NATIONAL INSTITUTE OF PUBLIC HEALTH AND THE ENVIRONMENT

BILTHOVEN, THE NETHERLANDS

RIVM-report no. 219101005

Statistics and the assessment of interlaboratory studies

R. Hoogerbrugge[1], G.H.M. Counotte[2],
C. Maas[3] *, J.F.N. Maaskant[4], R.C. Schothorst[1],
O.S.N.M. Smeets[5] and F.J. Spruit

November 1996

EURACHEM Nederland, task group: Statistics and assessment of interlaboratory studies (part of the working party on "Interlaboratory Studies").
1) National Institute of Public Health and the Environment (RIVM), Bilthoven
2) Animal Health Service Netherlands, Deventer
3) Inspectorate for Health Protection, Zutphen
4) Institute for Inland Water Management and Waste Water Treatment (RIZA), Lelystad
5) Laboratory of the Dutch Pharmacists, Den Haag

* Present address: Head Inspectorate for Health Protection, Rijswijk. Is replaced by: R. van Dijk, Inspectorate for Health Protection, Groningen (June 1993).

This research was carried out on behalf of the Directorate-General for Environmental Protection from the Ministry of Housing, Physical Planning and Environment (project 219101).
RIVM P.O Box 1 3720 BA Bilthoven, The Netherlands tel. 030 - 2749111, fax 030 - 2742971

MAILING LIST

1-3	Directoraat-Generaal Milieubeheer, Directie Stoffen, Veiligheid en Straling
4	Plv. Directeur Generaal Milieubeheer
5	Drs. R. Wismeyer (Actieprogramma ANVM, commissie Kwaliteit, secretaris)
6	Drs. H. Janssens (Alcontrol Henrici)
7	M. Lauwers (AOAC International)
8	Dr. I.C. Dijkhuis (Apotheek Haagse Ziekenhuizen)
9	A.M. Mos (C.N. Schmidt B.V.)
10	R. Vos (C.N. Schmidt B.V.)
11	Drs. E.P.Th. Ruwiel (DGM/Bodem)
12	Mw. Drs. M.N.E.J.G. Philippens (DGM/SVS)
13	Dr. G.J. de Groot (ECN)
14	Dr. E.W.B. de Leer (Eurachem Nederland, secretaris)
15	Drs. W. Oussoren (Eurachem wg-4, secretaris)
16	Bibliotheek (Gezondheidsdienst voor Dieren)
17	KAM-Dienst (Gezondheidsdienst voor Dieren)
18	Ir. B. Nijenhuis (Gist Brocades)
19	Dr. R.A. Baumann (HIGB-werkgroep OVR, secretaris)
20	R. van Dijk (IGB-Groningen)
21	Dr. W.P. Cofino (IVM)
22	Dr. M.L. Beekes (KEMA)
23	Drs. M.A.F.P.van Rooij (KIWA)
24	Dr.Ir. V.J.G. Houba (LUW)
25	Dr. M.A.J.van Montfort (LUW)
26	Dr. J. van der Meer (NIZO)
27	Dr. E.W.B de Leer (NMi Van Swinden Laboratorium B.V.)
28	Dr. T.M. Plantenga (NMi Van Swinden Laboratorium B.V.)
29-40	D. Hortensius (NNI, beleidscommissie Milieu, secretaris)
41	P.J. Jenks (Promochem)
42	R. D. Rucinski (R.T.C.)
43	Ir. H.A. Deckers (Raad voor Accreditatie)
44	Dr. Ir. J.G. Leferink (Raad voor Accreditatie)
45	Ing. W.J.G. Oosterveen (Raad voor Accreditatie)
46	Dr. J. de Boer (RIVO)
47	Drs. J.P. Baarsma (RIZA)
48	Bibliotheek RIZA
49	Ir A.B. van Luin (RIZA)
50	Dr. R. Visser (SGS-Ecocare)
51	Drs.R. Bosman (TNO)
52	Dr.J. Kragten (UvA)
53	Drs. G.H.J. Reimerink (VSL, secretaris)
54	Depot van Nederlandse publicaties en Nederlandse bibliografie
55	Directie RIVM
56	Ir. H .P. van Egmond
57	Ir. J.J.G. Kliest
58	Dr. H.A. van 't Klooster
59	Dr. W.H. Könemann
60	Drs. A.K.D. Liem
61	G.M. Overvliet

62 Dr. R.W. Stephany
63 Ir. H.J. van de Wiel
64 Dr. P. van Zoonen
65 Hoofd Bureau Voorlichting en Public Relations
66-90 Auteurs
90 Bureau Rapportenregistratie
91-92 Bibliotheek RIVM
93-123 Bureau Rapportenbeheer
123-150 Reserve exemplaren

CONTEXT AND ACKNOWLEDGEMENTS

After presentation of a draft version of this report, at the 1993 Eurachem meeting on proficiency testing (Noordwijkerhout) valuable comments were received, among others, from Ir. B. te Nijenhuis (Gist Brocades) and Dr. M.A.J.van Montfort (LUW) which were incorporated. On the basis of discussions in the Dutch working party on Interlaboratory Studies the report was expanded with examples. Three examples, on a particular subject, are presented in the annexes. The authors would like to acknowledge Dr. R. Visser (SGS-Ecocare) for the preparation of Annex 1. The publication of this paper as Eurachem document is in preparation. Parallel to this process the document is published as a report of the authors' institutes.

CONTENTS

ABSTRACT

A study of the various types of statistical methods used for interlaboratory studies is presented. Where possible reference is made to current standards and guidelines in this area. From the discussions of these standards and practical experiences it became apparent that there is no ideal statistical method for interlaboratory studies in general. Choices have to be made always and should be guided by the objective of the study. Apart from this main objective, the extraction and presentation of data obtained from the results, to support the participants in assessing and improving their activities, is always important. Hence, references to alternative calculation and presentation methods are included.

Many aspects, particularly the evaluation of the results of interlaboratory studies, have been discussed against this background. The use of outlier tests versus robust statistics and the presentation of results received particular attention.

SAMENVATTING

Dit rapport is een verslag van een studie naar de diverse vormen van de statistiek toegepast in interlaboratorium onderzoeken. Hierbij wordt zoveel mogelijk verwezen naar bestaande normen en guidelines op dit gebied. Bij de bespreking van deze normen in relatie tot ervaringen opgedaan in de praktijk bleek telkens dat de ideale statistiek voor interlaboratorium onderzoeken in algemene zin niet bestaat en dat telkens keuzes moeten worden gemaakt waarbij het doel van het onderzoek als een rode draad door de gemaakte keuzes loopt. Afgezien van dit concrete hoofddoel, wordt het extraheren en presenteren van informatie uit de resultaten, om de deelnemers zo goed mogelijk in staat te stellen om de kwaliteit van hun werk te beoordelen en te verbeteren, altijd van belang geacht. Derhalve zijn diverse verwijzingen naar alternatieve berekenings- presentatievormen opgenomen.

