Testing Scheme Report

ofi – Proficiency Test Series 2010

(Interlaboratory Comparison)

a) material testing

- plastics
- rubber, incl. non-vulcanized rubber

b) product testing

- selected products made of plastics

  (geotextiles and geosynthetics, plastic films and packaging materials, plastic pipes, plastic sheeting, sports surfaces, materials for car interiors, and rigid cellular plastics)
- liquids, surfaces and coatings

Vienna, April 2010 **Th. Karall**

For questions, comments or feedback
concerning this and future interlaboratory comparison tests, please, contact:

Dr. Thomas KARALL (phone ext. 433): pts@ofi.at

ofi Technologie & Innovation GmbH
Arsenal Objekt 213
1030 Vienna
AUSTRIA
www.ofi.at
Phone: +43 (0)1 798 16 01 – 433
Fax: +43 (0)1 798 16 01 – 977

Each participant receives reports containing results of individual tests and their statistical evaluation in the extent of the participant's choice of test methods submitted to *ofi* with the Registration form. Participants which were registered for a particular test but did not submit their results to this test receive the evaluation of the concerned test, too.

To maintain confidentiality of the identity of individual participants, an encoding scheme was employed. **Each participant knows only its own lab code number**, only the provider knows the encoding scheme.

SYMBOLS AND DECIMAL NOTATION

Symbols used in this Report:

$CV\%$... coefficient of variation in percent
n_i ..... number of repeat measurements made by an individual lab
n' ..... number of repeat measurements necessary to ensure a sufficiently low s_r in comparison with s^*
m ..... general mean of the test property (used in the additional check of the test method accuracy)
p^* .... number of reporting laboratories (including all outliers)
p ..... number of laboratories in the additional check for the test method accuracy (outliers eliminated)
r ..... repeatability limit
R ..... reproducibility limit
s ..... estimate of a standard deviation
s_i ..... within laboratory standard deviation
s_x ..... standard deviation in the inspected set of -values $\bar{x}_i$ (in the computation of the Grubbs' statistics only)
s^* ..... robust standard deviation
$\hat{\sigma}$ ..... standard deviation for proficiency assessment; for the purpose of this PTS: $\hat{\sigma} = s^*$
u_x ..... standard uncertainty of the assigned value X
x_i ..... test result (individual test result reported by laboratory i; depending on the test specification, x_i may be a result of a single measurement or a mean obtained by repeating measurements)
$\bar{x}_i$ ..... within laboratory mean (this symbol is written as x_i, i.e. not overlined, in all tables in the Report section EVALUATION due to settings in the software used in the statistical evaluation)
$\bar{x}_x$ ..... arithmetic mean of the inspected set of $\bar{x}_i$ -values (in the computation of the Grubbs' statistics only)
x^* ..... robust average
X ..... assigned value for proficiency assessment, for the purpose of this PTS: $X = x^*$
z ..... score used for proficiency assessment

Symbols used as subscripts:

i ..... identifier for a particular lab
k ..... identifier for an individual test result in a laboratory i obtained under repeatability conditions
L ..... between-laboratory (interlaboratory)
rel .... relative value (e.g. r_{rel} and R_{rel}; in percent of the general mean m)
r ..... repeatability
R ..... reproducibility
x ..... referring to $\bar{x}_i$, e.g., $\bar{x}_x$ is the arithmetic mean of all $\bar{x}_i$ - values in the given set of data

The bulk of the symbols used in this Report corresponds with the symbols used in ISO 13528:2005, ISO-Guide 43-1:1997 and ISO 5725-2:2002. In some cases, different symbols must be introduced to eliminate any confusion possibly caused by using symbols having different meanings in different documents.

Decimal notation:
In this Report, comma is used as decimal separator (as it is common in Middle Europe).

PART 1 - GENERAL REMARKS

Like in the past, with the present series of interlaboratory comparison tests (*ofi*-**PTS2010**) primarily testing methods were covered, for which, as a rule, a within-laboratory validation procedure is difficult to carry out and no certified reference materials are available. A couple of new methods were introduced in *ofi*-**PTS 2010**. From the originally offered 108 methods 45 were cancelled due to the too low number of registrations (concerned labs were informed). Thus, a total of **63** test methods, listed on the next page each with the referred number of participants were run in *ofi*-**PTS2010**.

