

# Comparison of results of selected PTs in the occupational hygiene sector

*Evaluation report of the 2<sup>nd</sup> comparison within CoEPT  
EU contract GTC1-2002-73002*

Adriaan M.H. van der Veen<sup>1</sup>, Michel van Son<sup>1</sup>, Piotr Robouch<sup>2</sup>, Barry Tylee<sup>3</sup>

<sup>1</sup>Nederlands Meetinstituut, Department of Chemistry, Thijsseweg 11, 2629 JA Delft, the Netherlands

<sup>2</sup>Institute for Reference Materials and Measurements, Retiesweg, B-2440, Geel, Belgium

<sup>3</sup>Health and Safety Laboratory, Harpur Hill, Buxton, SK17 9JN, United Kingdom

Delft, the Netherlands, 30 August 2005.

Deleted: 27 July 2005

## **Copyright Notice**

This work is protected under copyright by the enterprises engaged in the work on this project. The work is supported in part by the European Commission under contract number GTC1-2002-73002, "Comparability of the operation and evaluation protocols of European proficiency testing schemes".

Extracts of this report may be reproduced provided that the source is acknowledged and the extract is not taken out of context.

## Summary

The 2<sup>nd</sup> comparison in the project “Comparability of the operation and evaluation protocols of European proficiency testing schemes” (CoEPT) was carried out between February and November 2004. The aim of this 2<sup>nd</sup> comparison was to see how the statistical protocols of the participating PT-providers perform under practical conditions, which are as close as possible to the normal conditions of the respective proficiency tests. The aim of the 2<sup>nd</sup> comparison is to see whether the results of the 1<sup>st</sup> comparison are confirmed when using the same samples in the proficiency testing schemes of the participating PT-providers.

A freshly prepared PT-material was made available to the PT-providers participating in the project. This approach was chosen because there were no suitable certified reference materials (CRMs) available for trace metals on filters.

This 2<sup>nd</sup> comparison has clearly demonstrated that it is possible to make useful inferences about results obtained on the same sample measured by two different laboratories. The implied assumptions concerning metrological traceability of the results, which is a prerequisite for being able to make a useful comparison, have proved to be valid for the PTs involved in the project. That this conclusion can be extended to other PTs is likely, but subject to proof.

Generally there was good agreement between any pair of assigned values in a dataset for a given parameter. From this agreement, it can be concluded that there is generally good agreement between the PTs of the participating PT-providers. Larger differences exist in the standard deviations and uncertainties estimated from the data reported by the participating laboratories, but they simply reflect differences in (average) performance. In conclusion, there is good comparability across the participating PTs for trace metals on filters.

The biggest differences between the PT-providers are observed in the establishment of the standard deviation for performance rating and the evaluation of measurement uncertainty associated with the assigned value. The assigned values from the PT-providers agree usually with the reference value established for the certified reference material (CRM) used in this comparison. This fact not only leads to the conclusion that the approaches of the PT-providers for establishing an assigned value are valid, but also that the evaluation of uncertainty is at an acceptable level.

A common protocol, based on ISO 13528 has been used to evaluate the datasets for all parameters of all PT-providers. This evaluation revealed also differences between the datasets, in particular in the level of scatter. It also made clear that the different approaches for establishing the standard deviation for performance rating come *in addition to* different levels of scatter.

## Table of contents

<b>Summary</b>	<b>III</b>
<b>Table of contents</b>	<b>IV</b>
<b>List of figures</b>	<b>V</b>
<b>List of tables</b>	<b>V</b>
<b>List of symbols</b>	<b>V</b>
<b>1 Introduction</b>	<b>1</b>
<b>2 Protocol of the 2<sup>nd</sup> comparison</b>	<b>2</b>
<b>2.1 Samples</b>	<b>2</b>
<b>2.2 Comparison</b>	<b>2</b>
<b>2.3 Reporting requirements</b>	<b>2</b>
<b>2.4 Co-ordination and evaluation</b>	<b>3</b>
<b>2.5 Participating PT providers</b>	<b>3</b>
<b>3 Results</b>	Error! Bookmark not defined.
<b>4 Evaluation and discussion</b>	<b>9</b>
<b>5 Conclusions</b>	<b>18</b>
<b>6 References</b>	<b>21</b>

## List of figures

**Error! No table of figures entries found.**

## List of tables

Table 1: Points of contact	3
Table 2: List of PT providers	3

## List of symbols

$C$	Cochran's statistic
$CI$	Confidence interval
$d$	Error threshold ( $P$ -score)
$d_j$	Spiked amount ( $j = 1 \dots 2$ )
$G$	Grubbs' statistic
$LoD$	Detection limit
$LoQ$	Quantification limit
$MAD$	mean of the absolute deviations
$P$	$P$ -score
$p$	number of laboratories
$RSD_R$	relative standard deviation (of an interlaboratory study) under reproducibility conditions
$s$	standard deviation
$s^*$	robust standard deviation (ISO 13528)
$s_L$	between-laboratory standard deviation (ISO 5725-2:1994)
$sMAD$	standard deviation of the median, estimated as $\frac{MAD}{0.6745}$
$s_{PT}$	standard deviation used for performance assessment
$s_R$	reproducibility standard deviation (ISO 5725-2:1994)
$s_r$	repeatability standard deviation (ISO 5725-2:1994)
$u$	standard uncertainty
$U$	expanded uncertainty
$x_i$	result of laboratory $i$

$x_{cv}$	consensus value
$x_{med}$	median
$\bar{x}$	mean value
$x_{PT}$	assigned value
$x_{ref}$	reference value
$Z$	Z-score

# 1 Introduction

---

The 2<sup>nd</sup> comparison in the project “Comparability of the operation and evaluation protocols of European proficiency testing schemes” (CoEPT) was carried out between February and November 2004. The aim of this 2<sup>nd</sup> comparison was to see how the statistical protocols of the participating PT-providers perform under practical conditions, which are as close as possible to the normal conditions of the respective proficiency tests. The aim of the 2<sup>nd</sup> comparison is to see whether the results of the 1<sup>st</sup> comparison are confirmed when using the same samples in the proficiency testing schemes of the participating PT-providers. The objectives are:

- ◆ To evaluate the equivalence of reference and consensus values as normally obtained in the PTs
- ◆ To compare the performance and rating of the participating laboratories

In the 1<sup>st</sup> comparison, emphasis was put on the comparability and robustness of the statistical protocols used by the participating PT-providers. A selection of datasets from different PTs was provided for this purpose. One of the weaknesses of the approach of the 1<sup>st</sup> comparison is that relatively little context information concerning the data is available. As mentioned in the report from the 1<sup>st</sup> comparison [1], this lack of context information sometimes impaired the evaluation of the data in accordance with the statistical protocol of the PT-provider. This issue is not present in this 2<sup>nd</sup> comparison, as the PT-provider was provided with samples that were to be distributed to the participants in the PT, just as any other sample. In most cases, PT-providers opted for distributing the samples from this project as additional sample, rather than using it instead of one of the regular samples.

