

**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 3.1, November 2008)**

CONTENTS

1	INTRODUCTION	3
2	TYPES OF INTERLABORATORY STUDIES; BRIEF OVERVIEW	5
3	ORGANISATION	7
4	STATISTICAL PROCESSING OF THE TEST RESULTS.....	12
5	PERFORMANCE EVALUATION	16
6	REPORT CONTENTS	20
7	ANNUAL PROGRAM AND COSTS	21
8	LITERATURE REFERENCES	23

APPENDICES:

Appendix 1: Determination of Benzidine in Textile	24
Appendix 2: Determination of Aromatics in Jet Fuel A1	25
Appendix 3: Determination of D.V.P.E. (R.V.P.) of reformulated gasoline	26
Appendix 4: Determination of Aromatics in Naphtha	27
Appendix 5: Certificate of iis Reference Material: o-xylene	28
Appendix 6: Certificate of iis Reference Material: Jet Fuel A1	29

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. Studies are usually performed on commercially relevant products and involve testing on complete 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 600 participating laboratories from about 90 countries have been registered.

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 guide 43 [19] and ILAC-G13, (R007) [20], 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, may be taken into account by a National Accreditation Body with confidence.

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 registered in qualification records.

1.5 UNIQUE SET-UP

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

- Its world-wide set-up: more than 600 laboratories from about 90 countries have been registered.
- Its short turn-around time: normally, the complete time span from sample dispatch up to and including the distribution 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, three 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. i.i.s uses the z-score as the 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 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 September and finishing in June of the next year. The contents of its PT-program is discussed and laid down during the annual international laboratory managers meeting. The criteria for priority selection of products and tests for each years program, are chosen on the basis of an evaluation of commercial risks (claims, near-missers and complaints), findings in previous programs, requests from participants and technical developments in the laboratory area.

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.

Laboratories participating in 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, may be taken into account by a National Accreditation Body with confidence.

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 standard in the industry.

2.2 INTERLABORATORY STUDIES FOR PREPARATION OF REFERENCE MATERIALS

Proficiency tests are very useful tool for quality control in an analytical, but a normal frequency of PT's 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. Unfortunately, in practice there is a short on 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 following reference materials are available at the moment (see appendix 5 and 6 for an example of the certificates). The materials are all multipurpose and available in handy quantities. They can be ordered from iis directly.

Reference material	Characteristics
o-xylene	✓ High purity chemical ✓ all relevant impurities certified
Ethanol	✓ Purity and water certified ✓ Tested according to common industrial specification
Low sulphur automotive diesel fuel	✓ Modern automotive diesel fuel ✓ Selected high straight run character ✓ Low sulphur content ✓ Complies with EN590 ✓ Moderate winter quality
Fuel Oil	✓ Micro carbon residue certified
n-decane	✓ for 3 different flash point methods
Methanol	✓ High purity chemical ✓ All relevant physical tests certified
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: Overview of iis Reference Materials

Each reference material is accompanied with a certificate. Also, complete certification reports are available. For prices and availability see 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. That nonetheless minor differences may exist between the certificates is 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 during 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 render 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. Interlaboratory studies for method evaluation will reveal this information.

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 organises interlaboratory studies for the validation of 'in-house' developed methods.

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 representatives from participating laboratories support the annual PT-program.

