants and technical developments in the laboratory field.

Besides its annual PT-program, other interlaboratory studies are organized. These studies are initiated by the Institute itself or are tailor made and organized on request.

The actual PT-program and all (other) relevant information will be sent to interested laboratories on request. It can also be found on the Institute’s website www.iisnl.com.

2 TYPES OF INTERLABORATORY STUDIES; BRIEF OVERVIEW

2.1 INTERLABORATORY STUDIES FOR PROFICIENCY TESTING

A proficiency test (PT) is a special type of interlaboratory comparison to determine the performance of individual laboratories for one or more specific tests and to monitor laboratories' continuing performance. Participation in PT-schemes provides laboratories with an objective means of assessing and demonstrating the reliability of the data they are producing. So, proficiency tests allow laboratories to check their normal routine performance and to compare their results with those of other independent laboratories.

Participants of the world-wide laboratory PT-program of iis receive valuable information about the technical capability of its laboratory. This provides the lab (personnel, QA-manager and the management) and also its (potential) clients and accreditation bodies a good indication of its analytical competence. The responsible management can use the results and conclusions to diagnose and cure causes of deviating results if present. The program can be incorporated in the quality assurance systems of the laboratory to gain maximum profit. The performance of a laboratory participating in an iis proficiency test will be accepted with confidence by a National Accreditation Body.

Using strict protocols, the participating laboratories all analyse the same samples in the same period. Each laboratory uses its own routine procedures, generally validated standard methods, which are used in normal day-to-day practice. The results are collected by iis and statistically processed. The proficiency of each laboratory is expressed in a numerical parameter (z-score) and tested against the corresponding, internationally accepted test method, e.g. ISO, ASTM, EN, IP, IEC or another accepted industry standard.

The data which is gathered is intended to assess to performance. Some tests methods are very good, have fabulously narrow precision, and others are frankly very incomplete and/or poorly designed, we do not control that. What one is seeking is not some magic “ better than anyone could possibly imagine” data set, but a real data set, showing how the participating laboratories perform in real situations. If there are major problems, these will be experienced by their clients as well. This is an early warning system for the laboratories to use to try and avoid misunderstands, claims and complaints.

2.2 INTERLABORATORY STUDIES FOR PREPARATION OF REFERENCE MATERIALS

Proficiency tests are very useful as (independent) quality control tool, but the usual frequency of PTs seldom exceeds twice a year. Therefore, the day-to-day quality in a laboratory is measured in a much higher frequency by analyses of reference materials. With the use of reference materials the calibration of instruments can be verified even daily. Regretfully, in practice there is a shortage of suitable reference materials.

Considering above, the Institute for Interlaboratory Studies started preparation of Reference Materials in 1996. The Reference Materials are certified on the basis of the results of one or more interlaboratory studies. Preferably, the certification of values and uncertainties is combined with a proficiency test.

The reference materials are all multipurpose and available in handy quantities. They can be ordered from iis directly. Each reference material is accompanied with a certificate containing the certified reference values. Furthermore, certification reports are available. For all reference materials, actual prices and availability see the iis' website www.iisnl.com.

Also PT samples that are left over and retained after the PT can be purchased by the participating laboratories. Of course only when the order is made during the validity period of the samples.

2.3 INTERLABORATORY STUDIES FOR METHOD EVALUATION

Ideally, an analysis certificate of a commodity, issued by a laboratory should be similar to that issued by other laboratories that have analyzed the same commodity. Nonetheless minor differences may exist between the certificates, which are caused by the measuring uncertainties of the analytical methods. The measuring uncertainty of an analysis method is determined during its validation process. Many laboratories usually co-operate in the validation of a method by participating in an interlaboratory study. Once a method has been validated it can be expected that a good laboratory applying the method will find results within the measuring uncertainty.

A validated analysis test method (or standard) is not always available or its validity has been determined only for a limited number of products, matrices or concentration ranges. In general the 'official' test methods have not been validated for use with all kinds of products, at all levels of measurement. Matrix influences may have a negative effect on the reliability of the analysis method, as may differences in concentrations or measuring levels do. It is important that the Institute for Interlaboratory Studies generates this information and advises the trade community about unexpected risk implications.
Sometimes 'official' analytical methods are not available at all, are technically outdated or for other reasons not applicable, such as incompatible with the product matrix, time consuming, yielding too high uncertainties, requiring too much sample. Analytical methods developed 'in-house' fill this gap in methodology. The Institute for Interlaboratory Studies does organize interlaboratory studies for the validation of 'in-house' developed methods on request.

3 ORGANIZATION

The interlaboratory studies of iis are all based on the same standardized protocol. Slight modifications can be made for specific studies, based on the requirements or suggestions of e.g. the participants. Various international technical committees with experts and with representatives from participating laboratories support the annual PT-program.

