

Tang Luping
Björn Schouenborg

Methodology of Inter-comparison Tests and Statistical Analysis of Test Results

NORDTEST Project No. 1483-99

Methodology of Inter-comparison Tests and Statistical Analysis of the Test Results - Nordtest project No. 1483-99

Abstract

This report describes how to organise inter-comparison tests and how to statistically analyse the test results. In this report, relevant definitions, instructions, requirements and criteria from ISO, ASTM, EN standards, as well as national specifications are summarised and some examples of inter-comparison tests from practical applications are presented. The definitions of various terms commonly used in inter-comparison tests are given in Chapter 2. The complete procedures for organising inter-comparison tests are described in Chapter 3, while some useful statistical tools for analysing test results are given in Chapters 4 and 5. Practical applications of the precision test results are presented in Chapter 6 and, finally, the concluding remarks are given in Chapter 7.

Key words: Inter-comparison, test method, statistic

**SP Sveriges Provnings- och
Forskningsinstitut**
SP Rapport 2000:35
ISBN 91-7848-836-2
ISSN 0284-5172
Borås 2000

**SP Swedish National Testing and
Research Institute**
SP Report 2000:35

Postal address:
Box 857,
SE-501 15 BORÅS, Sweden
Telephone: +46 33 16 50 00
Telex: 36252 Testing S
Telefax: +46 33 13 55 02
E-mail: info@sp.se

Contents

Abstract	2
Contents	3
Preface	5
1 Introduction	7
2 Definitions Commonly Used in Inter-Comparison Tests	8
3 Organisation of Inter-Comparison Tests	10
3.1 Normative references	10
3.2 Planning of an inter-comparison test	10
3.3 Choice of laboratories	10
3.4 Recruitment of the laboratories	14
3.4.1 Invitation	14
3.4.2 Confirmation	14
3.4.3 Contact person	14
3.4.4 Training program	14
3.5 Test materials	14
3.5.1 Number of test levels	14
3.5.2 Number of replicates	15
3.5.3 Reserved specimens	15
3.5.4 Homogenisation of specimens	16
3.6 Dispatch of the specimens	16
3.6.1 Randomisation of specimens	16
3.6.2 Shipping	17
3.7 Instruction, protocol of the experiment	17
3.7.1 Instruction of the experiment	17
3.7.2 Protocol of the experiment	17
3.8 Precision test report	18
4 Standard Procedures for Statistical Analysis of the Test Results	19
4.1 Basic statistical model	19
4.2 Primary statistical calculations	20
4.2.1 Original data	20
4.2.2 Cell mean and cell standard deviation	21
4.2.3 General mean, repeatability and reproducibility	22
4.3 Consistency tests	24
4.3.1 Graphical technique	24
4.3.2 Numerical technique	25
4.4 Criteria for outliers and stragglers	27
4.5 Expression of precision results	28
5 Youden's Statistical Analysis – An Alternative Approach	29
5.1 Principles	29
5.2 A worked example	30
6 Applications of Precision Test Results	33
6.1 General	33
6.2 Comparing different test methods	33
6.2.1 Comparison of within-laboratory precision	33
6.2.2 Comparison of overall precision	34

6.2.3	Comparison of trueness	35
6.2.4	Comments on the results from comparison	36
6.3	Assessing the performance of laboratories	36
6.3.1	Types of assessment	36
6.3.2	Performance statistics and criteria	37
7	Concluding Remarks	40

Preface

This report has been written at SP on commission of the Nordtest.

The contents are based on international standards, guidelines and SP's experience from various commissions and many years of organizing and participating in inter-comparison tests.

The examples are taken from the field of building materials but can be used in most other fields too with minor modifications.

The financial support from the Nordtest is greatly appreciated.

Tang Luping and Björn Schouenborg

Borås, December 21, 2000

1 Introduction

Inter-comparison tests are usually used to evaluate the precision of test methods (precision tests) and to evaluate the performance of test laboratories (proficiency test). It is also a very powerful tool for maintaining and improving the laboratory competence.

A successful inter-comparison test is based on a good organisation of the test, including choice of test laboratories, choice of test materials, determination of test levels and number of replicates, plan of test schedule, distribution of test materials, instructions, design of test protocol, and collection of test results.

The collected test results should be statistically analysed for the precision of a test method or the performance of each test laboratory in accordance with generally acknowledged international standards/methods.

Since an inter-comparison test is often both costly and time-consuming, it is important to organise the test in a proper way so as to obtain valuable test results.

In general, due to unavoidable errors inherent in every test procedure, tests performed on presumably identical specimens under presumably identical test conditions do not yield identical results. The errors of test results have to be taken into account in the practical interpretation of test data. Different test methods or test procedures may produce different variations and errors of test results. Precision is a general term for the variability between repeated tests.

Apart from variations of supposedly identical specimens, the following factors may contribute to the variability of a test procedure:

- the operator;
- the equipment used;
- the calibration of the equipment;
- the environment (temperature, humidity, air pollution, etc.);
- the time elapsed between measurements.

In this report the methodology of inter-comparison tests and statistical analysis of the test results will be discussed. Some examples of inter-comparison tests on different materials will be given.

2 Definitions Commonly Used in Inter-Comparison Tests

The general principles and definitions are given in ISO 5725-1:1994(E) and ISO 3534:1993. In this chapter those terms that may cause confusion will be listed.

Accepted reference value: A value that serves as an agreed-upon reference for comparison. It is derived as

- a theoretical or established value based on scientific principles;
- an assigned or certified value based on experimental work of some national or international organisation;
- a consensus or certified value based on collaborative experimental work under the auspices of a scientific or engineering group;
- when the above are not available, experimentation of the measurable quantity, i.e. the mean of a specified population of measurements.

Accuracy: The closeness of agreement between a test result and the accepted reference value. When applied to a set of test results, it involves a combination of random errors and a common systematic error (“strictly systematic error” which is inherent in the test method because of some assumptions or approximations used to serve the method).

Bias: The difference between the expectation of the test results and an accepted reference value. It is the total systematic error contributed by one or several systematic error components (see 6.2.3 for an example).

Precision: The closeness of agreement between independent test results obtained under stipulated conditions. It depends only on the distribution of random errors and does not relate to the true value or specified value. Quantitative measures of precision depend critically on the stipulated conditions. Repeatability and reproducibility conditions are particular sets of extreme conditions.

Repeatability and reproducibility: Precision under repeatability and reproducibility conditions, respectively. They are quantitatively expressed by the repeatability standard deviation (σ_r or s_r) and reproducibility standard deviation (σ_R or s_R), or the repeatability limit (r) and reproducibility limit (R).

Repeatability/reproducibility conditions: The repeatability and reproducibility conditions are summarised in Table 2.1. There are some derivative conditions as defined in, e.g. EN 932-6 emanating from CEN TC 154 Aggregates. The repeatability or reproducibility obtained under such conditions is specially denoted with a subscript number, for instance, r_1 or R_1 . In this case sources of variation are different, as shown in Table 2.2, where

r_1 conditions: Repeatability conditions additionally with the laboratory sample reduction error.

R_1 conditions: Reproducibility conditions additionally with test portions of different laboratory samples of the same bulk sample.

R_2 conditions: Reproducibility conditions additionally with different bulk samples of a batch. Note that the subscript 2 indicates that sampling errors and sample reduction errors contribute to the variations measured under R_2 conditions.

Table 2.1. Repeatability and reproducibility conditions

Repeatability conditions	Reproducibility conditions
Using the same test method	
Measuring on identical material	
At the same laboratory	At different laboratories
By the same operator	By different operators
Using the same equipment	Using different equipment
Within a short interval of time	

Table 2.2. Sources of variation measured by repeatability conditions r or r_1 and reproducibility conditions R or R_1 or R_2 according to EN 932-6.

Sources of variation	r	r_1	R	R_1	R_2
Sampling error					+
Bulk sampling reduction error*				+	+
Laboratory sample reduction error		+		+	+
Between – laboratory testing variation			+	+	+
Within – laboratory testing variation	+	+	+	+	+

* Valid for particulate materials and other materials where the spread in the material is an important factor.

Repeatability/reproducibility standard deviation: The standard deviation of test results obtained under repeatability/reproducibility conditions (see section 4.2 for calculations).

Repeatability/reproducibility limit: The value less than or equal to which the absolute difference between two test results obtained under repeatability/reproducibility conditions may be expected to be with a probability of 95%. As a “rule of thumb”, the repeatability/reproducibility limit can be obtained by multiplying the repeatability/reproducibility standard deviation with a factor of 2.8, that is, $r = 2.8s_r$ and $R = 2.8s_R$.

Outlier: A member, of a set of values, which is inconsistent with the other members of that set. The consistency can be tested using graphical or numerical techniques (see sections 4.3 and 4.4).

Straggler: Similar to an outlier but with less inconsistency (see section 4.4 for criteria).

Uncertainty: An estimate characterising the range of values within which the true value is likely to (will probably) lie. When stating an uncertainty one should also state the probability level associated with the uncertainty.

3 Organisation of Inter-Comparison Tests

3.1 Normative references

The following references are normative even if they are not mentioned explicitly in the method. They are used as references for the structure of the method, for the planning of each inter-comparison programme and for the assessment of the results.