The 'Number of participants' listed on the next page gives the number of labs which submitted results for the evaluation (p^*), not the number of registrations in the particular test which was sometimes higher. The number of individual tests selected by a particular participant was not limited. In average >10,8 participants per method have been count in the present PT-scheme. The original requirement of at least 7 participants in each test was not fulfilled in 3 methods in the end. Before the samples were distributed, participants registered in less filled methods (<7 labs per method) were asked for the agreement whether the concerned test shall be cancelled or conducted in spite of the low number of participants. Such tests were only conducted if all participants agreed. In some cases, the number of participants was below the target because one registered participant did not report the results in the end. Statistical evaluations possess a weakened explanatory force when less than 7 or 6 labs are involved in the evaluation. Therefore, Youden plots (see below) were not implemented if $p^* < 6$.

A couple of methods yield more than one result in each test. Thus, e.g., E_f, σ_{fM}, and ε_{fM} were evaluated in the flexural test, or L^*, a^*, b^*, and ΔE_{ab}, ΔL^*, Δa^*, Δb^* were evaluated in the colorimetry.

A total of **159 testing laboratories from 29 countries** participated in this interlaboratory comparison test. The number of participants in different countries is given below. The participation was open for everybody who believed to be able to perform the selected test according to the given (standardized) or any other procedure suitable to deliver comparable results.

Number of participants	Country
38	Germany
34	Austria
10	Netherlands
9	Spain
8	United Kingdom
6	South Korea
5	Belgium, Italy
4	Czech Republic, France, Greece, Poland, Slovenia, Switzerland,
3	Portugal
2	Finland, Romania, USA
1	Australia, Iran, Japan, Latvia, Norway, Saudi Arabia, Singapore, Sri Lanka, Sweden, Thailand, Tunisia,

The testing according to a completely different 'in-house-method' did not occur but deviations from the standardized test procedure were not rare. All deviations, as far as reported, are denoted on the introductory page in the front of the concerned test methods in the section Evaluation.

After the possible correction, all results which gave rise to a z-score > 10 were discarded to prevent an unnecessary distortion of the PT evaluation. Robust statistics applied in the evaluation is in fact insensitive to outliers but, for all that, extremely biased results would cause a slight shift in the calculated assigned value.

Method No.	No. of partici-pants	Test Method
1	14	Tensile Test for modulus of elasticity
3	17	Tensile properties (type 1A specimens)
4	7	Injection moulding of test specimens and tensile test
5	17	Flexural properties
6	7	Charpy impact strength (1eU) at +23°C
7	7	Charpy impact strength (1eU) at -20°C
8	12	Charpy notched impact strength (1eA) at +23°C
9	8	Charpy notched impact strength (1eA) at -20°C
11	7	Puncture impact behaviour
12	8	Creep modulus in tension at +23°C
13	7	Ball indentation hardness
14	10	Hardness Shore D
15	10	Temperature of Deflection under Load
16	12	VICAT Softening Temperature
17	22	Content of carbon black (TGA)
18	16	Oxidation Induction Time (OIT)
19	18	Glass Transition Temperature (DSC)
20	25	Behaviour of Melting and Crystallization (DSC)
21	8	Coefficient of Linear Thermal Expansion
24	26	Melt flow rate (MFR)
25	8	Colorimetry
26	11	Specular gloss
27	7	Contact angle and surface energy
29	19	Density
31	10	Degree of crosslinking
32	6	Viscosity of a polymer solution
34	11	Ash Content
36	17	Hardness Shore A
38	8	Rubber Hardness IRHD
39	17	Tensile Test on rubber
40	13	Compression set
41	11	Density of rubber
43	10	Abrasion resistance (using a rotating drum)
44	9	Emission Properties of Plastics - VOC and FOG
45	7	Emission Properties of Plastics - Total Carbon Emission
46	7	Emission Properties of Plastics - Formaldehyde
48	12	Tensile Test on plastic film
50	7	Tear resistance - Elmendorf method
51	11	Film Thickness
53	7	Water Vapor Transmission Rate
55	8	Overall migration – 95 % ethanol
56	9	Overall migration – 3 % acetic acid
59	12	Internal pressure test
60	8	Ring stiffness
61	8	Tensile Properties of polyolefin pipes
63	7	Geotextiles - Thickness at specified pressure
64	7	Geotextiles - Characteristic opening size
65	10	Geotextiles - Wide-width tensile test
66	7	Geotextiles - Static puncture test
67	10	Geotextiles - Water permeation characteristics
68	8	Geotextiles - Dynamic perforation test
70	10	Compression behaviour of rigid cellular plastics
76	8	Haze for transparent materials
78	9	Determination of selected elements (XRF)
81	5	Thermoplastics pipes - creep ratio
88	9	Overall migration – iso octane
89	8	Geotextiles - Wide-width tensile test
95	8	Tear strength (angle test piece)
106	7	Differential scanning calorimetry (DSC) method - Measurement of the processing temperature
108	5	Weathering (EN 12224) with following tensile test
109	18	Determination of vertical ball behaviour
110	19	Determination of vertical deformation
111	20	Determination of shock absorption

