

Summary Report of the 1st Comparison in CoEPT

Summary Report of the 1st Comparison in CoEPT

EU contract GTC1-2002-73002

Deliverable D14

Adriaan M.H. van der Veen
Nederlands Meetinstituut
Department of Mass and Chemistry
Report number S-MC.04.12
Delft, 01 April 2004

Summary

Within the framework of the project “Comparability of the operation and evaluation protocols of European proficiency testing schemes” (CoEPT), a comparison of statistical protocols used for proficiency testing schemes is foreseen. The comparison is necessary to check that

- there is reasonable agreement between the statistical protocols of the various proficiency testing schemes
- the statistical protocols are well understood by the project consortium

In all sectors, for all parameters, the protocols used for processing the data are based on the assumption that data are normally distributed. Problems like bi-modality are dealt with quite differently, depending on the sector, the parameter, and the method. For example, in the food sector some PT providers make a method-specific evaluation, whereas others refer to the preferred (reference) method. In the water sector, some PT providers also use a method-specific evaluation, whereas others simply ignore such effects. Usually there is a plausible rationale for choosing a specific approach, but it is difficult to tell what approach is the “best”.

Values reported as “< x” (where x represents a number) are generally not rated, and not included in the establishment of the consensus value either. In fact, it is hard to tell whether a laboratory has any benefit from reporting a result in such a fashion. Many laboratories have reporting limits, which are certainly in their relationships with customers a useful instrument. For comparing results however, these reporting limits practically obstruct any processing from side of the PT provider.

The agreement between the protocols in each sector is generally good, tested on drinking water datasets is good. Laboratories with unsatisfactory performance in one PT are likely to get the same rating in another. When differences occur, they are usually due to differences in minimum requirements for acceptable performance, rather than due to differences in statistical protocols. In some sectors, like water [3] and food [5], the context of the data is very important, as it influences the decision taking process, and good attention should be paid to providing datasets for such a comparison.

The agreement in terms of rating the laboratories’ results is generally good: in 70% (water), 84% (food), 82% (soil), and 88% (occupational hygiene) of the cases, the rating was the same for all PT providers. Differences in rating, as they occur, are usually due to differences in the requirements set for rating the participants. In all cases, the differences can be explained from the statistical protocol.

The co-operation with the PT providers is generally good. The task of the PT providers is – despite the efforts put in the organisation- not always easy. Datasets from one PT are not necessarily ideal for the protocol of another PT. This conclusion leads to the recommendation to specify very clearly what data and context information is needed in the 2nd comparison from the laboratories.

Table of contents

Summary	II
Table of contents	III
List of figures	IV
List of tables	IV
1 Introduction	1
2 Methods	2
2.1 General	2
2.2 Comparison	2
3 Results	4
3.1 Participation	4
3.2 Nature of the statistical protocols	4
3.3 Water sector	4
3.4 Soil sector	7
3.5 Food sector	9
3.6 Occupational hygiene sector	10
4 Conclusions	12
5 References	13

List of figures

Figure 1: Means and standard deviations used for rating in the food sector	9
Figure 2: Means and standard deviations used for rating in the occupational hygiene sector (datasets #019 - #702-2)	10
Figure 3: Means and standard deviations used for rating in the occupational hygiene sector (datasets #702-3 - #705)	11

List of tables

Table 1: Means and standard deviations used for rating in the food sector	Error! Bookmark not defined.
Table 2: Means and standard deviations used for rating in the occupational health sector	Error! Bookmark not defined.

1 Introduction

Within the framework of the project “Comparability of the operation and evaluation protocols of European proficiency testing schemes” (CoEPT), a comparison of statistical protocols used for proficiency testing schemes is foreseen. This report summarises the evaluation of the results of the 1st comparison for four important sectors of proficiency testing in Europe. A protocol [1] for this exercise has been developed, stipulating the most important aspects for the comparison of the statistical protocols of the participating proficiency testing providers are compared. The comparison is necessary to check that

- there is reasonable agreement between the statistical protocols of the various proficiency testing schemes
- the statistical protocols are well understood by the project consortium

It is assumed that the proficiency testing schemes included in this project are run with well-established statistical protocols, but despite of the fact that there are several commonly used statistical approaches to the evaluation of proficiency testing data, there is hardly any indication whether these approaches, provided that they can be used, provide reasonably the same output. This assumption should of course be verified, and this aspect is the central focus in this part of the project.

A proficiency test is run on the basis of a (documented) design, hereafter referred as “statistical protocol”. This statistical protocol stipulates what data are gathered, for what purpose, and how they will be evaluated, and what type of conclusions can be drawn from the PT. The typical contents of such a protocol are given in ISO Guide 43-1 [2]. The basis for this comparison, as well as the protocol [1], is given in ISO Guide 43-1 [2].

This 1st comparison has the following objectives

- ◆ To evaluate (dis)similarities in statistical protocols
- ◆ To evaluate, whether these (dis)similarities lead to different results in terms of the assigned value [2], its uncertainty, and the establishment of the standard deviation (or other measure of dispersion) used for performance rating
- ◆ To evaluate, how statistical protocols cope with difficulties in datasets¹, such as results reported as “less than x ”, and bimodality or skewed distributions
- ◆ To compare the performance rating

The results of the four sectors water [3], soil [4], food [5], and occupational hygiene [6] have been published separately. This report covers the more general aspects and results of the 1st comparison and addresses some issues across sectors.