- ISO 5725, Accuracy (trueness and precision) of measurement methods and results;
- ASTM E 691, Standard Practice for Conducting an Interlaboratory Study to Determine the Precision of a Test Method;
- ASTM E 1301, Standard Guide for Proficiency Testing by Interlaboratory Comparisons;
- ISO Guide 43, Development and operation of laboratory proficiency testing;
- EN 932-6. Tests for general properties of aggregates - Part 6: Definitions of repeatability and reproducibility;
- ELA G6. WELAC Criteria for Proficiency Testing in Accreditation;
- ISO/REMCO N 231: Harmonized Proficiency Testing Protocol;
- Regulations by national accreditation bodies, e.g. Swedish regulation RIS 04: Rutin- och instruktionssamling. SWEDAC:s jämförelsprogram;
- Youden, W. J. Statistical Techniques from Collaborative Tests.

3.2 Planning of an inter-comparison test

Before starting an inter-comparison test, a **work group** (or called “panel” in ISO 5725) is needed with overall responsibility for the planning of the inter-comparison test including economics, number of laboratories, test materials, etc. The work group should consist of experts familiar with the test methods and their applications.

A **coordinator** should be chosen with the responsibility for the detailed planning, including invitation, sample preparation, answering sheets, results, preliminary assessment and direct contact with the participating laboratories.

One person with both adequate competence of the statistical methods and knowledge on the material itself should be appointed as a **statistician**. This person is responsible for the statistic design and analysis of experiments.

3.3 Choice of laboratories

It is our opinion that any laboratory considered qualified to run the test should be offered opportunity to participate in the inter-comparison test. A **qualified laboratory** should have

- proper laboratory facilities and testing equipment,
- competent operators,

- familiarity with the test method,
- a reputation for reliable testing work, and
- sufficient time and interest to do a good job.

Concerning the choice of laboratories, the standards present slightly different opinions:

The wording above comes from ASTM E 691. Without sufficient familiarity with the test method a familiarization process including a pre-test can be used. **“The importance of this familiarization step cannot be overemphasized”** “Many inter-laboratory studies have turned out to be essentially worthless due to lack of familiarization.” See more under 3.4.4 below.

ISO 5725-1:1994 states that “The participating laboratories should not consist exclusively of those that have gained special experience during the process of standardising the method. Neither should they consist of specialized “reference” laboratories in order to demonstrate the accuracy to which the method can perform in expert hands.

One can conclude from the above that, in order to minimize the risk of any bias, it is important to include all sorts of laboratories that use the method on a regular basis or has gained sufficient experience by a pre-test.

From a statistical point of view, those laboratories participating in an inter-comparison test to estimate the precision should have been chosen at random from all the laboratories using the test method. The number of laboratories participating in an inter-comparison test should be large enough to be a reasonable cross-section of the population of qualified laboratories. Thus the loss or poor performance of a few will not be fatal to the study.

In practice, the choice of the number of laboratories will be a compromise between availability of resources and a desire to reduce the uncertainty of the estimates to a satisfactory level. The number of laboratories and the number of test results (replicates) from each laboratory at each level of the test are interdependent. The uncertainty factor A at a probability level of 95% can be expressed by the following equations:

For repeatability

$$A_r = 1.96 \sqrt{\frac{1}{2p(n-1)}} \quad (3.1)$$

For reproducibility

$$A_R = 1.96 \sqrt{\frac{p[1+n(\gamma^2-1)]^2 + (n-1)(p-1)}{2\gamma^4 n^2 (p-1)p}} \quad (3.2)$$

where subscripts r and R represent repeatability and reproducibility, respectively;

p is the number of laboratories;

n is the number of test results;

γ is the ratio of the reproducibility standard deviation to the repeatability standard deviation.

The relationships between the uncertainty factor and the number of laboratories are illustrated in Figs. 3.1 and 3.2. It can be seen from the figures that a large number of

participating laboratories can always reduce the uncertainty of the estimates. The number of qualified laboratories for some new developed methods is, however, sometimes very limited. A way to solve the problem of the small number of participating laboratories is to increase the number of replicates, if the ratio of reproducibility to repeatability is not very large, e.g. less than 2. **It is common to choose the number of laboratories between 8 and 15.** If the precision of a test method is based on acceptable test results from fewer than 6 laboratories, the uncertainty of the estimates should be clearly mentioned in the final statement of precision!

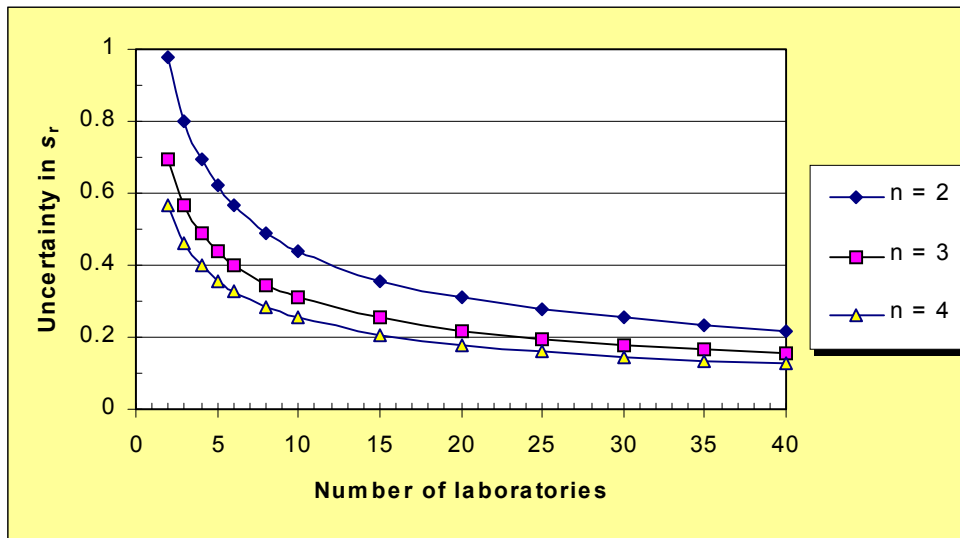


Fig. 3.1. Relationship between the uncertainty in the repeatability standard deviation and the number of laboratories.

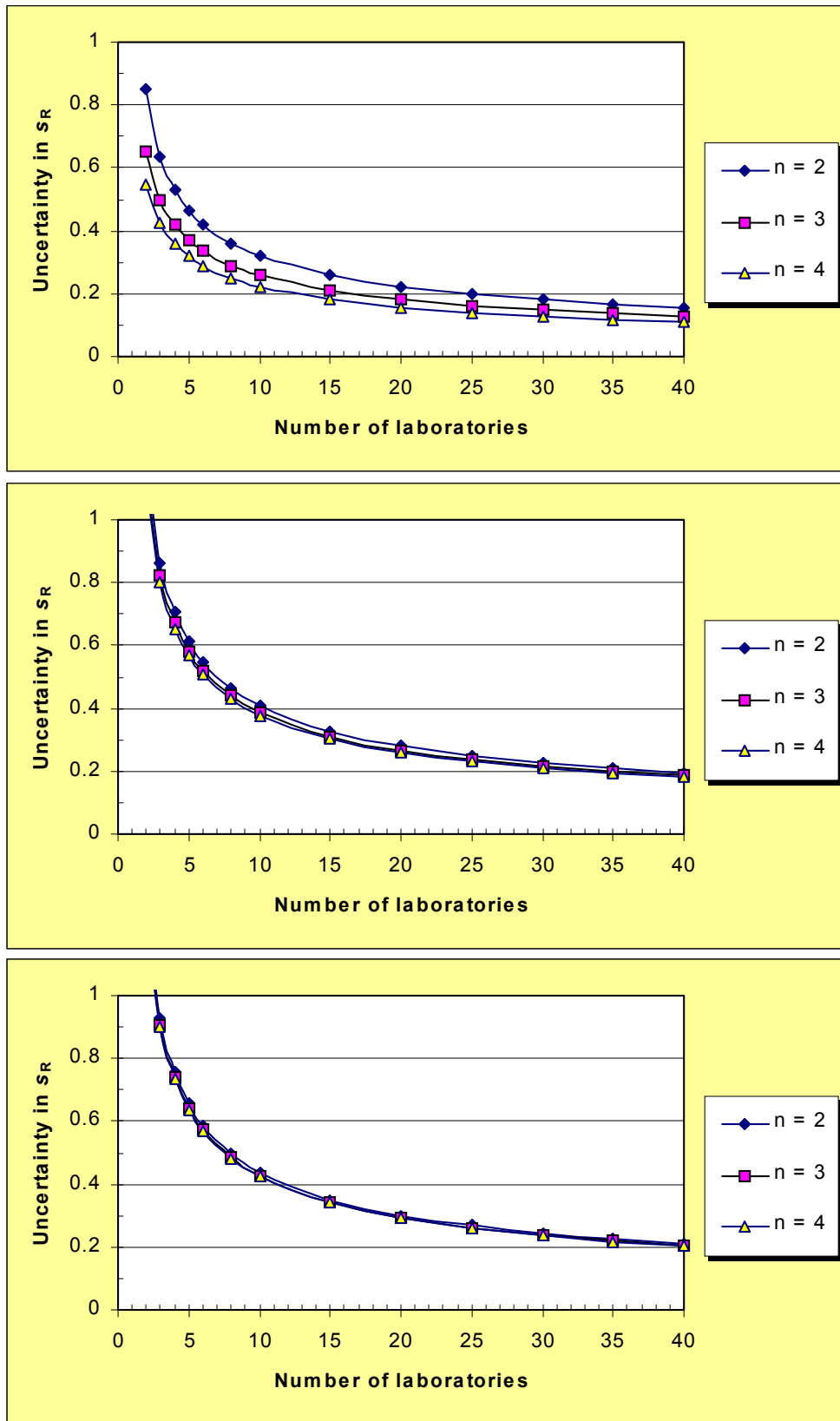


Fig. 3.2. Relationship between the uncertainty in the reproducibility standard deviation and the number of laboratories.