Additionally, the results of the performance assessment of all other participants would be too optimistic if definitely erroneous data would not be rejected. Wherever the rejection of submitted results was necessary,

this was noted in the respective sub-section of this Report. In this case, the rejected data was marked red and crossed through but left legible in the respective table and the colour of the corresponding bar in the z-score chart was converted to yellow. All other z-score-values (blue bars) were calculated after the extreme outliers (original $z > 10$) were rejected.

In all proficiency tests conducted in accordance with ISO-Guide 43-1, the **laboratory performance** is expressed by '**laboratory bias**', i.e., by the deviation of the laboratory result ($\bar{x}_i$ or x_i) from an assigned value (accepted reference value) X. In the ***ofi*-PTS2010**, the **assigned value** X was determined in accordance with ISO 13528:2005, Clause 5.6, as '**consensus value from participants**', namely as a **robust average** x^*. This is a standard procedure in ***ofi*-PTS200X** since years.

Computation of a "**z-score**" relating to the participating labs is a common way how interlaboratory comparisons for proficiency assessment are evaluated. The z-score is a measure of the distance of an individual result from the mean; the scale unit is the standard deviation. This approach has been applied in ***ofi*-PTS200X** since many years. The co called **standard deviation for proficiency assessment** $\hat{\sigma}$ is needed for the computation of the z-score. Like X, $\hat{\sigma}$ was determined in accordance with ISO 13528:2005, Clause 6.6, **from data obtained** as **robust standard deviation** ($\hat{\sigma} = s^*$).

As the robust estimates x^* and s^* are insensitive to **outliers**, extreme results need not be eliminated before the assigned value X is determined. The only exception in the ***ofi*-PTS2010** was the rejection of evidently erroneous data which is mentioned above.

Nevertheless, the identification of stragglers and outliers according to the Cochran's and Grubbs' tests using methods described e.g. in ISO 5725-2 was kept up in the ***ofi*-PTS2010** for the comparison with previous PT-schemes provided by ***ofi***. Outliers according to the **Grubbs'** test are extreme results with respect to the **deviation from the arithmetic mean of all results** ($\bar{x}_x$). Outliers according to **Cochran's** test are extreme results with respect to the **within-laboratory dispersion of the data** (Σs_i^2). In many other proficiency tests which do not utilize the data for the check of the test method accuracy, Grubbs' test is the only way how outliers are identified and later on excluded from the evaluation.

Each classical test for outliers (e.g. Grubbs' test) shows first only a single, the most serious outlier. The next outlier can be detected only if the outlier test is run second time after the outlier found in the previous run has been eliminated. In the ***ofi*-PTS2010**, we did not consequently search for Grubbs' and Cochran's outliers in this manner. Each lab which is a Grubbs' outlier has a z-score > 2 (but not vice versa). Therefore, the next Grubbs' outlier was identified – if present – during the data set preparation for the additional check of the test method accuracy (see bellow).

However, if two outliers (z-score >2) exist within a small data set no one may be identified as Grubbs' outlier due to above mentioned specific behaviour of the Grubbs' test. But as soon as the first (the worst) outlier is excluded from the data set, the next one is identified to be a Grubbs' outlier as a rule (and marked by asterisks in this Report). In contrast, all Grubbs' outliers are identified right from the start in large data sets. Cochran's outliers may not be identified in all cases due to the inconsequent search but each extraordinary within-laboratory dispersion is clearly evident in the graphical presentation of the PTS-results.