¹ A dataset is defined in this context as a series of laboratory results, for a single parameter, on a single matrix in a proficiency test.

2 Methods

2.1 General

For a comparison of statistical protocols of PT providers, datasets are needed. These datasets need be:

- representative for the field of measurement
- in a format that is compatible with the requirements of the PT provider
- complete, that is, all relevant context information is included

The first task in the comparison was the collection and selection of datasets. The main objective was to achieve a representative cross-section of the selected parameters for the four sectors. The only area where there were some problems was the occupational health sector [6], because the datasets delivered by the participating PT providers were hardly compatible with the needs expressed. A compromise could be worked out, that met the requirements stipulated in the protocol for this comparison [1].

For the other three areas, there was a good choice of datasets, and there were little problems in providing the PT providers with a collection of datasets that allowed assessing the performance of their statistical protocols. The format of the datasets was not always as desired. For example, in the water sector one PT provider used an uncertainty-based approach [3], and in case of datasets without uncertainty data, this organiser had to make significant compromises with respect to its statistical methods. In general, the PT providers were very flexible in this respect, and indicated the adaptations made to their protocols as appropriate. In some cases the PT providers reluctantly provided evaluation data, which is from a point of view of assuring a certain quality standard in the data evaluation fully justified.

With regard to the context information, such as identification of the methods (of measurement) used by the laboratories, the attitudes are widely different between PT providers, and there are also differences between sectors. For example, in the food sector there is clearly a reference method given in regulations, and this method is taken as reference by most PT providers [5]. In the soil [4] and [3] sector, it has become evident that there are differences between countries. In the soil sector, these differences are caused by both legislative differences as well as differences in levels of pollution, whereas in the water sector the legislative requirements are the most important factor.

2.2 Comparison

Like in many areas of proficiency testing, it is for a given dataset hard to tell what the “truth” is in terms of the key statistical parameters: mean, standard deviation, and rating of the laboratories. A comparison on the basis of consensus is therefore more appropriate, and in most cases there was a reasonable consensus across the PT providers for a given dataset. In the rating of the laboratories’ results, most differences were due to differences in the standard deviation used for rating, rather than discrepancies in the assigned value.

Reference values have been a problem in this comparison. In particular in the water [3] and occupational health [6] sectors, some PT providers rely quite heavily on the availability of reference values, and this caused some problems. Most PT providers have a foreseen the use of a consensus value should a reference value be unavailable, but in some cases this alternative was not seen as equivalent to the use of a reference value.

The main problem with the reference values was on one hand their availability, and on the other hand to mimic in a realistic way their establishment. In reality, it can occur that a reference value is deemed unreliable. Discrepancies in measurement data (e.g., from different methods), irregularities during measurements etc. can be reasons for such a decision. It is however very difficult to mimic this process and the decision/assessment processes in a realistic way.

3 Results

3.1 Participation

The participation of the PT providers was good. Most of them provided timely the requested information and usually in the requested format, which allowed quick processing of data. The number of participants per sector differed somewhat:

- water sector: 7 PT providers
- soil sector: 8 PT providers
- food sector: 5 PT providers
- occupational hygiene sector: 5 PT providers

3.2 Nature of the statistical protocols

In all sectors, for all parameters, the protocols used for processing the data are based on the assumption that data are normally distributed. Problems like bi-modality are dealt with quite differently, depending on the sector, the parameter, and the method. For example, in the food sector some PT providers make a method-specific evaluation, whereas others refer to the preferred (reference) method. In the water sector, some PT providers also use a method-specific evaluation, whereas others simply ignore such effects. Usually there is a plausible rationale for choosing a specific approach, but it is difficult to tell what approach is the “best”.

Values reported as “< x” (where x represents a number) are generally not rated, and not included in the establishment of the consensus value either. In fact, it is hard to tell whether a laboratory has any benefit from reporting a result in such a fashion. Many laboratories have reporting limits, which are certainly in their relationships with customers a useful instrument. For comparing results however, these reporting limits practically obstruct any processing from side of the PT provider.

The rating of results is in the vast majority based on a Z-score, with some important variants. Only one PT provider (from the water sector) uses E_n -scores, and in some cases other indices are used. The standard deviation used for rating the results usually differs appreciably more than the statistics computed from the datasets, which is an influence from other requirements put on the laboratories (for example, legislative performance requirements). Apart from these aspects, the protocols are very well comparable, even across sectors.

An important difference between the occupational health sector and the other three sectors is their reliance on different concentration levels in one proficiency test [6]. This reliance caused some problems with selecting datasets, and is not found to such an extent in other sectors. In some cases, Youden plots are used¹, for example in the water sector, but nevertheless the occupational health sector is in this respect different.