3.4 Recruitment of the laboratories

3.4.1 Invitation

The coordinator sends an invitation to the laboratories that are considered qualified. The invitation shall contain information about requirements on calibration of equipment, analysis procedures, starting date, deadline for reporting and costs for participating, if any. In the invitation the responsibilities of participating laboratories should be clearly stated. Finally, if necessary, the invitation shall declare which statistical model(s) for evaluation shall be used.

3.4.2 Confirmation

The laboratory that accepts the invitation shall send a written confirmation to the coordinator. In the confirmation the adequacy in equipment, time and personnel and the willingness of participation shall be stated. The formulation of a written confirmation can be prepared by the coordinator and sent together with the invitation to the laboratories. In some cases, the written confirmation may be replaced by a documented report from a telephone call, e-mail, fax, or similar.

3.4.3 Contact person

Each participating laboratory should appoint a contact person with an overall responsibility for the test.

3.4.4 Training program

If some laboratories have not had sufficient experience with the test method, a training program could be arranged to familiarise the operators in those laboratories with the test method before the inter-comparison test starts. The training program could be a course or seminar combined with a pilot inter-comparison test running with one or two test materials of known quality. The training program is of special importance for newly developed methods and standard methods, which are subjected to some radical modifications.

3.5 Test materials

3.5.1 Number of test levels

The materials to be used in an inter-comparison test should represent fully those to which the test method is expected to be applied in normal use. Different materials having the same property may be expected to have different property levels, meaning higher or lower values of the property. The number of test levels to be included in the inter-comparison test will depend on the following factors:

- the range of the levels in the class of materials to be tested and likely relation of precision to level over that range,
- the number of different types of materials to which the test method is to be applied,
- the difficulty and expenses involved in obtaining, processing, and distributing samples,
- the difficulty/length of time required for and expenses of performing the test,
- the commercial or legal need for obtaining a reliable and comprehensive estimate of precision, and

- the uncertainty of prior information on any of the above points.

In normal cases, at least 3 different materials representing 3 different test levels should be included in an inter-comparison test. If the purpose is to determine the precision of a test method, 5 or more levels should be included. If it is already known that the precision is relatively constant or proportion to the average level over the range of values of interest, or if it is the pilot investigation of a recently developed test method, the number of test levels can be reduced. Youden analysis (see Chapter 5) is an exception, which uses only two levels of similar qualities of materials.

3.5.2 Number of replicates

The number of replicates is usually three or four. Two replicates can be used in each of the following cases:

- the material is expected to be very homogeneous,
- it is already known that the reproducibility standard deviation is significantly larger than the repeatability standard deviation, i.e. γ value is larger than 2 (ref. Fig. 3.2),
- the number of laboratories is 15 or more.

If the material is expected to be very inhomogeneous, up to 10 replicates may be required depending on the level of uncertainty that is aimed at.

Example 1 - Determination of chloride content in concrete: The concrete powder was homogenised in one batch in a ceramic ball-mill and packed with the help of a rotary sample divider. So the material was expected to be very homogeneous. Thus two replicates were used in an inter-comparison test.

Example 2 - Determination of frost resistance of concrete: The property of concrete to resist frost attack is inhomogeneous when compared with other properties such as compressive strength. Thus 4 replicates are needed in the test.

3.5.3 Reserved specimens

It is wise to always prepare sufficient quantities of materials to cover the experiment and to reserve for accidental spillage or errors in obtaining some test results which may necessitate the use of complementary material. Depending on the type of material and the number of participating laboratories, the amount of material in reserve could be 10~50% as much as that to be directly used in the experiment. Any decision of using complementary specimens shall be taken by the coordinator in consultation with the working group.

Example 1 - Determination of chloride content in concrete: The amount of sample powder directly used in the experiment was less than 20 g per laboratory. For the sake of safety 30~40 g sample was sent to each laboratory. 9 laboratories participated the test. Thus at least 1 kg powder should be produced in one batch.

Example 2 - Determination of frost resistance of concrete: The property of concrete to resist frost attack is very sensitive to the age of concrete. There is no meaning to reserve specimens for testing errors. In this case 10-20% extra amount of material could be enough to cover the possible damages during preparation and transport of specimens.

3.5.4 Homogenisation of specimens

When a material has to be homogenised, this shall be done in the manner most appropriate for that material. A fluid or fine powder can be homogenised by, for instance, stirring. Particulate materials shall, when possible, be subdivided by aid of rotating sample divider. After the homogenisation, a statistical selection shall be tested for homogeneity. If this test will lead to a considerable delay in the test programme an alternative test may be used for an indirect assessment of homogeneity. If sample subdivision or reduction is a part of the test procedures, this should be clearly stated in the instructions.

Example 1 - Determination of chloride content in concrete: The concrete block was crashed with a jaw-breaker. The crashed particles were ground in small patches in a puck mill to powder. The powders from different patches were mixed and the mixed powder was homogenised in one batch in a ceramic ball-mill. The homogenised powder was packed (subdivided) with the help of a rotary sample divider.

Example 2 - Determination of particle size distribution of aggregate (Sieving Test): The test samples were prepared by proportionally weighing of different size-fractions, each of which was pre-homogenised in a rotating drum.

When the tests have to be performed on solid specimens that cannot be homogenised (such as concrete, stone block, metals, rubber, etc.) and when the tests cannot be repeated on the same test piece, inhomogeneity in the test material will form an essential component of the precision of the test method. It is important, in this case, to distribute the specimens to each laboratory with the property or inhomogeneity as similar as possible.

Due to the heterogeneity of concrete material, the variation between the specimens may sometimes be very large. This should be kept in mind when evaluating precision of measurement techniques.

Example 3 - Determination of frost resistance of concrete: The specimens are solid and the test cannot be repeated on the same piece. Thus the inhomogeneity in the test material is inevitable. In this case the coordinator should try his/her best to distribute the specimens with similar inhomogeneity. For instance, 40 concrete cubes will be divided into 10 groups (4 cube per group). To minimise the effect of casting order on the property of material, the cubes were numbered in the order of casting and then grouped by numbers 1-11-21-31, 2-12-22-32, ... and 10-20-30-40.

3.6 Dispatch of the specimens

3.6.1 Randomisation of specimens

All groups of specimens shall be labelled and randomised before the dispatch. Randomisation can be done by random number table, lottery or computerised random numbering.

For some inter-comparison tests it may be important to keep confidentiality about the results. In such cases a third party organisation may be contracted for randomising and coding.

Example 1 - A Swedish national inter-comparison test for evaluating the performance of accredited laboratories: Different groups of samples were prepared by the coordinator

and then randomised, coded and packaged by an official from SWEDAC (the Swedish Board for Technical Accreditation) before the dispatch to different laboratories.

3.6.2 Shipping

The specimens shall be properly packaged to prevent them from damages of impact or overload and to protect the stability and characteristics of material during transport. If the property of material is very sensitive to temperature, an air-conditioned container or truck should be arranged for transport.

On the package the receiver's name and address should be clearly and correctly labelled to ensure prompt delivery. If the package should be sent to a country where a custom declaration is needed, the coordinator should prepare sufficient documents, for instance, proforma invoices, attaching to the package. Preferably, the coordinator mails or faxes a copy of relevant documents for the custom declaration to the receiver. Problems arising in the custom declaration may sometimes greatly delay the delivery, resulting in a whole delay of the experiment.

After dispatch of the specimens, the coordinator should inform the receivers (contact persons of the participating laboratories) the dispatch including the dispatching date, name of the courier, shape and weight of the package, handling of the specimens on arrival, etc. The receivers should, in return, inform the coordinator about the arrival and conditions of the specimens.

3.7 Instruction, protocol of the experiment

3.7.1 Instruction of the experiment

It is of prime importance for the coordinator to write a clear instruction of the experiment including preparation, calibration and testing, especially when

- there are several options in the test method,
- the method is newly developed and some of the participants may not be experienced,
- there exist some unclearness in the current version of the standard test method, or
- the test is a modification of the standard test method.

In the instruction, the number of significant digits to be recorded should be specified. Usually, the number of significant digits in an inter-comparison test is, if possible, one more than required by the standard method. Where appropriate, using tables, figures and photo pictures can be of great assistance for the operators to understand the test procedure.

3.7.2 Protocol of the experiment

It is wise for the coordinator to prepare "standard" protocol sheets for recording the raw data and reporting the results. The measurement units and required significant digits can be expressed in the protocol. The coordinator's contact address could be cited in the protocol to facilitate the operator to contact him/her when any questions arise during the test. Sufficient space should be left in the protocol for the operator to make notes during the test. These notes could be useful information for interpretation of the inconsistency in test results, if any.

