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

Institute for Interlaboratory Studies
Spijkenisse, The Netherlands

Authors: R.J. Starink & R.G. Visser
Report: iis-protocol (version 3.3, April 2014)

CONTENTS

APPENDICES:

1 INTRODUCTION

1.1 THE INSTITUTE FOR INTERLABORATORY STUDIES (iis)

The independent Institute for Interlaboratory Studies (iis) organises 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 organises tailor made studies on request.

This report provides a comprehensive description of the organisation, 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 internet page at http://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 1100 laboratories from about 110 countries were actively participating in 2013.

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 known to a minimum of authorised 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 labcode number.

1.4 QUALITY

The Institute for Interlaboratory Studies in Spijkenisse, the Netherlands, is accredited in agreement with ISO/IEC 17043:2010 [28] since January 2000, by the Dutch Accreditation Council (Raad voor Accreditatie). See http://www.rva.nl for the actual accreditation scope.
The performance of a laboratory that participates in an iis proficiency test, will be accepted with confidence by a National Accreditation Body.

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 1100 laboratories from about 110 countries 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 two 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, DIN, ASTM, EN 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, DIN, ASTM, EN or other accepted industrial standards). Graphical tools are used to facilitate the interpretation of all data per test.

1.6 ANNUAL PT-PROGRAM AND INTERLABORATORY STUDIES ON REQUEST

iis works with an annual schedule, starting in August and finishing in July of the next year. The contents of its PT-program is discussed and decided upon during the advisory board meeting. 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 organised. These studies are initiated by the Institute itself or are tailor made and organised 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 internet page at http://www.iisnl.com.

2 TYPES OF INTERLABORATORY STUDIES; BRIEF OVERVIEW

2.1 INTERLABORATORY STUDIES FOR PROFICIENCY TESTING

Proficiency testing is the use of interlaboratory comparisons to determine the performance of individual laboratories for 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 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, DIN, ASTM or another accepted industry standard.

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.

Examples of reference materials which are available at the moment can be found in table 1. The materials are all multipurpose and available in handy quantities. They can be ordered from iis directly. See appendix 4 for an example of a certificate.

Reference material		Characteristics
o-xylene	√	High purity chemical
	√	all relevant impurities certified
Ethanol	√	Purity and water certified
	√	Tested according to common industrial specification
Ultra low sulphur B5 automotive diesel fuel containing	√	Modern automotive diesel fuel
	√	Selected high straight run character
	√	Low sulphur content
	√	Complies with EN590
	√	contains 5% FAME
Fuel Oil	√	Micro carbon residue certified
n-decane	√	for 3 different flash point methods
Lubricating Oil	√	14 wear metals certified
Monoethylene glycol	√	Regular ethylene glycol
	√	Purity and water certified
Jet A1 fuel	√	Regular aviation kerosine
	√	Complies with DERD 4294

Table 1: Examples of iis Reference Materials

Each reference material is accompanied with a certificate containing the certified reference values. Also, certification reports are available. For all reference materials, actual prices and availability see the iis' website at http://www.iisnl.com.

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 analysed 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 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 ORGANISATION

The interlaboratory studies of iis are all based on the same standardised 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.