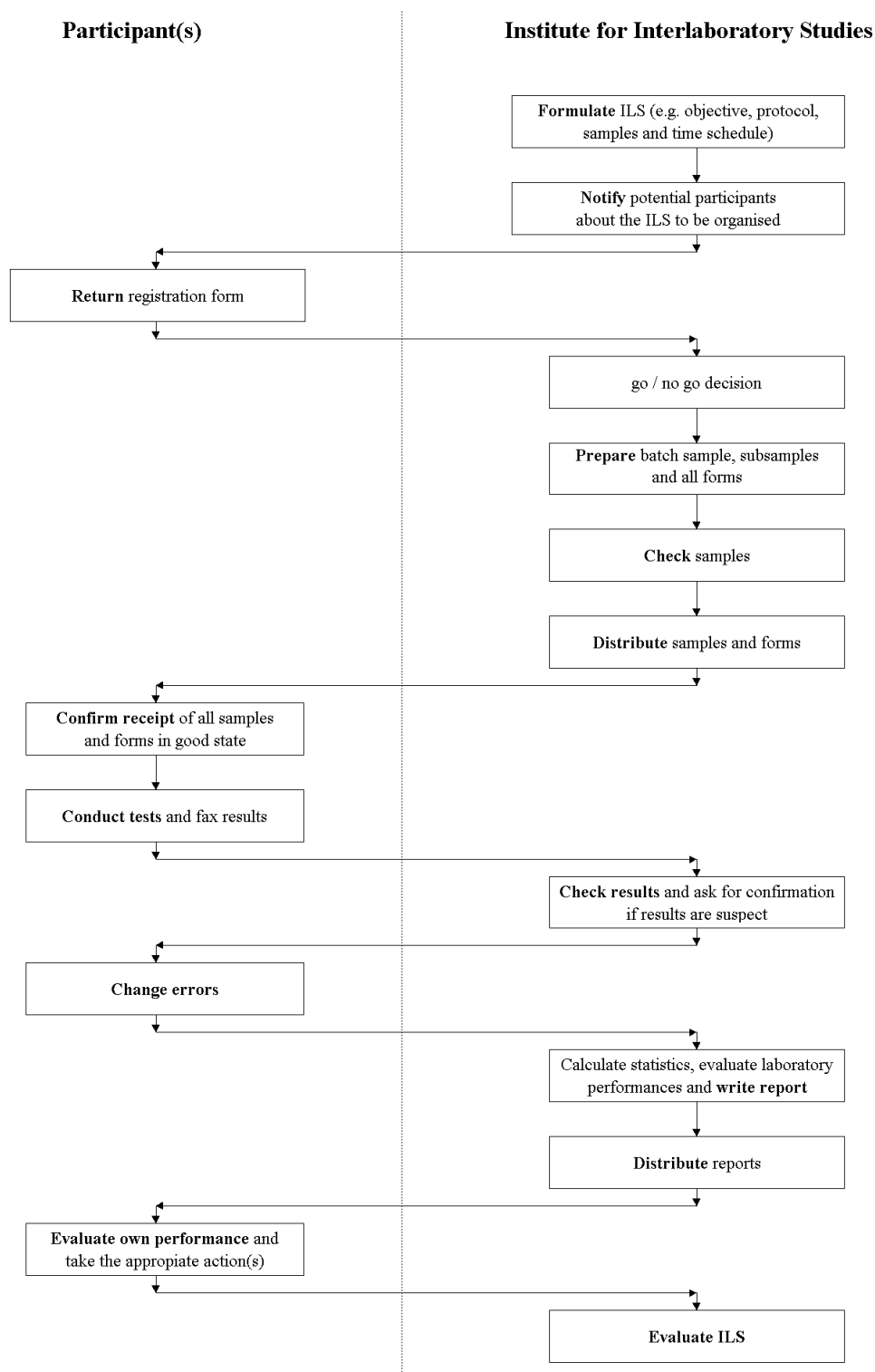


Figure 1: General procedure for the organisation of iis interlaboratory studies

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, new sample is sent.
11. The participants analyse the samples.
12. The results are collected in Spijkenisse.
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 anonymous final reports are 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 ISO13528:00 [21] the PT-guidelines ISO Guide 43 [19] and ILAC G13 requirements [20].

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 labs day-to-day performance the participant must treat the samples as if they were routine samples. So, it must use the analysis methods that it would use in

normal daily practice, no special attention must 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 straightforwardly. 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 5727	[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 all details.

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 can not 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 randomly checking one or more critical 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

Samples are distributed by road and/or air.

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, 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.

If necessary, a material safety data sheet (MSDS) is made and enclosed to the sample. Each MSDS is prepared according to relevant guidelines and examples (e.g. 'Chemiekaarten').

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 lab'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 report form is added to the samples. 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 complete 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 implied 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 have been programmed regularly.

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 implied are: AZO-dyes in textile and leather, allergenic dyes, heavy metals, free formaldehyde, pesticides and phenols in textile, cadmium, lead and chromium in plastics and phthalates in plastics.

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) the method number is sufficient, in other cases the key elements of the method are asked and 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. It is also asked for new results.

About one month after the deadline the final report is 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. A robust outlier test, Huber Elimination Rule, is used for this purpose. In case of erroneous results, the respective participant is notified immediately so it can take all necessary corrective actions. It is also asked for new results. The notification of deviating results is done shortly after the closing date for sending in the results, normally within 2 days after deadline.