3.8 Precision test report

A proposed contents list of the report from a precision test:

1. Abstract
2. Acknowledgements
3. Abbreviations
4. Symbols
5. Introduction
6. Definitions
7. The method used
8. The inter-comparison experiments
 - i. Laboratories
 - ii. Materials
 - iii. Data
 - iv. Averages and ranges
9. Assessment of the participating laboratories
 - i. Unusually large laboratory biases and ranges
 - ii. Repeatability checks
 - iii. Proficiency testing (a statement whether this is needed or not)
10. Precision
 - i. Repeatability and reproducibility
 - ii. Assessment of the reproducibility of determination of the property
 - iii. Assessment of the repeatability of the determination of the property
 - iv. Assessment of small test portions
11. Conclusions
12. References

Appendices should include:

1. Participating laboratories
2. Materials
3. Data
4. Average and ranges
5. Precision of the determination, tabled
6. Histograms
7. Graphs of laboratory averages, biases and ranges
8. Result that merits further investigations
9. Formulae used in the calculations
10. The test method

4 Standard Procedures for Statistical Analysis of the Test Results

4.1 Basic statistical model

The basic statistical model given in ISO 5725 for estimating the accuracy (trueness and precision) of a test method is

$$y = m + B + e \quad (4.1)$$

where y is the test result;
 m is the general mean (expectation);
 B is the laboratory component of bias under repeatability conditions;
 e is the random error occurring in every measurement under repeatability conditions.

The term B is considered to be constant under repeatability conditions, but to differ in value under other conditions. In general, B can be considered as the sum of both random and systematic components including different climatic conditions, variations of equipment and even differences in the techniques in which operators are trained in different places. Thus the variance of B is called the between-laboratory variance and is expressed as

$$\text{var}(B) = \sigma_L^2 \quad (4.2)$$

where σ_L^2 is the true value of the between-laboratory variance and includes the between-operator and between-equipment variabilities.

The term e represents a random error occurring in every test result. Within a single laboratory, its variance under repeatability conditions is called the within-laboratory variance and is expressed as

$$\text{var}(e) = \sigma_w^2 \quad (4.3)$$

where σ_w^2 is the true value of the within-laboratory variance. It may be expected that σ_w^2 will have different values in different laboratories due to differences such as in the skills of the operators, but it is assumed that for a properly standardised test method such differences between laboratories should be small. Thus a common value of within-laboratory variance for all laboratories using the test method could be established by the arithmetic mean of the within-laboratory variances. This common value is called the repeatability variance and is designated by

$$\sigma_r^2 = \overline{\text{var}(e)} = \overline{\sigma_w^2} \quad (4.4)$$

This arithmetic mean is taken over all those laboratories participating in the accuracy experiment which remain after outliers have been excluded.

According to the above model, the repeatability variance is measured directly as the variance of the error term e , but the reproducibility variance depends on the sum of the repeatability variance and the between-laboratory variance. As measures of precision, two

quantities are required: repeatability standard deviation σ_r and reproducibility standard deviation σ_R .

$$\sigma_r = \sqrt{\text{var}(e)} = \sqrt{\sigma_w^2} \quad (4.5)$$

and

$$\sigma_R = \sqrt{\sigma_L^2 + \sigma_r^2} \quad (4.6)$$

In the above equations σ represent the true value of a standard deviation with populations considered. In practice, these true standard deviations are not known due to limited populations. Estimates of precision values must be made from a relatively small sample of all the possible laboratories and within those laboratories from a small sample of all the possible test results. In statistical practice, the symbol σ is, replaced by s to denote that it is an estimate.

4.2 Primary statistical calculations

4.2.1 Original data

The original test results reported from different laboratories are collected in a work sheet and usually arranged in such a way that each column contains the data obtained from all laboratories for one test level of material, and each row contains the data from one laboratory for all test levels. An example is shown in Table 4.1.

Table 4.1. Example of the work sheet for collection of the original data.

Collection of the Original Data y_{ij}						
Decimal:	2	2	2	2	2	2
Laboratory	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6
Lab 1, test 1	3.52	4.65	7.51	9.78	12.55	15.11
test 2	3.57	4.83	7.32	9.09	11.67	16.29
Lab 2, test 1	3.55	4.24	6.55	8.96	11.98	15.65
test 2	3.54	4.12	6.58	8.49	11.19	14.52
Lab 3, test 1	3.73	4.82	7.11	9.67	12.42	15.91
test 2	3.66	4.57	7.06	9.78	12.37	16.04
Lab 4, test 1	2.99	4.29	6.57	7.80	9.76	13.64
test 2	3.12	5.54	7.17	8.79	11.31	14.62
Lab 5, test 1	3.29	4.85	6.28	8.91	10.90	14.19
test 2	3.11	4.49	6.67	8.93	11.07	15.34
Lab 6, test 1	3.56	4.44	7.54	10.10	12.71	15.64
test 2	3.44	4.81	7.48	9.14	11.56	15.25
Lab 7, test 1	4.04	4.86	7.55	9.76	12.61	16.18
test 2	3.95	4.86	7.18	9.74	12.73	16.24
Lab 8, test 1	3.45	4.77	7.25	9.10	12.16	15.28
test 2	3.48	4.37	6.94	8.75	11.55	15.66
Lab 9, test 1	3.63	4.62	6.59	8.89	11.63	13.89
test 2	3.39	4.37	6.84	8.54	11.09	14.27
Lab 10, test 1	3.56	4.54	7.22	9.17	11.83	14.33
test 2	3.51	4.60	6.77	9.19	11.67	14.82
Lab 11, test 1	3.31	3.79	7.06	8.86	13.00	15.26
test 2	3.22	3.87	6.64	9.22	11.89	15.37

4.2.2 Cell mean and cell standard deviation

The cell mean and the cell standard deviation for each test level are defined by the following equations:

$$\bar{y}_i = \frac{1}{n_i} \sum_{k=1}^{n_i} y_{ik} \quad (4.7)$$

and

$$s_i = \sqrt{\frac{1}{n_i - 1} \sum_{k=1}^{n_i} (y_{ik} - \bar{y}_i)^2} \quad (4.8)$$

where \bar{y}_i is the cell mean in the i -th laboratory;

n_i is the number of test results in the i -th laboratory;

s_i is the standard deviation of the test results in the i -th laboratory.

It should be noticed that every intermediate values in the calculation should be kept at least twice as many digits as in the original data. It is, of course, not a problem when a computer is used. Both the calculated values of \bar{y}_i and s_i should be expressed to one more significant digits than the original data. These calculated values should also be tabulated in the same way as in the original data. Some examples are shown in Tables 4.2. and 4.3.

Table 4.2. Example of the work sheet for collection of the cell mean (calculated from the data listed in Table 4.1).

Collection of the Cell Mean Y_{ij}						
	Experiment Level					
Laboratory	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6
Lab 1	3.545	4.740	7.415	9.435	12.110	15.700
Lab 2	3.545	4.180	6.565	8.725	11.585	15.085
Lab 3	3.695	4.695	7.085	9.725	12.395	15.975
Lab 4	3.055	4.915	6.870	8.295	10.535	14.130
Lab 5	3.200	4.670	6.475	8.920	10.985	14.765
Lab 6	3.500	4.625	7.510	9.620	12.135	15.445
Lab 7	3.995	4.860	7.365	9.750	12.670	16.210
Lab 8	3.465	4.570	7.095	8.925	11.855	15.470
Lab 9	3.510	4.495	6.715	8.715	11.360	14.080
Lab 10	3.535	4.570	6.995	9.180	11.750	14.575
Lab 11	3.265	3.830	6.850	9.040	12.445	15.315

Table 4.3. Example of the work sheet for collection of the cell standard deviation (calculated from the data listed in Table 4.1).

Collection of Cell Standard Deviation s_{ij}						
Laboratory	Experiment Level					
	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6
Lab 1	0.035	0.127	0.134	0.488	0.622	0.834
Lab 2	0.007	0.085	0.021	0.332	0.559	0.799
Lab 3	0.049	0.177	0.035	0.078	0.035	0.092
Lab 4	0.092	0.884	0.424	0.700	1.096	0.693
Lab 5	0.127	0.255	0.276	0.014	0.120	0.813
Lab 6	0.085	0.262	0.042	0.679	0.813	0.276
Lab 7	0.064	0.000	0.262	0.014	0.085	0.042
Lab 8	0.021	0.283	0.219	0.247	0.431	0.269
Lab 9	0.170	0.177	0.177	0.247	0.382	0.269
Lab 10	0.035	0.042	0.318	0.014	0.113	0.346
Lab 11	0.064	0.057	0.297	0.255	0.785	0.078

Collection of the Number of Replicates n_{ij}						
Laboratory	Experiment Level					
	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6
Lab 1	2	2	2	2	2	2
Lab 2	2	2	2	2	2	2
Lab 3	2	2	2	2	2	2
Lab 4	2	2	2	2	2	2
Lab 5	2	2	2	2	2	2
Lab 6	2	2	2	2	2	2
Lab 7	2	2	2	2	2	2
Lab 8	2	2	2	2	2	2
Lab 9	2	2	2	2	2	2
Lab 10	2	2	2	2	2	2
Lab 11	2	2	2	2	2	2

4.2.3 General mean, repeatability and reproducibility

The general mean \hat{m} is the average of the cell means, and expressed as

$$\hat{m} = \frac{\sum_{i=1}^p n_i \bar{y}_i}{\sum_{i=1}^p n_i} \quad (4.9)$$

where p is the number of laboratories.