3.3 Water sector

24 data sets were sent to the nine PT providers participating in the water sector study on 8 June 2003. The data sets covered the following analytes: Iron (x5), Mercury (x4), Calcium (x3), Nitrate (x2), Sulphate (x1), Lead (x1), Electrolytic conductivity (x2), Ammonium (x4), Chloride (x2). Four of these data sets were artificially generated using

¹ They have not been part of this comparison.

the Monte Carlo approach and the remainder were "real" PTS data sets provided by some of the PT providers.

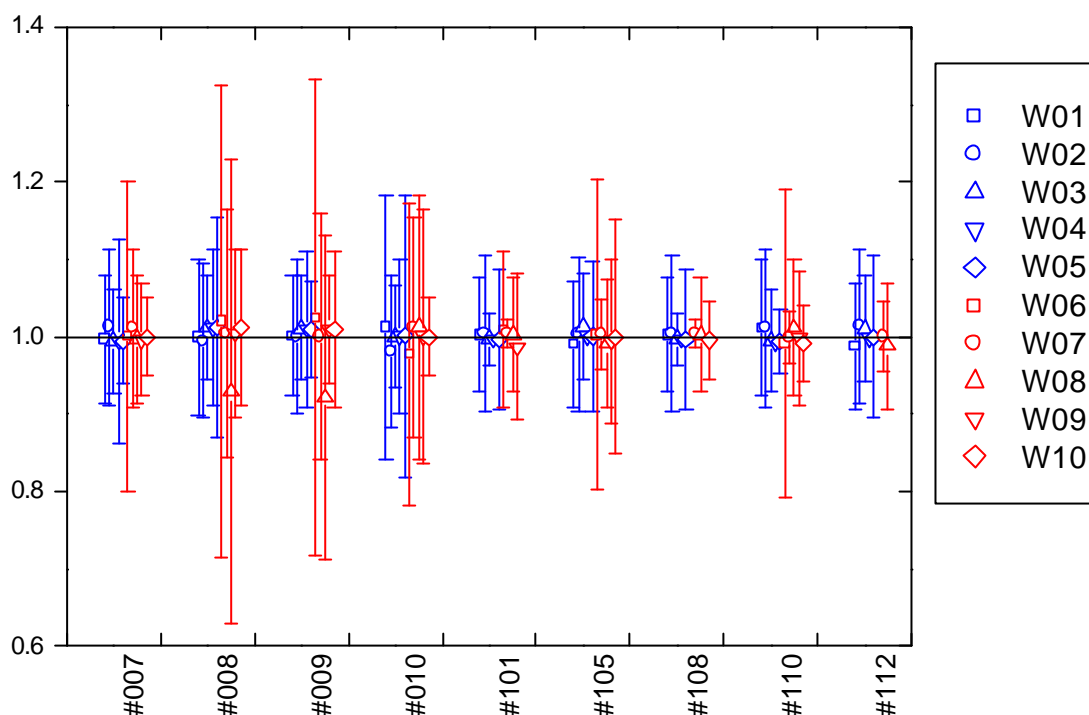


Figure 1: Means and standard deviations used for rating in the water sector (datasets #007 - #112)

The results from the water sector are summarised in figures 1-3. The assigned values agree usually very well among the PT providers. Notable exceptions form datasets #124 and #139.

In the case of dataset #124, a reference value was supplied that (deliberately) seriously deviated from the consensus value. The difference in the consensus values was about 12% (between 140 and 153 $\mu\text{g/l}$), whereas the reference value was set at 115 $\mu\text{g/l}$. Some PT providers indicated that they would rely on the reference value nonetheless, which makes that they can be clearly identified in figure 2. Dataset #139 had the same "feature" built in, and the 'separation' of the PT providers with regard to using the reference value (W02, W03, W06, and W07) and the consensus value (others) is clearly visible in figure 3.

The consistency of evaluation for the water sector was very good. The agreement between the assigned values was between 1 and 28%, and the assigned standard deviations between 5 and 40%, which is good, and where the agreement was less good, this was clearly reflected in the consistency of evaluations. Overall, the level of agreement between the providers in the evaluation of the data sets averaged approximately 70%, which was considered to be very encouraging, and indicative of the similarity of approach in this sector in Europe. For some of the data sets the agreement between providers was less good - typically between 10 and 40%. These data sets were of very low concentrations, where differences in statistical protocols have significantly greater influence. In these cases, two, and occasionally three, providers reported significantly different evaluations from the majority. If the two providers most consistently reporting different evaluations are ignored, the agreement between the remaining eight providers averages approximately 90%.