The corrected results replace the erroneous ones. In the final report the results sent in at first are mentioned as a remark.

4.2 CHECK ON NORMAL DISTRIBUTION OF THE TEST RESULTS

Prior to calculating statistical parameters, a check is done on the validity of the results to be used in the calculations: the type of distribution of the data and the presence of outlying results.

Many statistical procedures are only applicable to random samples from populations with a Gaussian distribution. Even the outcome of the most simple parameter 'mean', which should be a good estimate of the true value, depends strongly on the type of distribution of the data. For this reason, prior to using the data, the normality of the distribution of the data set per determination is checked by means of the Lilliefors-test [16], a variant of the Kolmogorov-Smirnov test.

4.3 DETECTION AND REMOVAL OF ERRONEOUS AND STATISTICALLY DEVIATING RESULTS

Also statistical outliers may affect the statistical parameters like the mean. 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 [9] 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, Grubbs Test 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. 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. Again, all data is screened for outliers and deviating results are removed prior to calculation of the certificate 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 not submitted to a Dixon and/or to a Grubbs outlier tests, but to Hubers 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.

Note on masking effect:

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 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 Huber Elimination Rule however is superior with a breakdown point of appr. 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.

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 \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 spread σ :

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

- * 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 34:

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

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 abnormal 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 [18]:

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

- * 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:

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

- * 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 .

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

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

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

Finally 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 done towards an external standard. Each laboratory receives a numerical indication (z-score, see par. 5.3)

In its 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 commonly 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 $z_{(\text{target})}$ -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 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, DIN, ASTM or another accepted standard in the industry.

The z-score calculation of iis results in a simple, straight forward comparison of a lab 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 in stead of 0.5 for ASTM D4052), 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 [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 cured 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 cure. 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 mean is presented by a continuous line. The reproducibility limits of the corresponding international accepted test method (e.g. ISO, DIN, ASTM) are presented by dotted lines parallel to the mean line (mean \pm reproducibility).

Two sample or Youden plot

In order to evaluate systematic deviations, two sample Youden plots [17] are made (see examples in appendices 3 and 4, pages 26 and 27).

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 participants points should be positioned if both results were accurate. Parallel to these two dotted lines for the means are continuous lines for the target reproducibilities at distances (mean - reproducibility) and (mean + reproducibility). The repeatability limits are represented by the 2 continuous lines with an angle of 45°. Both the reproducibility and 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. 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 always included:

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	Detail information	Summary of details from a determination
Appendix 3	List of Participants	List of the number of participants per country (no details)
Appendix 4	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 35-40 interlaboratory studies (mostly proficiency tests) are organised each year.

In iis proficiency tests, about 600 laboratories from about 90 countries are participating.

The actual PT-program and all (other) relevant information is 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.

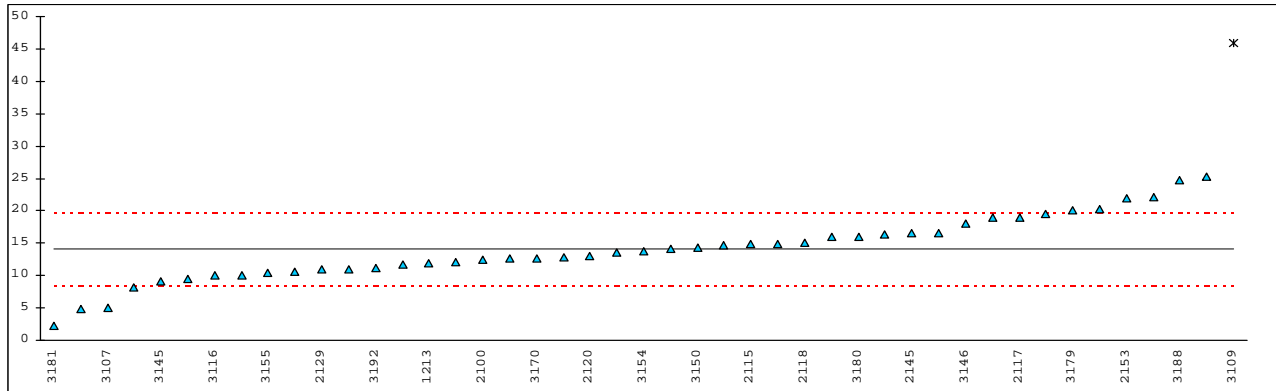
Costs for sample despatch are dependent of the sample type and the country where to it has to be sent and are therefore not included. These costs for sample package and despatch will be charged separately.