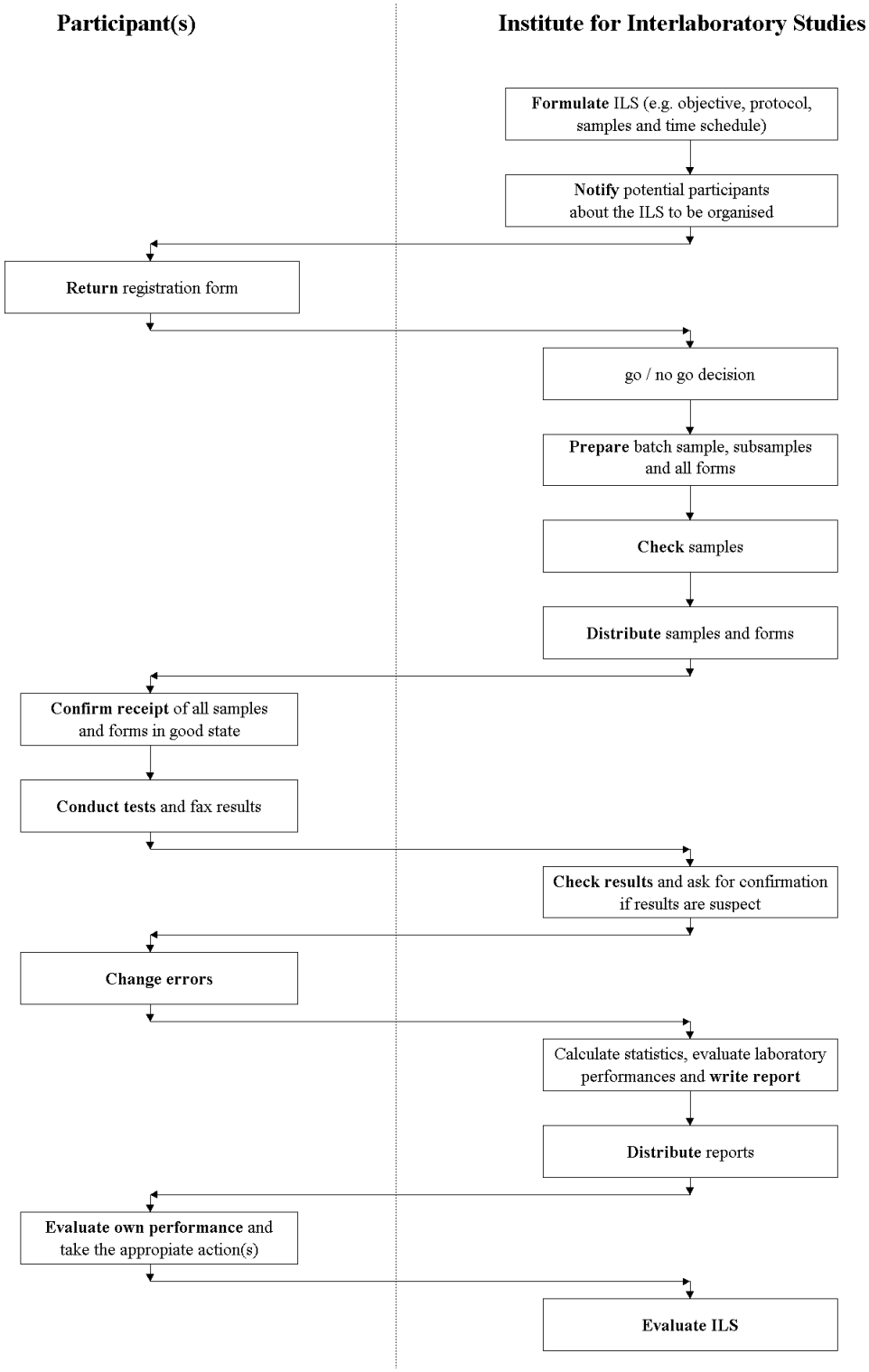


Figure 1: General procedure for the organisation of iis interlaboratory studies

The iis procedure for the organisation is described by the following steps:

1. The objective of the interlaboratory study to be organised 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. iis decides whether or not the planned interlaboratory study is organised.
5. The sample batch is prepared according to the protocol of sample preparation and checked for its fit for purpose.
6. The material is ensured to be stable during the proficiency test, based on critical parameters
7. The samples are bottled and the bottles are labelled.
8. The homogeneity of the bottled subsamples is checked.
9. All necessary samples are packed and distributed to the participants.
10. The participants report the sample receipt. If a package is not OK, a new sample is sent.
11. The participants analyse the samples.
12. The results are collected by iis.
13. After the deadline the results are checked for obvious errors and in case of erroneous results the participants are asked for confirmation or correction.
14. The complete dataset is analysed on normality and outliers are detected using the statistical protocol.
15. The statistical parameters are calculated, using the relevant protocol.
16. 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.
17. 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 in ASTM E1301[1], ISO 5725 [2,3], the J. AOAC [4] and ISO13528 [21], ILAC G13 [20] and ISO/IEC 17043 requirements [28]. Several of these references are obsolete and not readily available, but they are archived by iis.
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 analysed, 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 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, it should use the analysis methods that it

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.

Note for petroleum, liquid fuels and petrochemical laboratories:
Most of these laboratories are active in an ASTM affected market, yet the protocol in ASTM E691[5] is not followed to the letter. There are several reasons for this. This standard is only applicable for method validation tests as the title of ASTM E691 reads: 'Conducting an Interlaboratory Study to Determine the Precision of a Test Method'. This excludes the use for proficiency testing. Moreover ASTM E691 requires eight or more laboratories which should test at least three samples with different test levels and a minimum of replicates of three. This implicates that a participating laboratory should report at least nine results for each determination. In the literature more of such arrangements are found:

		minimum number of laboratories	minimum number of samples	minimum number of replicates
ASTM E1301	[1]	10		
ISO 5725	[2, 3]	8 - 15	1 - 6	2 - 3
AOAC	[4]	5	4	
ASTM E691	[5]	8 - 30	3 - 6	3 - 10
DIN 38402	[6]	7	3	3
IUPAC	[7]	5 - 8		
AMC	[8]	5	5	
ISO 4259	[9]	5	2 - 17	2

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 analyse.