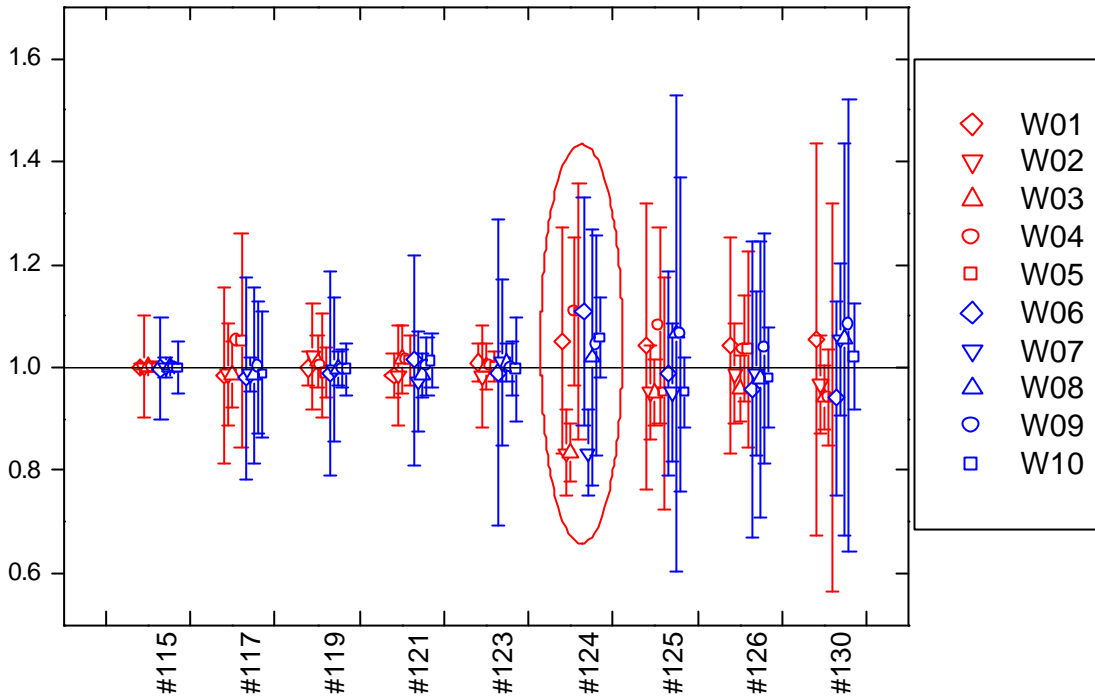


Figure 2: Means and standard deviations used for rating in the water sector (datasets #115 - #130)

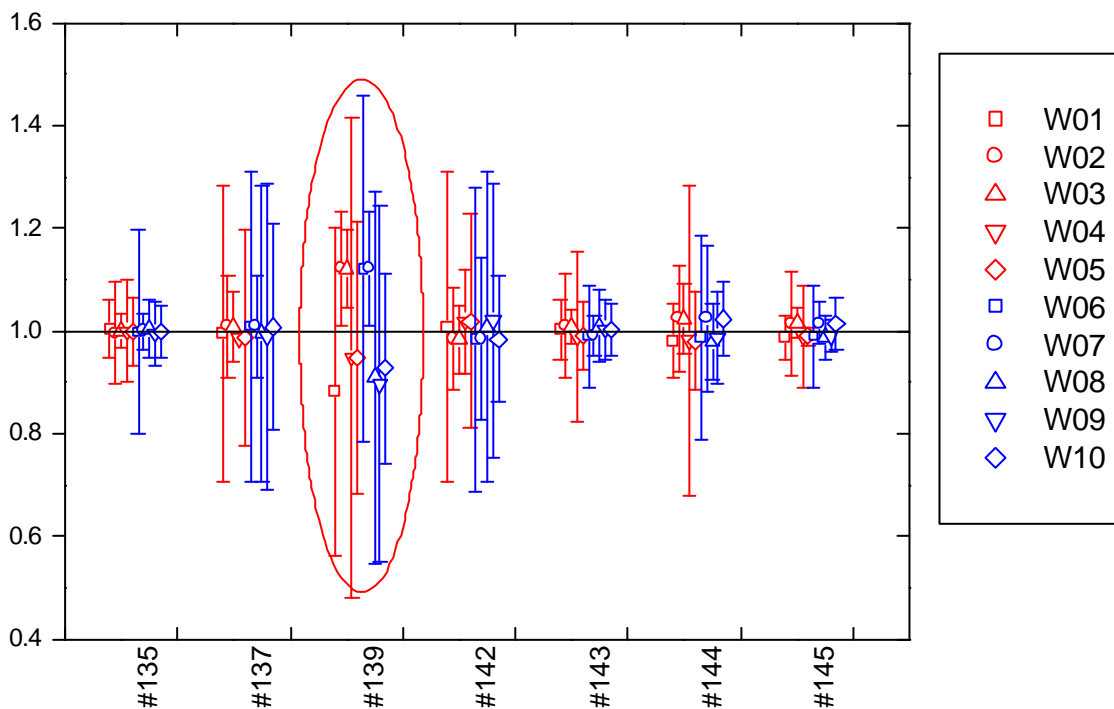


Figure 3: Means and standard deviations used for rating in the water sector (datasets #135 - #145)

The actual agreement in practice across proficiency tests will be better, simply because the datasets have been selected to be *representative* for the problems likely to occur in proficiency testing in the particular field. So, the problematic datasets are a bit 'over

represented' in the datasets used, but for a thorough assessment of the statistical protocols of PT providers, this is a necessity.

3.4 Soil sector

29 data sets were sent to the nine PT providers participating in the food sector on 30 June 2003. The data sets related to soil contaminated with organic pollutants and covered the 16 listed EPA Polyaromatic Hydrocarbons (PAHs). A full list is given in the evaluation report. Three of these data sets were artificially generated using the Monte Carlo approach and the remainder were "real" PTS data sets provided by some of the PT providers. The results for the soil sector are summarised in figures 4-6.