8 LITERATURE REFERENCES

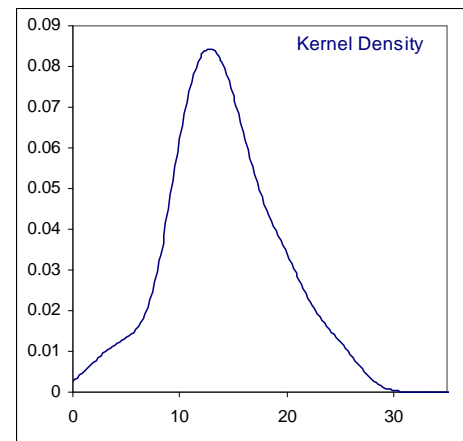
1. ASTM E1301-03 - Standard Guide for Development and Operation of Laboratory Proficiency Testing Programs
2. ISO5725:1986 - Precision of test methods - Determination of repeatability and reproducibility by interlaboratory tests
3. ISO5725:1994 - Accuracy (trueness and precision) of measurement methods 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
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 Method Committee, Recommendations for the conduct and the interpretation of co-operative trials, Analyst, 112, 679, (1987)
9. P.L. Davies, Fr. Z. Anal. Chem, 331, 513, (1988)
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
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
20. ILAC -G13:2007 - ILAC requirements for the competence of providers of proficiency testing schemes
21. ISO 13528:2005 – Statistical methods for use in proficiency testing by interlaboratory comparisons
22. Analytical Methods Committee Technical brief, No4 January 2001.
23. The Royal Society of Chemistry 2002, Analyst 2002, 127 page 1359-1364, P.J. Lowthian and M. Thompson (see <http://www.rsc.org/suppdata/an/b2/b205600n/>).
24. R.G. Visser, W. Oussoren, Accred Qual Assur (1998) 3:497–498
25. R.G. Visser, Accred Qual Assur (2006) 10: 521–526

APPENDIX 1: DETERMINATION OF BENZIDINE IN TEXTILE

Example taken from the iis proficiency test Azo Dyes (Aromatic Amines) of April 2003



Normality	OK	
n	44	
Outliers	1	
mean (n)	14	mg/kg
st.dev. (n)	5	mg/kg
R(calc.)	14	mg/kg
R(LMBG 82.02-2)	5	mg/kg



Comments:

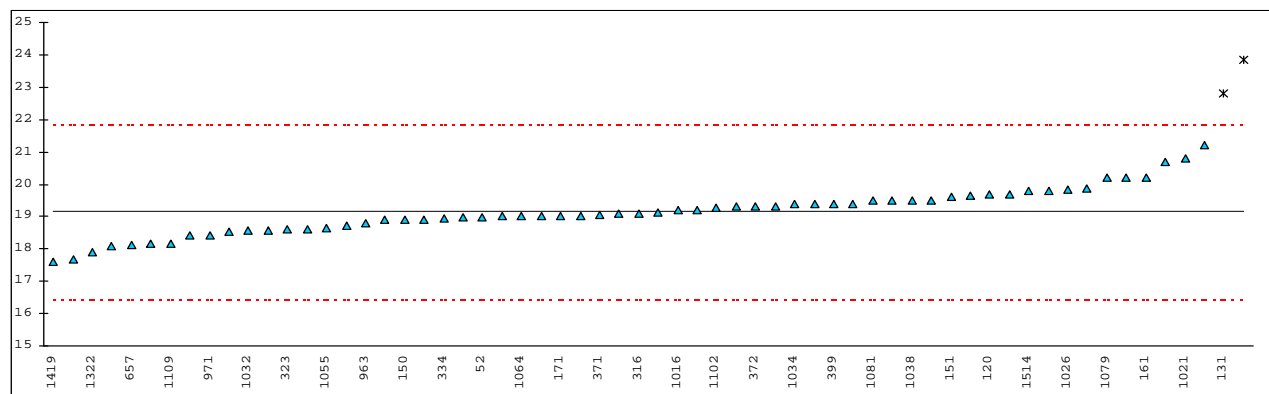
The Gauss plot shows that the results of most participants lie between the reproducibility limits of the relevant standard (LMBG 82.02-2). Apparently, the analytical performance of most of these laboratories for the analyses benzidine in textile is satisfactory or even good.