Met deze achtergrond zijn vele aspecten van met name de evaluatie van de resultaten van interlaboratorium onderzoeken behandeld. Hierbij zijn de toepassing van uitbijtertesten versus robuuste statistiek en de presentatie van de resultaten belangrijke onderwerpen.

1. INTRODUCTION

1.1 The objective of interlaboratory studies

The relevance of the results of chemical analyses depends greatly on its quality. The most appropriate method to assure the quality of the results is to establish an operational quality system which is assessed by an independent body. Such an approach is essential to prevent errors and to limit the intralaboratory dispersion of the results in particular. However, to estimate uncertainties comparison with standards or with the results of other laboratories are necessary. A number of instruments to determine and improve this comparability exist; however they always depend on the results of interlaboratory studies. Instruments are: reference samples, validated methods and interlaboratory proficiency studies. Each of these instruments imposes its own particular requirements on the interlaboratory studies. Thus, there is no one optimum solution for certain problems and choices; this will often depend on the objective of the study such as laboratory performance, material certification and method performance studies.

1.2 Aims of Eurachem and Eurachem-Nederland

Eurachem is:

- a cooperative network of European laboratories;
- a forum for discussions on common problems, solutions and strategies;
- a framework for laboratories to develop agreements and to compare results;

to promote:

- the reliability of chemical analyses and the international acceptance of analysis results;
- interest in quality issues in laboratories;
- the application of internationally accepted quality standards;
- the development and application of validated methods;
- traceability of measurement results, based on reference materials (standards);
- the development of mutually recognised methods for the quality management of analytical methods, e.g. by participation in national and international interlaboratory studies;
- the awareness of users of chemical analysis information of the objectives (including relevance, sensitivity, precision and uncertainty).

Eurachem was established in 1989 and its current membership is spread over 17 countries. Eurachem-Nederland was established in 1990 and now includes over 100 delegates who provide a good reflection of the chemical analysis sector in the Netherlands. Many members are active in working groups on publicity and information, education, reference materials, calibration and traceability, provision of information and quality management, sampling, validation, harmonisation of quality systems and interlaboratory studies. The working group on interlaboratory studies has set up a task group on "statistics and the assessment of comparative studies". The now reported study was carried out by that task group.

1.3 Background and motivation

Participation in interlaboratory studies is important to maintain and improve the quality of chemical analyses. A set of documents was written on the organization of interlaboratory studies. These could, or should, be used as guidelines [1-5]. Guidelines have also been published on the statistical evaluation of results. However, in the founding meeting of the Eurachem-Nederland working party "Interlaboratory Studies" on 20 May 1992 it was felt that the statistical interpretation of the results does not always reveal all relevant information present in the results. This may be due either to the statistical methods used or to the presentation of the results. A specialised and rather sensitive element of this is the assessment of the proficiency of laboratories on the basis of the results of a comparative study. Initially two task groups were set up within the working group; one

on statistical analysis and one on assessment. When defining the programme of activities it was found that a good evaluation depends greatly on a range of parameters such as the objective of the study, quality of the results, number of participants and type of comparative study. In this light the separation outlined above did not appear to be attractive and was therefore not implemented.

To structure the study the initial activity was the collection and comparison of the relevant characteristics of a number of interlaboratory studies. The criteria for the selection of these comparative studies were that they had to be carried out regularly and have a fair number of participants (>20). The members of the task group immediately provided a reasonable distribution across a wide range of analytical and clinical chemistry. A number of premises on interlaboratory studies were defined based on in-house experience of interlaboratory studies and literature information. These premises supplement the current extensive guidelines.

As a result, during the discussions, it appeared that various organisers had regularly stated their intentions to modify the structure, evaluation or reporting of interlaboratory studies. In addition to statistical aspects this concerned other elements associated with the organization of interlaboratory studies. Thus, this report covers a wider range of subjects than suggested by the title.This report does not pretend to be a standard or specification. Its primary aim is to provide a source of information to organisers of and participants in interlaboratory studies.

The items in this report are classified into relevant to:
- the organization of interlaboratory studies;
- statistical analysis and assessment;
- reporting on interlaboratory studies.

The discussion of statistical matters in particular is based many references which may assist the organisers of interlaboratory studies to find alternative methods to reveal the information hidden in the results of their work.

2. ORGANIZATION OF INTERLABORATORY STUDIES

This chapter covers some aspects of the organization of interlaboratory studies. As some of these are not related to the primary objective of the comparisons they are described shortly.

2.1 Objective (type)

The studies may have one of three objectives:

- Proficiency Testing (PT), [1,5] or laboratory performance study.
- Certification of Reference Materials (RM), [2] or material certification study
- Method Validation (MV), [3] or method performance.

They may also be classified as:

- a series of interlaboratory studies to improve quality;
- a single study to determine comparability (these two options may apply to all three types);
- a special type of proficiency testing: interlaboratory studies to determine the quality of laboratories and to use this information, e.g. for a series of assignments.

Recommendation: When introducing a study or inviting participants the objective should be stated clearly. From the onset it should be clear whether the intention of the initiator is to organize a series of interlaboratory studies or the results will be used to assess the quality of the laboratories and it's consequences. The type of study should also be indicated in the name. Especially for reference material certification and method validation, the analyses should be carried out by qualified analysts. In case of method validation, data about the robustness of the method should be available.

2.2 Homogeneity

The protocol should describe who will prepare the sample and how this will be done. The way in which homogeneity is assured should also be described. If this is not known from earlier studies or from the literature, homogeneity tests will have to be undertaken. For all types of interlaboratory studies the organisers protocol should include criteria for homogeneity [3].

2.3 Stability

The protocol should define the stability criteria applied. Stability tests will have to be carried out if the stability is not known from earlier studies or the literature. The outcome of the stability tests should be used as a basis for measures affecting the transport of the samples. Samples for PT and MV comparative studies should remain stable at least from the time of preparation through to analysis (based on the above criterion). Samples for RM should remain stable until the claimed date of expiration.