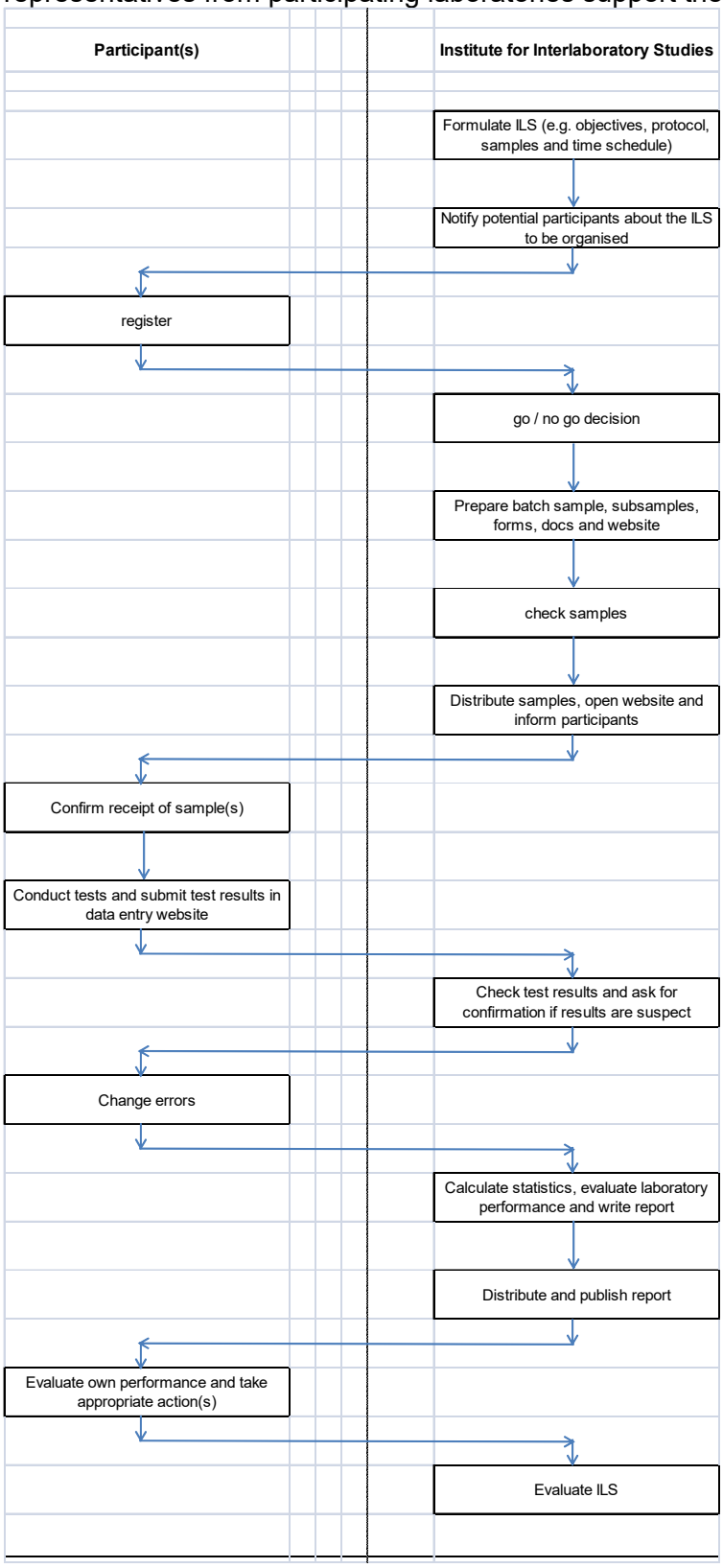

Figure 1: General procedure for the organization of iis interlaboratory studies

The outline of the iis procedure for the organization are described by the following steps:

1. The objective of the interlaboratory study to be organized is formulated, the general protocol is chosen and the samples are defined.
2. The full time schedule is made.
3. All potential participants and other relevant laboratories are notified. They receive at least a summary of the planned interlaboratory study and also a registration form.
4. All registrations are confirmed by email upon receipt.
5. iis decides whether or not the planned interlaboratory study is organized.
6. The sample batch is prepared according to the protocol of sample preparation and checked for its fit for purpose.
7. The material is ensured to be stable during the proficiency test, based on critical parameters.
8. Sub samples are prepared and the units (e.g. bottles or bags) are labelled.
9. The homogeneity of the sub samples is checked on critical parameters.
10. All necessary samples are packed and distributed to the participants.
11. The participants report the sample receipt.
12. The participants analyse the samples.
13. The results are submitted by the participants.
14. After the deadline the results are checked for obvious errors and in case of erroneous results the participants are asked for confirmation or correction.
15. The dataset is analysed on normality and outliers are detected using the statistical protocol.
16. The statistical parameters are calculated, using the relevant protocol.
17. The performance on each test is evaluated as well as the performance per laboratory and the performance of the total group, using the evaluation protocol.
18. The anonymized final report is sent to the participants.

The details of this procedure may vary upon the type of interlaboratory study.

3.1 PROTOCOL

The iis interlaboratory studies are conducted according to a well defined protocol. This protocol is based on the guidelines as described ISO 5725 [3], J. AOAC [4] and ISO 13528 [5] and ISO/IEC 17043 requirements [1,2].

It is generally acknowledged that the number of participating laboratories and the number of test results are interdependent. This implies that the fewer samples are analyzed, the more replicates or the more participants are needed to enable appropriate evaluation of random errors. Therefore, for the large scale proficiency tests and for the small scale method validation tests, different protocols are used.