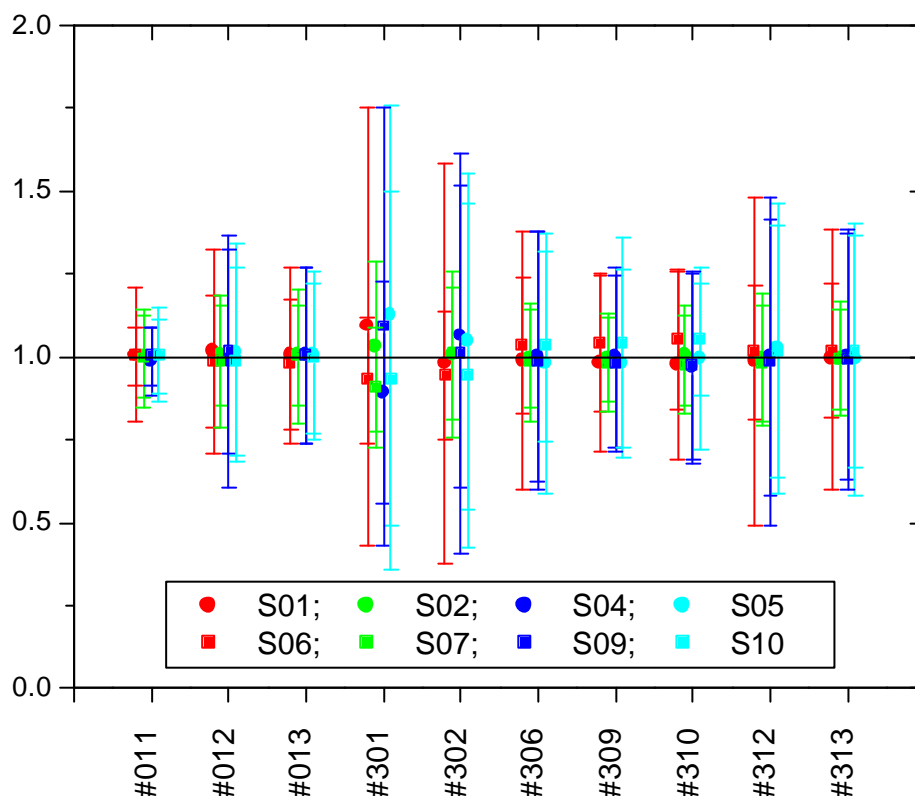


Figure 4: Means and standard deviations used for rating in the soil sector (datasets #011 - #313)

In the vast majority of the cases, the agreement with regard to the assigned values is very good, with a few possible exceptions: datasets #301, #302, #336, and #370. Datasets #301, #302, and #336 are simply of poor quality (a phenomenon, not uncommon in PAH analysis in soil), which caused the relatively high scatter in the assigned values. Dataset #370 had 8 "< x" values, and was -apart from this problem- also relatively poor. All four "exceptions" have in common that the concentration levels were very low.

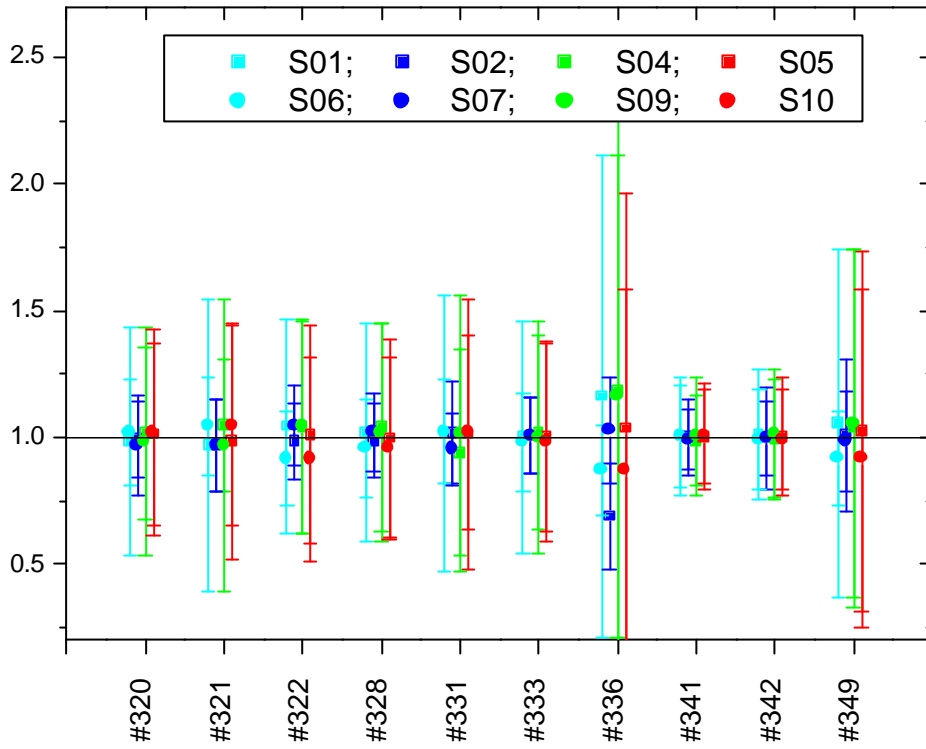


Figure 5: Means and standard deviations used for rating in the soil sector (datasets #320 - #349)

