Accreditation and Quality Assurance (2019) 24:289–296
https://doi.org/10.1007/s00769-019-01383-x

PRACTITIONER'S REPORT

# Fifteen years of proficiency testing of total petrol hydrocarbon determination in soil: a story of success

**Roland Becker**[1] · **Andreas Sauer**[1] · **Wolfram Bremser**[1]

Received: 10 December 2018 / Accepted: 4 April 2019 / Published online: 24 April 2019


## Abstract

The total petrol hydrocarbon (TPH) content in soil is determined by gas chromatographic separation and flame ionisation detection according to ISO 16703 in routine laboratories for about 20 years. The development of the interlaboratory variability observed with this analytical procedure over 15 years in a proficiency testing scheme conducted annually with more than 170 participants is evaluated in detail. A significant improvement of the reproducibility standard deviation among participants is observed over the years and attributed to an increasing familiarity with the procedure. Nevertheless, the determination of TPH in the environmentally relevant mass fraction range between 500 mg/kg and 10 000 mg/kg in soils or sediments is far from reaching the reproducibility standard deviations predicted by the *Horwitz* curve. It is seen that laboratories with sporadic participation tend to report higher bias, while a core group of laboratories participating on a regular basis arrived at reproducibility standard deviations below 20 %. Results from a given laboratory obtained on two different samples tend to be highly correlated in the same PT round indicating a sound repeatability. Expectedly, the within-laboratory correlation between results from consecutive rounds was considerably lower. However, results from consecutive rounds with a temporal distance of 1, 2 or 3 years revealed largely similar correlations which suggests that the within-laboratory reproducibility adjusts to a constant level at least after a period of 1 year.

**Keywords** ISO 16703 · Gas chromatography · Flame ionisation detection · Soil · Interlaboratory comparison · Reproducibility

## Introduction

Interlaboratory comparisons are an important tool for the external assessment of the proficiency of routine laboratories and are operated in nearly all areas of analytical chemistry [1–3]. Successful participation in proficiency testing (PT) schemes is often a prerequisite to obtain or maintain accreditation for a respective field of analysis or may be part of a contractor selection for a specific task. Most PT schemes provide interlaboratory comparison rounds on a given matrix/analyte combination and on a repetitive basis, while the respective intervals between rounds may differ largely between schemes. Especially in cases with larger intervals between rounds, an idea about the persistence of the performance over consecutive rounds should be helpful for interval planning.

The objectives of this paper are:

- to look at the development of the reproducibility among laboratories in successive rounds of a major PT scheme hosting a large number of participants. Is there a significant development towards smaller reproducibility standard deviations in consecutive interlaboratory comparison rounds?
- to see whether there is a tendency for a given laboratory to similarly perform on different samples within one round and over consecutive rounds and if so, how long does such a tendency last?
- to define whether there are groups of participants with constantly better or worse performance than would be

**Electronic supplementary material** The online version of this article (https://doi.org/10.1007/s00769-019-01383-x) contains supplementary material, which is available to authorized users.

✉ Roland Becker
roland.becker@bam.de

[1] Bundesanstalt für Materialforschung und -prüfung (BAM), Berlin, Germany

expected by a random distribution of laboratory results within each round and how large these groups are.

Especially, the questions concerning the improvement of reproducibility over consecutive rounds and the maintenance of the performance level of a given laboratory are best investigated with analytes that have been adopted by routine laboratories only relatively recently.

Thus, the determination of total petrol hydrocarbons (TPHs) in soil using gas chromatography–flame ionisation detection (GC-FID) was chosen as the method/analyte combination. Apparently, up to the date, nothing has been reported regarding interlaboratory comparisons for this analyte except an early report on preliminary experiences [4].

## Background of the PT scheme

The proficiency testing scheme "Contaminated Sites" is operated by BAM since 1996 [4] and covers a range of organic and inorganic trace analytes relevant on contaminated sites with a historic industrial or military background. The scheme aims at routine laboratories seeking accreditation. Apart from aqua regia extractable heavy metals, polycyclic aromatic hydrocarbons, polychlorinated biphenyls, chlorinated pesticides, adsorbable organically bound halogens (AOX) [5] and total cyanides, it routinely included the sum parameter total petrol hydrocarbons, one of the most often encountered pollutants on sites with a history of military use. Its determination according to ISO 16703 [6] prescribes the summary quantitation of hydrocarbons after a specified extraction and clean-up procedure followed by GC-FID and summary integration of the peaks within the retention time range between decane ($C_{10}$) and tetracontane ($C_{40}$). ISO 16703 and its preceding documents were introduced to routine environmental analysis since the late 1990s as a substitute for the abandoned infrared spectroscopic method ISO/TR11046 [7] because the latter was based on the use of meanwhile banned halogenated solvents. The PT scheme is conducted following the rules laid down in ISO/IEC 17043 [8], ISO 13528 [9], and DIN 38402-45 [10].