2.4 Instruction

When the samples are mailed, instructions will have to be given on at least the following aspects: reporting of the results (how many, reporting of results below the detection limit, how many decimal places, should outliers be reported?) sample treatment and where relevant the analysis method (depending on the type of study). The objective of the study should also be reiterated. Instructions should correspond to the objective of the study. Preferably results should be entered on a standardised form. In MV studies the described method should be strictly applied. When this method contains certain degrees of freedom the choices are preferably reported to enable a statistical analysis on the justification of these degrees of freedom.
In PT a clear choice should be made between reporting on a standard form or reporting according to the laboratories procedures.
If the study includes a learning stage the organiser can provide pilot samples. The protocol should

identify the results which will be included in the statistical analysis and what will be done with results not included in the analysis.

2.5 Assessment per laboratory

An interlaboratory study could be organised with the objective of comparing laboratories on the basis of their results (proficiency testing). Such an assessment might have few consequences and could be used to inform laboratories of incorrect results. However, the assessment might can also have important consequences, for example, if the results are used by the initiator to select a contract laboratory or when the initiator is in a position of authority over a group of laboratories.

Recommendation: Before the study commences it should be indicated clearly whether or not there will be an assessment. If there is to be an assessment its consequences (e.g. assignments from a customer), criteria and who decides in the event of problems need to be defined.

2.6 Discussion

By organising meetings with participants the initiator will have an opportunity to discuss matters relating to the interlaboratory study in detail. They can be very useful for all types of interlaboratory studies. The frequency of these meetings will depend on the need for them. They could be held before the samples are mailed or after the initial or final evaluation of the results.

2.7 Confidentiality (distribution)

It is recommended that those not directly involved in the study should not have access to its results [5]. In RM the concluding results including the methods applied should be reported to describe the quality of the material and it's assignment. In MV future users of the method should have access to a summary of the results. In PT an agreement should be made on whether a summary is to be public.

2.8 Participants

It is recommended that the laboratories addressed by the interlaboratory study are clearly identified. It should also be clear whether or not the study is open to any laboratory in the target group.

2.9 Period available for the analysis

For planning reasons a period should be set for each interlaboratory study which should be referred to in the letter accompanying the sample. Alternatively, annual plans could be developed.

2.10 Availability of materials

Also for PT and MV it is recommended that the organiser of the interlaboratory studies should have a sufficient quantity of the material available to supply samples after the study. This would provide laboratories whose results are incorrect with an opportunity to trace the cause of the errors.

3. STATISTICAL ANALYSIS

This section covers some aspects relevant to the statistical analysis of interlaboratory studies.

3.1 Screening

Are the analytical results received by the organisers of the interlaboratory study screened on obvious discrepancies? Is the lab in question informed if evident discrepancies occur? Will the laboratory be given the opportunity to improve the results of the analysis? Will any corrections be included in the final report?

Recommendation: Sometimes the body organising the interlaboratory study will screen the results of the analysis on major errors. Common major errors include:
- conversion of the results to the units used in the interlaboratory study;
- transcription errors.

The extent to which correction of errors is permitted or desirable depends greatly on the objective of the study. For example, no corrections at all may be permitted in the strictest form of proficiency testing, which includes the required reports. However, when reference materials are being compiled correction or exclusion of almost all major errors referred to above would be desirable. Where applicable it is advisable to include the following in reports on interlaboratory studies:
- reasons for discarding results;
- results which were corrected.

3.2 Normality test and outliers

Most of the common statistical analysis methods are based on the assumption that the data has a normal distribution. There are several methods to look for non-normality and to promote normality:
- test on the distribution of the results
- outlier test (in contrast with a normal distribution)
- transformation to normality (e.g log).

Recommendation: In classical statistics extreme values are removed on the basis of outlier tests. In robust statistics the influence of extreme values is limited by the statistical method. In both cases it has to be considered whether, in view of the objective of the study, it is justified to minimise the influence of extreme values. In MV studies, extreme values may be due to failure to follow the analysis protocol exactly. The unjustified minimisation of the influence of extreme values results in an impression of the performance of the method or the laboratories which is too optimistic. The strategy on outlier testing is preferably chosen before the samples are distributed especially when on the basis of the results contract laboratories will be selected.

When the influence of extreme values is reduced the quality of the laboratories will appear to be higher, particularly the quality of the results which could be obtained. As a result laboratories which differ from the majority stand out. This may be decisive if the objective of the interlaboratory study or series of interlaboratory studies is to improve comparability.

3.2.1 Normality test

When normality of the results is expected the normality of the distribution can be tested for example with the Kolmogorov-Smirnov-Lilliefors or the Shapiro and Wilk test.

Recommendation: Specify the action which will be taken when it is concluded that the data does not have a normal distribution. A series of normality tests and associated reactions has been defined in the form of an expert system [6].

3.2.2 Outliers

An outlier is a result or set of results that is markedly different from the majority of the other results.

Two types of tests are used when assessing the results of analyses of identical samples:

1. Assessing the mean of the results from one laboratory by comparison with the overall mean.

2. Assessing whether the variance within a laboratory differs from the variances of the other laboratories.

ISO 5725-2 (1994) [4] specifies the use of outlier-tests: Grubbs for means and Cochran for variances. The grubbs test has the masking effect as a negative point.

IUPAC/AOAC [1] refers to [11]. The organiser should choose between Dixon and Grubbs outlier tests or the application of robust statistics, i.c. Huber [14]. The Huber technique is one of the many techniques which behave well even for long tailed distributions.

In the event that outlier tests are employed the following assumptions are made:
- The data has a normal distribution.
- The intralaboratory variances are assumed to be homogeneous, see e.g. [7].

Extreme values can be identified and deleted on this basis.

The following objections can be made against these assumptions:
- The data will seldom have a normal distribution. The number of observations in the tail of the distribution is generally higher than would be expected given a normal distribution.
- Observations will be skewed, particularly near the detection limit.
- Intralaboratory variances may not be homogeneous.

The reliability of the outlier tests is doubtful given these objections.

Recommendation: After applying the outlier tests described above it is advisable to verify whether the assumptions are true, for example by:
- testing for normality,
- testing the homogeneity of the variances (Bartlett) [7].

The outliers should be rehabilitated when the remaining data is not behaving according to the assumption of a normal distribution or homogeneity of variances.
The fact that an outlier has been skipped, always has to be reported (and the background for getting an outlier has to be investigated).

3.2.3 Robust statistics

The use of robust statistical methods provides an alternative to the deletion of outliers. The influence of extreme values is reduced without actually removing them; see [11-14].