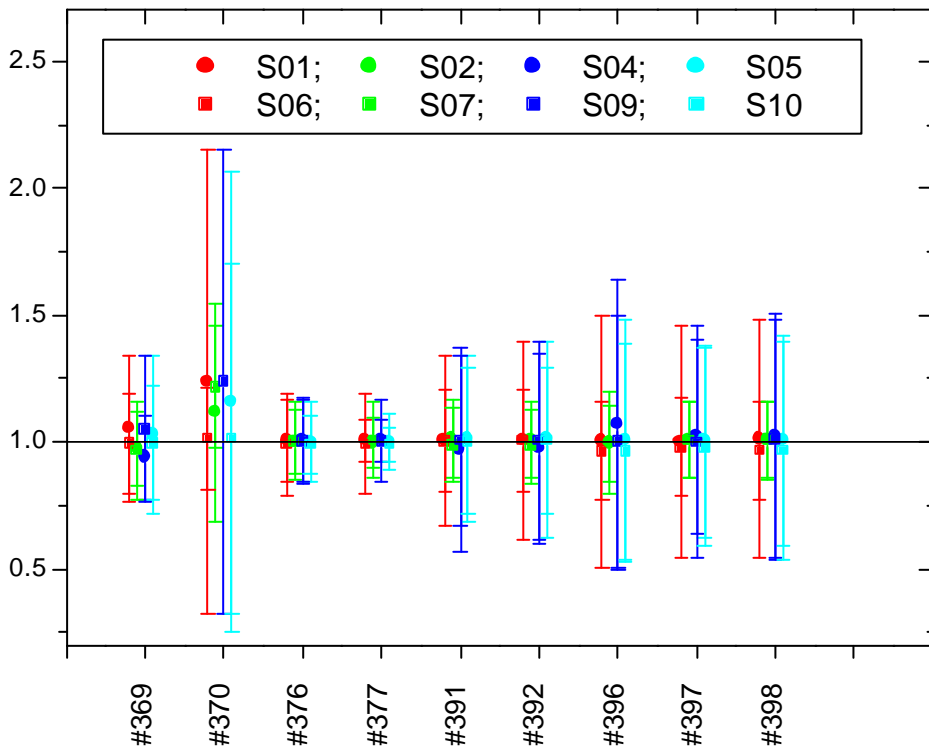


Figure 6: Means and standard deviations used for rating in the water sector (datasets #369 - #398)

The consistency of evaluation for the soil sector was again very good: 82% of data sets were evaluated the same by all providers. The major variation was in the treatment of

results that were not evaluated - these may have been below a threshold value, or reported as less than a detection or reporting limit - but reflected accurately the procedures of these providers. The agreement between the assigned values was excellent (between 1 and 11%), but the assigned standard deviations varied more (up to 50%), and where the agreement was less good, this was clearly reflected in the consistency of evaluations. This could have been predicted from studying the statistical protocols of each provider. For the 18% of data sets where the agreement was less good, in more than 50% of cases only one PT provider disagreed. This PT provider is just starting to operate in the contaminated soil sector and is still developing their statistical protocol. This was considered to be very encouraging, and indicative of the similarity of approach in this sector in Europe.

3.5 Food sector

14 data sets were sent to the five PT providers participating in the food sector on 30 June 2003. The data sets related to milk powder and covered the following analytes: fat (x5), moisture (x3), protein (x3), ash (x2) and lactose (x1). Three of these data sets were artificially generated using the Monte Carlo approach and the remainder were "real" PTS data sets provided by some of the PT providers.

The results for the food sector are summarised in figure 7. Apart from datasets #015 and #512-2, the agreement with respect to the assigned values is very good. The standard deviations used for rating are sometimes widely different between PTs, and they are also mainly responsible for differences in rating.