According to equations (4.5) and (4.6), the estimate of the repeatability standard deviation is

$$s_r = \sqrt{s_w^2} \quad (4.10)$$

and the estimate of the reproducibility standard deviation is

$$s_R = \sqrt{s_L^2 + s_r^2} \quad (4.11)$$

The value of $\overline{s_W^2}$ is calculated by

$$\overline{s_W^2} = \frac{\sum_{i=1}^p (n_i - 1) s_i^2}{\sum_{i=1}^p (n_i - 1)} \quad (4.12)$$

It is obvious that, if the number of results is uniform through all participating laboratories, the repeatability standard deviation can be simply expressed as

$$s_r = \sqrt{\frac{\sum_{i=1}^p s_i^2}{p}} \quad (4.10')$$

The value of s_L^2 is calculated by

$$s_L^2 = \begin{cases} s_m^2 - \frac{s_r^2}{\hat{n}} & s_m^2 > \frac{s_r^2}{\hat{n}} \\ 0 & s_m^2 \leq \frac{s_r^2}{\hat{n}} \end{cases} \quad (4.13)$$

where s_m is the standard deviation of the cell means,

$$s_m^2 = \frac{1}{p-1} \cdot \frac{\sum_{i=1}^p n_i (\bar{y}_i - \hat{m})^2}{\hat{n}} \quad (4.14)$$

$$\hat{n} = \frac{1}{p-1} \left[\sum_{i=1}^p n_i - \frac{\sum_{i=1}^p n_i^2}{\sum_{i=1}^p n_i} \right] \quad (4.15)$$

If the number of results is uniform through all participating laboratories, s_m^2 can be simplified as

$$s_m^2 = \frac{1}{p-1} \sum_{i=1}^p (\bar{y}_i - \hat{m})^2 \quad (4.14')$$

Thus the reproducibility standard deviation becomes

$$s_R = \sqrt{s_m^2 + s_r^2 \frac{n-1}{n}} \quad (4.11')$$

4.3 Consistency tests

Many techniques can be used for testing the consistency of the test results reported from various laboratories. ASTM E 691 recommends a graphical technique (Mandel's k and h statistics), while ISO 5725 recommends both graphical (Mandel's k and h statistics) and numerical techniques (Cochran's statistic and Grubbs' statistic).

4.3.1 Graphical technique

Mandel's k -statistic is a "within-laboratory consistency statistic" which indicates a measurement deviation in one laboratory when compared with the pooled within-cell standard deviation (repeatability standard deviation). The k -value is defined as

$$k_i = \frac{s_i}{s_r} \quad (4.16)$$

Mandel's h -statistic is a "between-laboratory consistence statistic" which indicates a deviation of the cell mean measured from one laboratory when compared with the general mean (quasi-true value if the true value is unknown) obtained from the all laboratories in a round-robin test. h -value is defined as

$$h_i = \frac{\bar{y}_i - \hat{m}}{s_m} \quad (4.17)$$

An advantage of Mandel's statistics is that the results of consistency test can be graphically illustrated, as shown in Figs. 4.1 and 4.2.

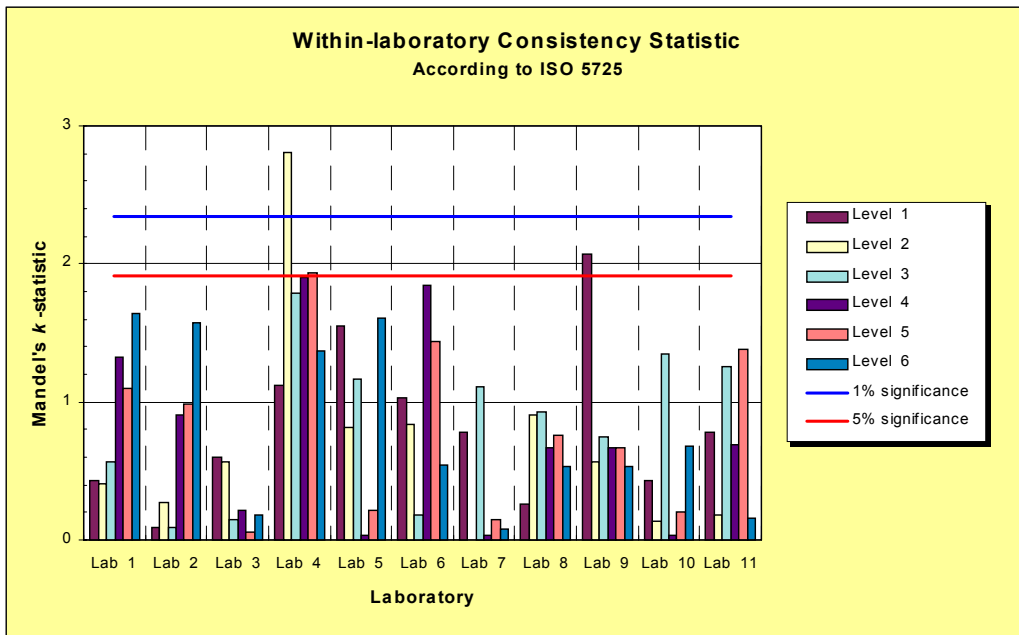


Fig. 4.1. Example of Mandel's k -statistic graphics (calculated from the data listed in Table 4.1).

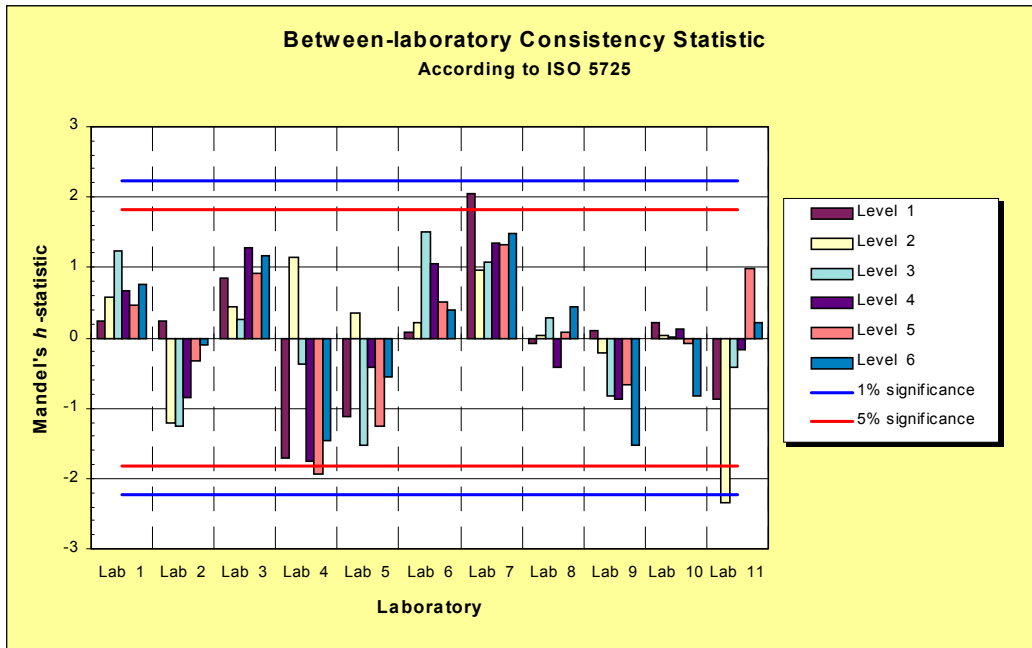


Fig. 4.2. Example of Mandel's *h*-statistic graphics (calculated from the data listed in Table 4.1).

4.3.2 Numerical technique

In ISO 5725-2, Cochran's test is recommended for examining the within-laboratory consistency and Grubbs' test for examining the between-laboratory consistency. Grubbs' test can also be used to test the consistency of results measured in one laboratory on the identical material.

Cochran's statistic C is defined as

$$C = \frac{s_{\max}^2}{\sum_{i=1}^p s_i^2} \quad (4.18)$$

where s_{\max} is the highest standard deviation in the set of standard deviations tested. Cochran's test is a one-sided outlier test because it examines only the highest value in a set of standard deviation.

In Grubbs' test, the data to be tested should be arranged in ascending order, that is, for a set of p values x_j ($j = 1, 2, \dots, k$), x_1 is the smallest and x_k is the largest. To examine the significance of the largest value, Grubbs' statistic G_k is

$$G_k = \frac{x_k - \bar{x}}{s} \quad (4.19)$$

where

$$\bar{x} = \frac{1}{k} \sum_{j=1}^k x_j \quad (4.20)$$

$$s = \sqrt{\frac{1}{k-1} \sum_{j=1}^k (x_j - \bar{x})^2} \quad (4.21)$$

To examine the smallest value, Grubbs' statistic G_1 is

$$G_1 = \frac{\bar{x} - x_1}{s} \quad (4.22)$$

To examine whether the two largest values may be outliers,

$$G_{k-1,k} = \frac{s_{k-1,k}^2}{s_0^2} \quad (4.22)$$

where

$$s_0^2 = \sum_{j=1}^k (x_j - \bar{x})^2 \quad (4.23)$$

$$s_{k-1,k}^2 = \sum_{j=1}^{k-2} (x_j - \bar{x}_{k-1,k})^2 \quad (4.24)$$

$$\bar{x}_{k-1,k} = \frac{1}{k-2} \sum_{j=1}^{k-2} x_j \quad (4.25)$$

Similarly, to examine the two smallest values,

$$G_{1,2} = \frac{s_{1,2}^2}{s_0^2} \quad (4.26)$$

where

$$s_{1,2}^2 = \sum_{j=3}^k (x_j - \bar{x}_{1,2})^2 \quad (4.27)$$

$$\bar{x}_{1,2} = \frac{1}{k-2} \sum_{j=3}^k x_j \quad (4.28)$$

When Grubbs' test is applied to the cell means (for testing the between-laboratory consistency), $x = \bar{y}$ and $k = p$, while when applied to the original data (for testing the consistency of n number of results in one laboratory, $x = y$ and $k = n$).