For **proficiency testing**, only one sample sent to the participants can be sufficient, because the number of participants in the proficiency tests is large and enough data can be collected for meaningful statistical calculations. In iis proficiency tests however, often more than one unit of a sample is sent to the participants, because the number of analyses in one interlaboratory study is normally quite large and otherwise not enough sample would be present to perform all analyses.

In order to get a good idea about a laboratory's day-to-day performance, the participant should treat the samples as if they were routine samples. So, the laboratory should use the analysis methods that this laboratory would use in normal daily practice. No special attention should be paid to the samples and no extra work or testing should be carried out.

For the **reference material certification** studies, two or more samples are sent to the participants. This is necessary to verify the quality of the results produced by the participants in the interlaboratory study.

For the **method evaluating** interlaboratory studies, two or more samples are sent to the participants. This is required as the number of participants in these interlaboratory studies is usually much smaller than in the proficiency tests. The participants have to follow the prescribed analysis method under evaluation in detail.

3.2 SAMPLES

iis aims to use natural matrix materials as samples in its interlaboratory studies. This guarantees a close resemblance between the test items in the interlaboratory study and the samples the participating laboratories normally analyze. The samples are correlation samples.

One needs to keep in mind that, in the real, commercial world, the samples entering any of our participating labs are subject to the same laws of physics as the samples we send out. In fact, PT samples are intended NOT to be treated as something extraordinary or special, but as any other commercial samples, as the participating labs are looking to assess lab performance as their clients would experience it. They should be completely routine samples, passed through the lab in the normal way. If it is treated as some sort of " special" sample (and we know it often is) the simulation is false. We experience that this is often missed by laboratories. A good performance is considered more important than controlling the real way the sample is analyzed in daily situations.

The entire batch is thoroughly homogenized (and if necessary stabilized) and tested for suitability in the interlaboratory study. Sometimes, suitable matrix samples cannot be found and additives are added to a natural matrix or a complete synthetic sample is prepared.

The batch is divided in subsamples, which will be sent to the participating laboratories. Prior to distribution the homogeneity of the subsamples is tested by checking one or more critical and sensitive key parameters by a laboratory that has performed the tests in accordance with for ISO/IEC 17043 relevant requirements of ISO/IEC 17025 on a sufficient amount of stratified randomly selected subsamples.

A sufficient number of samples is prepared, estimated from the participation in the previous round, plus samples needed for homogeneity, plus spare samples and samples to cover 10% increase in participation. The samples that are not used during the PT, can be purchased by laboratories that are actively participating in one or more iis PT schemes after the finalization of the PT.

The minimum number of participants needed to meet the objectives of the statistical design is determined based on the considerations for small numbers of participants as described in ISO 13528, paragraph 5.4.2. Considering the statistical analysis methods used in iis PTs and the use of a consensus value instead of an assigned value, the minimum number of participants to approximate a population is eight.

Note for petroleum, liquid fuels and petrochemical laboratories:
iis is purchasing large quantities of straight run product cuts at the distillation unit at one time. Preferably this stable and fresh material is used as a basis for interlaboratory study material. As certain product grades can not be obtained in this way (for instance RFG and other gasolines), in such cases day-to-day samples are combined to produce sufficient quantities of material. Sometimes additives are added to obtain the desired physical or chemical properties, like cold properties for gas oils, desired levels of sulfur, detectable quantities of trace impurities, etcetera.

In the case of a method validation study, more than one sample is prepared with the analytes at different levels. Standard addition may be used to create the various concentration levels.

3.3 SAMPLE DISTRIBUTION

In case of special requirements or dangerous goods (low flash point, corrosive, toxic) the sample distribution is being performed by a specialized party (SGS DGCC, Spijkenisse, The Netherlands). This highly qualified shipping department has been awarded with the E-status by the Dutch Authority of Civil Aviation. Packaging is done strictly according to UN rules and dangerous goods declarations comply with the IATA rules.
Small sized, non dangerous goods samples are distributed by courier, such as DHL.

When necessary or on request, additional documents are enclosed to the sample, e.g. (material) safety data sheet ((M)SDS), certificate of origin, (pro forma) invoice, certificate of quality, etc.

One has to understand that sample transport is a matter of pushing (by the sender) and pulling (by the recipient). Without the assistance of the recipient it is impossible to deliver a sample, no matter how hard this is tried by the sender. Therefore the (pro-active) assistance of the receiving company is of utmost importance in order to get the sample at the laboratory within an acceptable time frame. For example import licenses need to be arranged prior to the sample dispatch date.

The participants are requested to confirm the sample receipt and to report in case the samples are damaged or otherwise not fit for testing. In case a sample is lost or arrives damaged, a new sample will be sent.

3.4 ANALYSES

In the proficiency tests the participants are urged to use the methods that they use in normal circumstances. In order to get a good idea about a laboratory's day-to-day performance the participant has to treat the samples as if they were routine samples. However, the participants are requested to report as much significant figures as possible. And they are also requested not to report 'less than' test results, which are above the detection limit, because such test results cannot be used for meaningful statistical calculations.

In a method evaluating interlaboratory study the participants have to follow the prescribed analysis method under evaluation in all details.