PART 2 - CONTENTS OF TABLES AND CHARTS

TABLES

In the PART 4 "**EVALUATION**", individual results reported by all participants are listed in the **first table** headed **'Results submitted by …'** on the left of each test-result-related double-page sub-section. A statistical evaluation of all reported results (means $\bar{x}_i$ and standard deviations s_i, robust statistics for x^* and s^*, as well as the consistency analysis, i.e. analysis for outliers, for the sufficient repeatability and sufficiently low standard uncertainty u_x) is included in this table. One or two **asterisks** in the columns headed with "Cochran" and "Grubbs" mark the corresponding **stragglers** and **statistical outliers**, respectively.

A check of the **test method accuracy** is reported in the **second table** on the left page in each double page sub-section. In contrast to the proficiency testing based on the robust statistics, **general mean** m used here must be calculated from data freed from outliers. The second important condition is that the data should be **normally distributed** (see below). In the strict sense, the calculation of parameter which characterize the test method accuracy is only correct when the input data comes from a normal distribution. It was shown in the ***ofi*-PTS2004** that if test results deviated from x^* by more than $2s^*$ were eliminated, most of the remained data was recognized to come from a normal distribution. In contrast to this, if statistical outliers and stragglers according to the – obviously less rigorous – Grubbs' test were eliminated, only a few remaining data sets were recognized to come from a normal distribution.

Outliers according to the definition "z>2" are marked by an '**X**' in the corresponding column in the **first table** in the sub-section 'Outliers'. It happened sometimes that new outliers according to the Grubbs' and Cochran's test occurred in the reduced set of data when z>2-outliers were eliminated from the original data set. These additional Grubbs' and Cochran's outliers are marked with asterisks in the respective table as described above, too. In some cases, these "new" outliers occurred in the data sets also if no Grubbs' and/or Cochran's outliers were identified within the original data. The reasons are explained above.

A **repeating of the tests** (2 times to 5 times; k = 2 to 5) introduced in the most methods in the ***ofi*-PTS2004** was kept in the ***ofi*-PTS2010**. This additional work is necessary to obtain a sound base for the assessment of the **test method accuracy** and to get a s_r-value which is utilized in the section **'determination of laboratory performance'**. This repeatability standard deviation s_r shall not be too large in comparison with s^* and the repeating of the tests can decrease s_r. A corresponding remark ("NOT OK") concerning the number of the tests repetitions appears in the table 'Test results' if the ratio s_r / s^* is too high.

The **Anderson-Darling** test was applied on the data sets freed from outliers in the check for normal distribution of the data. This test was described by Stephens, M. A. in "EDF Statistics for Goodness of Fit and Some Comparisons", Journal of the American Statistical Association, Vol. 69, (1974), pp. 730-737. In the ***ofi*-PTS2010**, outliers (outlier laboratories) were eliminated on the basis of their extreme mean values $\bar{x}_i$ ($z > 2$-outliers). However, the test for the normal distribution of the data was conducted with all remaining individual results (x_i-values), not only with the respective averages $\bar{x}_i$.

If the test for normal distribution of the data is performed with a rather limited number of the data, the result is frequently positive (this gives "YES" in the corresponding table concerning the accuracy of the examined test method). As the number of data increases, the test response to the outlying data becomes more selective and the non-conformity of the data distribution with the assumption (data normally distributed) is indicated with an increased sensibility. Therefore, the significance of the Anderson-Darling test for the normal distribution of the data shall not be overestimated particularly if the number of data is rather low.

Taking the statistical nature of the data into account, it can be supposed that the **laboratory performance** has been **assessed properly** without any curtailment if the amount of data was high enough for a reliable and sufficiently sharp statistical evaluation. For this reason, **an internal limit number of participants (7) was set**. Really no doubts about the assessment reliability and correctness exist if additional collateral conditions were also met, i.e.:

- ✓ if the number of repeat measurements was high enough (to get a reasonable value of s_r)
- ✓ if the standard uncertainty of the assigned value $u_x \leq 0{,}3 \cdot s^*$, i.e. if the number of participants $p^* \geq 16$ (remark "NOT OK" concerning the standard uncertainty u_x appears in the table 'Test results' if $p^* < 16$); this requirement represents a considerable raising of the above mentioned limit
- ✓ if the data came from the normal distribution

On the other hand, the result of the **proficiency assessment** is **not very reliable** if the threshold number of participants (7) was not reached, and it is **slightly diminished**

- if the number of measurement repetitions was too low and the resulting ratio s_r / s^* too high or
- if the number of participants $p^* < 16$ or
- if the data did not come from a normal distribution

In such cases, laboratories could receive (false) warning signals (higher values of z-score) because of inaccuracy in the determination of the assigned value, not due to procedural flaws within the laboratories.