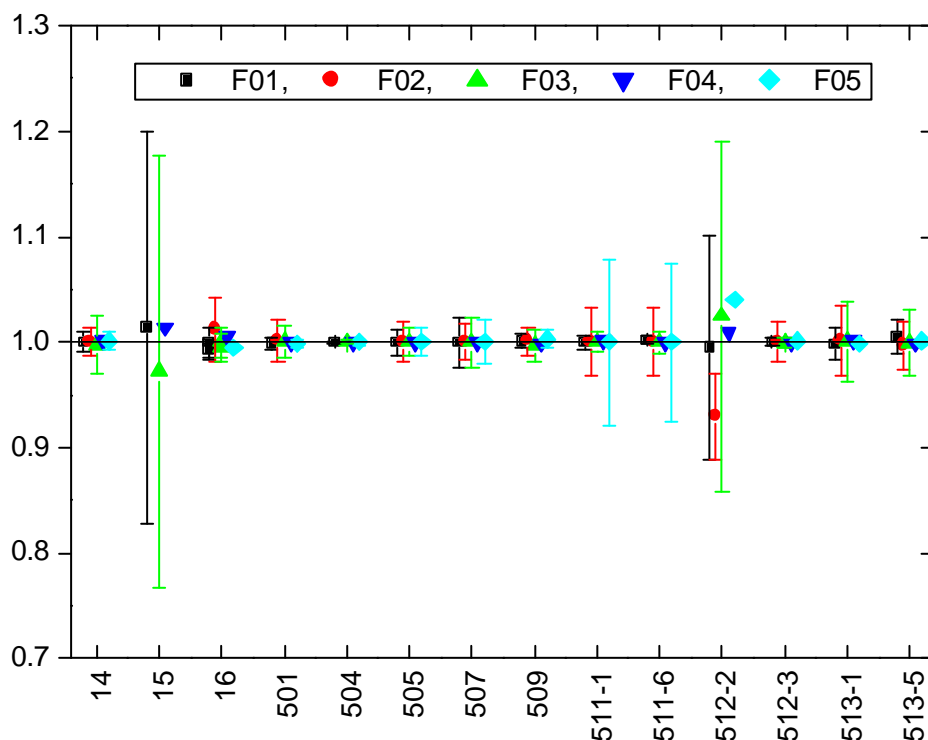


Figure 7: Means and standard deviations used for rating in the food sector

The consistency of evaluation for the food sector was again very good, although some of the data sets gave a more disagreement in the evaluation than others. In particular, when a data point was assigned to a specific test method, this inconsistency was highlighted. The agreement between the assigned values was between 1 and 12%, but the assigned standard deviations, although good overall, were more varied than in the water sector,

varying by up to 200% in some cases. Where the agreement was less good, this was clearly reflected in the consistency of evaluations. It should be noted that the agreement between assigned values varied more significantly in this sector than the others. Although one of the five PT providers did not evaluate performance using the z-score approach, the assigned values of all the providers are very similar. For the four providers using z-scores, others the level of agreement between the providers in the evaluation of the data sets was approximately 84%. Of the 16% of data sets evaluated differently, half of these were due to one PT provider. For one data set results were given which were method-specific. The evaluations for these results showed no agreement, as for three of the PT providers their protocol does not cover this method, which is only followed in a few specific European countries. The final report is available as a separate deliverable, and incorporates the issues raised at the workshop [5].

3.6 Occupational hygiene sector

7 datasets, each consisting of three (concentration) levels were provided to the PT providers. Two artificial datasets (#019 and #020) supplemented the 5 'real' ones, comprising the following elements: Cobalt, Copper, Lead, Manganese and Nickel. The result of this sector are shown in figures 8 and 9.

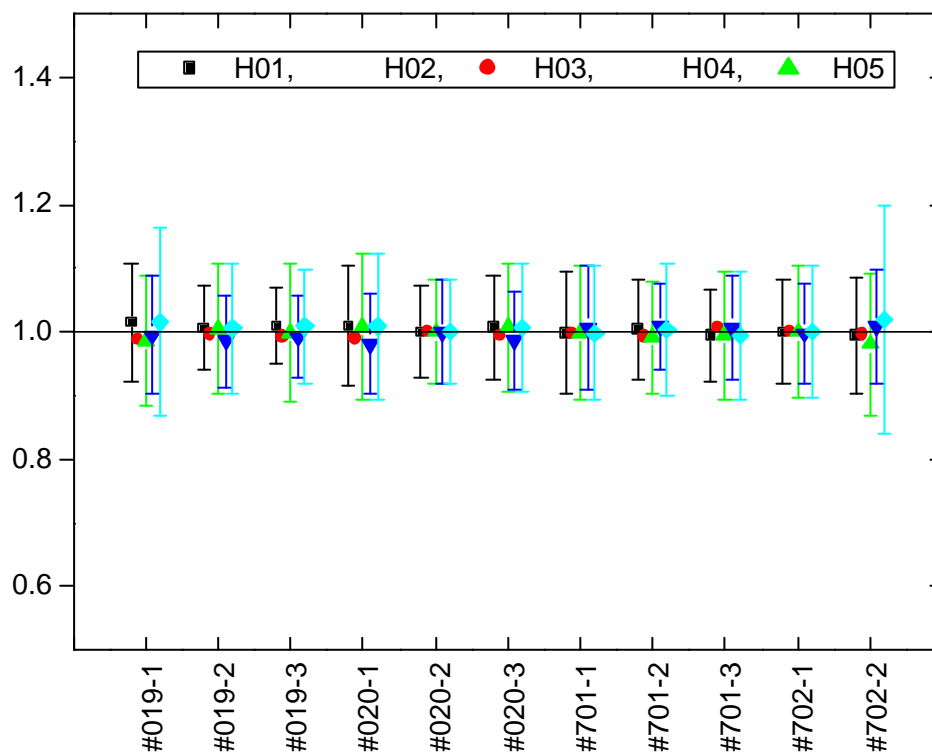


Figure 8: Means and standard deviations used for rating in the occupational hygiene sector (datasets #019 - #702-2)