To get comparable test results, a detailed report form and a letter of instructions are prepared. On the report form the reporting units are given as well as the reference test methods that will be used during the evaluation. The detailed report form and the letter of instructions are made available on the data entry portal, of which a link can be found on the iis website www.iisnl.com.
The participating laboratories are also requested to confirm the sample receipt on this data entry portal. The letter of instructions can also be downloaded from the iis website.

3.5 METHOD INFORMATION

In most cases, iis asks for information about the test method used by the participants in the interlaboratory studies. The descriptions (or summaries) are included in the report. In case of standard test methods (e.g. ISO, ASTM, EN, IP, IEC, …) the method number is sufficient, in other cases the key elements of the method may be asked to be reported.

Participants in iis PT schemes are requested to perform the test methods that they would routinely do. If a method is suitable to measure a particular parameter in a given matrix, test results according to this method will be allowed by iis for this parameter and matrix. In the many years that iis has organized PTs, it has been found that the differences between methods that measure the same parameter are small in general. If a significant bias is known or observed data between methods, the group data will be investigated to determine whether one method or the methods should be evaluated separately or in the future be requested per method (provided there is a sufficient number of participants to make conclusions based on the group data).

3.6 TIME SCHEDULE

During five weeks after sample distribution, the results of the individual laboratories are collected on the data entry portal.

Directly after the deadline for reporting results, the received results of the participating laboratories are checked for obvious errors. In case of erroneous results, the respective participant is notified immediately so it can take all necessary corrective actions.

About six weeks after the closure of the PT, the PT report is published and a copy sent to the participants by email.

4 STATISTICAL PROCESSING OF THE TEST RESULTS

4.1 DETECTION OF OBVIOUS ERRORS

The test results of the participating laboratories are checked for obvious errors, like unit errors or typing errors. A robust outlier test, Huber Elimination Rule, is used for this purpose. In case of clear erroneous results, the respective participant is notified immediately so it can take all necessary corrective actions. The notification of deviating results is done shortly after the closing date for reporting the test results, normally within 2 days after closure of the PT. The revised test results will replace the erroneous ones. However, in the PT report the originally reported test result is mentioned under ‘remarks’, next to the revised test result. Prior to calculation of the statistical parameters, a check is done on the validity of the reported test results to be used in the calculations: the distribution of the data is checked as well as the presence of outlying test results.

4.2 CHECK ON NORMAL DISTRIBUTION OF THE TEST RESULTS

Many statistical procedures are only applicable to random samples from populations with a Gaussian distribution. Even the outcome of the simplest statistical parameter 'mean', which should be a good estimate of the true value, may depend on the type of distribution of the data. For the assignment of a property value (the consensus value ‘mean’), the assumption of a Gaussian distribution function is less critical than for outlier testing. However, as the descriptive statistics used is based on a normal distribution, it is checked whether the distribution of the data agrees reasonable with the normal distribution prior to use of the data. There are more than 30 tests of normality available in the literature [6]. The tests of normality can be sub-divided into three categories which are graphical methods, descriptive statistics and theory-driven methods. Skewness (3^{rd} moment) and (excess) kurtosis (4^{th} moment) coefficients [7] are categorized as descriptive statistics. Theory-driven methods include the normality tests such as Shapiro-Wilk [8], Kolmogorov-Smirnov [9], Lilliefors [10,11] and Anderson-Darling [12], the last two being improved versions of the Kolmogorov-Smirnov test. Each test of normality has its (dis)advantages. For example, the Shapiro-Wilk test is a very powerful test [13], but only up to 50 values and it works very well if every value is unique, and it performs less when several values are identical and the power is low for small sample size. The Lilliefors test always outperforms the Kolmogorov-Smirnov test. Some of the tests on normality can only be applied under a certain condition, i.e. a minimum sample size. Moreover, different tests may produce different results i.e. some tests reject while others fail to reject the assumption of normality. Therefore, the investigation on the data distribution is given thorough attention as part of the PT evaluation.
The normality of the distribution of the data per determination is checked by means of the Lilliefors-test, a variant of the Kolmogorov-Smirnov test and by the calculation of skewness and excess kurtosis [14]. Evaluation of the three normality indicators in combination with the visual evaluation of the graphic Kernel density plot [15,16], leads to judgement of the normality being either ‘unknown’ (for <9 values), ‘OK’, ‘suspect’ or ‘not OK’.

4.3 DETECTION AND REMOVAL OF ERRONEOUS AND STATISTICALLY DEVIATING RESULTS

The presence of statistical outliers will affect statistical parameters like mean and standard deviation. Therefore the detection and treatment of outliers is given thorough attention before the actual PT evaluation.

In the literature no consensus is found whether outliers should be rejected or not. The Analytical Methods Committee [17] recommends that outliers must be retained. Reason for this is that an occasional overestimation of the variability is safer than a consistent underestimation of the variability. This is considered to happen frequently. In this vision only transcription errors may be corrected. Davies [18] criticizes the use of outlier tests and proposes a different evaluation procedure. Theoretically, it is possible that the majority of results is incorrect, whilst the 'aberrant' result is the only correct value. ISO/IEC 17043:2023 [2] states that the influence of outliers shall be minimized by using an appropriate statistical approach. It suggests removing outliers prior to calculation and refers to ISO 13528:2022 [5]. In ASTM E178 [19], a procedure is given for handling data with possible outliers. If the physical reason for the outlier is known, the observation should be corrected or rejected. If the physical reason is unknown a statistical test should be used to correct or to reject the observation or to utilize statistical calculations on restricted observations. For the detection of outliers various techniques can be used, such as Dixon Test [20], Grubbs Test [3], Rosner's generalized ESD test [21] and/or Tietjen-Moore Test [19].