Median

The **median** (MED) and **median of absolute differences** (MAD) are a robust alternatives to the mean and standard deviation. In the large international schemes on analysis in plants and in soil [8-10] the calculation of MED and MAD is applied.

Advantages:
- low sensitivity to extreme values (25% of the data at each side have no influence on MED and MAD);
- easily calculated.

Disadvantage:
- sensitive to minor errors e.g. extensive rounding in the middle value or middle two values have an effect on the median;

Class of redescending estimators:
Both an estimate of the assigned value and a measure of the dispersion are measured. The influence of extreme values on the final assigned value is reduced by a weighing factor. Examples include the Andrews sine wave, Tukey's biweight, Hampel, Tanh estimator.

Advantages:
- low sensitivity to extreme values;
- insensitive to distributions where observations "in the tail" are over-represented;
- the information contained in the data is used effectively.

Disadvantages:
- the iterative calculation is a minor disadvantage since modern computer facilities enable such calculations quite comfortably.

Huber estimator:
The application of the Huber estimator (c = 1.5) for interlaboratory studies is described in [11]. In these two papers the application of these type of robust statistics is compared with traditional methods including notable remarks on the interpretation. The calculation of the expected value and variance on page 1699 is incorrect. A sound procedure is described by Huber [14].
Advantages:
- very effective use of the information in the data;
- limited sensitivity to extreme values.
- handles long tailed distributions well.

Disadvantage:
- the iterative calculation is a minor disadvantage since modern computer facilities enable such calculations quite comfortably. Calculating the mean and standard deviation iteratively can result in a unrealistic low estimate of the standard deviation (Cofino, personal communication).

3.2.4 Logarithmic transformation
A logarithmic transformation can be used when a log-normal distribution of the results is expected. In that case confidence intervals can be calculated for the interpretation of the result. The logarithmic mean and standard deviation can also be calculated. The standard deviation is an approximation of the coefficient of variation of x.

3.3 Standards
It is obvious that the organization of the interlaboratory study and the statistical parameters and their calculation have to be decided in advance. Generally an accepted standard or recommended method will be followed. This could be ISO 5725 or another defined standard.

Recommendation: The use of a recognised and defined standard is recommended, e.g. ISO 5725 [4] or IUPAC/AOAC [1]. It will have to be demonstrated that the protocol used corresponds with the objective of the interlaboratory study (see paragraph 3.1) .

Comparison between ISO 5725 and IUPAC/AOAC

ISO 5725 draft 1990	IUPAC/AOAC	Comments
Single and multiple analyses	This protocol does not provide information about multiple analyses	
Method evaluation	Laboratory evaluation	
Assessment of intralaboratory variance: - Mandels k - Cochran	not applicable	Mandels k and Cochran are algebraically related
Assessment of the measurements of a laboratory: - Mandels h - Grubbs	 z score Grubbs, Dixon or Huber	z score with statistically calculated X and S corresponds to Mandels h Grubbs is preferred over Dixon [15]. In algebraic terms Grubbs and Mandels h are identical.

IUPAC/AOAC allows the option of calculating the z score on the basis of a reference value and a target variance. When assessing laboratories it is then assumed that in 5% of all cases $|z|>2$ and in 0.3% of cases $|z|>3$. This is only valid if the statistically calculated expected value and variance are used and if the data has a normal distribution.

3.4 Repeatability and reproducibility

The following items are generally applied to describe the result of an interlaboratory study [3]:

S_w intralaboratory standard deviation, calculated separately for each laboratory

S_r repeatability expressed as standard deviation, calculated by combining the S_w values

S_L interlaboratory standard deviation, based on the laboratory means with a correction for S_r

S_R reproducibility expressed as standard deviation, calculated by: $(S_R)^2 = (S_r)^2 + (S_L)^2$

Relative standard deviations, RSD, are calculated as the standard deviation divided by the average

3.4.1 Repeatability
The value below which the absolute difference between two single measurements obtained under identical circumstances affecting the laboratory, operator, equipment and time interval is expected with a probability of 95% (ISO 5725). The repeatability, r, is 2.8 * S_r in case of normality.

3.4.2 Reproducibility
The value below which the absolute difference between two single measurements obtained under different circumstances affecting the laboratory, operator, equipment and time interval is expected with a probability of 95% (ISO 5725). The reproducibility, R, is 2.8 * S_R in case of normality.

3.4.3 Reproducibility test
- Assessment criteria may be based on the objective of the analysis or statutory or other regulations. They could be specified as target values.
- The reproducibility could be compared with reproducibility in similar studies. Very general correlations such as Horwitz [16] or specific historical data as used in Grol plots [17] can be employed.

Recommendation: The reliability of the calculated repeatability and reproducibility depends greatly on the number of participating laboratories. This was demonstrated by Karpinski [18]. Clearly, this will have to be borne in mind when setting the targets. In addition to comparing the reproducibility with a target, the reproducibilities of different analysis methods could be compared. A subtle MV example of this was described by Martin and Soliman. The only difference between the methods studied by them was the number of points included in the calibration line [19]. On this basis a statistically supported recommendation could be made to reduce the number of points on this line from 10 to 5, without loss of accuracy.

3.4.4 Relation between repeatability and reproducibility
A analysis of variance is generally carried out to calculate the intralaboratory variance and interlaboratory variance. The F-test is used for inference or the existence of an interlaboratory variance. If the between laboratory variance is significant then results of laboratories are expected to differ seriously. If the between laboratory variance is not significant it is, in statistical terms, irrelevant which laboratory carries out the analyses. The F-test can be applied if the distributions are normal and the intralaboratory variances are homogeneous. A robust version has been described which can be used to estimate intralaboratory and interlaboratory dispersion [20].

An alternative method for variance analysis, which can be applied to unbalanced data sets (e.g. due to missing data) was described by Crowder [21].

Alternatively a randomisation test could be used [22]. This method does not depend on the statistical distribution of the data.

The link between repeatability and reproducibility can also be studied by using Intermediate Precision Measures (IMP) [4-part 3, 23]. This takes the number of factors (operator, equipment and time) which have changed into account. Another interesting intermediate form is the use of samples with approximately, but not exactly, the same concentration. If the results associated with both samples are plotted against each other in a Youden plot [24] a fairly good indication of the random or systematic nature of the errors of a given laboratory is obtained.