The entire batch is thoroughly homogenised (and if necessary stabilised) 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.

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 gasoils, desired levels of sulphur, detectable quantities of trace impurities, etcetera.

In the case of a method validation study, more samples are prepared at different levels of standard addition.

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 specialised party (Sample and Dangerous Goods Management, SGS Nederland BV, 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.

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

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.

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

In order to ensure that all results will be reported in the same units, a detailed report form is made available. On the same form the analytical methods applied are reported by the participants.

Note for petroleum, liquid fuels and petrochemical laboratories:
The main part of the PT-annual program focuses on petroleum products, liquid fuels and petrochemicals. These products are normally investigated on full product profiles. Only in case of special interest, products are analysed on a single or a few analyte(s): market influences, problem areas or method evaluation purposes.
The petroleum products regularly programmed are: gasoil (automotive diesel profile), gasoline (also RFG profile), jet fuel (DERD profile), fuel oil, crude oil, gas condensate, lubricating oil, hydraulic oil, naphtha and the biofuels B100, B5 and biogasoline.
The programming in petrochemicals has more variation: methanol (IMPCA profile), ethanol, MTBE, ETBE, MEG, MPG, styrene, mixed xylenes, o-xylene, glacial acetic acid, acetone and benzene/toluene.

Note for consumer product laboratories:
The other part of the PT-program focuses on consumer products. These products are mainly investigated on banned components (RoHS). The consumer products programmed are: AZO-dyes in textile and leather, allergenic disperse dyes, heavy metals, formaldehyde in textile and leather, pesticides and phenols in textile, cadmium, lead, mercury and chromium in plastics, flame retardants, PFOA/PFOS, phthalates in plastics and specific and overall migration on food contact materials.

3.5 METHOD INFORMATION

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

3.6 TIME SCHEDULE

During four weeks after sample distribution, the results of the individual laboratories are collected.

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 one month after the deadline the final report is published and a copy sent to the participants.

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 deadline.
The revised test results will replace the erroneous ones. However, in the PT report the originally reported test results are mentioned under ‘remarks’.
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 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 [31]. 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 [27] are categorized as descriptive statistics, whereas theory-driven methods include the normality tests such as Shapiro-Wilk [32], Kolmogorov-Smirnov [34], Lilliefors [16, 33] and Anderson-Darling [35], 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 [36], 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.
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 [37]. Evaluation of the three normality indicators in combination with the visual evaluation of the graphic Kernel density plot [22,23], 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.

In the literature no consensus is found whether outliers should be rejected or not. The Analytical Methods Committee [8] 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 [30] criticises 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 Guide 43 [19] states that when participants' results are used to determine assigned values, techniques should be in place to minimise the influence of extreme results. It suggests removing outliers prior to calculation and refers to ISO 5725 [3]. In ASTM E178 [10] 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 utilise statistical calculations on restricted observations. For the detection of outliers various techniques can be used, such as Dixon Test [3], Grubbs Test [3], Rosner's generalized ESD test [26] and/or Tietjen-Moore Test [3, 10].

Most procedures for detection of outliers will only work properly if the data have a normal distribution and if enough data 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, Dixon and/or Grubbs outlier tests (<21 test results) are used. For larger data sets (> 20 test results) Rosner's Generalized ESD outlier test [26] 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. 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, DIN, ASTM, EN) 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.

In iis interlaboratory studies for **method validation** the number of participants usually is relatively low (10 - 20). Because of this, Gaussian statistics, assuming a normal distribution of the data, can not always be used [11]. Also, when using normal statistical calculations, the detection of outlying data is much less meaningful, since only few data are available and the distribution of these data is not known. For these reasons the results per determination and per sample are <u>not</u> submitted to a Dixon and/or to a Grubbs outlier tests, but to Huber's Elimination Rule, a robust outlier test [13]. This test does not suffer from the weaknesses of the normal outliers tests, like the so-called masking effect.

> *<u>Note on masking effect:</u>*
> The masking effect occurs when the number of actual outliers is larger than the number on which the test is based. Most outlier tests have small breakdown points and therefore cannot be relied upon to detect outliers. The Dixon Test and the Grubbs Test mentioned in ASTM E178 and ISO 5725 clearly

have this disadvantage. For instance a data set may contain only two large outliers which are not detected by the Dixon Test. The Grubbs Test is somewhat less worse: it will detect four to five outliers in a set of 40 data. The Generalized Extreme Studentized Deviate Test by Rosner is less sensitive for the masking effect and may detect more outliers in one run than the Grubbs Test. Howver, the Huber Elimination Rule is superior with a breakdown point of approx. 50%. Even 19 outliers in a set of 40 data can be detected.