Most procedures for detection of outliers will only work properly if the data have a normal distribution and if enough data (numerical test results) are present. Rejection of the outlying data will reduce the number of data for the necessary calculations and therefore is only allowed if the total number of data is sufficiently large. For iis proficiency tests both conditions are usually met.
In the iis procedure for proficiency tests, outliers are detected prior to calculation of the mean, standard deviation and reproducibility. For small data sets (<21 test results), Dixon and/or Grubbs outlier tests are used. For larger data sets (> 20 test results) Rosner's Generalized ESD outlier test is used. The decision whether or not to remove deviating results (e.g. outliers) is not made on statistical grounds solely. Other information (e.g. reported test methods, consistency analysis, max. percentage of outliers) is also used in order to make a sound decision.
The above procedure provides a fair basis of comparison between the reproducibilities found in other interlaboratory studies (e.g. from ISO, ASTM, EN, IP, IEC,…) and those found in the iis studies.

iis certifies **reference materials** on the basis of the results of one or more interlaboratory studies. The procedure for the detection and removal of erroneous and statistically deviating results is similar to the procedure applied in proficiency tests. All data is screened for outliers. Deviating and otherwise suspect test results are removed prior to calculation of the reference values.

4.4 CALCULATION OF THE SUMMARY PARAMETERS

The Institute of Interlaboratory Studies decided to use the consensus value of the reported test results, based on all reported test results for a particular analyte without outliers or other suspect data as explained in paragraph 4.3 as assigned value for the following reasons:

1. The PT samples used by iis are not formulated and therefore the assigned value cannot be calculated as per 7.3 of ISO 13528:2022 [5].
2. CRM are not used as PT samples by iis. The use of CRM in a PT is considered to be misuse of valuable materials. Also the scope of the available CRM is limited and insufficient to run all iis PTs.
3. The assigned values are not based on the results from one or more (expert) laboratories. For many of the tests evaluated in the iis PT, no expert laboratories are available in Western Europe. The use of expert laboratories to determine the assigned values would increase the costs significantly, which would result in a decrease of the number of participants. Also, the quality of the assigned values would not be sufficiently guaranteed.

In iis **proficiency tests** the *normal statistical* parameters are calculated, after rejection of non-valid or suspect results and/or the statistically deviating results:

* The mean $\bar{x}$, as best estimate of the true value μ:

$$\bar{x} = \frac{\sum_i x_i}{n}$$

* The standard deviation s_R, as measure of the variation σ:

$$s_R = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{(n-1)}}$$

* The reproducibility R, as measure of the interlaboratory variation [2]:

$$R = 2 \times \sqrt{2} \times s_R = 2.8 \times s_R$$

The statistics for certification of **reference material** are very much the same as for proficiency tests. Deviating results (e.g. outliers) are detected and removed. In case data distribution is normal, the traditional mean and standard deviation (see above) are calculated. In case the data distribution is not normal, robust statistics (see beneath) are used. The uncertainties of the certified values are calculated acc. to ISO Guide 35:1989 [22]:

$$confidence\ interval\ at\ 97{,}5\% = \mu \pm \frac{t \times s}{\sqrt{n}}$$

where: μ = estimate of the 'true value'
t = 0.975 fraction of Student distribution with (n-1) degrees of freedom
s = standard deviation
n = number of observations

For the certification of **reference materials** often *robust statistic* is used instead of traditional statistics. In the case of robust statistics [23,24] a normal distribution of the data is not required and no information is lost due to data reduction as the outlying data are not rejected. Furthermore robust statistics is insensitive to gross errors and will usually produce sensible values even in the presence of a fair proportion of suspicious results. Hence, robust statistics is used for the statistical calculations of certified values in the case of an anormal distribution and in the case of the relatively small method validation interlaboratory studies.

The robust estimate of the true value μ of an analyte is calculated as the so-called 'Tukey biweight mean': as best robust estimate of μ. The median is taken as estimate for the mean and consecutively the outlying data are replaced by so-called 'pseudo-values'. In this iterative process the 'biweight mean' T_{bi} is calculated [25]:

$$T_{bi} = \frac{(x_i - T_{bi})}{c \times s_{bi}}$$

* The robust standard deviation $s_{R(DoD)}$ [26], as measure for the variation, is also calculated without prior removing of stragglers and outliers. The calculation is based on all absolute differences, hence the name DoD (deviation of differences):

$$s_{R(DoD)} = Y_{\left(\left[q_{s_R} \times n\left(n-\frac{1}{2}\right)\right]+1\right)}$$

* The reproducibility *R*, as measure of the interlaboratory variation [3b]:

$$R = 2 \times \sqrt{2} \times s_{R(DoD)} = 2.8 \times s_{R(DoD)}$$

4.5 UNCERTAINTY

For each assigned value the uncertainty u_X is determined in accordance with ISO 13528:2022 [5]:

$$u_x = 1.25 \times \frac{s^*}{\sqrt{p}}$$

where: p = number of observations
s* = robust standard deviation of the observations

Subsequently the calculated uncertainty u_X is evaluated against the respective requirement based on the target reproducibility in accordance with ISO 13528. When the uncertainty passes the evaluation, no remarks are made in the report. However, when the uncertainty fails the evaluation it is mentioned in the report and it will have consequences for the evaluation of the test results.