If a series of parameters is measured of all samples in an interlaboratory study it may be very informative to study the results associated with these parameters as a group rather than individually. Several multivariate techniques are available for this [25,26]. An example is the Principal Component Analysis which offers the opportunity for a optimal graphical presentation of the laboratories with respect with each other. In a separate plot the differences between the laboratories can be interpreted in terms of which parameters are relatively high for which laboratories. This technique is for example applied on interlaboratory data set on PCB analysis [27].

3.5 Reference value

This is the content of the substance being determined in the matrix used to assess the results of the various laboratories. This may be a true value (seldom if ever known), an assigned value (the concentration determined independently of the interlaboratory study which provides the best possible approximation of the true concentration) or the average concentration across all laboratories (consensus value). The average concentration could be determined by several methods and procedures to obtain a better approximation of the true concentration.

Even if the reference is obtained from the interlaboratory study itself, e.g. when greater priority is given to comparability than to accuracy, it may still be possible to obtain an external assessment of the reference.

Recommendation:

When organising interlaboratory studies it is advised to use samples whose concentration is known before the study commences. A single sample whose true concentration is known is preferred. If the true content is not known the concentration should be determined independently before the start of the study. According to [1] the results of the participating laboratories should be assessed to this concentration (target value, using a target variance).

Particularly during MV studies it is advisable to test the correctness of the reference, independently of the interlaboratory study. If certified material is used there is no need to make a new determination. However the certification method should be specified in the documents accompanying certified material.

In RM studies a variety of different techniques should be used to prevent method related biases. The best estimate of the true value of the substance being analyzed will be the (weighted) mean.

In PT studies application of a variety of techniques is also recommended. Despite these precautions to prevent bias one should realize that reference values always contain some uncertainty.

There are several options to compare the measured values with the reference:

- application of a t-test;
- application of a randomisation test;
- graphical method. It may be practical to plot both the mean, the externally determined true value and the individual measurements in a single graph.

3.6 Comparison of methods

In interlaboratory studies intended mainly for proficiency testing of laboratories, each using their own methods and procedures, it may be useful to use statistical methods to assess the differences between the methods.

Recommendation:
It is recommended that information about the methods used is requested from the participants. It should be tried to identify critical steps in the methods before the study commences. One problem with the analysis may be that the number of combinations of critical settings easily becomes unmanageably large. Variance analysis (ANOVA or MANOVA) is a statistical tool to deal with this problem. It permits simultaneous tests on multiple variables to determine whether there are significant differences in the data sets [28]. Anova and Manova are applicable if the condition of a orthogonal design is met. Some comparative studies provide information about the methods used in the form of codes. At present this is not normally subjected to statistical analysis.

A suitable strategy may be to request certain types of raw data when the analysis results are submitted. Normally this data will not be used, unless significant differences are later found between the various laboratories. If the organization also aims for quality improvement the best possible expertise in the area concerned will be needed.

3.7 Laboratory assessment

The criteria used to assess laboratories have to be known in interlaboratory studies used to provide information about the performance of laboratories. These could include an assessment on the basis of accuracy (i.e. trueness and precision).

Recommendation:
The assessment of the trueness (deviation from the true value) should be based on reference material. This should be a material with a concentration similar to that to be determined and with the same matrix. The precision (accuracy) should preferably be expressed as repeatability, intralaboratory reproducibility and reproducibility. If there is a correlation between the concentration and the accuracy this correlation should be described. If no suitable reference material was available, the way in which the reference value was determined should be specified. For all types of reference values the uncertainty in the value should be considered in the laboratory assessment.

If a comparative study is organised to determine, on behalf of a future customer, which laboratories are capable of undertaking certain analyses, it is recommended that clear agreements about the criteria to be applied are made before the start of the study. Given the potential major financial stakes the responsibilities of the customer and those of the organiser should be defined clearly. Normally the organiser should be independent of both the customer and the participants and should report the findings to the customer who will decide independently on the basis of these findings.

4. REPORT

4.1 Reporting on results below the lower detection limit

When analyzing parameters close to the lower detection limit some results may be reported as being below the lower detection limit. To interpret parameters such as the mean, dispersion and outliers it is important to know how these results were dealt with.

Recommendation:
The report on the original results should clearly indicate those reported as "below the limit" or which were considered as such by the organiser. Analysis reports just below the lower detection limit which would normally be reported as "below the lower detection limit" when reporting individual results, can still be informative. To assess the quality of such results the participants could be asked to report the calculated result as well as "below the limit". The reason for this is that an individual result may be subject to other quality requirements than a series of results.

4.2 Reporting on individual results

The primary parameters of interest to the organisers and participants should be calculated on the basis of the original results reported.

Recommendation:
It would be very desirable to report the actual original results, for example, to have the opportunity to check that all results were correctly entered in the organiser's data set. Except for very large data sets this requirement is also included in [1]. Another advantage of including the analysis results is that the participants can, if they wish, calculate alternative parameters or interpretations. This may assist in finding the cause of systematic deviations.

4.3 Reporting on outliers

The report should clearly identify the outlier tests used, the method used to deal with outliers and whether or not outliers can be recognised as such.

4.4 Reporting method

By reporting the methods of analysis (including pre-treatment) used by the participants the participants can determine whether there is any pattern in the errors observed.

Recommendation:
Such a report enables the participants in PT studies to trace major method-dependent differences. In MV-studies this option is useful when the method allows choices. When statistical methods are applied to study the difference between various analytical methods the results should be described in terms understandable to participants.

4.5 Reports on the comments of participants

Where possible the comments of participants, which may be relevant to the interpretation of the results, should be included.

5. RECOMMENDATIONS AND CONCLUSIONS

When faced with any of the many choices during the organization and evaluation of an interlaboratory study the *objective* of the study should always be considered.

Even if the objective referred to is different from that of a given study it is always advisable to provide the participants with the greatest amount of information possible which may help them to improve the quality of their analyses.

A wide range of statistical tools has been described, only some of which are used to extract information from the results of interlaboratory studies. Clearly, this information will only be useful if it is presented in a way the participants and other interested parties can understand.

6. REFERENCES

[1] Thompson M. and Wood R., "The international harmonised protocol for the proficiency testing of (chemical) analytical laboratories". ISO/IUPAC/AOAC, Pure & Appl. Chem. 65 2123-2144 (1993) (also ISO/REMCO N280, august 1993).

[2] ISO 35, "Certification of reference materials - General and statistical principles" (February 1985)

[3] ISO 5725, "Precision of test methods", International Organization for Standardization, (1986)

[4] ISO 5725-(1,2,3,4,6), "Accuracy (Trueness and Precission) of Measurement Methods" (1994)

[5] WELAC Criteria for Proficiency testing in Accreditation, WELAC WGD 4, European Laboratory Accreditation Publication ELA-6. (september 1993).