CHARTS AND PLOTS

In the diagram presented in the upper part of the right page of each double page test-result related subsection, all **mean values** $\bar{x}_i$ and the respective **standard deviations** s_i are plotted against the LabCodeNo. The **robust average** x^*, i.e., the assigned value X, is displayed by a red horizontal line and the band width of **±1 robust standard deviation** s^* is marked by two blue dotted lines in this chart. Additionally, general mean m obtained in the additional check of the test method accuracy (outliers eliminated) is displayed by a thin green line for comparison with X.

In the second diagram on this page, the **z-scores** obtained are plotted against the LabCodeNo. Additionally, the $z = 2$-level which helps to identify the outliers is displayed with an orange line in the z-score chart. The problem of the too high ratio u_x / s^* caused by the too low number of participants could be solved if the lab proficiency would be assessed by means of z'-score instead of z-score. Nevertheless, the more common z-score was used in all tests in the ***ofi*-PTS2010** regardless of the ratio u_x / s^* resulting in the particular test. Drawbacks of the additional complexity seem to outweigh the advantage of the z'-score correctness. The only benefit of using z'-scores would be a slightly better performance, i.e., less number of 'warning signals' (z' > 2) and 'action signals' (z' > 3), in a few cases. In such situations ($p^* << 16$) the participants are well-advised to look if the difference between their own result and the assigned value X is acceptable from the practical point of view, not (only) from the statistical point of view.

Generally, a **split level sampling concept** was adopted in ***ofi*-PTS2010** to obtain an additional information with respect to the cause of the bias in some laboratories. This also enabled a multiple check of the test method accuracy as well as a multiple check of the lab's performance. For this purpose, always two samples, A and B, on more or less different levels were tested using the same method and the same apparatus. In this situation, a so-called **Youden plot** (W. J. Youden: Industrial Quality Control, 15, (1959) pp. 24-28) could be implemented for the first time in the ***ofi*-PTS2005**.

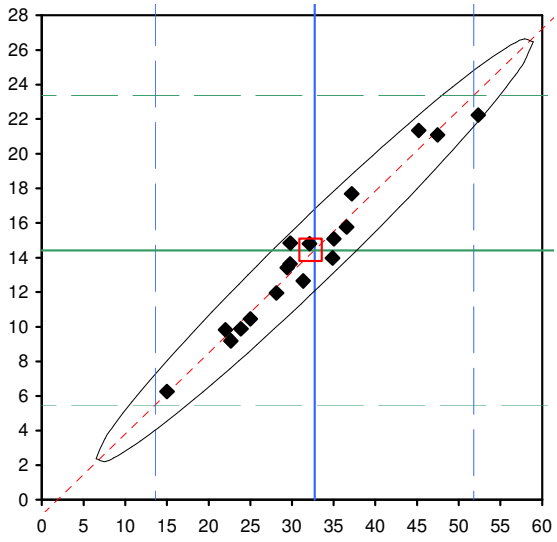
For this plot, mean values $\bar{x}_A$ and $\bar{x}_B$ across all $\bar{x}_{iA}$ and $\bar{x}_{iB}$ and corresponding standard deviations (s_A and s_B) are calculated; $\bar{x}_{A,B} \neq m_{A,B}$ in many cases (if at least one z-score $z > 2$ occurred in the data set

and the corresponding lab was eliminated from the test method accuracy check based on m as reference value). The Youden plot is formed as follows:

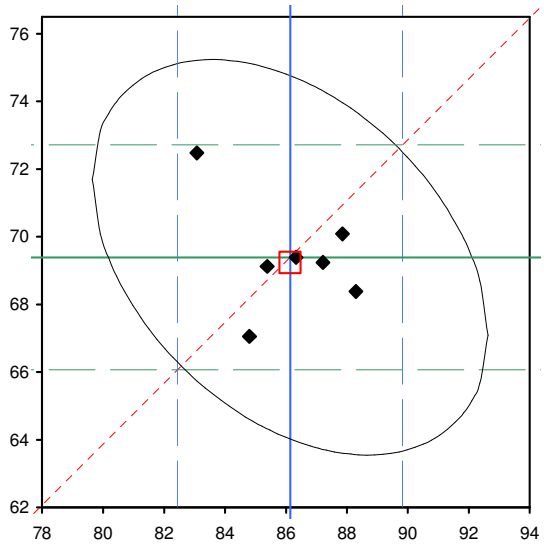
- horizontal axis: level 1 response value (Sample A)
- vertical axis: level 2 response value (Sample B)

A green horizontal line at $\bar{x}_B$ and a blue vertical line at and $\bar{x}_A$ divide the diagram area into four quadrants. A nearly 45-degree reference line (dotted, red) is drawn through the crossing point of the horizontal and vertical line ($\bar{x}_A$; $\bar{x}_B$). The intervals $\bar{x}_A \pm 2s_A$ and $\bar{x}_B \pm 2s_B$ are displayed by thin dashed lines, blue and green, respectively. The axes are frequently unequally scaled for the sake of approx. rel. equality of the uncertainty measures ($2s_A$ and $2s_B$) on both axes. A data point corresponding with the robust averages (x_A^* ; x_B^*) is included in the plot, too. Due to the individual scaling of the axes, $\bar{x}_A$ and $\bar{x}_B$ need not to be placed precisely in the middle of the corresponding axes, and the quadrants are then not the same size.

Examples of different forms of the Youden plot:



Data set which exhibits a minimum random error and a strong influence of a systematic error



Data set which exhibits a strong influence of the random error

This scatter plot is a simple but effective method for comparing both the within-laboratory variability and the between-laboratory variability. It indicates possible causes for biased data. Each lab generates one plot symbol (point) in the diagram area. As shown above, a lab point movement away from the best estimate along the 45-degree reference line (bottom left / top right) indicates an increasing systematic error. Its movement perpendicular to this reference line (movement in the direction top left / bottom right) indicates an increasing random error. Where Youden plot is included in this Report, strongly biased labs (LabCode-No.) are labelled near to their data points for better orientation.

A **95% control ellipse** (confidence ellipse for the mean) is included in the Youden plot. The ellipse is centred on the mean values ($\bar{x}_A$; $\bar{x}_B$). A randomly selected lab is included inside the ellipse in 95% of all cases. The size of the ellipse is strongly influenced by the number of participants: the larger the number of participants, the tighter the ellipse, even though the spread of data may be large. This plot does not make much sense if the number of data is too low, since the tendency in data scatter cannot be estimated in this case, and the 95% control ellipse has a low information content. Therefore, the Youden plot is not included in this Report if the number of points in the diagram would be less than 6.

As already stated, the axes scales in the Youden plots are frequently not equal (cf. diagram on the left side above on this page). The re-adjusting of the scales was often necessary especially if $\bar{x}_A >> \bar{x}_B$ or vice versa. The aim was always to get the principal axis of the 95% control ellipse in parallel with the diagram diagonal. If the axes scales are not equal, the shape of the ellipse is more or less "distorted". For this reason, it can happen that a particular type of error (random or systematic) predominates in a particular test but for all that the ellipse is nearly a circle. Therefore, it is important to have a look at the axes scales when the predominating type of error shall be quickly assessed from the shape of the ellipse.

COMMENTS

On the first page of each test method related sub-section, **participants' remarks** were collected, i.e., additional information stated in the Reports (particularly concerning the **test conditions** and **testing equipment**) and remarks concerning results, **properties of samples** and **comments** to any other corresponding matter of general interest. **Provider's comments** concerning the respective method were also put down here if it was necessary from the point of view of the ***ofi*-PTS**-team. In some cases, the samples submitted to the participants were too far from optimal, homogenous and stable material. This was mostly caused by the fact that "materials from the real life" were used as testing samples. However, it was also caused by deficits in the sample preparation in some cases. Such problems are – if they occurred – discussed in the comments to the concerned method.