In the case of robust statistics, the outliers are not excluded before calculation of the Robust statistical parameters [11,12].

4.4 CALCULATION OF THE SUMMARY PARAMETERS

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

* The mean x, as best estimate of the true value μ:

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

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

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

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

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

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 normal mean and standard deviations (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 [29]:

$$\text{confidence interval} = \mu \pm \frac{t \times s}{\sqrt{n}}$$

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 data

For the certification of **reference materials** and **method validation** studies often *robust statistic* is used instead of traditional statistics. In the case of robust statistics [11,12] 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 distibution 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 [18]:

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

* The robust standard deviation $s_R(DoD)$ [14], as measure of the spread, is also calculated without prior removing of stragglers and outliers. The calculation is based on all absolute differences:

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

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

$$R = 2 \times \sqrt{2} \times s_R\,(DoD)$$

For **method validation** studies with a two-level design, split-level calculations are used. The calculations (from ISO 5725) are applied to those parameters that are not influenced by the (small) additions made to create different analyte levels. Besides the normal statistical parameters also the repeatability standard deviation s_r is calculated. This is a measure for the intralaboratory spread σ.

$$s_r = \sqrt{\frac{\sum_i \left[\left(x_i - y_i\right) - \left(\overline{x_i - y_i}\right)\right]^2}{2 \times (n-1)}}$$

From the standard deviation s_r, the repeatability *r* is calculated:

$$r = 2 \times \sqrt{2} \times s_r$$

Finally, the confidence limits (95 %) of the consensus values for μ are calculated according to IP 367/84 [15]:

$$\text{confidence limit} = \mu \pm \frac{R}{\sqrt{2n}}$$

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 has produced. The evaluation of the accuracy is in principle done towards an external standard when available. Each laboratory receives a numerical indication (z-score, see par. 5.3) 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, DIN, ASTM or another accepted standard in the industry). This allows a straight forward and easy evaluation of performance 'over time' [24].

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, DIN, ASTM 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. Four types of evaluations have been implemented in the PT-program.

- Indicators calculated **per test** can be used for detailed inspection of the test results per laboratory in a round.
- Indicators **per test** that measure the bias of a laboratory compared to the group.
- Indicators **per profile** are a measure of the integral proficiency of a laboratory, for instance in testing a certain product on multiple analytical parameters (so-called profile).
- Indicators **per group** of laboratories can be used to compare the proficiency of a whole group of laboratories with the official analytical standards (optional).

In addition to numerical performance parameters, the graphic representation of the results is another simple tool to evaluate the results (see paragraph 5.5).

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. 5.3). This most common indicator compares the bias with a standard error. The bias is calculated as the difference between the reported result of laboratory i (x_i) and the assigned value (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, DIN or ASTM.

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

$$z_i = (x_i - X) / \sigma$$

where: x_i = the **result** of laboratory i for that specific test.
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 analyse. The items do not have a known composition (e.g. concentrations or amounts). The mean of all valid lab results is used as the consensus 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, DIN, ASTM 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 parameter since, with normally distributed results, the 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 lab (personnel, QA-manager and the management) and also its (potential) clients and accreditation bodies a good indication of its analytical competence.

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 method that may be well applicable, but has a very different reproducibility. When the reproducibility of the method used is higher than the reference method (e.g. 1.5 for ASTM D1298 instead of 0.5 for ASTM D4052, average of two determinations on lub oil, distillate or basestock), 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 it z-score(s) in accordance with the calculation of paragraph 5.3 to get the correct impression of its performance [25].

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 recognised test methods have been validated and values for the reproducibilities have been established.

Deviating reproducibilities may be due to a number of laboratories that produce strongly 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.

In iis **method validation** interlaboratory studies, the results per test of two samples are presented in a two-sample or Youden plot.

One sample or Gauss plot
In order to visualise the data against the required reproducibilities, Gauss plots using the sorted data for one determination, are made (see examples in appendices 1 and 2, pages 24 and 25).

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, DIN, ASTM).

Two sample or Youden plot
In order to evaluate systematic deviations, two sample Youden plots [17] are made (see examples in appendice 3, page 26)

On the X axis the results from sample one are plotted against the results of the other sample on the Y-axis. Therefore, each participant is presented by one point in the graph. Accepted data are presented as a triangle; outliers and other data, which were excluded from the calculations, are presented as a cross. The means of the results are presented by the dotted lines. The intersection of these lines is the target value, where the participant's points should be positioned if both results were accurate. Parallel to the dotted lines for the means are continuous lines for the target reproducibilities. The repeatability limits may be represented by the continuous lines with an angle of 45°. Both the reproducibility and the repeatability are taken from the relevant standards (e.g. DIN, ISO, ASTM). Sometimes, not all lines are visible in the plot.