[6] Danzer K., Wank U. and Wienke D.,"An Expert system for the evaluation and interpretation of interlaboratory comparisons', Chem. and Intel. lab syst. 12 (1991) 69-79.

[7] Dixon W.J. , Massey F.J., Introduction to statistical analysis, McGRAW-HILL, 1969.

[8] Houba V.J.G. "International Plant-Analytical Exchange", Wageningen Agricultural University, Annual Reports.

[9] Houba V.J.G. "International Soil-Analytical Exchange", Wageningen Agricultural University, Annual Reports.

[10] Montfort M.A.J. van, "Statistical remarks on round robin data of IPE& ISE", Wageningen Agricultural University, Technical Note 92-02 (1992).

[11] Analytical Methods Committee, Analyst 114, (1989) p1693-1702. Part 1. Basic Concepts, Part II. Interlaboratory Trials.

[12] Hampel F.R., ea, "Robust Statistics. The approach based on influence functions" John Wiley & Sons, (1986)

[13] Lischer P., Statistik und Ringversuche, FAC Forschungsanstalt für Agrikulturchemie und Umwelthygiene, 3097 Liebefeld-Bern (in German).

[14] Huber P.J. Robust Statistics, John Wiley & Sons, (1981)

[15] Davies P.L., "Statistical evaluation of interlaboratory tests", Fresenius Z. Anal. Chem. 331 (1988) 513- 519.

[16] Horwitz W., Kamps L.R. and Boyer K.W. J. AOAC 63 (1980) 1344

[17] Groennou, J.Th. and Olrichs Sj.H.H.,"Statistische aspecten aangaande de verwerking van vergelijkend interlaboratorium onderzoek resultaten", KIWA SWE-354, Nieuwegein 1981.

[18] Karpinski K.F. "Reliability of Repeatability and Reproducibility Measures in Collaborative Trials" J. AOAC 72 (1989) 931-935.

[19] Martin J.I. and Soliman A.M., "Interlaboratory Study of Decreasing the Number of Standard Points in the Official Iron Standard Curve" J. AOAC 75 (1992) 384-385.

[20] Thompson M., Mertens B. and M. Kessler, "Efficacy of Robust Analysis of Variance for the Interpretation of Data from Collaborative Trials", Analyst, 118 (1993) 235-240.

[21] Crowder M., "Interlaboratory Comparison: Round Robins with Random Effects", Appl. Statist. 41 (1992) 409-425.

[22] Vankeerberghen P., Vanderbosch C., Smeyers- Verbeke J. and Massart D.L. Chemometrics and Intelligent Laboratorium Systems 12 (1991) 3-13

[23] Miyazu T. and Yamamoto H. in Parkany M. (ed.) Quality Assurance for Analytical Laboratories, The Royal Society of Chemistry, Cambridge, (1993).

[24] Youden W.J. and Steiner E.H., "Statistical manual of the AOAC", Arlington, VA, 1975.

[25] Massart D.L., Vandeginste B.G.M., Deming S.N., Michotte Y. and Kaufmann L., "Chemometrics: a textbook", Elsevier Amsterdam (1988).

[26] Misra R.K., Uthe J.F. and Misial C.J., "Multivariate Analysis of a Round-robin Study on the Measurement of Chlorobiphenyls in Fish Oil", Analyst 117 (1992) 1085-1091.
[27] Boer J. de, Duinker J.C. de, Calder J.A. and Meer J. van de, "Interlaboratory study on the analysis of chlorobiphenyl congeners", J.AOAC 75 (1992) 1054-1062.
[28] Jones N.E., "Multiway Analysis of Variance for the Interpretation of Interlaboratory Studies", Anal. Chem. 62 (1990) 1531-1532.
[29] BCR/48/93 - Guidelines for the production and certification of BCR reference materials.
[30] BCR/139/90 - The certification of the contents of chlorobiphenyls 28, 101, 118, 153 and 180 in waste mineral oil (CRM 420).
[31] A.K.D. Liem, R. Hoogerbrugge, G.U.M. Lindström and J.R. Startin, "Simultaneous Determination of Toxicologically Relevant Chlorobiphenyl Congeners Occurring in Foods: Results from the First Part of a Collaborative Study. To be submitted to Pure & Appl. Chem.

ANNEX 1: HOMOGENEITY AND STABILITY

Ad.2.2 Homogeneity

As the goal of a RM is to produce a reference material the homogeneity assurance here is fundamental. The BCR did put a lot of work in this aspect of the interlaboratory studies. For example the following guidelines are from BCR [29]:

A measurement method must be selected to test the homogeneity. The sensitivity and the repeatability of this method are the relevant selection criteria. Both the bias of the method and the traceability are not very important in this case. The test method should be applied under best repeatability conditions.

In any case the final, packaged form of the material should be sampled. Where inhomogeneity is detected additional sampling at intermediate stages should lead to understanding of the physical phenomena, which caused the inhomogeneity.

The number of samples should allow to demonstrate the within- and between-units homogeneity of the material using a reasonable number of units and measurements. Within-unit homogeneity is more readily detected with lower sample size. It is therefore recommended to use the lowest sample size which allows sufficiently repeatable measurement results with tile chosen test method.

If there is one single batch of material it is sufficient to analyze 5 or 6 sub-samples from each of 2 units to test the within-units homogeneity. To test the between-units homogeneity of the same batch the recommended number is 10 to 30 depending on the repeatability of the test method.

The BCR-protocol does not give any details of the statistical evaluation, but Thompson & Wood recommend the use of one-way analysis of variance, without exclusion of outliers. A complete worked-out example is to be found in appendix II of lit. [1]

Ad.2.3 Stability

One may estimate the stability of the material with the Arrhenius model based on some short-term measurements (often at higher temperatures) and subsequently one may mention an expiry date on the sample material. As the goal of a RM is to produce a reference material the stability aspect in this type of interlaboratory study is very important. The BCR does not want to use an expiry date due to the uncertainty in the predictions. As alternative BCR uses continuous monitoring during the entire lifetime of the CRM [29].