PART 3 - PERFORMANCE STATISTICS AND TEST METHOD ACCURACY

Only short basics concerning the corresponding calculation are presented in this chapter. Therefore, look at the specific literature if more detailed information is of interest. The terms specified below were used in the statistical evaluation of the interlaboratory comparison. They are generally known or are defined, among other sources, in ISO 13528:2005 or ISO 5725-2:1994 including Technical Corrigendum 1:2002, as follows. See the list of symbols at the beginning of this Report for the symbol explanation if the explanation included here does not seem to be sufficiently clear.

Arithmetic mean, average ($\bar{x}$):

Quotient of the sum of independently identified individual values (in this test x_i) and their number n:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Note 1:

In the present proficiency test, the robust average x^* (i.e. assigned value for the proficiency test) is not an average value; it is derived from median of all $\bar{x}_i$ and calculated using the algorithm described in ISO 5725-5 and in ISO 13528:2005, Annex C.

Note 2:

In contrast to x^*, the general mean m used in the additional check of the test method accuracy is an average value.

Note 3:

In the computing of the Grubbs' statistics, over-all average value $\bar{x}_x$ is used:

$$\bar{x}_x = \frac{1}{p^*}\sum_{i=1}^{p^*} \bar{x}_i$$

Variance (s^2):

Quotient of the sum of squares of deviations of the individual values from the arithmetic mean and ($n-1$), i.e. number of degrees of freedom:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Standard deviation (s):

Positive value of the root of the variance of a series of measured values:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Note 1:

The robust standard deviation s^* is not calculated by the above simple formula but using the algorithm described in ISO 5725-5 and in ISO 13528:2005, Annex C.

Note 2:

In the computing of the Grubbs' statistics, standard deviation of the original results s_x is used:

$$s_x = \sqrt{\frac{1}{p^* - 1}\sum_{i=1}^{p^*}(\bar{x}_i - \bar{x}_x)^2}$$

Coefficient of variation ($CV\%$):

Dispersion of individual results expressed as quotient of the standard deviation and arithmetic mean in percent.

Repeatability conditions:

Conditions where independent test results are obtained with the same method on identical test items in the same laboratory by the same operator using the same equipment within short intervals of time.

Reproducibility conditions:

Conditions where test results are obtained with the same method on identical test items in different laboratories with different operators using different equipment.

Repeatability variance (s_r^2):

Arithmetic mean of s_i^2 taken over all those labs taking part in the accuracy experiment which remained after outliers have been eliminated

$$s_r^2 = \frac{\sum_{i=1}^{p}(n_i - 1)s_i^2}{\sum_{i=1}^{p}(n_i - 1)}$$

Between-laboratory variance (s_L^2):

Term including between-operator and between-equipment variabilities, relating to experiments with single level and equal or unequal number of measurements in all labs (cf. ISO 5725-2:2002; Clause 7.4.5.2)

$$s_L^2 = \left\{\frac{1}{p-1}\left[\sum_{i=1}^{p} n_i(\bar{x}_i - m)^2\right] - s_r^2\right\} \quad / \quad \left\{\frac{1}{p-1}\left[\sum_{i=1}^{p} n_i - \left(\sum_{i=1}^{p} n_i^2 / \sum_{i=1}^{p} n_i\right)\right]\right\}$$

Where, owing to random effects, a negative value for s_L^2 was obtained from the calculations for a particular data set, the value was assumed to be zero (cf. ISO 5725-2:2002, Clause 7.4.5.4). In these cases, $r = R$ and $s_r = s_R$ results from the corresponding calculation, although this is generally not true (usually $R>r$).

Reproducibility standard deviation (s_R):

The standard deviation of test results obtained under reproducibility conditions:

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

Repeatability limit (r):

A value less than or equal to what the absolute difference between two test results obtained under repeatability conditions may be expected to be with a probability of 95%:

$$r = 2{,}8 \cdot s_r$$

Note:

Two test results obtained under repeatability conditions shall be judged not equivalent if they differ by more than the "r". Vice versa, two test results obtained under repeatability conditions shall be judged to be equivalent if they differ by less than the "r". Any such judgment would have an approx. 95 % probability of being correct. This may be an important perception particularly in accredited laboratories which are obliged to know the measurement uncertainty of applied testing methods (cf. ISO / IEC 17025) and in the assessment of obtained test results for compliance with specified limit values.