5 PERFORMANCE EVALUATION

5.1 OBJECTIVES OF EVALUATION

A laboratory that participates in a proficiency test will primarily be interested in the accuracy of the test results that it reported. The evaluation of the accuracy is in principle done towards an external standard (Reference Test Method) when available. Each laboratory receives a numerical indication (z-score, see par. 5.3) in principle for each numerical reported test result.

In iis proficiency tests z-scores are calculated with the use of a fixed standard deviation taken from the corresponding, internationally accepted test method (e.g. ISO, ASTM, EN, IP, IEC or another accepted Standardized Test Method in the industry). This allows a straight forward and easy evaluation of performance 'over time' [27].

In the proficiency tests of iis the obtained accuracy of the laboratories is compared with the imposed accuracy target as defined by the corresponding, internationally accepted test method, e.g. ISO, ASTM, EN, IP, IEC or another accepted standard in the industry. This parameter is essential in reviewing the performance of the group in relation to accepted standards in the industry.

5.2 PERFORMANCE MEASURED IN NUMERICAL PARAMETERS

Simple performance indicators will provide the laboratory management a quick tool to identify problem areas. Various types of evaluations have been implemented in the PT-program.

- Indicator **per test** per laboratory: the z-score is a measure for the bias of the laboratory test result.
- Indicator **per test** for the group of participating laboratories: the calculated reproducibility is a measure to compare the proficiency of a group of participating laboratories with the official analytical standards.

In addition to above numerical performance parameters, graphic representations of the test results are other simple tools to evaluate the test results (see paragraph 5.5).
And on each of the data entry websites (for links see iis website www.iisnl.com) a graphical tool is available that will give a presentation of all z-scores for a selected test or group of tests during a certain time period. All underlying data can be downloaded as Excel document for off-line calculations.

5.3 INDIVIDUAL TEST RESULTS: THE $z_{(TARGET)}$-SCORE

The international accepted z-score is used as an indication of the performance of a participant (see par. 5.2). This most common indicator compares the bias with a standard error. The bias is calculated as the difference between the reported test result (x_i) of laboratory i and the assigned value ($\bar{x}$). This difference is divided by a standard deviation, thus resulting in a normalized z-score.
In the calculation of the z(target)-score, for the standard error, literature requirements are taken, e.g. calculated from the reproducibilities of ISO, ASTM, EN, IP, IEC….

For each parameter the $z_{(target)}$-score of laboratory i is calculated as:

$$z_i = \frac{(x_i - \overline{x})}{\sigma}$$

where: x_i = the test **result** of laboratory i for that specific parameter.
$\bar{x}$ = the **assigned value**, an estimate for the 'true value'. iis aims to use in its proficiency tests real samples. This guarantees a close resemblance between the PT-test items and the samples the participating laboratories normally analyze. The items do not have a known composition (e.g. concentrations or amounts). The mean of all valid laboratory results is used as the assigned value.
σ = the **target standard deviation** of the reproducibility. This value is derived - if possible - from the corresponding, internationally accepted test method, e.g. ISO, ASTM, EN, IP, IEC or another accepted standard in the industry.

This z-score calculation does result in a simple, straight forward comparison of a laboratory test results with the reproducibility stated in the corresponding international accepted test method. It indicates how many times the standard deviation the reported result deviates from the 'true value'.

The z-score is a convenient performance indicator. With normally distributed test results, the z-scores can easily be interpreted as follows:

$\|z\| < 1$	"Good": will occur in about 68% of all cases
$1 < \|z\| < 2$	"Satisfactory": will occur in about 27% of all cases
$2 < \|z\| < 3$	"Questionable": but will occur in about 5% of all cases
$\|z\| > 3$	"Unsatisfactory": will only occur in about 0.3% of all cases

The z-score provides each laboratory (personnel, QA-manager and the management) and also its (potential) clients and accreditation bodies a good indication of its analytical performance.

Based on the z-score performance in a PT, the participating laboratories may receive a Certificate of Excellence or a Certificate of Participation, when ordered before the start of the PT.
When all z-scores in a PT have a value less than 3 and larger than -3, the laboratory is entitled to receive a Certificate of Excellence. In all other cases the laboratory will receive a Certificate of Participation.

However, in some cases the z-scores may not give a proper presentation of the laboratory's performance. This is the case when the laboratory did not use the reference method, but an alternative test method that may be applicable, but has a very different reproducibility than the reference test method. When the reproducibility of the used test method is higher than the reference test method (e.g. 1.5 for ASTM D1298 instead of 0.5 for ASTM D4052, average of two determinations on lube oil, distillate or base stock), the calculated corresponding z-scores will be too high (e.g. 3 times as in the density example).