Systematic errors as well as the random errors are visualised in the Youden Plot. If the results from the different laboratories vary entirely because of random error, the results will fall randomly round the average and approximately equal numbers of points in each of the four quadrants of the plot will be present (see appendix 3, page 26). If, however, systematic errors are the main cause of the variation, one can expect that a laboratory obtaining a high value for sample 1 would also tend to obtain a high value of sample 2. This will lead to a predominance of points in the top right and the lower left quadrants of the plot (see appendix 4, page 27). Thus, in principle 95% of all results should fall within the inner section of the hexahedron that is formed by the reproducibility and repeatability limits.

In practice, since random errors are always present to some extent, the points will fall within an ellipse that has the 45° diagonal as its major axis. The length of the perpendicular from an individual point to this diagonal gives a measure of the random error, and the perpendicular intersects the diagonal at a point at a distance from the centre which is related to the systematic error of that laboratory.

On basis of these Youden plots and the ratio R/r, the observed differences are discussed per determination. In case the results of both samples are correlated, a line through (0,0) with an angle of 45° will be observed and R/r will be > 3. If random errors are dominant, about the same amount of points in each quadrant of the plot will be present and R/r < 3.

Kernel Density plot

In order to visualise the distribution of the data a Kernel Density plot [22, 23] 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 standardised format. The following paragraphs are included in principle:

Paragraph	Title	Contents
1	**Introduction**	The proficiency test is summarised.
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	Stability	A reference is made to the fit-for-use/homogeneity/stability sheet.
2.6	Analysis	A summary is given of the analyses that had to be performed by the participants on the distributed 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	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, DIN, ASTM, EN).
4.3	Comparison with previous PT's	The proficiency test and the participants' results are compared with the previous rounds of the PT.
Appendix 1	Data and statistical results	Per test, the 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 repeatability and reproducibility stated in the appropriate standard (e.g. ISO, DIN, ASTM) 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.

7 ANNUAL PROGRAM AND COSTS

7.1 ANNUAL PROGRAM

About 80 interlaboratory studies (mostly proficiency tests) are organised each year.

In iis proficiency tests, about 1100 laboratories from about 110 countries are participating.

The actual PT-program and all (other) relevant information are sent to interested laboratories on request. This and additional information can also be found on the Institute's internet page at http://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 valid, independent on the number of tests performed. Costs for sample despatch are dependent of the sample type and the country where to it has to be sent and are therefore not included. These costs for sample package and despatch will be charged separately.