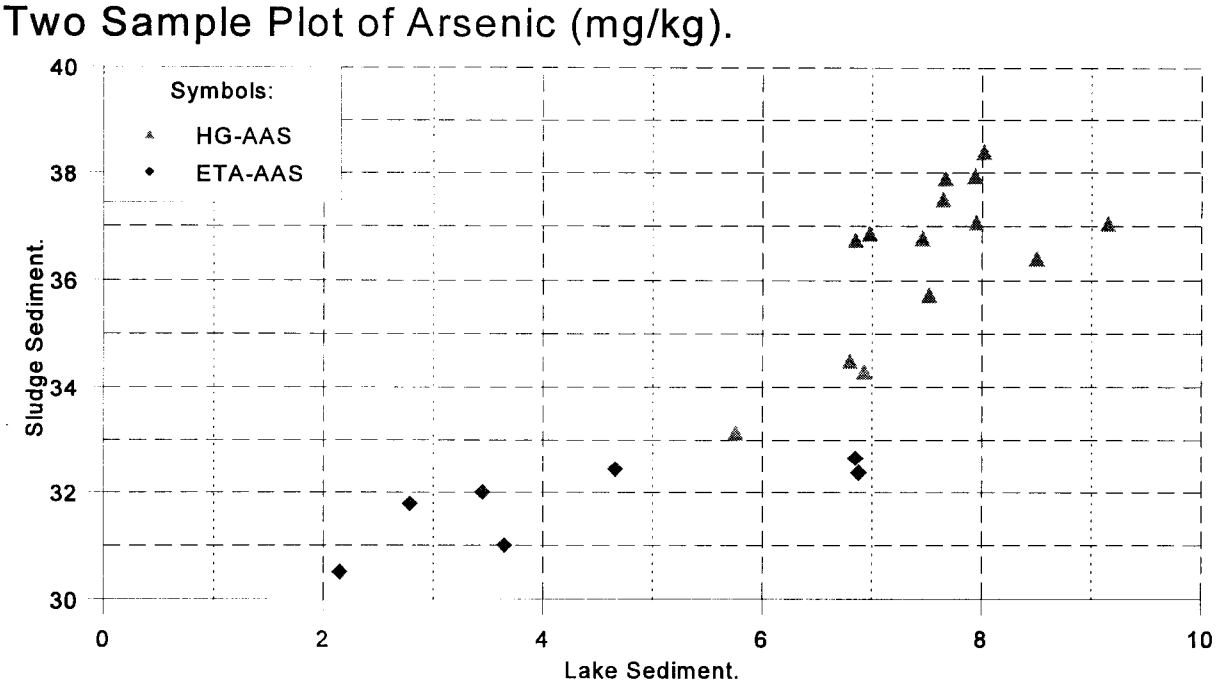
The stability of the material sometimes has to be tested in order to check for its suitability for use. This can be done by storing the material at various temperatures. E.g. the mineral oil used for CRM 420 was stored at -20 °C, +18 C and +35 C [30]. After one, three and six months five subsamples were taken and analysed for PCB's. The mean values and the standard deviations of the various analytes for each temperature and for each period were compared.

ANNEX 2: THE USE OF HOTELLING T^2 AS A MULTIVARIATE STATISTICAL TOOL

A certified reference material is developed in a cooperation between RIZA (Netherlands), Promochem GmbH (Germany), RTC (USA) and CN-Schmidt BV (Netherlands). The material (CRMPR 9472) is a sewage sludge from a publicly owned treatment works, that is representative for an residential area with industrial influence, located in the Western United States.

Assessment of the results for the element Arsenic.
Arsenic was not given a certified value from the RIZA-results. Reproducibility in respect to the repeatability was to large in relation to the other elements. From the observations it is clear that there is a systematic difference between the results of the graphite furnace (ETA-AAS) and the hydride generation atomic absorption technique (HG-AAS). This difference is highly significant according to the t-test ($t = 4.2$, $p = .0038$).

Additional a Youden two sample plot is used in which the individual laboratory results are plotted for the sludge and a sediment sample. The different measuring techniques are given a different symbol. This is shown in the picture below



To test if the mean results for both measuring methods are also statistically different a bivariate approach is used. The Hotelling-T^2-statistic is used to test the null hypothesis:
The mean of each measuring method obtained for each matrix is equal to the mean of the other, where the variance/covariance of both methods are assumed equal but unknown.
The first diagonal element of the variance/covariance matrices is the variances of the results around the method-mean for the lake sediment sample and the other diagonal element for the sludge sediment sample. The covariance elements contain the correlation between the differences of the individual results with respect to the method-means of both samples.
The T^2 statistic is calculated as follows:

$$T^2 = \frac{N_1 N_2}{N_1 + N_2} \boldsymbol{D}^T \boldsymbol{S}^{-1} \boldsymbol{D}$$

in which $\boldsymbol{D}$ is the difference vector for the measuring methods between the different matrices involved. $\boldsymbol{S}$ is the pooled covariance matrix. The critical T^2 is:

$$\frac{q\,(N_1 + N_2 - 2)}{N_1 + N_2 - 2 - q} F_{(q,\, N_1 + N_2 - q - 1)}$$

in which q=2 is the number of samples. The calculated T^2-statistic results to 74.3 ($p \ll 0.001$). The difference between the two measuring techniques is statistically significant. The significant lower analytical results for the graphite furnace technique is probably due to matrix-effects, which inhibit the response for this element.

ANNEX 3: MULTIVARIATE DATA ANALYSIS

The Working Group Halogenated Hydrocarbon Environmental Contaminants (IUPAC; Applied chemistry division; Food chemistry) has initiated a study on the quality of methods for the simultaneous determination of the toxicologically relevant chlorobiphenyl congeners (CBs) in foods. This example shows results from the first exercise of a step-wise designed collaborative study in which seventeen laboratories from eleven countries participated. Results are used of 16 tri- to octachlorinated CB congeners in a standard mixture. Target concentrations were chosen in the range between 10-200 ng/mL, depending on reported levels in 'real' food samples. A detailed protocol was send to all participants including both mandatory and optional procedures for handling of the ampouled standards and instructions for the reporting of data.

Univariate results

Reported data were discussed in a meeting of representatives from the participating laboratories on the basis of a statistical evaluation of within-laboratory repeatability and between laboratory reproducibility. A relative standard deviation for within-laboratory repeatability (RSD_r) of less than 5% was found for most of the congener-specific data reported by the participating laboratories. A relative standard deviation of 20-30% was found for the between laboratory reproducibility (RSD_L). A discussion of the great variety in analytical approaches resulted in some proposals to modify some of the applied conditions. However, differences in the handling of standards was suggested to be the major contributor to the systematical errors observed.