## Organisation of the scheme and preparation of the test samples

Laboratories seeking accreditation for trace analysis on contaminated sites were invited and took part mostly on a repetitive basis, while a certain fluctuation occurred. Test samples were prepared from real-case soils or freshwater sediments collected on various sites with a reported history of contamination dating back for decades. Fortification was only done in one case where a wet farmland soil was spiked with commercial diesel fuel, homogenised, air-dried and then stored for two years in the dark at outside temperatures until further processed as outlined in the following for all starting materials. Bulk soils or freshwater sediments taken from the field were air-dried to constant weight, and portions above 250 µm were removed by sieving. Further sieve fractionising yielded optionally fractions < 63 µm, 63 µm – 125 µm and 125 µm – 250 µm. Sieving fractions with a suitable TPH content were homogenised using a drum hoop mixer. In some cases, fractions with similar particle size ranges and bulk densities were blended to adjust a desired level of TPH content. A material with defined particle size range, bulk density and homogeneous TPH content was used as test material. At least three test materials displaying different TPH contents were bottled for each round with a spinning riffler and using a system of partitioning and back-mixing [11]. A typical test material consisted of 80–160 units of a constant mass in the range between 50 g and 80 g of soil or sediment bottled per unit. Different test materials intended for the same PT round contained similar masses per unit. Directly after bottling, test materials were stored at −20 °C until shipment to the participants. The homogeneity of the TPH content was tested on selected units of each test material, and the acceptance criterion from ISO 13528 was prerequisite for use in the PT scheme. From both own and literature experience [11], it is known that under the applied storage conditions TPH in dry soils or sediments are stable within time spans much longer than those elapsing from preparation of the material until collection of results from all PT participants. However, long-term storage over years may reveal a certain TPH degradation. For a report including considerations on this issue, see [11]. For a PT, this effect is fully negligible. The determination of TPH content as a sum parameter and complicated measurand involves extraction, clean-up, flame ionisation detection of the hydrocarbons, and integration in the pre-defined range for TPH as laid down in ISO 16703 as the stated reference.

Each participant received two test items which came always from two different test materials displaying different concentration levels. At least three different test materials were used to avoid consultation of participants on the results during the interlaboratory comparison period. Also for this reason participants were asked to report results of a duplicate determination on each sample within 20 days after delivery. Data received from the participants were handled anonymously. Statistical evaluation of results and assessment of participant proficiency were carried out according to DIN 38402-45 [10] using the *Hampel* estimator and the *q*-estimator [12] as robust estimates of the average $X$ and reproducibility standard deviation $s_R$. The average $X$ was taken as assigned value $X_{ref}$ for the proficiency assessment along with the standard deviation for proficiency assessment $\hat{\sigma}$ selected by perception according to ISO 13528. The proficiency assessment of each participant was based on the asymmetric $z_u$-scores according to DIN 38402-45 [13].

## Over the rounds assessment of the PT scheme results

Since the PT scheme was open to all laboratories willing to participate, any given PT round may include participants with little experience leading, in some cases, to biased results that do not represent the actual scatter of results among laboratories reasonably familiar with the method. This required a certain policy of restrictions made on the full data set in order to arrive at realistic reproducibility estimates. One option is to use robust statistics to eliminate the effect of outlying observations, and the alternative is to exclude laboratories with few and irregular participation over the rounds. Both options should produce data sets that represent best estimates of the interlaboratory scatter at the respective point in time.

For the assessment of the PT scheme, the following questions have to be answered:

- Do participating laboratories experience an improvement of their skills (local improvement)?
- Can it be proven that after a longer period of participation, the overall performance of laboratories is increasing (global improvement)?
- Is there a number of laboratories performing (on average) better than the rest?
- Does a successful participation in one round of a PT guarantee similar performance in the future?