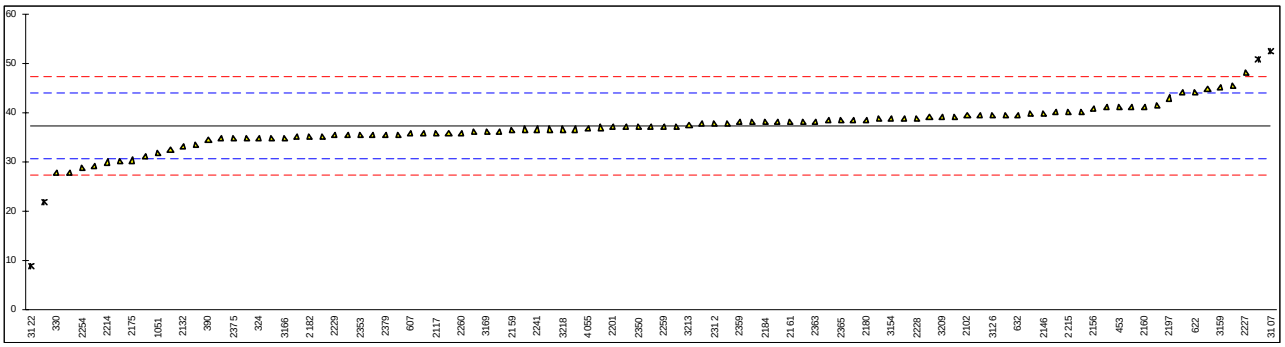
8 LITERATURE REFERENCES

1. ASTM E1301-03 - Standard Guide for Development and Operation of Laboratory Proficiency Testing Programs (withdrawn)
2. ISO5725:1986 - Precision of test methods - Determination of repeatability and reproducibility by interlaboratory tests (withdrawn)
3. ISO5725:1994 - Accuracy (trueness and precision) of measurement methods and results, parts 1-6
4. M. Thompson and R. Wood, J. AOAC Int, 76, 926, (1993), also ISO/REMCO N280 (1993) and also Pure & Appl. Chem., 65, 2123, (1993)
5. ASTM E691-05 - Standard practice for conducting an interlaboratory study to determine the precision of a test method (replaced by E691-13)
6. DIN 38402-T 41 - General information - Interlaboratory tests; planning and organization (A 41)
7. IUPAC - Nomenclature for interlaboratory analytical studies - recommendations (1992)
8. Analytical Methods Committee, Recommendations for the conduct and the interpretation of co-operative trials, Analyst, 112, 679, (1987)
9. ISO4259:1992 - Petroleum products -- Determination and application of precision data in relation to methods of test
10. ASTM E178-02 - Standard Practice for Dealing with Outlying Observations
11. R. Hoogerbrugge et al, "Statistics and the assessment of interlaboratory studies", Eurachem Netherlands, (December 1996)
12. M. Thompson et al, Analyst, 118, 235, (1993)
13. J.N. Miller, Analyst, 118, 455, (1993)
14. W. Beyrich et al, KfK 4721, EUR 11398 EN (1990)
15. IP 367/96 – Petroleum products - Determination and application of precision data (replaced by IP367/07, identical to ISO4259:1992)
16. W.J. Conover, Practical Nonparametric Statistics, Wiley, NY, 302, (1971)
17. W.J. Youden and E.H. Steiner, 'Statistical Manual of the AOAC', (1975)
18. K. Kafadar, J. Amer. Stat. Assoc., 77, 378, 416, (1982)
19. ISO Guide 43-1:1997 - Proficiency testing by interlaboratory comparisons - Part 1: Development and operation of proficiency testing schemes (withdrawn)
20. ILAC -G13:2007 - ILAC requirements for the competence of providers of proficiency testing schemes (withdrawn)
21. ISO 13528:2005 – Statistical methods for use in proficiency testing by interlaboratory comparisons
22. Analytical Methods Committee Technical brief, No 4 January 2001.
23. The Royal Society of Chemistry, Analyst 2002, 127, p 1359, P.J. Lowthian and M. Thompson, see also The AMC Technical brief no. 23 of March 2006, fee downloadable from www.rsc.org/Membership/Networking/InterestGroups/Analytical/AMC/TechnicalBriefs.asp
24. R.G. Visser, W. Oussoren, Accred Qual Assur (1998) 3:497–498
25. R.G. Visser, Accred Qual Assur (2006) 10: 521–526
26. Bernard Rosner, Percentage Points for a Generalized ESD Many-Outlier Procedure, Technometrics, 25(2), 165-172, (1983)
27. R.B. D'Agostino., 1986, Goodness-of-fit techniques, D'Agostino & Stephens eds., Marcel Dekker, New York, p. 367
28. ISO/IEC17043:2010 – Conformity Assessment – General requirements for proficiency testing
29. ISO Guide 35:2006 – Reference Materials - General and statistical principles for certification
30. P.L. Davies, Fr. Z. Anal. Chem, 331, 513, (1988)

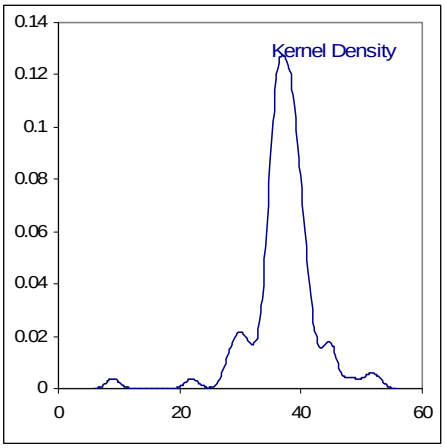
31. J.M. Dufour, A. Farhat, L. Gardiol, L. Khalaf, Simulation-based Finite Sample Normality Tests in Linear Regresssions, Econometrics Journal, Vol. 1, p.154, (1998)
32. S.S. Shapiro, M.B. Wilk, An analysis of variance test for normality (complete samples), Biometrics, 52, p. 591 (1965)
33. H.W. Lilliefors, On the Kolmogorov_smirnov Test for Normality with Mean and Variance Unknown, Journal of the American Statistical Association Vol 62, 318, p. 399 (1967)
34. W.W. Daniel, Biostatistics: a foundation for analysis in the health sciences, p. 603, 4th Ed. John Wiley & Sons, Inc. New York (1987)
35. Anderson, T.W. and Darling, D.A., A Test of Goodness-of-Fit. Journal of the American Statistical Association Vol 49: p. 765, (1954)
36. 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)
37. 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)