4.4 Criteria for outliers and stragglers

The critical values of k -statistic and h -statistic can be expressed by the following equations:

$$k_c = \sqrt{\frac{pF\{\alpha, f_1, f_2\}}{F\{\alpha, f_1, f_2\} + (p-1)}} \quad (4.29)$$

and

$$h_c = (p-1)t\{\alpha, f\}\sqrt{p(t^2\{\alpha, f\} + p-2)} \quad (4.30)$$

where $F\{\}$ is the inverse of the F -distribution with the degrees of freedom $f_1 = (n-1)$ and $f_2 = (p-1)(n-1)$;
 $t\{\}$ is the inverse of the Student's two-tailed t -distribution with the degree of freedom $f = (p-2)$;
 α is the significance level.

The critical values of Cochran's test and Grubbs' test were given in ISO 5725-2.

Normally, the critical values at the significance levels $\alpha = 1\%$ and $\alpha = 5\%$ are used as criteria for outliers and stragglers, respectively. In other words,

- if the test statistic is greater than its 1% critical value, the item tested is called a statistical outlier;
- if the test statistic is greater than its 5% critical value but less than or equal to its 1% critical value, the item tested is called a straggler;
- if the test statistic is less than or equal to its 5% critical value, the item tested is accepted as correct.

If an outlier or a straggler is found from the consistency test, the laboratory that produced this outlier or straggler should be consulted for possible technical error (e.g. clerical or typographical error, wrong sample, improper calibration, a slip in performing the experiment, etc.). If the explanation of the technical error is such that it proves impossible to replace the suspect item, the outlier should be rejected (discarded) from the study unless there is a good reason to remain it, while the straggler can be retained in the study.

If several stragglers and/or outliers in the k -statistic or Cochran's test are found at different test levels within one laboratory, this may be a strong indication that the laboratory's within-laboratory variance is exceptionally high (due to unfamiliarity to the test method, different operators, improper test equipment, poor maintenance of test environment, or careless performance), and the whole of the data from this laboratory should be rejected.

If several stragglers and/or outliers in the h -statistic or Grubbs' test are found at different test levels within one laboratory, this may be a strong indication that the laboratory's bias is exceptionally large (due to large systematic errors in calibration, or equation errors in computing the results). In this case the data may be corrected after redoing the calibration

or computation. Otherwise the whole of the data from this laboratory should also be rejected.

After rejection of outliers the consistency test may be repeated but the repeated application of the consistency test should be used with great caution, because this repeated process sometimes leads to excessive rejections.

Example 1: It can be seen from Fig. 4.1 that the k -statistic of the result at Level 2 from Lab 4 is greater than its 1% critical value and this result is classed as an outlier, while the k -statistic of the result at Level 1 from Lab 9 is in the range between its 5% and 1% critical values, thus this result is classed as a straggler.

4.5 Expression of precision results

After rejection or rejections of outliers the remaining data are used for calculating the final precision results, including general mean, the repeatability standard deviation and the reproducibility standard deviation. Other information may include the number of valid laboratories (after rejection of the outlier laboratories), the ratio of s_R/s_r , the repeatability limit $2.8s_r$, and the reproducibility limit $2.8s_R$. The information about rejection of outliers may also be included in the expression of the precision results. An example of the precision results is given in Table 4.4.

Table 4.4. Example of the expression of precision results (calculated from the data listed in Table 4.1, after rejection of the outliers).

Precision Results						
Test level j	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6
Number of valid laboratories ρ	11	9	11	11	11	11
General mean m	3.483	4.601	6.995	9.121	11.802	15.159
Repeatability variance s_r^2	0.00675	0.0335	0.05587	0.13574	0.32228	0.25746
Between-laboratory variance s_b^2	0.05959	0.01962	0.08876	0.15221	0.26336	0.37035
Reproducibility variance s_R^2	0.066	0.053	0.145	0.288	0.586	0.628
Repeatability std. dev. s_r	0.082	0.183	0.236	0.368	0.568	0.507
Reproducibility std. dev. s_R	0.257	0.23	0.381	0.537	0.766	0.792
Repeatability COV(s_r)	2.4%	4.0%	3.4%	4.0%	4.8%	3.3%
Reproducibility COV(s_R)	7.4%	5.0%	5.4%	5.9%	6.5%	5.2%
$\gamma = s_R/s_r$	3.13	1.26	1.61	1.46	1.35	1.56
Repeatability limit $r = 2.8s_r$	0.23	0.512	0.661	1.03	1.59	1.42
Reproducibility limit $R = 2.8s_R$	0.72	0.64	1.07	1.5	2.14	2.22
Number of excluded outliers	0	2	0	0	0	0
Outlier type*		kk, hh				
Outlier laboratories		Lab 4, 11				
Number of excluded stragglers	0	0	0	0	0	0
Straggler type						
Straggler laboratories						

* kk - according to Mandel's k -statistic test

hh - according to Mandel's h -statistic test

5 Youden's Statistical Analysis – An Alternative Approach

5.1 Principles

For repeatedly performed inter-comparison tests and for expensive tests, the “Youden approach” (reference) may provide a useful alternative to a full scale ISO 5725 - designed test.

The simple principle is to test only two nearly identical samples and compare the difference in results. Both systematic and random errors are clearly detectable by a graphical diagnosis.

The pair of test results is used to plot one point in a diagram with the X- and Y-axes representing the range of A- and B-test results. The general means (or median values) for A and B results are used to draw a horizontal and a vertical line, respectively.

In the ideal situation with random errors only, all points are equally distributed in the four quadrants of the diagram (see Figs. 5.1 and 5.3). If the random errors are small a circle, drawn with its centre in where the two median lines cross, will also be small (Fig. 5.1).

If the random errors are large, the circle will be large (Fig. 5.3).

Systematic errors will cause the points to be drawn out along a 45° line through the median cross. This is because a systematic error, by definition, will produce either a value smaller or larger than the general mean (see Fig. 5.2). Such a pattern should lead the assessor to investigate possible errors in the method/procedure.

The precision can be estimated by measuring the distance from each point to the 45° line (here called the perpendicular). Multiplying the mean length of all the perpendiculars with $\sqrt{\pi/2}$ gives an estimate of the standard deviation of one single measurement.

Using this standard deviation as a radius one can produce a circle with the centre in the cross of the medians. The size of the circle depends on the standard deviation and number of points (laboratories) that we want to include. Under a normal distribution, about 95 % of all points should be included within the circle if its radius is 2.5 times the standard deviation. This is deducted from a probability table. A radius of 3s will have about 99 % of the points included within the circle.

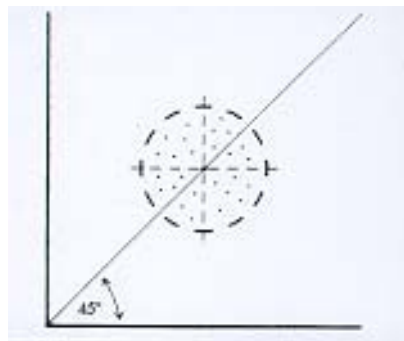


Fig. 5.1. Normal and small variation. All points of the laboratories' results are equally divided in the 4 quadrants. The test is accurate.

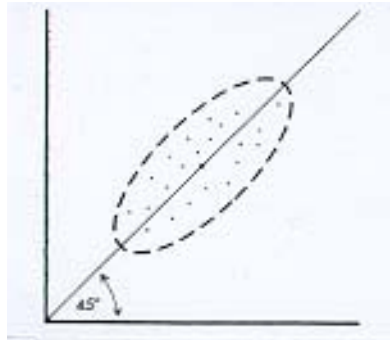


Fig. 5.2. A variation along the 45 ° line is a result of laboratory effects.

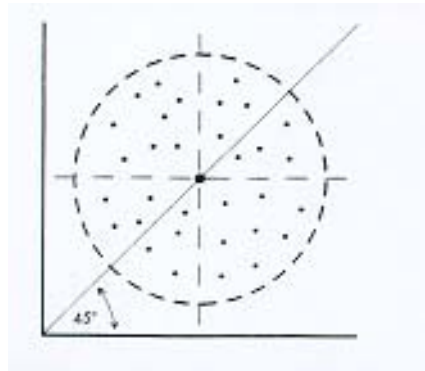


Fig. 5.3. Normal and large variation. The variation is large, which means that the test is inaccurate.

5.2 A worked example

As an example we present data from an inter-comparison test concerning an aggregate test according to SS-EN 1097-2: Methods for determination of resistance to fragmentation. The Los Angeles value (LA) was determined on two samples with similar fragmentation properties.