One laboratory is marked as outlying and another ten labs also exceed the reproducibility limits. These laboratories clearly have a problem and should take corrective actions.

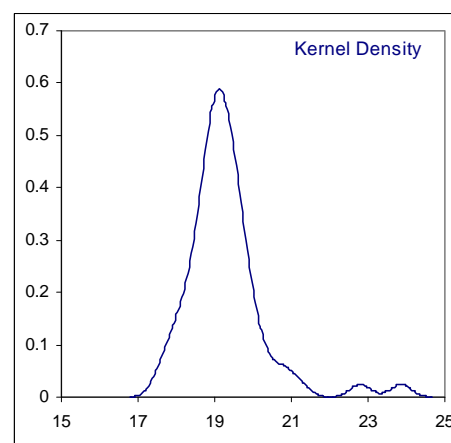
The group of participants as a whole is not able to meet the requirements of LMBG 82.02-2. The calculated reproducibility R(calc) exceeds the literature reproducibility R(LMBG 82.02-2) significantly.

APPENDIX 2: DETERMINATION OF AROMATICS IN JET FUEL A1

Example taken from the iis proficiency test Jet Fuel A1 of October 2003



normality	OK	
n	60	
outliers	2	
mean (n)	19.15	%V/V
st.dev. (n)	0.716	%V/V
R(calc.)	2.00	%V/V
R(D1319)	2.71	%V/V



Comments:

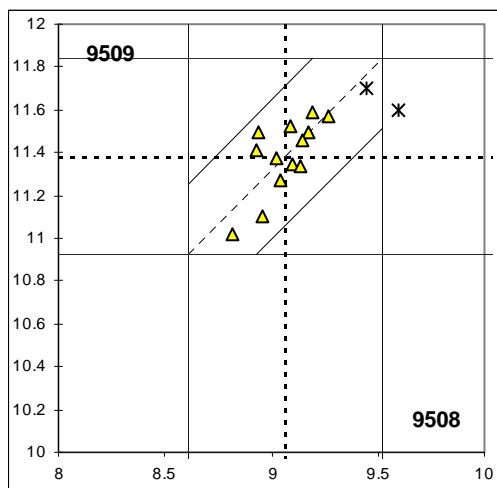
The Gauss plot shows that only two laboratories clearly have a problem. The other laboratories do not deviate strongly from the mean result and lie within the interval $\text{mean} \pm \text{reproducibility}$.

For this test the borderlines of the group of participating laboratories ($R = 2.00$) are smaller than the requirements mentioned in the literature ($R = 2.71$), in this case the ASTM D1319. In other words the group of participating laboratories is able to match the reproducibility of the standard.

The test results do have a normal (Gaussian) distribution.

APPENDIX 3: DETERMINATION OF D.V.P.E. OF GASOLINE

Example taken from the iis proficiency test Gasoline of July 1995



Normality	OK	OK	
N	13	13	
Outliers	2	2	
Mean (n)	9.06	11.39	psi
Stdev (n)	0.126	0.171	psi
R(calc)	0.353	0.479	psi
R(D5191)	0.440	0.470	psi

Comments:

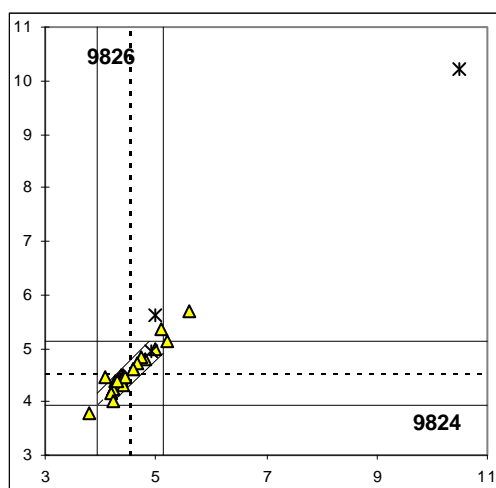
This two-sample Youden plot shows results for two different samples that scatter all over the plot. This indicates that the errors made by the participating laboratories are of the random type. The two excluded laboratories used a significant different method (ASTM D4953), whilst all other labs used ASTM D5191.

If the magnitude of the scale of random errors is small it means that the reproducibility between the labs is good and that the reproducibility is in the order of the repeatability.