This report covers the results of the Occupational Health Sector, in which 5 PT-providers participated. The draft protocol for this comparison was discussed during the 2<sup>nd</sup> Workshop in Geel (October 2003) and provided as draft in June 2004. The final version was made available in December 2004. The changes between the draft and final protocol were solely editorial.

## 2 Protocol of the 2<sup>nd</sup> comparison

---

### 2.1 Samples

The PT programmes and schemes, which participated in phase 1 of the COEPT project (UK, Spain, Poland, Germany and Hungary), continued into this phase. At the second workshop it was agreed that the currently available CRMs in this area, were not particularly suitable, or were too expensive for the likely budget available. It was therefore agreed that a special batch of samples would be made for the project, and that they would be loaded with the determinants chosen by the sector, within the concentration ranges normally used by the PT programmes.

In the spring of 2004 the WASP (UK) scheme prepared the samples (chromium, lead, cadmium and nickel on 25mm mixed ester membrane) and the validation process was started (according to the relevant CEN standard). However this process, using a number of recognised analytical techniques by a number of national institutes would take about two years, much longer than the timescale of the project, and a strategy was devised to assess the results after an extensive internal validation (using ICP OES/MS) and a limited external one. The analysis of the data at this point showed that the samples were fit for their purpose, and the despatch date then agreed with each scheme organiser. The samples were duly despatched, together with relevant blanks, over the summer of 2004 together with a certificate showing the relevant analytical details and the uncertainty budget calculations.

### 2.2 Comparison

For each of the sectors, a selection has been made of the parameters. The list of parameters was chosen to be the same as for the 1<sup>st</sup> comparison.

In the occupational hygiene sector, a membrane filter has been chosen as matrix, and the following trace elements:

Chromium, lead, cadmium, and nickel<sup>1</sup>

### 2.3 Reporting requirements

The following information was requested from the participating PT-provider:

1. Results from data evaluation as normally undertaken by the PT-provider;
2. Assigned value (consensus and reference value), and its associated uncertainty;
3. Results from outlier detection (if applicable); indication of ground for rejection
4. Performance rating of the laboratories; if such a rating is not foreseen, please indicate on the report forms how the laboratories are instructed to interpret the results

The evaluation of the data, as described in this report will include *inter alia* the following:

1. A review of the assigned values, in view of their uncertainties<sup>2</sup>; this review includes a comparison among the proficiency tests, where appropriate complemented by a comparison with estimates obtained by other statistical techniques
2. Differences between reference and consensus values, if the latter are used when a reference value is considered to be invalid or otherwise unusable for the purpose
3. Direct comparison of the performance ratings; in particular laboratories that seem to have performed poorly will be looked at

---

<sup>1</sup> Not all PT providers regularly have nickel in their programme. Therefore, this parameter will not be reported by all PT providers.

<sup>2</sup> This estimate of uncertainty can be the one provided from the PT provider, or an uncertainty estimated by the coordinator.

4. Comparison of performance requirements; if there are differences, these will be considered as *facts* rather than as subject to debate; the underlying deliberations are not at debate in CoEPT, but for the project objectives, these differences are relevant
5. Comparison of data screening methods and results
6. Compilation of stated uncertainties of assigned values, and a comparison with values obtained from some commonly applied techniques for deriving these uncertainties from comparison data; no assessment on the impact of these differences on performance rating, as with the current rating schemes, there is no impact

## 2.4 Co-ordination and evaluation

The comparison in the occupational health sector is coordinated by the Health and Safety Laboratory (HSL). The evaluation of the 1<sup>st</sup> intercomparison is done jointly by NMi VSL, IRMM and HSL (table 1).

**Table 1: Points of contact**

Organisation	Contact person	Address	E-mail
IRMM	Mr Piotr Robouch	Retieseweg, BE-2440 Geel, Belgium	<a href="mailto:piotr.robouch@irmm.jrc.be">piotr.robouch@irmm.jrc.be</a>
NMi VSL	Mr Adriaan van der Veen	Department of Mass and Chemistry, Postbus 654, 2600 AR Delft, The Netherlands	<a href="mailto:avdveen@nmi.nl">avdveen@nmi.nl</a>
HSL	Mr Barry Tylee	Harpur Hill, Buxton, SK17 9JN, UK	<a href="mailto:Barry.Tylee@hsl.gov.uk">Barry.Tylee@hsl.gov.uk</a>

## 2.5 Participating PT providers

The following 5 PT providers are project partners and contribute to the 1<sup>st</sup> comparison (table 2).

**Table 2: List of PT providers**

Acronym	Contact person	E-mail
HSL	Barry Tylee	<a href="mailto:Barry.Tylee@hsl.gov.uk">Barry.Tylee@hsl.gov.uk</a>
INSHT	Carmen Arroyo	<a href="mailto:mcarrovo@mtas.es">mcarrovo@mtas.es</a>
BIA	Dietmar Breuer	<a href="mailto:dietmar.breuer@hvbv.de">dietmar.breuer@hvbv.de</a>
NOFER	Jan Gromiec	<a href="mailto:jpgrom@imp.lodz.pl">jpgrom@imp.lodz.pl</a>
NCPH	Miklos Naray	<a href="mailto:naraym@fjokk.hu">naraym@fjokk.hu</a>

## 2.6 Interpretation of results

The first parameter to be compared is the assigned value. For each parameter, there is a reference value and an associated uncertainty available. The difference between the assigned value of a particular PT and the reference value can be regarded as a measure for the comparability of the PT with this independent reference. The difference is to be compared with its associated uncertainty, and as a criterion the following condition can be formulated

$$|x_{PT} - x_{ref}| \leq k \sqrt{u^2(x_{PT}) + u^2(x_{ref})} \quad (1)$$

or, equivalently,

$$\frac{|x_{PT} - x_{ref}|}{k\sqrt{u^2(x_{PT}) + u^2(x_{ref})}} \leq 1 \quad (2)$$

This ratio is commonly known as  $E_n$ . When  $E_n > 1$ , there can be two possible reasons

1. the difference  $x_{PT} - x_{ref}$  is too large
2. the uncertainty  $u(x_{PT})$  is too small

or of course a combination thereof. The presumption is made that the uncertainty associated with the reference value is credible.