For answering these questions, one would consider from a statistical point of view

- a regression for certain (long-term participating) laboratories of their accuracy
- plots of biases versus precision for different time periods, and the development of the total variability over the time of running the scheme
- *t* test and *F*-test (shrinking to core laboratories)
- the correlation coefficient between the performance attained at a certain instance in time and the one attained one or two PT rounds later.

## Results and discussion

### General overview

Figure 1 shows the development of the overall performance data on TPH. Relative between-laboratory reproducibilities $s_{R,rel}$ in %, also known as coefficients of variation CV, are presented for each test materials in terms of robust *q*-estimators [12]. The individual data can be found in the supplementary Table S1.

The naked eye suggests a strong improvement of robust $s_R$ after the initial years until 2004. This may be attributed to the growing experience acquired in the laboratories. It can also be seen that after the period until 2013 (assessed

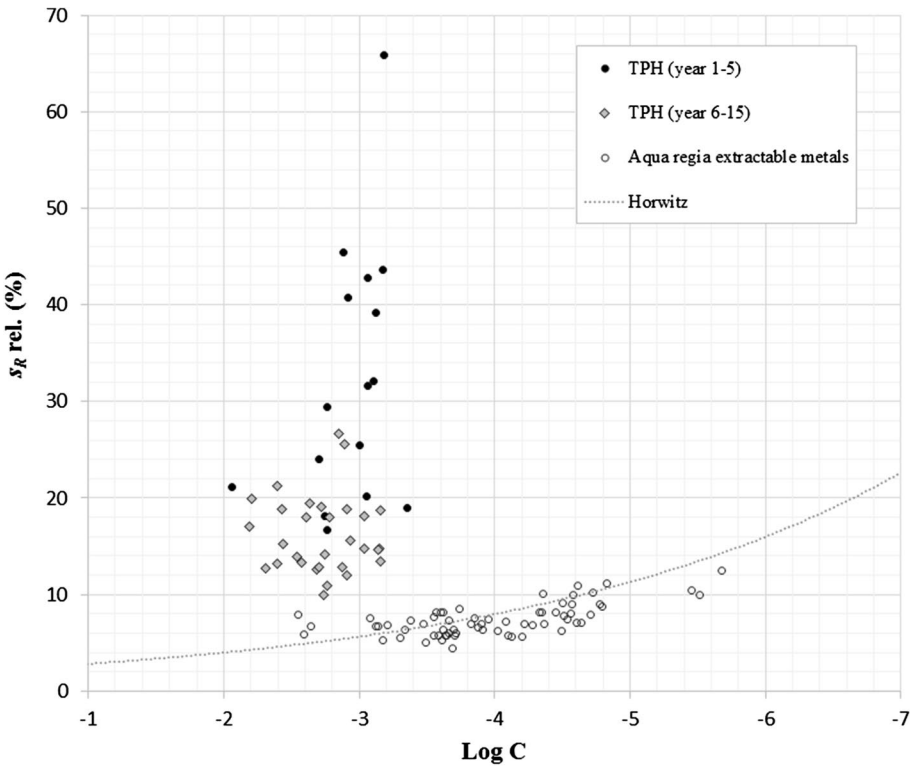


**Fig. 1** Coefficients of variation (CV) over TPH content C from PT rounds before and after 2004

in more detail in the following) no significant further improvement became evident. Similar rapid improvements in the initial rounds of PT have been reported for blood [14] and olive oil analysis [15]. Figure 1 displays for comparison the relative $s_R$ as observed for *aqua regia* extractable metals (As, Cr, Cu Hg, Ni, Pb, Zn) in the same PT scheme. It is seen that the metals follow, in good agreement, the prediction of the *Horwitz* curve [16] that provides a benchmark for the interlaboratory variability valid for many matrix/analyte combinations. Clearly, the analyte TPH does not follow this rule and the current state outlined in this report does not suggest that this will change in the future. It should be noted that the alternative TPH quantification using IR spectrometry [7] and withdrawn because of the use of a banned fluorochlorohydrocarbon as extraction solvent did follow the *Horwitz* prediction [4].