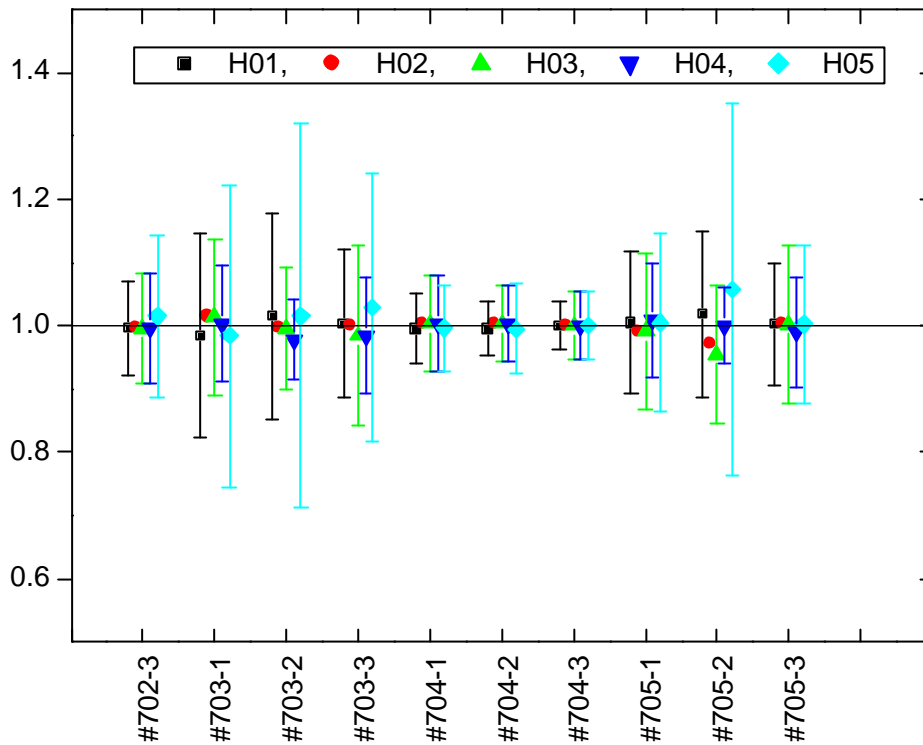


Figure 9: Means and standard deviations used for rating in the occupational hygiene sector (datasets #702-3 - #705)

The consistency of evaluation for the occupational hygiene sector was again very good, despite the two different groupings of statistical evaluation protocol, and one provider did not use the z-score approach. The agreement between the assigned values and the assigned standard deviations (where these were used) was very good, with the assigned values varying by between 1 and 3%. This resulted in a level of agreement between the providers in the evaluation of the data sets of approximately 88%, which was considered to be extremely encouraging. However, the agreement between the assigned values from the providers and the reference values provided by the task leader for some of the datasets was quite poor. Six of the data sets were responsible for 50% of the dissimilar evaluations. Overall, this is indicative of the co-operation between providers in this sector in Europe, and a common protocol for all providers in the sector has been discussed but not yet agreed.

4 Conclusions

The agreement between the protocols in each sector is generally good, tested on drinking water datasets is good. Laboratories with unsatisfactory performance in one PT are likely to get the same rating in another. When differences occur, they are usually due to differences in minimum requirements for acceptable performance, rather than due to differences in statistical protocols. In some sectors, like water [3] and food [5], the context of the data is very important, as it influences the decision taking process, and good attention should be paid to providing datasets for such a comparison.

The agreement in terms of rating the laboratories' results is generally good: in 70% (water), 84% (food), 82% (soil), and 88% (occupational hygiene) of the cases, the rating was the same for all PT providers. Differences in rating, as they occur, are usually due to differences in the requirements set for rating the participants. In all cases, the differences can be explained from the statistical protocol.

The co-operation with the PT providers is generally good. The task of the PT providers is - despite the efforts put in the organisation- not always easy. Datasets from one PT are not necessarily ideal for the protocol of another PT. This conclusion leads to the recommendation to specify very clearly what data and context information is needed in the 2nd comparison from the laboratories.

5 References

- [1] Van der Veen A.M.H., Final protocol 1st comparison in CoEPT, Growth Project GTC1-2002-73002 CoEPT, Deliverable D3, NMI VSL, Delft (NL), 8 June 2003
- [2] International Organization for Standardization, "ISO/IEC Guide 43-1:1997, Proficiency testing by interlaboratory comparisons - Part 1: Development and operation of proficiency testing schemes", ISO Geneva
- [3] Van der Veen A.M.H., "Comparison of PT protocols used in the water sector", Growth Project GTC1-2002-73002 CoEPT, Deliverable D5, NMI VSL, Delft (NL), March 2004
- [4] Van der Veen A.M.H., "Comparison of PT protocols used in the soil sector", Growth Project GTC1-2002-73002 CoEPT, Deliverable D7, NMI VSL, Delft (NL), March 2004
- [5] Van der Veen A.M.H., "Comparison of PT protocols used in the food sector", Growth Project GTC1-2002-73002 CoEPT, Deliverable D6, NMI VSL, Delft (NL), March 2004
- [6] Van der Veen A.M.H., "Comparison of PT protocols used in the occupational health sector", Growth Project GTC1-2002-73002 CoEPT, Deliverable D8, NMI VSL, Delft (NL), March 2004