APPENDIX 1: DETERMINATION OF METALS IN PLASTICS

Example taken from the iis proficiency test metals in plastics of September 2009



			Only EN1122 data:
Normality	not OK		OK
n	95		35
outliers	4		1
mean (n)	37.26	mg/kg	37.92
st.dev. (n)	3.707	mg/kg	3.428
R(calc.)	10.38	mg/kg	9.60
R(EN1122:01)	9.32	mg/kg	9.48



Comments:

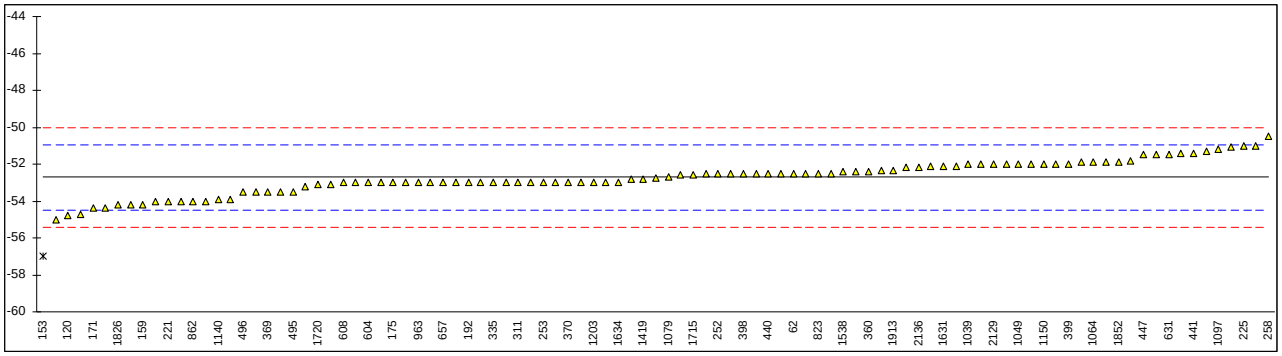
The Gauss plot shows that the results of most participants lie between the 3s reproducibility limits of the relevant standard (EN 1122:01). Apparently, the analytical performance of most of the participating laboratories for the analyses of metals in plastics is satisfactory.

Four laboratory results are marked as statistically outlying and one other laboratory result also exceeds the 3s reproducibility limits. The five respective laboratories clearly have a problem and should take corrective actions to improve their quality.

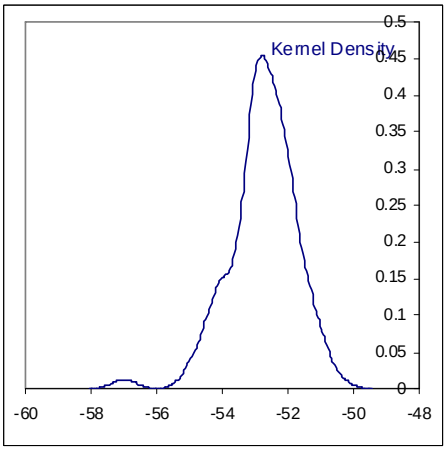
The group of participants as a whole is not able to meet the requirements of EN 1122:01. The calculated reproducibility R(calc) exceeds the literature reproducibility R(EN 1122:01). When the 36 EN 1122 test results are evaluated separately, the calculated reproducibility R(calc) is in full agreement with the literature reproducibility R(EN 1122:01).

APPENDIX 2: DETERMINATION OF FREEZING POINT OF JET FUEL A1

Example taken from the iis proficiency test Jet Fuel A1 of September 2009



normality	not OK	
n	98	
outliers	1	
mean (n)	-52.72	°C
st.dev. (n)	0.901	°C
R(calc.)	2.52	°C
R(D2386:06)	2.50	°C



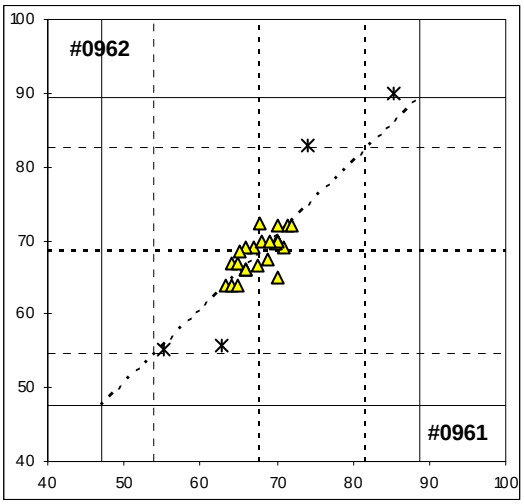
Comments:

The Gauss plot shows that only one laboratory clearly has a problem as its test result has a z-score smaller than minus three (-3) and this result appears to be the only statistical outlier (Grubbs, with 99% certainty).
The test results of the other laboratories do not deviate significantly from the mean result and all lie within the reproducibility 3s limits.