## Correlation of within-laboratory results over the rounds

In order to assess whether laboratories tend to retain their performance level over the rounds, the correlation of results from within the rounds was compared with results obtained by the same laboratories in consecutive rounds with intervals of one, two, and three years. Thus, the TPH results were sorted such that all data reported by the same group of participants on two specific test materials could be associated with the respective *Pearson's* correlation coefficients. Then, the correlations of any participant subgroup on two given test materials within the rounds and between consecutive rounds may be obtained. This principle is clarified in the supplementary information along with the test of correlations for significance. The synopsis in Fig. 2 reveals a strong correlation of laboratory results within one round, while the number of significant correlations ($p < 0.05$) over consecutive rounds is much smaller and independent of the interval between rounds (one, two or three years). This suggests that the correlation between independent results obtained on two different test samples at different points in time is reduced already after one year to a level that does not decrease further with longer intervals.

This finding should not to be confused with observations that either the reproducibility among laboratories did not improve with the increase in PT round frequency [17] or the reproducibility among laboratories improved over time in other cases [18, 19] including the scheme reported here. The reproducibility among laboratories may improve over consecutive rounds regardless if individual laboratory results in terms of bias from the respective averages tend to display a certain correlation over the rounds or not.

## The performance of the participants over consecutive rounds

Performance evaluation of the participants was (mainly) conducted on the basis of the Q score which is

$$Q = \left(x_{\text{lab}} - X_{\text{ref}}\right)/X_{\text{ref}}$$

expressed in per cent. With widely varying data distributions, it may be assumed that the Q score is normally distributed. The above-mentioned synopsis already revealed an improvement of compatibility of results from participants over the rounds. However, the robust evaluation procedure is designed to minimise the impact of outlying results and tends to provide best $s_R$ estimates for the likely "outlier-free" data sets. Furthermore, only the respective participants including those without regular attendance contribute to the $s_R$ estimates. In order to assess the results obtained by laboratories with a repeated participation over the rounds, the data from laboratories with less than three participations were excluded. Thus, the total number of 351 laboratories having participated in the scheme occasionally was reduced to 151 regular participants which indeed participated at least three times. Figure 3 displays the average mean Q score and the standard deviation of the attained Q scores of the selected laboratories, assessed over their individual time of participation in the PT scheme. Figure 4 depicts the relative standard deviation over the bias of the arithmetic mean from the reference value for the named laboratories. The reference value was always the robust consensus mean of all laboratories having analysed the respective test material. Obviously, the reproducibility among this truncated group increased with time and interestingly the bias between reference value derived from all participants and the arithmetic mean of this group diminished with time. This is consistent with the concept that even seldom participating laboratories tend to arrive on average at the arithmetic mean of the more regular participants. Figure 5 depicts the average standard deviation observed among the regular participants in the respective round.

The latter leads to a slow but constant improvement of performance (at least improved reproducibility) as can be seen from the graph, depicting the performance of 22 laboratories constantly participating in the PT scheme, as shown in Fig. 6. The envelope describing the reduction in the between-laboratory variability (or reproducibility according to ISO 5725-2) is constantly decreasing over the time of participation, approximating the attainable precision and trueness of the method itself.