The uncertainty associated with  $x_{PT}$  depends, when the assigned value is a consensus value, on the number of participants, among others. Generally speaking, the larger the number of participants, the smaller the uncertainty  $u(x_{PT})$  becomes. For the interpretation of the difference  $x_{PT} - x_{ref}$ , this influence is not an issue – it can be expected that the agreement between the reference value and the assigned value becomes better when the number of participants increases. The assumption is then made that all measurement results are metrological traceable, which is of course subject to verification.

With regard to the agreement of the uncertainty  $u_{PT}$ , it must first be recalled that the uncertainty associated with an uncertainty is typically between 10 – 30% relative. Furthermore, differences may come from

1. differences in the number of participants
2. differences in the (average) performance of the participants
3. evaluation method

The evaluation methods have been compared in the 1<sup>st</sup> comparison [ 1].

## 3 Protocols of the PT-providers

---

### 3.1 Summary of statistical protocol of PT H01

Participants analysing metal analytes are sent one sample containing 3 analytes once a year. The Grubbs test at the 95% level is used to determine outlying results. For the metals in dust the true value is determined using the certified value. Participants are scored using Z scores. Satisfactory performance is determined as  $Z < \text{or equal } 2$ .

#### 3.1.1 General statistical parameters

The means are taken of the three single results per harmful substance listed in the analytic report. A single mean value is computed as follows:

$$x_{jk} = \frac{1}{n_{jk}} \sum_{i=1}^{n_{jk}} x_{ijk} \quad (3)$$

where  $x_{ijk}$  denotes a single observation,  $x_{jk}$  the average of  $n_{jk}$  replicate measurements,  $i$  the index over the replicates,  $j$  that over the laboratories, and  $k$  that over the values.

The calculation of the total mean value  $x_k$  follows, which is needed for additional statistical evaluations, such as the total standard deviation  $s_k$ , Grubbs' test, and Z-score in the round-robin tests for solvents, hydrocarbons, PAHs, and inorganic acids. The index value for the round-robin tests for metals and chromium(VI) compounds is determined with the help of reference laboratories (see also: index value). Total mean value:

$$x_k = \frac{1}{N_k} \sum_{j=1}^{N_k} x_{jk} \quad (4)$$

and the total standard deviation

$$s_k = \sqrt{\frac{1}{N_k - 1} \sum_{j=1}^{N_k} (x_{jk} - x_k)^2} \quad (5)$$

where  $N_k$  is the mean of the laboratory means for level  $k$ .

#### 3.1.2 Outlier test according to Grubbs

It is generally assumed that the data material is subject to normal distribution. The outlier test is then conducted at the 95% level (on both sides  $\alpha = 2.5$ ).

By eliminating the outliers, a mean value is finally found that is very close to the "true value" of the sample.

With the help of the outlier-free mean value and the outlier-free standard deviation, the Z-score of all participating laboratories – including those eliminated as outliers – is calculated. The Z-score can be considered as the quality characteristic of the mean value of the single laboratories (see below).

In the outlier test conforming to Grubbs, the procedure is such that all laboratory means are first established, and then the total mean value and the total standard deviation are determined. Following this, the difference between the individual mean values and the total mean value is established, while the individual mean value with the largest difference from the index value / total mean value is marked with a \* and entered into the formula for the Grubbs' outlier test. The sample quantity  $t$  is

compared with the table value for a significance level of  $\alpha = 2,5\%$  as viewed from both sides (see above).

If the suspicion is confirmed that it is an outlier in question, this is then removed from the collective data and the total mean value and total standard deviation are recalculated. The difference between the new total mean value, or the index value and the single mean values are established anew, and the largest difference is put into the GRUBBS formula. If the suspicion is once again confirmed that it is an outlier in question, this, too, is removed.

This cycle of calculations is continued until no more outliers can be eliminated.

If the data are too scattered, to the point that no outliers can be determined by the GRUBBS test, then the outliers are defined by the deviation from the index value. An outlier can then be recognised if the deviation is  $\pm 36\%$ .

### 3.1.3 Z-score evaluation

In evaluating the Z-score, all the single mean values are considered, including the values recognised as outliers by the GRUBBS test. Yet the total standard deviation and the total mean value / index value necessary for the calculation are adjusted for outliers! The z-score evaluation is based on the following equation:

$$Z = \frac{x_i - x_{PT}}{s_{PT}} \quad (6)$$

Normally, the standard deviation of the data material adjusted for outliers is used for the standard deviation for performance rating  $s_{PT}$ . However, the maximum permissible deviation  $s_{PT}$  for calculating the Z-score is 10% of the index value.

The individual results are now assessed according to the following scheme:

Classification of Z-scores has been done as follows:

$ Z  \leq 1$	good
$1 <  Z  < 2$	satisfactory
$2 <  Z  \leq 3$	questionable
$ Z  > 3$	unsatisfactory

A result with  $|Z| \leq 2$  is considered satisfactory, meaning that the round-robin test is considered "passed". If the Z-score is above two, a test of the analytic methods used is advised.

Using the z-scores, still other statistically important characteristic quantities can be identified, but these will not be further discussed here.

## 3.2 Summary of statistical protocol of PT H02

This scheme has organised PTs for metals (Pb, Cd, and other metals on membrane filter):

for 20 labs Pb and Cd on membrane filters,

Samples are supplied at 5 conc. level+ 2 blank

Conc. range: for Pb 3-35 ug/filter, for Cd 2-20 ug/filter ,

Evaluation: Q-score ( for acceptance: < 20 % ), Z-score,

Results : CV% for Pb : 4 - 6 % ( at conc. range: 8-35 ug/filter)

for Cd: 8 - 12 % ( at conc. range: 8-20 ug/filter)

### 3.3 Summary of statistical protocol of PT H03

PICC-MET Metals Lead and Chromium spotted onto cellulose ester filters. Four filters supplied four times a year.

The 'true value' is determined using the participant mean, outliers excluded on by the Dixon test. Limits for acceptable results are +/- 26%.

(For lead results +/- 0% of the median are excluded for the calculation of the participant mean)

The proportion of the uncertainty attributed to the preparation is less than 3% for the metals on filters.

### 3.4 Summary of statistical protocol of PT H04

The results of control testing in each round are subject to statistical analysis in order to determine:

- Assigned value
- Acceptable error limits
- Quality performance in a given laboratory

The assigned value of the concentration of a substance determined in a control sample is calculated as an arithmetic mean of a set of determination results for that sample in all the laboratories, having rejected the outliers. After the outliers have been rejected the Z-score is calculated from the following formula:

$$Z = \frac{x_i - x_{PT}}{s_{PT}} \quad (7)$$

Where  $x_i$  denotes the determined result of the laboratory,  $x_{PT}$  is the arithmetic mean (assigned value), after rejecting outliers,  $s_{PT}$  is the standard deviation (taken as standard deviation for performance rating).