For this determination the reproducibility of the results of the group of participating laboratories (R_{calc} = 2.52) nicely equals the requirement mentioned in the standard test method that was used as target (R_{target} = 2.50), in this case ASTM D2386:06.
In short, the group of participating laboratories is able to match the reproducibility of the standard.

APPENDIX 3: DETERMINATION OF ETHYLBENZENE IN STYRENE

Split level example taken from the iis proficiency test Styrene of September 2009



Normality	OK	not OK	
n	28	28	
Outliers	2	2	
Mean (n)	67.85	68.63	mg/kg
Stdev (n)	2.893	5.475	mg/kg
R(calc)	8.10	15.33	mg/kg
R(D5135:07)	19.39	19.52	mg/kg

Comments:

This two-sample Youden plot shows results for two different samples that scatter all around the middle of the plot, indicating that the errors present in these results are of the random type. Another four results - that all appeared to be statistical outliers - are far away from the central data cloud and show a linear relationship between the analysis results on the two samples. In other words: two laboratories reported consistently high results and two other laboratories reported consistently low results for both samples. A systematic difference between the laboratories is found. Two of the 4 excluded laboratories used an in house method, which may be significantly different from ASTM D5191, that was used by all other laboratories.

APPENDIX 4: EXAMPLE OF CERTIFICATE OF REFERENCE MATERIAL

COMMERCIAL GRADE JET A1 FUEL

Certificate of Analysis

Reference Material JF-011097

Jet Fuel A1

Reference Material JF-011097 consists of a 260 ml bottle, containing approximately 250 ml of regular aviation kerosine (type Jet Fuel A1). This RM is intended primarily as a quality control material for use in the determination of Freezing Point, Density, MSEP, Naphtalenes, Smoke Point, Sulphur and some distillation properties.

Certified Property Values
The certified values are given in table 1. The certified values in table 1 have been derived from the results obtained from 2 independent international interlaboratory studies in which respectively 15 and 63 laboratories participated. The results of these interlaboratory studies are presented and discussed in the iis report iis97J02-RM.

Table 1. Certified values[b] for JF-011097.

Parameter	Certified value[a]	
Aromatics, %V/V	23.6	± 0.3
Density @ 15°C, kg/L	0.80455	± 0.00005
Freezing Point, °C	- 49.3	± 0.3
Kinematic viscosity @ -20°C, mm^2/s	3.554	± 0.014
MSEP, %	98.1	± 0.4
Naphthalenes, %V/V	2.99	± 0.04
Smoke Point, °C	21.2	± 0.5
Sulphur, % W/W	0.0157	± 0.0010
i.b.p., °C	147.8	± 1.0
50% recovered, °C	191.9	± 0.4
f.b.p., °C	262.5	± 0.8

a) The estimated uncertainty is given as 95% confidence limits
b) The following values were also determined for this RM. These values are not certified, but for indication only:
Distillation: 10% recovered, °C 165.8 ± 0.5 and 90% recovered, °C 239.9 ± 0.6;
Mercaptane Suphur, %W/W 0.00015 ± 0.0005; Specific Energy, MJ/kg 43.113 ± 0.005;
Total Acidity, mgKOH/g 0.0019 ± 0.005

NOTICE AND WARNINGS TO USERS

Shelf life: The preparation of this RM was finished February 27, 1998. When stored properly and unopened, the expiring date of this RM is March 2016. I.I.S. regularly checks the validity of the RMs in stock. If there is any doubt about the validity of the RM you are advised to contact iis (iisnl@sgs.com).

Storage: Bottles should be stored in a dark and cool place, preferably at a temperature between 0 °C and + 10 °C.

Suggested procedure for use of the RM as quality control sample:
Before opening a bottle and taking a sample for analysis, the contents must be mixed to ensure homogeneity. Once the bottle has been opened, the material is susceptible to contamination (e.g. laboratory dust or vapours) or losses. Certified values are not applicable to bottles stored after opening, even if resealed.

Safety handling instructions: Kerosine is inflammable. The flash point of the material of this RM is 41 °C; therefore, care should be exercised during handling and use. Use proper methods for disposal of waste.

Spijkenisse, The Netherlands
April, 2014 (version 12)

dr. R.G. Visser
Institute for Interlaboratory Studies