9 laboratories participated in the inter-comparison test. Some of them were not very familiar with the test since the standard is not compulsory yet. The test results are listed in Table 5.1 and the Youden plot is shown in Fig. 5.4.

Table 5.1. Original data from the determination of the LA-value.

Laboratory No.	Sample A	Sample B
1	11.6	6.9*
2	11.7	11
3	12.5	13.5
4	11.6	12.7
5	8.4*	11.9
6	11.2	12.4
7	10.4	9.7
8	11.5	12.6
9	11.1	13.2

* Suspected Outliers.

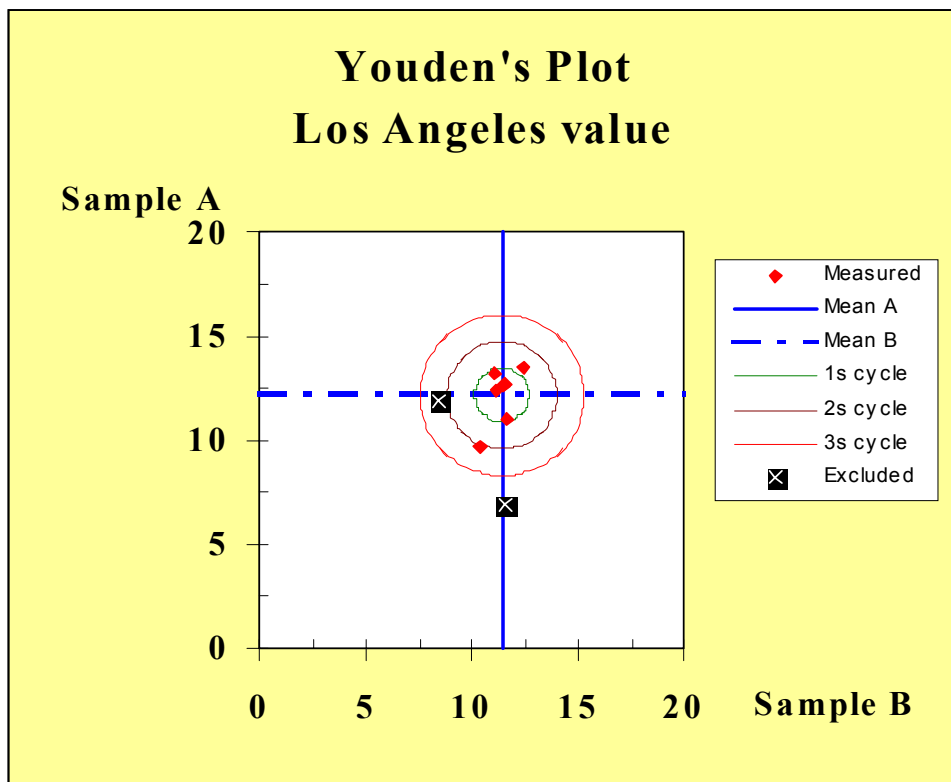


Fig. 5.4. Youden's plot from the LA-test.

From Fig. 5.4 it can clearly be seen that two laboratories are outside the 2s and 3s circles, respectively, and should be considered suspect as outliers. They were consequently excluded from the calculation of the final precision, as shown in Table 5.2.

Table 5.2. Calculation results according to the principles of Youden.

<i>Level</i>	<i>LA-value</i>
Number of valid lab p	7
Mean value of Test A, m_A	11.4
Mean value of Test B, m_B	12.2
Std. dev. of Test A, s_A	0.64
Std. dev. of Test B, s_B	1.34
Std. dev. of total data, s_d	1.29
Std. dev. of precision, s_r	0.74
Std. dev. of bias, s_b	0.75
Number of excluded outliers	2
Outlier criterion	2s
Outlier laboratories	Lab 1, 5

6 Applications of Precision Test Results

6.1 General

Precision test results are usually used for comparing different test methods and for assessing the performance of test laboratories. In addition, the precision data are also used for quality control and quality assurance. For instance, checking the acceptability of test results or monitoring stability of test results within a laboratory. The practical applications of the precision data are described in detail in ISO 5725-6. In this chapter some application examples will be given in comparing different test methods and assessing the performance of test laboratories.

6.2 Comparing different test methods

An example will be given in comparing three test methods for determining total chloride content in concrete. Reference samples containing a known content of chloride (0.071% by mass of sample with an expanded uncertainty of $U = 0.004\%$ by mass of sample, the coverage factor $k = 2$) were tested in different laboratories using three different methods, namely Method A, B and C, respectively. Method B is a standardised one, Method A is a modified alternative, and Method C is a simplified test suitable for a rapid determination of chloride in concrete. The precision test results are given in Table 6.1.

Table 6.1. Precision data from the example.

	Method A	Method B*	Method C
Number of valid lab p	6	7	11
Number of replicates n	2	2	2
Precision data			
General mean m	0.0696	0.0649	0.0583
Repeatability variance s_r^2	0.00000061	0.00001192	0.00000850
Between-lab variance s_L^2	0.00000359	0.00003697	0.00006284
Reproducibility variance s_R^2	0.00000420	0.00004888	0.00007134
Repeatability std. dev. s_r	0.00078	0.00345	0.00292
Reproducibility std. dev. s_R	0.00205	0.00699	0.00845

* Established standard method.

Since Method B is a standardised one, it will be used as a reference method, that is, other two methods will be compared with this one.

6.2.1 Comparison of within-laboratory precision

The following equations are applicable for comparison of within-laboratory precision:

$$F_r = \frac{s_{rx}^2}{s_{rB}^2} \quad (6.1)$$

where subscribe x represents either Method A or C,

$$F_{r_low-limit} = F\{(1 - \alpha/2), f_1, f_2\} \quad (6.2)$$

and

$$F_{r_high-limit} = F\{\alpha/2, f_1, f_2\} \quad (6.3)$$

where $F\{\}$ is the inverse of the F -distribution with the degrees of freedom $f_1 = p_x(n_x - 1)$ and $f_2 = p_B(n_B - 1)$, and the significance level $\alpha = 0.05$ is recommended in ISO 5725-6.

- If $F_{r_low-limit} \leq F_r \leq F_{r_high-limit}$, there is no evidence that the methods have different within-laboratory precisions;
- If $F_r < F_{r_low-limit}$, there is evidence that the method x has better within-laboratory precision than Method B;
- If $F_r > F_{r_high-limit}$, there is evidence that the method x has poorer within-laboratory precision than Method B.

The calculated results are listed in Table 6.2.

Table 6.2. Results from comparison of within-laboratory precision.

Comparison of within-laboratory precision			
	Method A	Method B*	Method C
Degree of Freedom f_r	6	7	11
F_r	0.051	1	0.713
$F_{r_low-limit}$	0.176	0.200	0.266
$F_{r_high-limit}$	5.119	4.995	4.709
Remarks	Better	-	Similar

* As reference.

6.2.2 Comparison of overall precision

The following equations are applicable for comparison of overall precision:

$$F_R = \frac{s_{Rx}^2 - \left(1 - \frac{1}{n_x}\right) s_{rx}^2}{s_{RB}^2 - \left(1 - \frac{1}{n_B}\right) s_{rB}^2} \quad (6.4)$$

The limit values of the inverse of the F -distribution $F_{R_low-limit}$ and $F_{R_high-limit}$ are similar to equations (6.2) and (6.3), respectively, but with the degrees of freedom $f_1 = p_x - 1$ and $f_2 = p_B - 1$.

- If $F_{R_low-limit} \leq F_r \leq F_{R_high-limit}$, there is no evidence that the methods have different overall precisions;
- If $F_r < F_{R_low-limit}$, there is evidence that the method x has better overall precision than Method B;
- If $F_r > F_{R_high-limit}$, there is evidence that the method x has poorer overall precision than Method B.

The calculated results are listed in Table 6.3.

Table 6.3. Results from comparison of overall precision.

Comparison of overall precision			
	Method A	Method B*	Method C
Degree of Freedom f_R	5	6	10
F_R	0.091	1	1.563
$F_{R_low-limit}$	0.143	0.172	0.246
$F_{R_high-limit}$	5.988	5.820	5.461
Remarks	<i>Better</i>	-	<i>Similar</i>

* As reference.

6.2.3 Comparison of trueness

The following equations are applicable for comparison of trueness:

$$\delta = |\mu - \hat{m}| \quad (6.5)$$

where μ is the true value, in this case $\mu = 0.071\%$ by mass of sample,

$$\delta_{cr} = \frac{2\sqrt{s_{Rx}^2 - \left(1 - \frac{1}{n_x}\right)s_{tx}^2}}{p_x} \quad (6.6)$$

- If $\delta \leq \delta_{cr}$, the difference between the general mean and the true value is statistically insignificant;
- If $\delta > \delta_{cr}$, the difference between the general mean and the true value is statistically significant. In such a case,
 - if $\delta \leq \delta_m/2$, there is no evidence that the test method is unacceptably biased;
 - if $\delta > \delta_m/2$, there is evidence that the test method is unacceptably biased

where δ_m is the minimum difference between the expected value from the test method and the true value. In this example, it is assumed that $\delta_m = U = 0.004\%$ by mass of sample. The calculated results are listed in Table 6.4.

Table 6.4. Results from comparison of trueness.