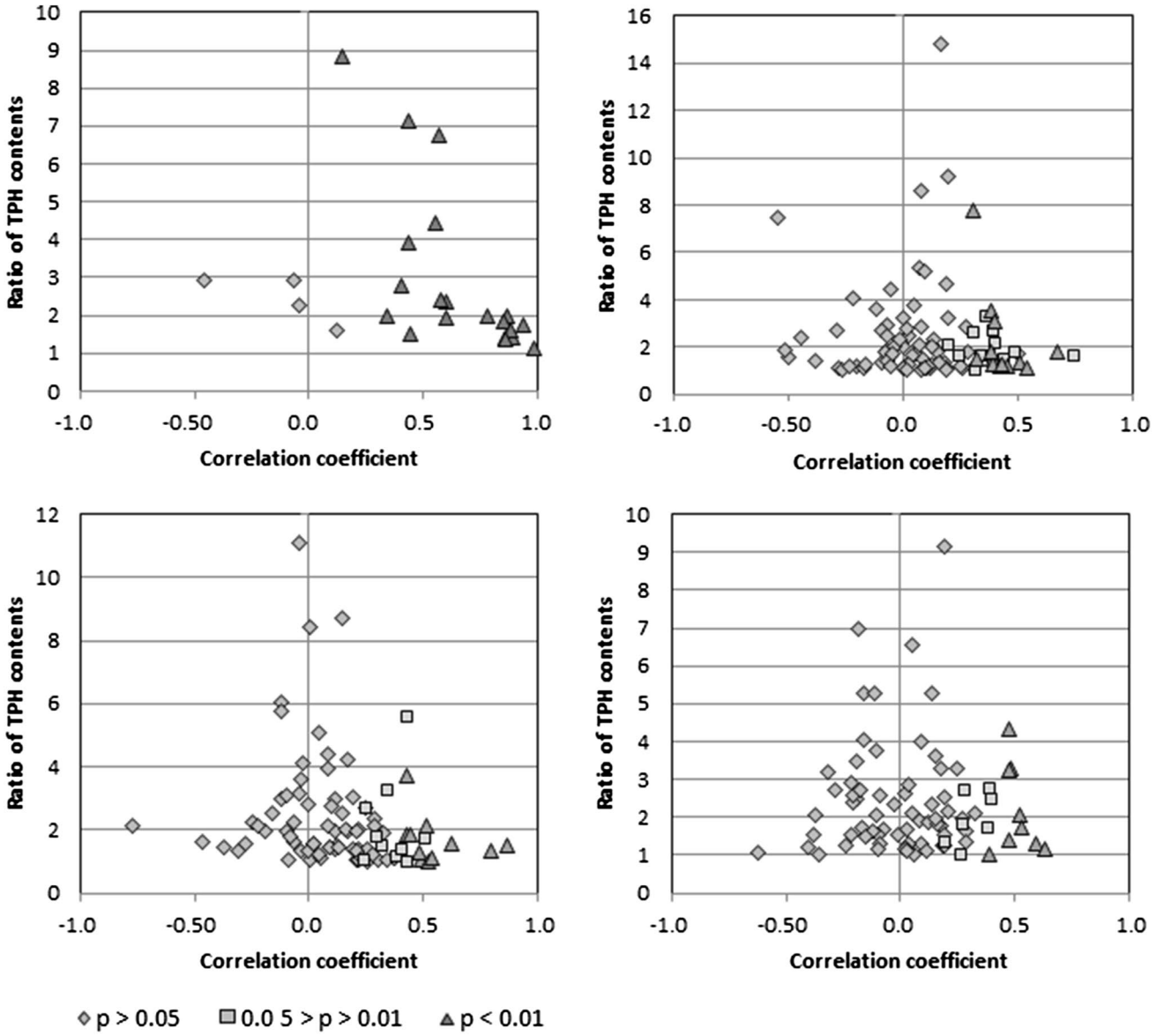


**Fig. 2** Correlation within a round (upper left) and between one, two and three consecutive rounds (clockwise). Each data point represents the correlation of results obtained on two different test materials by the same group pf laboratories. The *y*-axis represents the TPH mass fraction ratio of the respective two test materials

Within this group, some of the laboratories perform better or even much better than the rest, and Fig. 6 visualises three of those. Regression coefficients, slope uncertainty, residual scatter and the significance level of the regression for the three laboratories highlighted in Fig. 6 are listed in Table 1. Note that the regressions in Table 1 span 15 years and a minimum of 10 participations within the period of interest. While laboratories 009 and 052 reveal no trends and, on average, remain at the same slight bias of 5 % at maximum, laboratory 010 started from a quite high bias (see intercept) reducing, over time, to some 7 %–8 % (see the negative slope). All laboratories are "true in itself", i.e. reveal a long-term reproducibility much better than even the full core of the longer-participating laboratories. Reproducibilities of the named laboratories are within the range of 7.5 % to 10 %, significantly better than the 20 % for the core group of regular participants, let alone randomly involved participants. However, also an excellent in the long-term run laboratory cannot fully avoid occasional underperformances.