In such cases the participating laboratory should recalculate its z-score(s) in accordance with the calculation of paragraph 5.3 to get the correct impression of its performance [28].

5.4 GROUP PERFORMANCE: REPRODUCIBILITY TESTING

The reproducibilities obtained in the proficiency testing studies of iis are compared - if possible - with those defined by the official standards. These officially recognized test methods have been validated and values for the reproducibilities have been established.

Deviating reproducibilities in a PT may be due to a number of laboratories that produce deviating results, whereas the majority of the laboratories produce acceptable results. This situation can be improved by corrective actions in the laboratories concerned.
However, it may also be the case that the variance in the group of laboratories is too high, without laboratories scoring extreme results within the group. This situation is more difficult to solve.
It may for instance also indicate that a certain test standard has not been validated properly for a specific type of product.

5.5 GRAPHIC EVALUATION TOOLS

The graphical presentation of the results used in iis reports depends on the type of interlaboratory study.

The **proficiency test** reports can have different types of graphs. The results of a single sample are presented in a Gauss plot. For the results of two samples, a Youden plot is made. To visualize the distribution of the reported results a Kernel Density plot usually is prepared.

iis **Reference Materials** are certified on the basis of the results of one or more interlaboratory studies. In the certification report a reference is made to the corresponding PT report and no additional graphs are included.

<u>One sample or Gauss plot</u>
In order to visualize the data against the required reproducibilities, Gauss plots using the sorted data for one determination, are made (see examples in appendices 1 and 2).

On the Y-axis the test results are plotted. The corresponding laboratory numbers are on the X-axis. The valid results of the participants are presented by triangles; outliers and other data, which were excluded from the calculations, are presented by crosses. The consensus value is presented by a continuous line. Four striped lines, parallel to the consensus value line, show the +3s, +2s, -2s and -3s target reproducibility limits of the selected standard test method (e.g. ISO, ASTM, EN, IP, IEC).

<u>Kernel Density plot</u>
In order to visualize the distribution of the data a Kernel Density plot [15,16] is used. This is a statistical calculation method for producing a smooth density approximation to a set of data that avoids some problems associated with histograms. A normal Gauss curve is projected over the Kernel Density Graph for reference. The advantage over the non-graphic Kolmogorov-Smirnov test for the determination of the distribution is apparent.

6 REPORT CONTENTS

The proficiency test reports of iis have a standardized format. The following paragraphs are included in principle:

Paragraph	Title	Contents
1	**Introduction**	The proficiency test is summarized.
2	**Set-up**	
2.1	Quality system	The accreditation status is explained
2.2	Protocol	A reference is made to this protocol. Deviations from the protocol are mentioned.
2.3	Confidentiality statement	A confidentiality statement is given
2.4	Samples	A description of the sample preparation, the homogeneity check and its results are presented.
2.5 (Petro)	Stability	A remark is made to the stability study performed.
2.5 (CRS) 2.6 (Petro)	Analyses	A summary is given of the analyses that may be performed by the participants on the PT samples.
3	**Results**	
3.1	Statistics	A summary of the relevant statistics in this protocol is given.
3.2	Graphics	A summary of the relevant graphics in this protocol is given.
3.3	z-scores	The procedure to calculate the z-scores is explained.
4	**Evaluation**	
4.1	Per test OR per component	The test results are discussed one after one and a summary of the main conclusions per test is given. Problems encountered in the analyses are mentioned and - where possible - suggestions for quality improvement are formulated.
4.2	For the group of laboratories	For each test a comparison is made between the results of the group of participants and the requirements given by the relevant standard (e.g. ISO, ASTM, EN, IP, IEC).
4.3	Comparison with previous PTs	The proficiency test and the participants' results are compared with the previous rounds of the PT. When no previous PT is available, this paragraph is deleted or replaced by 'Overview of the PT'.
Appendix 1	Data, statistical results and graphics	The test results and the analytical methods reported by the participants are tabulated. Also, the z-scores calculated by iis for each participant are tabulated. The calculated summary (e.g. mean, standard deviation, reproducibility) is given. Also, the relevant requirements such as reproducibility stated in the appropriate standard (e.g. ISO, ASTM, EN, IP, IEC) are mentioned.
Appendix 2	List of Participants	List of the number of participants per country (no details)
Appendix 3	Abbreviations and literature	All abbreviations used in the report are explained. A list of relevant literature is given.
Last page	Contact details	Contact details of iis are listed, like address, telephone number, email address and website.

7 ANNUAL PROGRAM AND COSTS

7.1 ANNUAL PROGRAM

About 100 interlaboratory studies (mostly proficiency tests) are organized each year.

In iis proficiency tests, about 1900 laboratories from about 120 countries are participating.

The actual PT-program and all (other) relevant information are sent on request to potential participants (analytical laboratories) and other interesting parties (e.g. accreditation bodies). This and additional information can also be found on the Institute's website at www.iisnl.com.

7.3 COSTS INVOLVED

Participation in the iis schemes is open for all laboratories. However, participation is not free of costs. Per round a participation fee is charged, independent on the number of tests performed.
Costs for sample dispatch are dependent on the sample type and the country where to it has to be sent and are therefore not included. These costs for sample package and dispatch will be charged separately. Additional costs may be applicable, e.g. for a Certificate of Origin. A special costs overview is available. This overview is sent to each new customer.