Comparison of trueness			
	Method A	Method B*	Method C
Absolute difference δ	0.0014	0.0061	0.0127
Critical difference δ_{cr}	0.0007	0.0019	0.0015
Minimum difference δ_m	0.0040	0.0040	0.0040
Remarks	<i>Insignificant bias</i>	<i>Significant bias</i>	<i>Significant bias</i>

* Established standard method.

6.2.4 Comments on the results from comparison

From Tables 6.2 to 6.4 it can be seen that Method A has better within-laboratory and overall precisions than the established standard method (Method B), and its general mean is closest to the true value. Thus this method can absolutely be used as an alternative to the standard method.

According to the results, Method C has precisions similar to Method B. From Table 6.4, it can be seen, however, that both these methods show significant bias (underestimated results of chloride content), especially Method C. After having studied the possible causes, it was found that, for Method B, too little amount of water was specified in the standard for washing the remainder in the filter paper, which may contain significant amount of chloride ions, resulting in an underestimated chloride content. For Method C, it was found that its test procedure involves a strictly systematic error. With a modified test procedure, Method C functions well as a rapid test.

6.3 Assessing the performance of laboratories

One of the important applications of the inter-comparison test is to assess the performance or proficiency of laboratories, especially of the accredited laboratories.

The assessment of laboratory performance works in both ways. There is a possibility for a second (customer) or third party (e.g. accreditation body) to assess the performance in relation to other laboratories. It is also possible to compare uncertainty statements of the laboratory versus the results from participation in inter-comparison tests. A customer needs to know the uncertainty in the test results and take this into account when he makes a statement of product tolerance.

The benefits for the participating laboratory are obvious. Participation in inter-comparison tests is a strong tool to maintain and improve the quality in the laboratory and also to make improvements in the method. The type of error, as described above, can in many cases be deducted to a certain type of problem in the laboratory or with the method. Här kan man se på graf från kontinuerligt deltagande i jämförelseprovning som visar på en gradvis förbättring (jag har sådant från Holland). Slutsatser från Youden graf. H-statistics may give information about inherited errors in the method by producing results systematically above or below the 0-line.

6.3.1 Types of assessment

There are three types of assessment depending on the existence of reference materials for the method or of a reference laboratory:

- Individual assessment by using reference materials. When reference materials exist on an adequate number of levels, the assessment may take place with the participation of the individual laboratory only.
- Individual assessment by comparing with a reference laboratory. When there exist no reference materials, the laboratory has to be compared with a high-quality laboratory which is widely recognised as providing an acceptable benchmark for the assessment.
- Continued assessment of laboratories by participating in inter-comparison tests. This is the case for annual or biennial assessment of the accredited laboratories.

For the individual assessments the precision data must be known in advance, for instance, from the previous inter-comparison tests, while for the continued assessment the

precision data can be obtained from the same inter-comparison test if the test is organised in such a way that it meets the requirements for precision determination, or from the previous inter-comparison tests if the scale of test is limited.

6.3.2 Performance statistics and criteria

The proficiency test results often need to be transformed into a performance statistic to aid interpretation and to allow comparison with defined goals. Depending on the type of assessment, different performance statistics can be used.

For the individual assessment by using reference materials, the performance statistic could be an absolute error δ . According to ISO 5725-6, the acceptance criterion is

$$\delta = |\bar{y} - \mu| \leq 2\sqrt{\sigma_R^2 - \left(1 - \frac{1}{n}\right)\sigma_r^2} \quad (6.7)$$

For the individual assessment by comparing with a reference laboratory, the performance statistic could be an absolute difference D . According to ISO 5725-6, the acceptance criterion is

$$D = |\bar{y}_1 - \bar{y}_2| \leq 2\sqrt{2}\sqrt{\sigma_R^2 - \left(1 - \frac{1}{n}\right)\sigma_r^2} \quad (6.8)$$

For the continued assessment by participating in an inter-comparison test, the most commonly used statistic is so called Z-scores:

$$Z_i = \frac{x_i - X}{s} = \frac{\bar{y}_i - \hat{m}}{\sqrt{s_R^2 - \left(1 - \frac{1}{n}\right)s_r^2}} \quad (6.9)$$

where x_i is the test result from the i -th laboratory;
 X is the group average;
 s is the group standard deviation.

It should be noticed that the precision data in the above equation must be reliable, that is, based on enough observations to allow outlier detection and have low standard error. The significant outliers must be excluded prior to calculation. According to ASTM E 1301-95, the following criteria are appropriate:

- if $|Z| \leq 2$, the performance is satisfactory;
- if $2 < |Z| \leq 3$, the performance is questionable;
- if $|Z| > 3$, the performance is unsatisfactory.

An example of Z-scores performance statistic is shown in Fig. 6.1. It can be seen that the performance of Lab 1 is questionable while the performance of Lab 6 is unsatisfactory. These two laboratories should be required for extensive follow-up and corrective action.

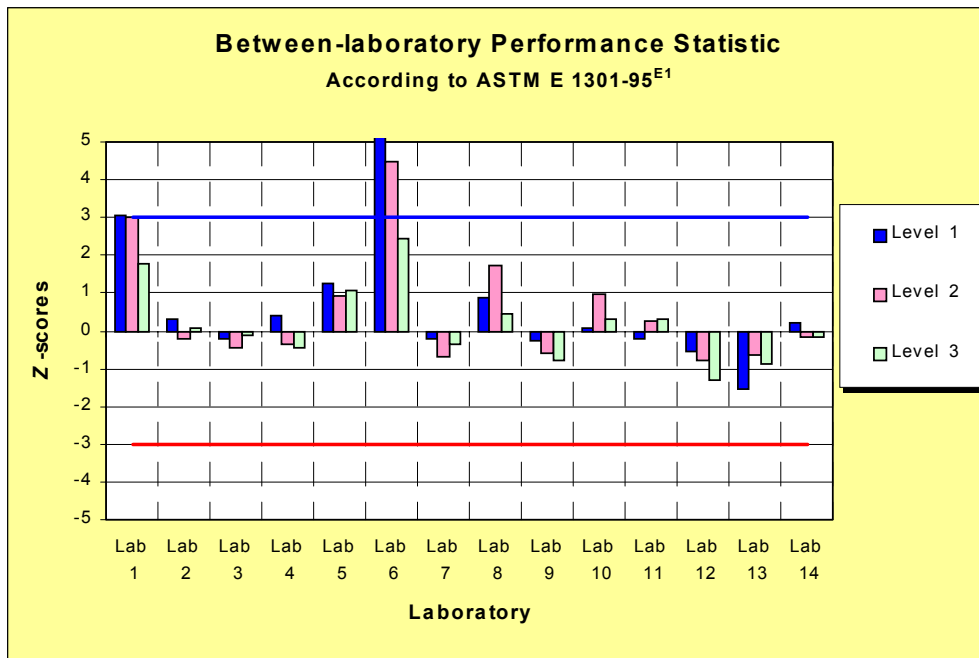


Fig. 6.1. Example of Z-scores performance statistic (data from a Swedish national inter-comparison test for frost resistance of concrete).

6.4 Establishing specification limits for materials

The results from precision tests should be used when specification limits are established in product standards.

In order to minimise the risk of having test results on either sides of a specified quality class, the size of a class should be given in proportion to the reproducibility of the test method. According to ISO/CD 10 576 “Specifications: Statistical principles for the construction of limiting values and the comparison of test results therewith”, the size of each quality class interval should be either 12 times the reproducibility standard deviation or 4 times the reproducibility.

As an example we have included the SS-EN 1097-9 - Studded tyre test (e.g. Determination of the resistance to wear by abrasion from studded tyres). The precision of this test method was established by a Nordic inter-comparison test (reference). The relationship between the Nordic Abrasion value (A_N) and the reproducibility (R) is the following:

$$R = 0.14A_N + 0.27 \quad (6.10)$$

The proposed quality categories are listed in Table 6.5.

Table 6.5. Quality categories of studded tyres.

Class/interval	Mean of the proposed interval	Recommended interval	Out of statistical point of view
<6			
6-9	7.5	3	3.2
9-13	11	4	5
13-18	15.5	5	7.4

The solution almost always has to be a compromise between true statistics and what is practically possible and also, in many cases, politically acceptable.

7 Concluding Remarks

There are several strong reasons for participating and organizing inter-comparison tests. Besides the basic reason to establish precision data of test methods, many other fields of application can be described.

Participating in inter-comparison tests should be an essential part of quality assurance activities for any testing laboratories. It should also be the security requirement for any commissioning company that has its products tested in a laboratory.

However, it is strongly recommended that every inter-comparison test should be performed in accordance with internationally accepted rules and guidelines. The omission of pre-testing the homogeneity of the dispatched test samples has led to many erroneous conclusions when assessing the results. Clearly, unambiguous instructions will minimise the risk of large variations in test results. It is the authors' hope that this report will give the reader an increased overall understanding and of all necessary measures of precaution in order to organise a valid inter-comparison test.

To facilitate a free trade within the European Union and to ensure trust in test results it is a necessity for laboratories to participate in international inter-comparison tests. Only then can their results be compared with an accepted quality level. For this purpose several operators have tried to establish themselves as PT-providers. Some of them are more serious than others.

EPTIS is the European Proficiency Testing Information System available in English on internet (www.eptis.bam.de). The web site contains information about PT-organizing companies and consortia for various materials and products in various countries.