At a given round the laboratories are evaluated and classified according to the number of unacceptable errors made during an analysis of a given control sample set. The following marks are assigned:

- very good – if the number of all analysis fall within the range of the acceptable error
- satisfactory – if the number of results with the acceptable error does not exceed 20%
- unsatisfactory – if more than 20% of results has the unacceptable error.

Classification of laboratories after n-the control test

The participating laboratories are listed in ascending standardised Z-scores. The top laboratory receives 'n' points where 'n' is the number of participants who performed the determination of a given sample within a given round. The next laboratory in the list receives 'n-1' points and the subsequent one the preceding value minus 1. The lowest laboratory receives '0' points.

### 3.5 Summary of statistical protocol of PT H05

Four samples (with differing analyte levels) and blanks are despatched to participants four times a year. For each analyte (determinant) reported by a participant, a Performance Index (PI) is calculated for that round. The basis of the performance index for a given laboratory, round and analyte is the calculation of the sum of the squared deviations from 1.00 of the standardised results, divided by 4. This is equivalent to the variance of the standardised results about 1.00. To avoid using small numbers, this basic performance index is arbitrarily multiplied by 10,000 to give the reported Performance Index that will be rounded to the nearest integer. The lower the index, the better the performance.

In addition, a Running Performance Index (RPI) is calculated by averaging the performance indices for the best four of five rounds of the scheme. Should a laboratory for any reason fail to report one round for an analysis it normally carries out, the latest 4 rounds' results will be used to calculate an RPI; however, more than one such 'gap' in a 5-round period will mean that no RPI is calculated. The square root of this score is an estimate of standard uncertainty in terms of the percentage standard deviation (%RSD) over the analytical range used. Laboratories are advised to compare this figure with their own in house uncertainty estimates.

The calculations of sample means, standardised results, PIs and RPIs will be carried through to their conclusion with full precision, but will be rounded to realistic levels on the report sheets. This may on occasion give rise to small apparent discrepancies if the calculations are checked through by participants themselves. These will not be significant, and the reported values will represent the best estimates of the various parameters.

The report form includes the performance index for the latest round and the running performance index (once at least 4 rounds have been completed). To give an indication of performance relative to participants as a whole, the mean PI and RPI (averaged over all participants for that analyte exclusive of any outlier results) and the ranking order of the laboratory in terms of RPI are also provided.

#### Performance Categories

Because the quantity of data going into the single round PI is small (only four results), the uncertainties in using the PI as a measure of underlying performance are rather high. Therefore, only the RPI, with 16 results contributing, is used to assign laboratories to a performance category for each analyte. Note that categorisation is on an analyte-by-analyte basis: no attempt is made to combine results from more than one analyte into a global classification.

For each analyte, a running performance index reference value is set, based initially on data available from performances of laboratories in the previous Analytical Quality Assurance (AQUA) scheme, the NIOSH Proficiency Analysis Testing (PAT) scheme. These reference values may need to be changed if experience in WASP indicates that any are widely inappropriate.

The upper and lower 95% points of the theoretical distribution of running performance indices produced by a laboratory operating consistently at the reference level are used to divide PRIs into three categories, designated 1, 2 and 3. Details are given in Appendix 1, paragraph 10. A laboratory's performance is categorised according to where its calculated RPI falls relative to these boundaries.

The performance category is indicated on the report sheet (Figure 1), the category boundaries and the laboratory's historical performance in relation to them are also plotted

#### Further Statistical Information

Additional parameters for the mean deviation and within round variance (DELTA and SSW) are provided on the report sheet to enable further statistical tests for bias, etc. to be made by participants if they so wish.

## 4 Evaluation of datasets using a common approach

---

### 4.1 Common protocol

The discussions during the 3<sup>d</sup> CoEPT workshop held in Ede, the Netherlands, on 8 February 2005 indicated that there was a need for evaluating the data with a common statistical protocol, to assess broadly the equivalence between the PTs participating in the project.

The protocol selected was based on ISO 5725-5:1998 [2], algorithm A, which is also found in ISO 13528 [3]. The consensus value and the robust standard deviations have been calculated using the following algorithm:

Denote the  $p$  items of data, sorted into increasing order, by:

$$x_1, x_2, \dots, x_i, \dots, x_p$$

Denote the robust average and robust standard deviation of these data by  $x^*$  and  $s^*$ . Calculate initial values for  $x^*$  and  $s^*$  as:

$$x^* = \text{median of } x_i \quad (8)$$

$$s^* = 1.483 \times \text{median of } |x_i - x^*| \quad (9)$$

Update the values of  $x^*$  and  $s^*$  as follows. Calculate:

$$j = 1.5s^* \quad (10)$$

For each  $x_i$  calculate

$$x_i^* = \begin{cases} x^* - j & x_i < x^* - j \\ x^* + j & x_i < x^* + j \\ x_i & \text{otherwise} \end{cases} \quad (11)$$

Calculate the new values of  $x^*$  and  $s^*$  from:

$$x^* = \frac{1}{p} \sum_{i=1}^p x_i^* \quad (12)$$

and

$$s^* = \frac{1}{p-1} \sqrt{\sum_{i=1}^p (x_i^* - x^*)^2} \quad (13)$$

The robust estimates  $x^*$  and  $s^*$  may be derived by an iterative calculation, i.e. by updating the values of  $x^*$  and  $s^*$  several times, until the process converges. Convergence may be assumed when there is no change from one iteration to the next in the third significant figure of the robust standard deviation and of the equivalent figure in the robust average.

Furthermore, it was assumed  $Z$ -scores be used, with the following interpretation:

$|Z| \leq 2$ : satisfactory performance (in graphs: green area)

$2 < |Z| < 3$ : questionable performance (in graphs: yellow area)

$|Z| \geq 3$ : unsatisfactory performance (in graphs: red area)

An alternative interpretation of the colours in the graphs is as follows:

Green = 95% CI<sup>1</sup>

Green + yellow = 99% CI

Where a  $t$ -distribution is assumed. Furthermore, it is assumed that the datasets are symmetrical, which is not always the case for the parameters discussed in the next section.