If the magnitude of the scale of random errors is large it means that a poor reproducibility is accompanied by a poor repeatability, in other words the laboratories are not able to apply the respective method in a proper way. This situation can be improved by evaluation of the method (or its application) by the participating laboratories.

APPENDIX 4: DETERMINATION OF AROMATICS IN NAPHTHA

Example taken from the iis proficiency test Naphtha of May 1998



normality	not OK	
n	24	
outliers	3	
mean (n)	4.536	%V/V
stdev (n)	0.4102	%V/V
r(calc)	0.225	%V/V
R(calc)	1.159	%V/V
r(D5443)	0.164	%V/V
R(D5443)	0.596	%V/V

Comments:

This Youden plot shows an almost linear relationship between the analysis results on two different samples. In other words: some labs are reporting consistently high results, other laboratories are reporting low results for both samples. A systematic difference between the laboratories is found, in this case probably in the way of instrument calibration. Since the variance is due to systematic errors between the participating laboratories, remediation can be achieved by for instance supplying calibration samples to all participating laboratories.

The test results do not have a normal (Gaussian) distribution. Therefore, the calculated parametric statistics as mean, standard deviation and reproducibility have to be interpreted with caution. This situation, an anormal distribution, is not uncommon in proficiency tests.

APPENDIX 5: CERTIFICATE OF IIS REFERENCE MATERIAL

O-XYLENE

Certificate of Analysis

Reference Material DSOX-041096

o-Xylene for purity determination by GLC

Reference Material DSOX-041096 consists of a 60 ml crimp capped vial, containing approximately 55 ml of high purity o-xylene. This RM is intended primarily as a quality control material which should be used in gaschromatographic methods for determination of the purity of o-xylene.

Certified Concentrations of Impurities

Certified concentrations of some impurities in %m/m are given in table 1. The certified values in table 1 are derived from the gas chromatographic results obtained from an international interlaboratory study in which 21 laboratories participated. The results of this interlaboratory study are presented and discussed in the iis report iis96C03-RM.

Table 1. Certified values^b for DSOX-041096.

<u>Compound</u>	<u>Concentration (%m/m)^a</u>
p-xylene	0.136 ± 0.004
m-xylene	0.778 ± 0.023
ethylbenzene	0.0127 ± 0.0009
isopropylbenzene	0.426 ± 0.012
styrene	0.0083 ± 0.0008
ethyltoluenes	0.0117 ± 0.0006
n-propylbenzene	0.0121 ± 0.0009

a) The estimated uncertainty is given as 95% confidence limits

b) Also the following compounds are present in this RM. The concentrations of these compounds are not certified, but for indication only:

o-xylene 98.56 ± 0.03 ; other aromatics 0.0016 ± 0.0011 and nonaromatics 0.046 ± 0.009

NOTICE AND WARNINGS TO USERS

Shelf life: When stored properly and as long as the vial cap is undamaged, the expected shelf life of this RM is two years after preparation and may be extended after reanalysing and is reapproved on August 15, 2007. The shelf life is reset to August, 2009

Storage: Sealed vial, as received, should be stored in the dark at a temperature between 10-30 °C.

Suggested procedure for preparing a quality control sample: The following procedure provides purity determination in accordance with ASTM D3797:1995.

1. Allow vial to equilibrate at a temperature of 23 ± 3 °C and shake for one minute.
2. Open the vial and transfer an amount of RM into a volumetric flask of 50 ml.
3. Adjust the volume to 50 ml with RM from the vial.
4. Add a known amount of the internal standard to be used into the volumetric flask and homogenise the mixture well.

Toxicity: This RM consists of o-xylene, which is considered to be harmful; therefore, care should be exercised during handling and use. Use proper methods for disposal of waste.

Spijkenisse, The Netherlands
Reapproved: September, 2007

dr. R.G. Visser
Institute for Interlaboratory Studies

APPENDIX 6: CERTIFICATE OF IIS 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, Napthalenes, 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.

<u>Parameter</u>	<u>Certified value^a</u>	
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 Sulphur, %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: When stored properly and as long as the vial cap is undamaged, the expected shelf life of this RM is two years after preparation and may be extended after reanalysing and is reapproved on August 15, 2007. The shelf life is reset to August 2009

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
 Reapproved: September, 2007

dr. R.G. Visser
 Institute for Interlaboratory Studies