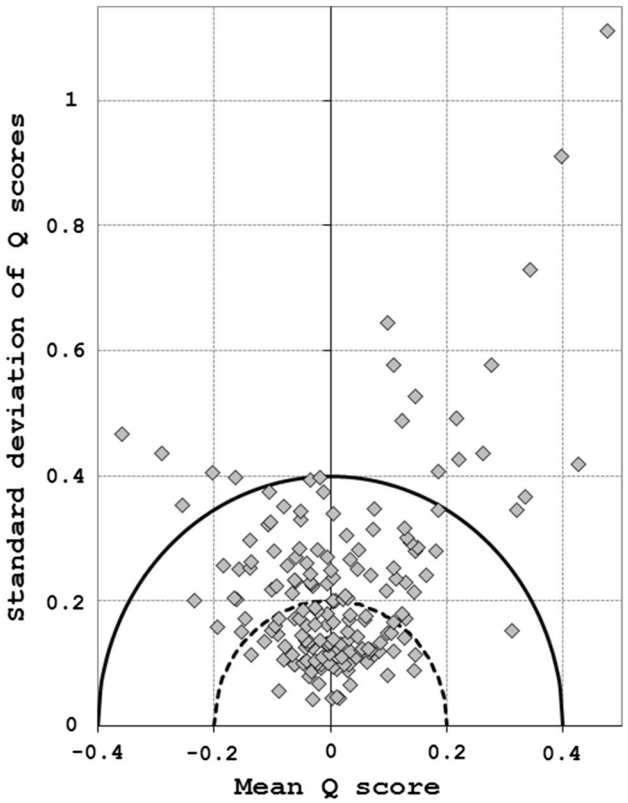


**Fig. 3** Average mean $Q$ score and the standard deviation of the attained $Q$ scores of 174 selected laboratories, assessed over their individual time of participation in the PT scheme

## Conclusions

The determination of TPH in soils and sediments experienced an improving degree of equivalence among routine laboratories since its implementation about 20 years ago. The demonstrable improvement in method domination is largely due to fact that this analytical procedure was introduced in routine laboratories as exchange for the banned chlorofluorocarbon-based IR spectroscopic procedure only at the beginning of the time covered by this report. In this case, the increasing degree of equivalence among laboratories occurring over the first years of wider application among routine laboratories, and a subsequent approximation towards the maximum attainable performance of the method could be observed. The reproducibility standard deviations at around 20 % for TPH appear to represent state of the art. However, the *Horwitz* model which provides a sound estimate for the coefficient of variation for many matrix/analyte combinations [14] does not apply even after the transition period. It should be assumed that the substantial complexity of the analytical procedure and the relatively high LOD of several hundred mg/kg led to this discrepancy. There are obviously a group of laboratories that reported constantly lower variability of results, significantly better than the rest, and better than expected from the corresponding distribution attained in the year of execution. Such groups form on a long basis of participation in PTs and corresponding, ongoing

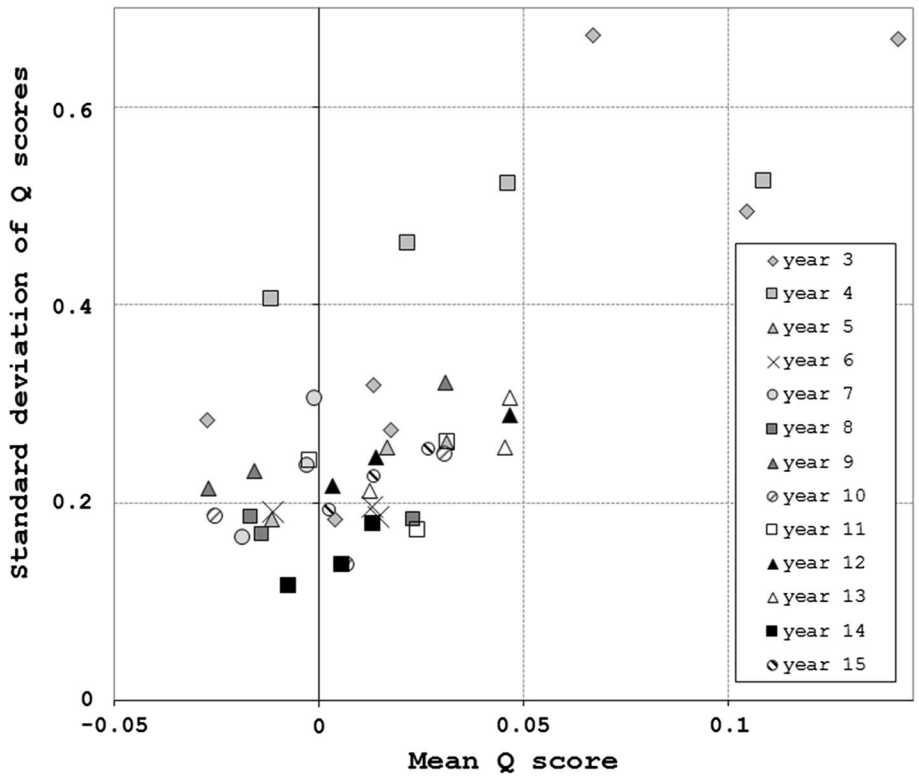


**Fig. 4** Mean $Q$ score and the standard deviation of the attained $Q$ scores per PT round and level, averaged over the actual number of participants