## 4.2 Data evaluation of selected data sets

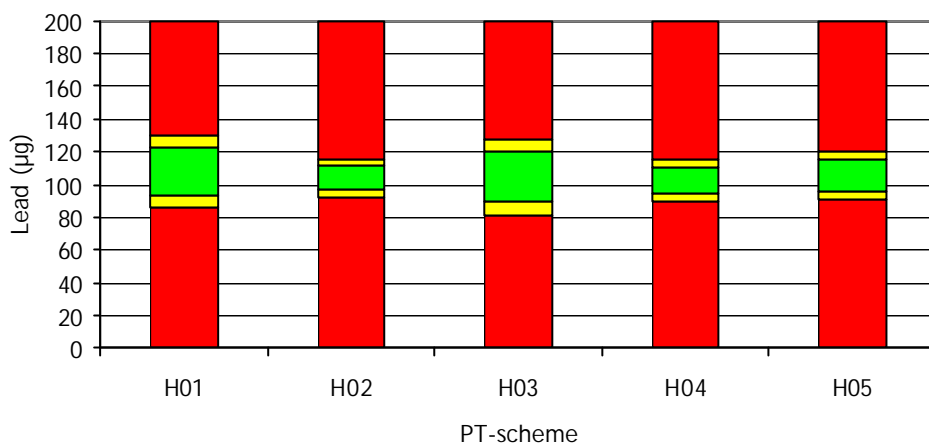
Table 3 summarises the results of selected parameters.

**Table 3: Robust means and standard deviations in accordance to ISO 13528, algorithm A**

		<b>H01</b>	<b>H02</b>	<b>H03</b>	<b>H04</b>	<b>H05</b>
Pb	$x_{cv}$	108.4	104.1	104.7	102.5	105.7
	$s^*$	7.36	3.81	7.69	4.16	5.01
Ni	$x_{cv}$	62.24	60.34	59.52		
	$s^*$	3.07	3.46	5.98		
Cr	$x_{cv}$	173.52	165.29	163.29	155.67	164.66
	$s^*$	10.23	6.88	15.83	15.75	11.02
Cd	$x_{cv}$	19.97	19.67	19.47	19.04	19.54
	$s^*$	1.68	0.58	1.53	0.76	0.92

<sup>1</sup> CI = Confidence interval

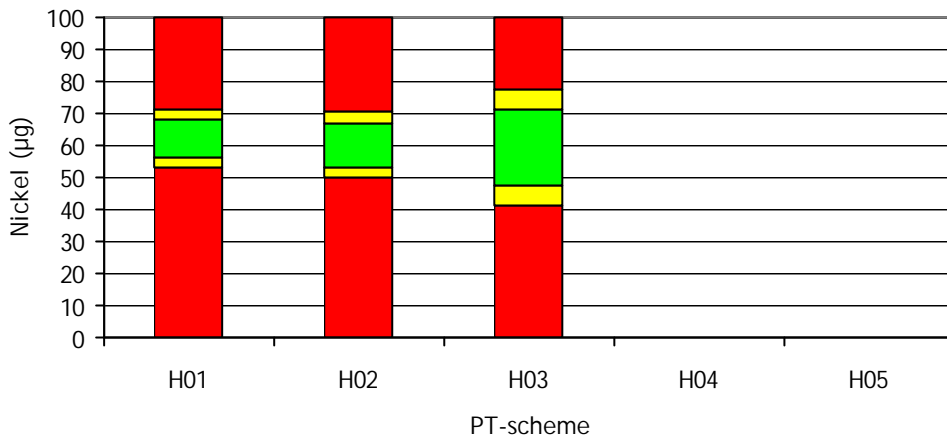
### CoEPT -- ISO 13528 evaluation



**Figure 1: Robust consensus means and standard deviations for lead**

The results of the evaluation of the data using the protocol outlined in section 4.1 for lead are given in figure 1. The data obtained from PT-providers H02 and H04 shows the smallest scatter. There is a good agreement among the consensus values. The agreement among the estimated robust standard deviations is less good, and reflects the differences in levels of scatter in the results.

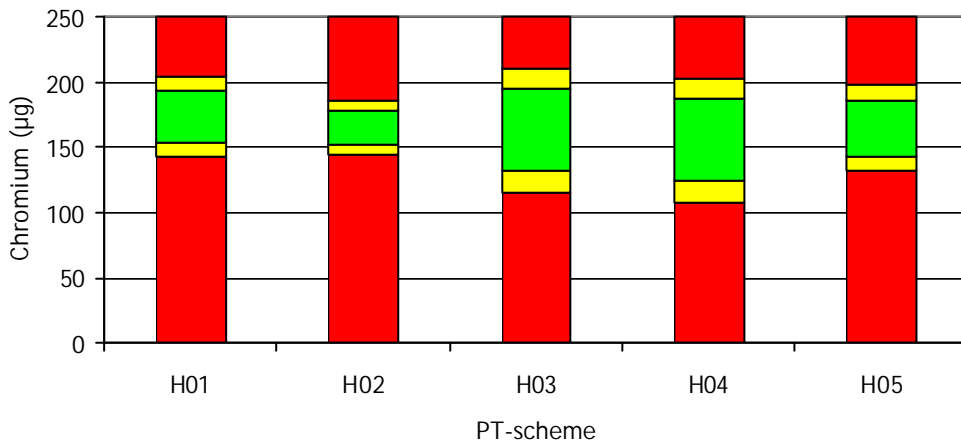
CoEPT -- ISO 13528 evaluation



**Figure 2: Robust consensus means and standard deviations for nickel**

The results of the evaluation using the common protocol for nickel are given in figure 2. PT-providers H04 and H05 do not include nickel in their schemes. The consensus values agree well, the dataset from PT-provider H03 shows somewhat more scatter than that of the other PT-providers.

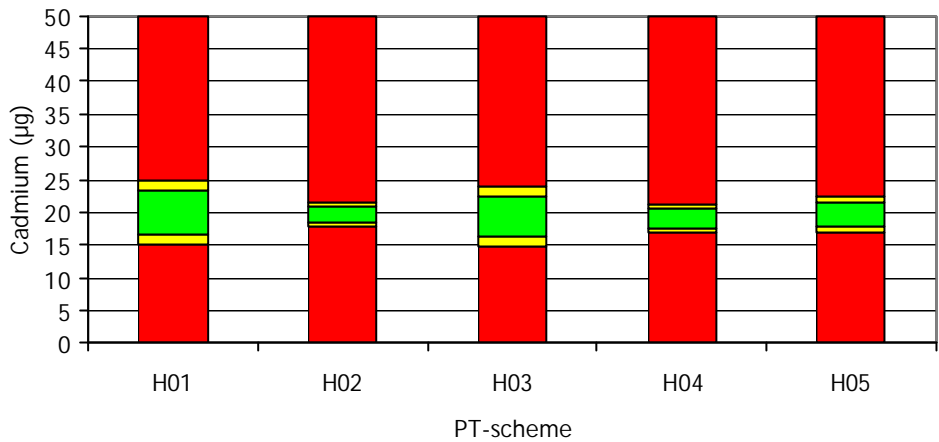
CoEPT -- ISO 13528 evaluation



**Figure 3: Robust consensus means and standard deviations for chromium**

The evaluation of the results using the common protocol for chromium (figure 3) shows appreciable differences in the robust standard deviations, reflecting appreciable differences in the levels of scattering in the datasets. The agreement of the consensus values is good.