Reproducibility limit (R):

A value less than or equal to what the absolute difference between two test results obtained under reproducibility conditions may be expected to be with a probability of 95%:

$$R = 2{,}8 \cdot s_R$$

Note:

Two test results obtained under reproducibility conditions shall be judged not equivalent if they differ by more than the "R". Vice versa, two test results obtained under reproducibility conditions shall be judged to be equivalent if they differ by less than the "R". This may be an important perception particularly if results obtained in two or more labs are compared. Any such judgment would have an approx. 95 % probability of being correct.

Outlier according to Grubbs' test:

With this test, the extreme values of $\bar{x}_i$ ($x_{extr} = \bar{x}_{\max}$ or $\bar{x}_{\min}$) are tested to be an outlier ("outlier regarding the mean value")

$$Grubbs \quad criterion \quad G = \frac{\left| \bar{x}_x - \bar{x}_{extr} \right|}{s_x}$$

where

$\bar{x}_x$ = arithmetic mean of the inspected set of data $\bar{x}_i$
$\bar{x}_{extr}$ = extreme value of $\bar{x}_i$
s_x = standard deviation of the inspected set of data $\bar{x}_i$
and G-values for statistical outliers (probability 99%) and possible outliers, i.e. stragglers, (probability 95%) are listed in the corresponding literature.

Outlier according to Cochran's test:

With this test, the within-laboratory variances are tested for homogeneity ("outliers regarding standard deviations"):

$$Cochran\ \ criterion\ \ C = \frac{s_{\max}^2}{\sum_{i=1}^{n} s_i^2}$$

where

s_{max} = highest value of s_i

and C-values for statistical outliers (probability 99%) and possible outliers, i.e. stragglers, (probability 95%) are listed in the corresponding literature.

Standard uncertainty of the assigned value (u_x):

When the robust average x^* is the assigned value, the standard uncertainty of the assigned value is estimated as:

$$u_x = \frac{1{,}25 \cdot s^*}{\sqrt{p^*}}$$

***z*-score:**

In the calculation of performance statistics the z-score is a commonly used variability measure.

$$z = \frac{(\bar{x}_i - X)}{\hat{\sigma}}$$

This score is used in different variants depending on the selection of $\hat{\sigma}$ and X values. The robust standard deviation s^* was set for $\hat{\sigma}$ and the robust average x^* was set for X in this PTS.

A z-score >2 denotes that the result of the respective laboratory deviates by more than $\pm 2\hat{\sigma}$ from the accepted reference value for the proficiency assessment X (= 'warning signal'). Approximately 95% of all results may lie in the interval $X \pm 2\hat{\sigma}$ if data is normally distributed. A z-score >3 shall be considered to give an "action signal", i.e., the respective laboratory shall start up to look for reasons of its extreme bias immediately.

The resulting data is assessed as follows:

$|z| \leq 1$.....the performance of the laboratory is **very good**

$1 < |z| \leq 2$.....the performance of the laboratory is **satisfactory**

$2 < |z| \leq 3$.....the performance of the laboratory is **questionable**

$|z| > 3$.....the performance of the laboratory is **unsatisfactory**

PART 4 -

EVALUATION

In all testing methods covered by this PT-scheme, the evaluation is presented in two tables on the left page and two diagrams on the right page of each double page sub-section concerning a particular test parameter (result) quoted above the 1st table. In the majority of cases, one or more Youden plots are presented at the end of the test method section.

a) Test results submitted and calculated values (1st table)

[input data, within laboratory means and standard deviations, outliers, assigned value (robust average) and results of the data analysis for uncertainty and good conditions of repeatability]

b) Graphical presentation of the test results (Line plot)

[within laboratory means, within laboratory standard deviations, robust average (red line), robust standard deviation (doted blue lines) and general mean after outliers were eliminated (green line)]

c) Evaluation of the laboratory performance (Bar chart)

[z-score and limit value for the acceptance of data for the additional check of test method accuracy (z = 2,0; orange line)]

d) Additional check of the test method accuracy (2nd table)

[check of the normal distribution of the data, general mean after outliers were eliminated and evaluation of the data for repeatability and reproducibility characteristic values]

e) where applicable Youden plot (Scatter diagram)

[only in cases where two samples, A and B, were tested by the same procedure;
to check on the contribution of the random error and systematic error to the lab bias;
strongly biased labs are labelled with their LabCodeNo. in this plot]