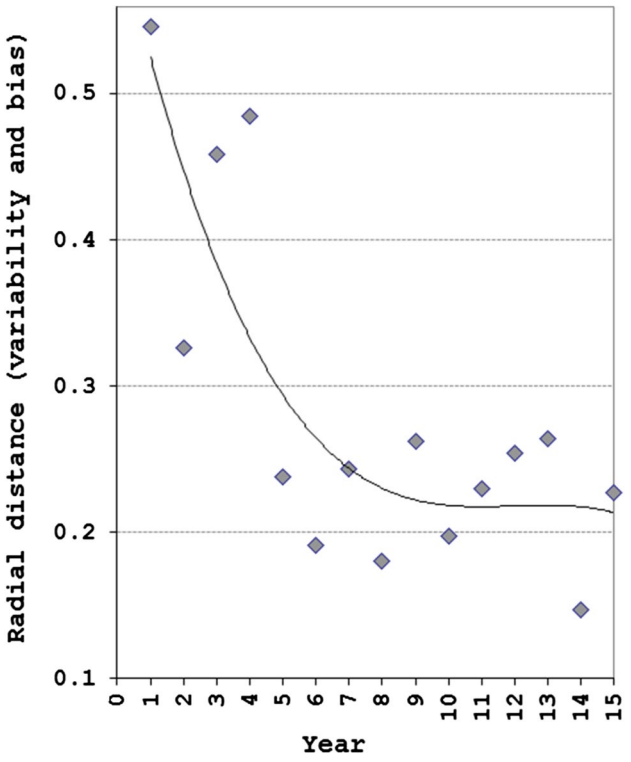


**Fig. 5** Distance from the origin of the data points in Fig. 4, representing both the variability and the average bias of data attained in a PT round, by year. The fit is third-order polynomial and for illustration purposes only

**Table 1** Regression parameters for three of the well-performing laboratories

| Laboratory | Intercept | Slope | *u* (slope) | Residual scatter | *p* value |
|---|---|---|---|---|---|
| L009 | −0.051 74 | 0.000 05 | 0.007 95 | 0.098 97 | 0.995 |
| L010 | 0.335 37 | −0.023 70 | 0.005 67 | 0.082 05 | 0.002 |
| L052 | 0.022 97 | −0.005 88 | 0.005 51 | 0.074 33 | 0.308 |

application of the method under discussion. Although it has been found that routine laboratories establish their "personal" performance profile after a period of time of one to two years, a membership in a better-than-the rest group requires long-term experience and constant quality assurance measures. Groups with documented repeatedly greater equivalence may be employed for defining reference values, for example, for PT rounds or certification of reference materials. Apart from the general concept of external proof of proficiency, PT remains on ongoing task, because punctual outliers may occur also in excellent laboratories and strongly deviating results are more likely with laboratories not participating on a regular basis.