CoEPT -- ISO 13528 evaluation



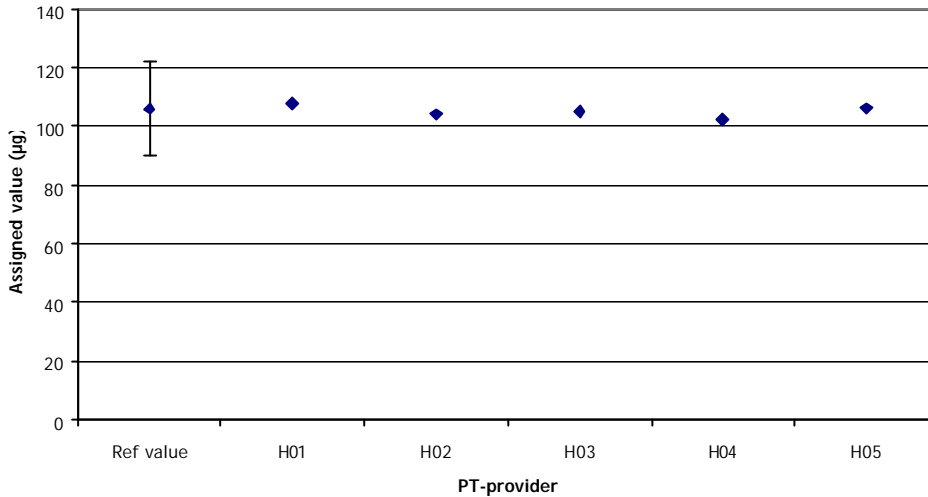
**Figure 4: Robust consensus means and standard deviations for cadmium**

The evaluation of the data for phenanthrene using the common protocol (figure 4) shows good agreement for the consensus values, but appreciable differences for the robust standard deviations.

## 5 Results

### 5.1 Lead

The assigned values from the PT-providers and the reference value established for the material for lead are shown in figure 5. Although no uncertainty data are available, it can be concluded that the assigned values from the PT-providers agree well with the reference value; they fall all well within the 95% confidence interval.



**Figure 5: Assigned values for lead**

In table 4, the consensus values and standard deviations for performance rating are given for lead. The values printed in italic have been calculated from the data (these characteristics did not come from the PT-provider). There is quite some difference in the standard deviations for performance rating ( $s_{PT}$ ).

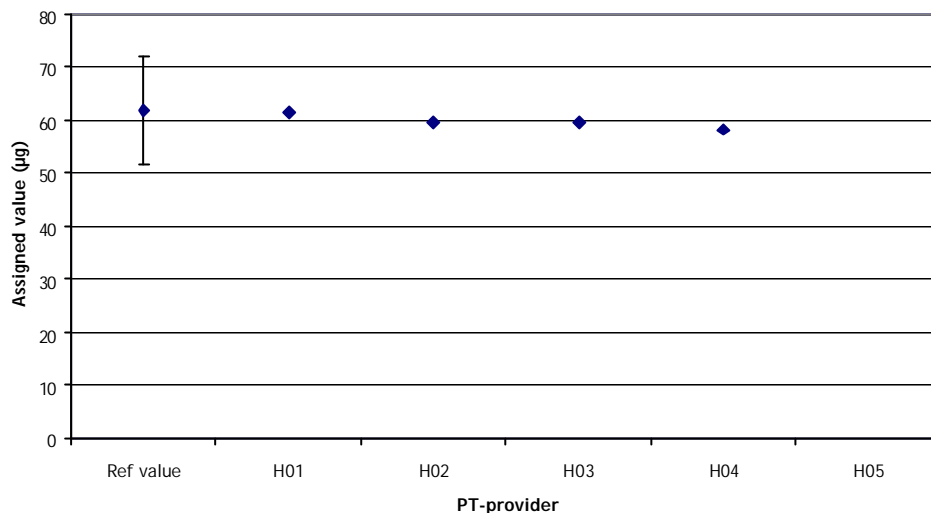
**Table 4: Consensus values, assigned values, and standard deviations for lead**

PT	$\bar{x}_{cv}$	$U(\bar{x}_{cv})$ ( $k=1.96$ )	$\bar{x}_{av}$	$U(\bar{x}_{av})$ ( $k=1.96$ )	$s$	$s_{PT}$
H01	107.8		107.8		8.5	8.5
H02	104.1		104.1		4.1	4.1
H03	<i>103.6</i>		<i>103.6</i>		<i>11.6</i>	<i>11.6</i>
H04	102.3		102.3		3.7	3.7
H05	106.1		106.1		7.3	7.3

### 5.2 Nickel

The assigned values from the PT-providers and the reference value established for the material for nickel are shown in figure 6. PT-provider H05 did not include nickel in its programme. Although no

uncertainty data are available, it can be concluded that the assigned values from the PT-providers agree well with the reference value; they fall all well within the 95% confidence interval.



**Figure 6: Assigned values for nickel**

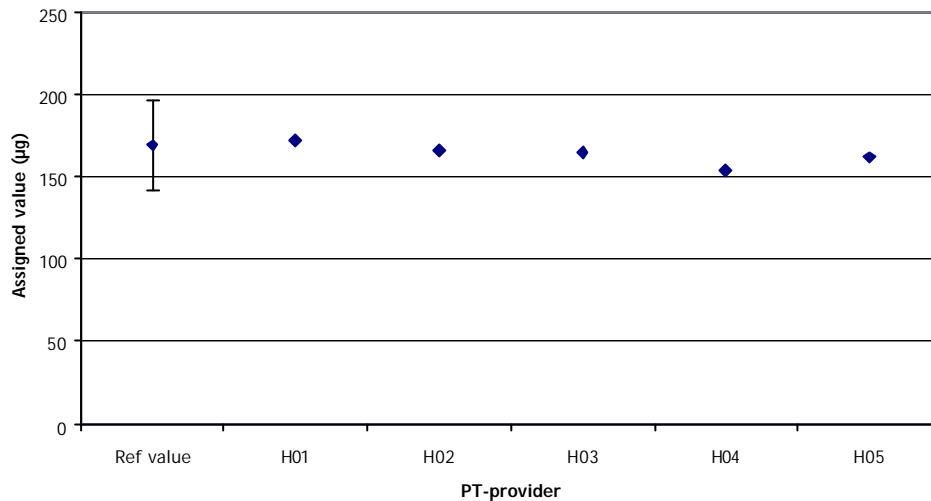
In table 5, the consensus values and standard deviations for performance rating are given for nickel. The values printed in italic have been calculated from the data (these characteristics did not come from the PT-provider). There is quite some difference in the standard deviations for performance rating ( $s_{PT}$ ), even when the re-calculated standard deviation of PT-provider H03 is not considered.