8 LITERATURE REFERENCES

1. ISO/IEC 17043:2010 – General requirements for proficiency testing
2. ISO/IEC 17043:2023 – General requirements for the competence of PT providers
3. ISO 5725 parts 1-6
   a. ISO 5725-1:2023 General principles and definitions
   b. ISO 5725-2:2019 Basic method for the determination of repeatability and reproducibility
   c. ISO 5725-3:2023 Intermediate precision and altern. designs for collaborative studies
   d. ISO 5725-4:2020 Basic methods for the determination of the trueness
   e. ISO 5725-5:1998 Alternative methods for the determination of the precision
   f. ISO 5725-6:1994 Use in practice of accuracy values
4. M. Thompson and R. Wood, International Harmonized Protocol for Proficiency Testing of (Chemical) Analytical Laboratories, J. AOAC Int, 76, 926, (1993), also ISO/REMCO N280 (1993) and also Pure & Appl. Chem., 65, 2123, (1993)
5. ISO 13528:2022 - Statistical methods for use in proficiency testing by interlaboratory comparison
6. J.M. Dufour, A. Farhat, L. Gardiol, L. Khalaf, Simulation-based Finite Sample Normality Tests in Linear Regresssions, Econometrics Journal, Vol. 1, p.154, (1998)
7. R.B. D'Agostino., 1986, Goodness-of-fit techniques, D'Agostino & Stephens eds., Marcel Dekker, New York, p. 367
8. S.S. Shapiro, M.B. Wilk, An analysis of variance test for normality (complete samples), Biometrics, 52, p. 591 (1965)
9. W.W. Daniel, Biostatistics: a foundation for analysis in the health sciences, p. 603, 4th Ed. John Wiley & Sons, Inc. New York (1987)
10. W.J. Conover, Practical Nonparametric Statistics, Wiley, NY, 302, (1971)
11. W. Lilliefors, On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown, Journal of the American Statistical Association Vol 62, 318, p. 399 (1967)
12. Anderson, T.W. and Darling, D.A., A Test of Goodness-of-Fit. Journal of the American Statistical Association Vol 49: p. 765, (1954)
13. N.M. Razali, Y.B. Wah, Power comparisons of Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors and Anderson-Darling tests, Journal of Statistical Modeling and Analytics, Vol.2, 1, p.21, (2011
14. R.B. D'Agostino, A. Belanger, R.B. D'Agostino Jr., A suggestion for using powerful and informative tests of normality, The American Statistician, Vol. 44, 4, p.316 (1990)
15. Analytical Methods Committee Technical brief No 4, Representing data distributions with kernel density estimates, January 2001, revised 2006
16. P.J. Lowthian and M. Thompson, The Royal Society of Chemistry, Bump-hunting for the proficiency tester – searching for multimodality, Analyst 2002, 127, p 1359
17. Analytical Methods Committee, Recommendations for the conduct and the interpretation of co-operative trials, Analyst, 112, 679, (1987)
18. P.L. Davies, Statistical evaluation of interlaboratory tests, Fresenius Z Anal Chem, 331, 513, (1988)
19. ASTM E178:21- Standard Practice for Dealing with Outlying Observations
20. Robert B. Dean and Wilfrid J. Dixon (1951) "Simplified Statistics for Small Numbers of Observations". Anal. Chem., 1951, 23 (4), 636–638
21. Bernard Rosner, Percentage Points for a Generalized ESD Many-Outlier Procedure, Technometrics, 25(2), 165-172, (1983
22. ISO Guide 35:1989 – Certification of reference Materials - General and statistical principles

23. Hoogerbrugge et al, "Statistics and the assessment of interlaboratory studies", RIVM report 219101005, (November 1996)
24. M. Thompson et al, Efficacy of robust analysis of variance for the interpretation of data from collaborative trials, Analyst, 118, 235-240, (1993) Efficacy of robust analysis of variance for the interpretation of data from collaborative trials DOI https://doi.org/10.1039/AN9931800235
25. K. Kafadar, A Biweight Approach to the One-Sample Problem,J. Amer. Stat. Assoc., vol 77, issue 378, blz 416-424, (1982)
26. W. Beyrich et al, The DoD method, Kernforschungszentrum Karlsruhe, KfK 4721, EUR 11398 EN (1990) https://publikationen.bibliothek.kit.edu/270029141/16534512
27. R.G. Visser, R., Oussoren, W. The influence of different evaluation techniques on the results of interlaboratory comparisons Remarks on Linsinger TPJ, Kandler W, Krska R, Grasserbauer M. Accred Qual Assur 3, B497–B498 (1998). https://doi.org/10.1007/PL00022944
28. R.G. Visser, Interpretation of interlaboratory comparison results to evaluate laboratory proficiency, Accred Qual Assur (2006) 10: 521–526 DOI: 10.1007/s00769-005-0051-2

Address: Malledijk 18, P.O. Box 200, 3200 AE Spijkenisse, The Netherlands
Telephone number: +31 (0)88 214 45 41
Email address: nl.iis@sgs.com
Website: www.iisnl.com