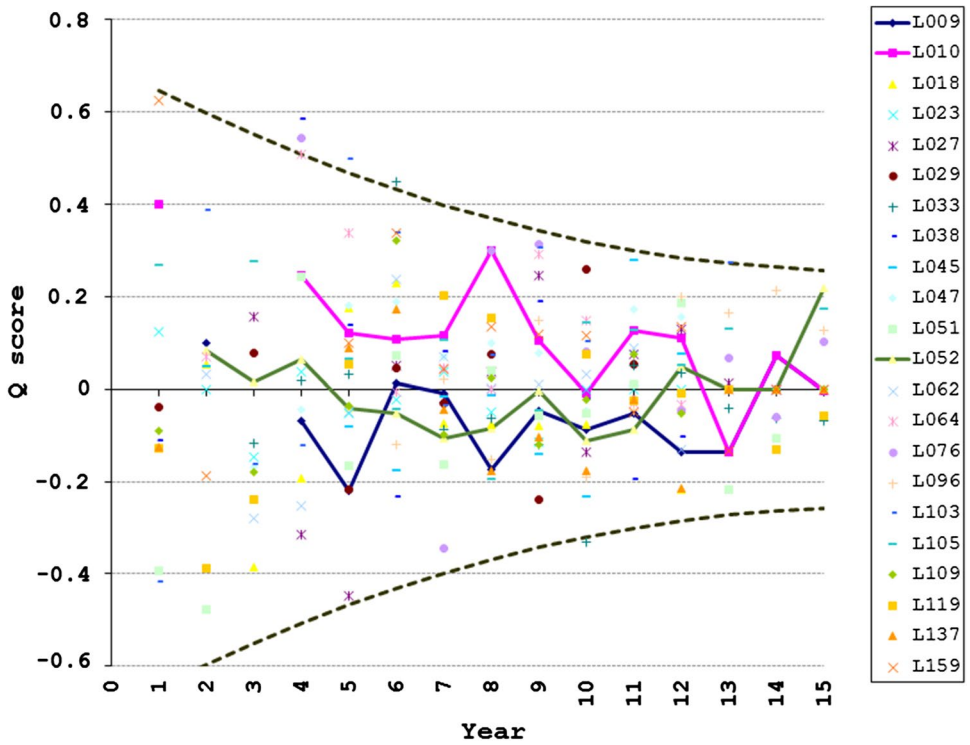


**Fig. 6** Performance development of laboratories participating between 9 and 12 times, over the years (*x* axis). The measure of performance is the *Q* score attained by the laboratory in the corresponding PT round, and selected laboratories performing better than the rest

## References


1. Taverniers I, De Loose M, Van Bockstaele E (2004) Trends in quality in the analytical laboratory. II. Analytical method validation and quality assurance. Trac-Trend Anal Chem 23:535–552
2. Hund E, Massart DL, Smeyers-Verbeke J (2000) Inter-laboratory studies in analytical chemistry. Anal Chim Acta 423:145–165
3. European Proficiency Testing Information System EPTIS. www.eptis.bam.de
4. Becker R, Koch M, Wachholz S, Win T (2002) Quantification of total petrol hydrocarbons (TPH) in soil by IR-spectrometry and gas chromatography—conclusions from three proficiency testing rounds. Accred Qual Assur 7:286–289
5. Becker R, Buge HG, Nehls I (2007) The determination of adsorbable organically bound halogens (AOX) in soil: interlaboratory comparisons and reference materials. Accred Qual Assur 12:647–651
6. ISO 16703 (2004) Soil quality—determination of content of hydrocarbon in the range $C_{10}$ to $C_{40}$ by gas chromatography
7. ISO/TR 11046 (1994) Soil quality—determination of mineral oil content—method by infrared spectrometry and gas chromatographic method
8. ISO/IEC 17043 (2010) Conformity assessment—general requirements for proficiency testing
9. ISO 13528 (2015) Statistical methods for use in proficiency testing by interlaboratory comparison
10. DIN 38402-45 (2014-06) German standard methods for the examination of water, waste water and sludge—general information (group A)—Part 45: interlaboratory comparisons for proficiency testing of laboratories (A 45)
11. Becker R, Buge HG, Bremser W, Nehls I (2006) Mineral oil content in sediments and soils: comparability, traceability and a certified reference material for quality assurance. Anal Bioanal Chem 385:645–651
12. Wilrich PT (2007) Robust estimates of the theoretical standard deviation to be used in interlaboratory precision experiments. Accred Qual Assur 12:231–240
13. Uhlig S, Henschel P (1997) Limits of tolerance and z-scores in ring tests. Fresenius J Anal Chem 358:761–766
14. Zhong K, Zhao Y, Xiao YL, Wang W, He FL, Wang ZG (2015) 8-year review of laboratory performance on blood lead level external quality control assessment surveys 2006–2013 in China: continual improvement. Accred Qual Assur 20:25–28
15. Generali T, Stefanelli P, Girolimetti S, Barbini DA (2015) Proficiency tests on olive oil organized by the Italian National Reference Laboratory for pesticides: long-term performance of laboratories. Accred Qual Assur 20:247–253
16. Horwitz W (1982) Evaluation of analytical methods used for regulations of food and drugs. Anal Chem 54:67A–76A
17. Thompson M, Lowthian PJ (1998) The frequency of rounds in a proficiency test: does it affect the performance of participants? Analyst 123:2809–2812
18. Dorgerloh U, Becker R, Lutz A, Bremser W, Hilbert S, Nehls I (2012) How to improve reliability in groundwater analysis: over a decade of experience with external quality control in field campaigns on volatile halogenated compounds. J Environ Monit 14:217–223
19. Whetton M, Finch H (2009) Analytical performance is improved by regular participation in proficiency testing: an analysis of data from the Aquacheck proficiency testing scheme. Accred Qual Assur 14:445–448