**Table 5: Consensus values, assigned values, and standard deviations for nickel**

PT	$x_{cv}$	$U(x_{cv})$ ( $k=1.96$ )	$x_{av}$	$U(x_{av})$ ( $k=1.96$ )	$s$	$s_{PT}$
H01	61.4		61.4		4.4	4.4
H02	59.6		59.6		4.8	4.8
H03	<i>59.2</i>		<i>59.2</i>		<i>8.4</i>	<i>8.4</i>
H04	58.2		58.2		2.5	2.5
H05						

### 5.3 Chromium

The assigned values from the PT-providers and the reference value established for the material for chromium are shown in figure 7. As is the case for the other parameters, no uncertainty data are available. Nevertheless, it can be concluded that the assigned values from the PT-providers agree well with the reference value; they fall all well within the 95% confidence interval.



**Figure 7: Assigned values for chromium**

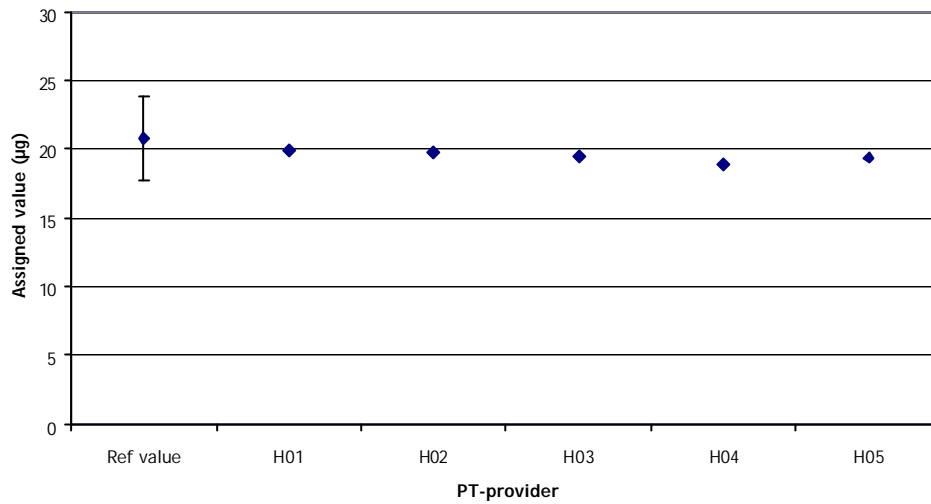
In table 6, the consensus values and standard deviations for performance rating are given for chromium. The values printed in italic have been calculated from the data (these characteristics did not come from the PT-provider). There is quite some difference in the standard deviations for performance rating ( $s_{PT}$ ).

**Table 6: Consensus values, assigned values, and standard deviations for chromium**

PT	$x_{cv}$	$U(x_{cv})$ ( $k=1.96$ )	$x_{av}$	$U(x_{av})$ ( $k=1.96$ )	$s$	$s_{PT}$
H01	172.2		172.2		12.6	12.6
H02	165.8		165.8		7.5	7.5
H03	<i>159.7</i>		<i>159.7</i>		<i>30.0</i>	<i>30.0</i>
H04	153.4		153.4		10.7	10.7
H05	161.8		161.8		22.3	22.3

## 5.4 Cadmium

The assigned values from the PT-providers and the reference value established for the material for chromium are shown in figure 8. As is the case for the other parameters, no uncertainty data are available. Nevertheless, it can be concluded that the assigned values from the PT-providers agree well with the reference value; they fall all well within the 95% confidence interval.



**Figure 8: Assigned values for cadmium**

In table 4, the consensus values and standard deviations for performance rating are given for cadmium. The values printed in *italic* have been calculated from the data (these characteristics did not come from the PT-provider). The standard deviation for performance rating ( $s_{PT}$ ) agree reasonably well, not taking into consideration the re-calculated figure of PT-provider H03.

**Table 7: Consensus values, assigned values, and standard deviations for cadmium**

PT	$x_{cv}$	$U(x_{cv})$ ( $k=1.96$ )	$x_{av}$	$U(x_{av})$ ( $k=1.96$ )	$s$	$s_{PT}$
H01	19.9		19.9		1.7	1.7
H02	19.8		19.8		1.3	1.3
H03	<i>19.0</i>		<i>19.0</i>		<i>3.0</i>	<i>3.0</i>
H04	18.9		18.9		1.0	1.0
H05	19.4		19.4		1.5	1.5

## 6 Discussion

---

### 6.1 Evaluation of datasets using the common protocol

The use of a common statistical protocol as described in 4.1 as an aid to interpret the datasets from the 10 PT-providers has proved to be very useful. Although in the first comparison it was concluded that there existed no big differences in the statistical protocols of the participating PT-providers, in particular the interpretation of the various standard deviations calculated from data and used has proved to be sometimes difficult [1]. The statistical protocol chosen has been based on algorithm A of ISO 5725-5 [2] and can also be found in the new standards ISO 13528 [3]. Furthermore, it has been assumed that Z-scores be used, which for this sector is usually the preference of the PT-providers too. The common protocol should however not be misinterpreted as being better, preferred, or otherwise superior to the statistical protocols of the PT-providers as described in chapter **Error! Reference source not found.**

The validity of the approach of using a common protocol lies in the following observations:

1. Calculating a robust consensus value and a robust standard deviation is possible from any dataset in this sector
2. All consensus values obtained can be compared, if necessary taking into consideration their respective uncertainties (which can be calculated as described in, e.g., ISO 13528 [3])
3. All robust standard deviations can be compared
4. There is no 'noise' in the interpretation of consensus values or standard deviations.
5. The approach is relatively insensitive to reporting errors, straight blunders, and outliers.