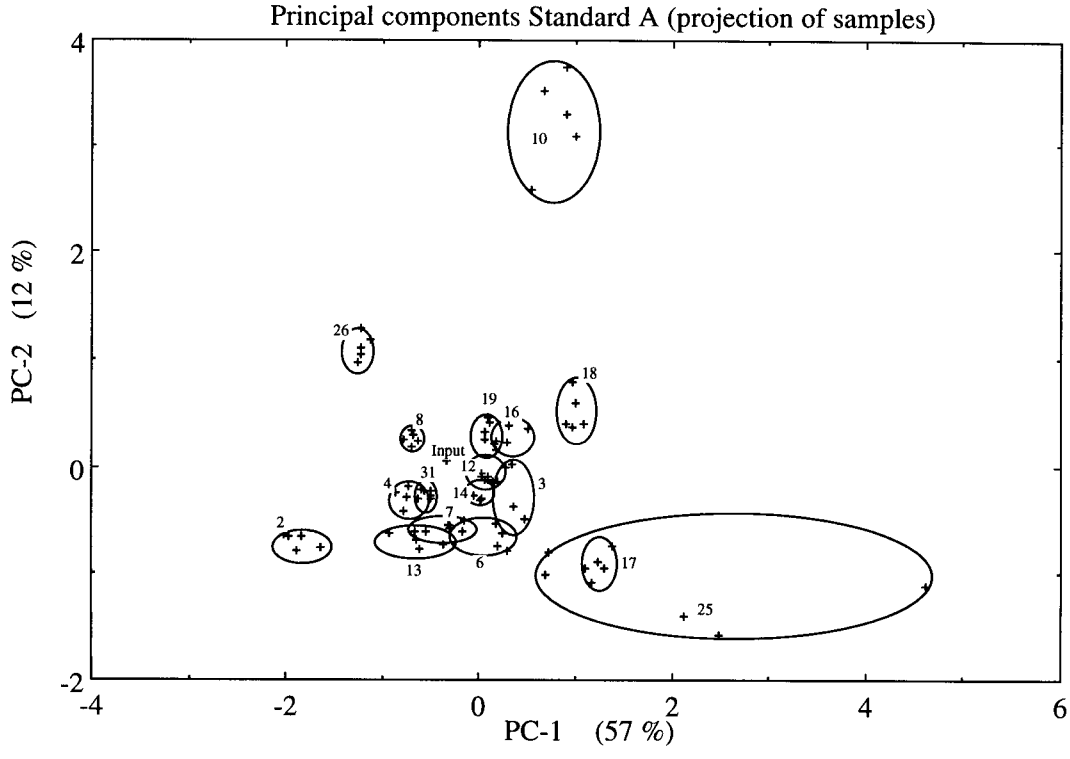


Figure 1. Projection of the results of each laboratory on the first two PC's (see text).

Multivariate approach
In this example data set obviously the standard deviation between laboratories is much larger then the standard deviation of repeatability. Under such circumstances, which frequently occur, a study for systematic behaviour, or correlations, is expected to be fruitful. A technique which is often applied to study correlations of these type is principal component analysis (PCA).
The first principal component is defined as the linear combination of variables, in this example PCBs, which describes the maximum amount of variance present in the data set. The second PC describes the maximum amount of the remaining variance etc.
Figure 1 shows the projection of the results of each laboratory on the first 2 principal components (PC's). In order to interpret this picture also the projection of the PCBs on the first 2 PCs is shown in figure 2. Figure 2 shows that all PCBs have a strong positive projection on the first PC indicating that this direction, describing 57 % of the variation in the data set, has a very systematic origin.

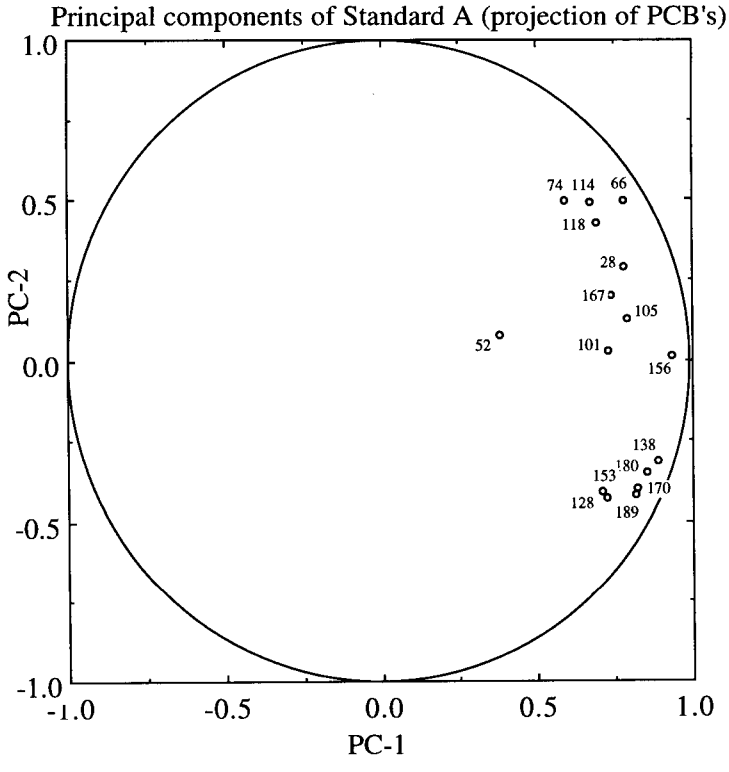


Figure 2. Projection of the PCB's on the first two PC's (see text).

From a comparison of figure 2 with figure 1 it can be concluded that the results of Lab 2 are low with respect to the assigned value (34% lower) and that the results of Lab 25 and 17 were high with respect to the assigned value (Lab 17: 17% higher). The second PC shows a slight discrimination between PCBs with low numbers, positive side, and PCBs with high numbers on the negative side. This discrimination apparently is present in the results of Lab 10.
In Figure 1 also the repeatability of the results is visualized showing that systematic variations are much more important than the deviations introduced by repeating the measurements for all laboratories except Lab 25.

Discussion
In this data set the advantage of using a multivariate projection, or data reduction, technique is that in a single picture the majority of the variation in the data can be shown and interpreted. Inspection of the individual data is still useful to find sources of error which are not correlated. An example of such a sources of error is a single standard for a single laboratory which has a very different concentration or where standards are mistaken one for the other in the chromatogram [31]. An important conclusion from this example data set from the multivariate data analysis is that the large systematic deviations indicate that a major source of deviation can found in the laboratory itself since deviation caused by suppliers will not necessarily correlate that perfect over the laboratories.