### 6.2 Comparing results across PTs

The evaluations of the datasets of all PT-providers for all parameters revealed that the consensus values agree, and therefore the results of the ('satisfactory' performing) PT-participants. The degree of agreement between two laboratories in a single PT is characterised by the reproducibility as defined in ISO 5725-2 [4],

$$R = 2\sqrt{2}s_R \quad (14)$$

where in this instance for the reproducibility standard deviation  $s_R$  the robust standard deviation of the common protocol should be substituted. The reproducibility can be interpreted as the maximum deviation between any two results of two PT-participants that is still (statistically) insignificant at a given level of confidence, in the case of ISO 5725, 95%. If two PTs have (significantly) different levels of scatter, characterised in the case of the common protocol by two different values for the robust standard deviation, the expression for the maximum difference across the proficiency tests becomes

$$R_{12} = 2\sqrt{(s_1^*)^2 + (s_2^*)^2} \quad (15)$$

It can easily be verified that if both robust standard deviations are equal, then the expression for the reproducibility as taken from ISO 5725-2 is obtained. Equation (15) is only valid when the consensus values are equal. Furthermore, it is assumed that the metrological traceability of the results of the PT-provider meets elementary requirements, such as those stipulated in clause 5.6 of ISO/IEC 17025 [6]. This 2<sup>nd</sup> comparison in the CoEPT project is unique in its approach that in different PT-schemes, the same samples have been used, so that the condition concerning the consensus values is reasonably met. The assumption concerning the metrological traceability is harder to verify, but if the PT-

participants regularly participate in PTs in which the provider gives attention to this aspect in the design of its protocol, it is reasonable to assume that this assumption is also valid.

The agreement among the consensus values is exactly for this reason so important, that it gives evidence that something useful can be said about the comparability of the results of two PT-participants participating in two different PTs, even if no direct link has been established. Assessments like the one made here are also possible in practice, but it is not to be expected that it will happen on the scale as done in this project.

In reality, one may want judge whether the difference between two measurement results (on the same sample) from two laboratories participating in different PTs (if they participate in one PT, this case is covered by ISO 5725-2 [4]) can be explained from the results of participating in the respective PTs. Before modelling, some assumptions must be made:

1. The standard deviation obtained from the proficiency test is a good measure for the actual standard uncertainty of the measurement under study
2. Elementary requirements of metrological traceability are met
3. The standard deviations obtained in the proficiency tests are relevant for the range in which the results are to be compared

Most of the assumptions are self-explanatory and it is quite obvious why they are made. The third assumption is possible the most problematic one in practice. In many cases, the uncertainty associated with a measurement result is dependent on its value; it is in analytical chemistry often assumed that the relative uncertainty is more or less constant over a large range. Many performance characteristics (e.g., repeatability and reproducibility) in written standards indicate that this assumption is often, but certainly not always valid.

The first step in the comparison is to calculate the difference between the two measurement results  $x_1$  and  $x_2$

$$\Delta x_{12} = x_1 - x_2 \quad (16)$$

This difference is to be compared with the uncertainty associated with  $x_1$  and  $x_2$  respectively. From the proficiency tests in which the laboratories participate, there are standard deviations  $s_1$  and  $s_2$ , which can be assumed to be a "good measure" of the standard uncertainties  $u_1$  and  $u_2$ , associated with results  $x_1$  and  $x_2$  respectively. Consequently, the difference  $\Delta x_{12}$  is to be compared with

$$R_{12} = 2\sqrt{(u_1)^2 + (u_2)^2} \quad (17)$$

setting

$$u_1 = s_1 \text{ and } u_2 = s_2 \quad (18)$$

### 6.3 Results from the 5 PTs

The results from the 5 PTs generally indicate good comparability among participating laboratories in a PT, and across PTs. For the occupational hygiene sector as a whole, the assumptions made in 6.2 hold, which makes that in reality a comparison of results of laboratories *across* proficiency tests is possible, and the outcome is meaningful. One may argue that in practice it is not always possible to demonstrate on a case-by-case basis metrological traceability of the assigned values in the two PTs concerned, but –as appropriate– it is always possible to take appropriate measures when an actual sample is measured. The results of the PTs can aid interpreting the differences observed when two different laboratories measure the same sample.

## 7 Conclusions

---

The 2<sup>nd</sup> comparison in the CoEPT project has clearly demonstrated that it is possible to make useful interferences about results obtained on the same sample measured by two different laboratories. The implied assumptions concerning metrological traceability of the results, which is a prerequisite for being able to make a useful comparison, have proved to be valid for the PTs involved in the project. That this conclusion can be extended to other PTs is likely, but subject to proof.

Generally there was good agreement between any pair of assigned values in a dataset for a given parameter. From this agreement, it can be concluded that there is generally good agreement between the PTs of the participating PT-providers. Larger differences exist in the standard deviations and uncertainties estimated from the data reported by the participating laboratories, but they simply reflect differences in (average) performance. In conclusion, there is good comparability across the participating PTs for trace metals on filters.

The biggest differences between the PT-providers are observed in the establishment of the standard deviation for performance rating and the evaluation of measurement uncertainty associated with the assigned value. The assigned values from the PT-providers agree usually with the reference value established for the PT-material used in this comparison. This fact not only leads to the conclusion that the approaches of the PT-providers for establishing an assigned value are valid, but also that the evaluation of uncertainty is at an acceptable level. Certainly the publication of ISO 13528 [3] is important to mention here, as it provides a reasonable approach to calculate the uncertainty associated with a consensus value.

A common protocol, based on ISO 13528 has been used to evaluate the datasets for all parameters of all PT-providers. This evaluation revealed also differences between the datasets, in particular in the level of scatter. It also made clear that the different approaches for establishing the standard deviation for performance rating come *in addition to* different levels of scatter.

## 8 References

---

- [1] Van der Veen A.M.H., Robouch P., Tylee B., “Comparison of PT protocols used in the occupational health sector – Evaluation report of the 1st comparison within CoEPT, EU contract GTC1-2002-73002”, Delft, 23 March 2004
- [2] International Organization for Standardization, “ISO 5725-5:1998 Accuracy (trueness and precision) of measurement methods and results -- Part 5: Alternative methods for the determination of the precision of a standard measurement method”, ISO Geneva, 1998
- [3] International Organization for Standardization, “ISO 13528:2005 Statistical methods for use in proficiency testing by interlaboratory comparisons”, ISO Geneva, 2005
- [4] International Organization for Standardization, “ISO 5725-2:1994 Accuracy (trueness and precision) of measurement methods and results - Part 2: Basic method for the determination of repeatability and reproducibility of a standard measurement method”, Statistical methods for quality control, Vol. 2 (1994), pp 30-74
- [5] International Organization for Standardization, “ISO 5725-1:1994 Accuracy (trueness and precision) of measurement methods and results - Part 1: General principles and definition”, Statistical methods for quality control Vol. 2 (1994), pp 9-29
- [6] International Organization for Standardization, “ISO/IEC 17025:1999 General requirements for the competence of testing and calibration laboratories”, ISO Geneva