Accred Qual Assur (2006) 10: 521–526
DOI 10.1007/s00769-005-0051-2

Rob G. Visser

Interpretation of interlaboratory comparison results to evaluate laboratory proficiency

Received: 26 July 2005
Accepted: 9 October 2005
Published online: 24 January 2006

R. G. Visser
Institute for Interlaboratory Studies,
Spijkenisse, The Netherlands

Abstract Guidelines are given for the evaluation of proficiency test (PT) results in order to increase the effectivity of PT participation. For better understanding, some statistical background is given along with some examples to show the effects of the choices made by the PT provider. The calculation method of the assigned value and the selection of the standard deviation both affect the *z*-score that is used by the participating laboratory to judge the quality of its performance in the PT. Therefore, the participating laboratory is advised to use the PT results with care and, if necessary, to recalculate the *z*-scores. Finally, advice is given on how not to follow up bad PT results along with some valuable steps that could be part of an effective follow-up procedure.

Keywords Proficiency test · Assigned value · Standard deviation · *z*-score · Root cause analysis · Corrective action · Effectivity

Introduction

As an organiser of proficiency tests, the author of this article has received many questions from participating laboratories about bad results from the proficiency tests that have been organised. Many questions originate from external auditors, e.g. from the national accreditation organisation. Also when visiting laboratories as an auditor, the author has had many discussions with laboratory managers about procedures for following up bad results from proficiency tests. Often, as a first step, a re-analysis of the sample is performed. Sometimes investigations are undertaken and a reason for the aberrant result is found. However, only in rare cases are structural quality measures taken to prevent re-occurrence. More often, only a short-term corrective action is done that will only solve the current problem for the moment. In this way, the effectivity regarding participation in proficiency tests is unnecessarily low.

Although relevant literature is available on this subject [1], in this article some new views are discussed and enriched with examples in order to motivate the laboratory community to improve the follow up of proficiency test results and, consequently, improve the effectivity of proficiency test participation.

Background

Already, for many decades, the laboratory community has used interlaboratory studies as an external quality control. During recent years, the importance of interlaboratory studies in particular proficiency tests (PTs) has become an instrument of the accreditation organisations for evaluating the laboratory performance in an objective manner. In some countries, participation in PTs is mandatory for accredited laboratories, while in other countries, participation is strongly advised by the accreditation organisations.

Since the introduction of ISO17025 in 1999, this quality control instrument has become even more widely used, although the respective clause in ISO17025 [2] is not very strict about this.

> ISO 17025, clause 5.9: “The laboratory shall have quality control procedures for monitoring the validity of tests.” This monitoring shall be planned and reviewed and may include, but not be limited to, the following:

1. Regular use of certified reference materials......
2. Participation in interlaboratory comparison or proficiency-testing programmes;

In Europe, EA (European co-operation for Accreditation), the organisation of EAL (European co-operation for Accreditation of Laboratories) and EAC (European Accreditation of Certification) published more detailed guidelines [3, 4] that can be downloaded from the EA web site www.european-accreditation.org/:

> EA-02/10, clause 4.2: "Accreditation bodies need to fully document their policies and procedures in relation to PT activities. In particular, they must be able to evaluate, through the accreditation process, that the participation in PT activities of laboratories accredited by them is effective and that corrective actions are carried out when necessary"
>
> EA-03/04, clause 7.1.7: (Under actions for Accreditation Bodies and Assessors) "Check that laboratories have a written procedure in the Quality Manual (QM) or in laboratory instructions covering participation in proficiency testing, including how the performance in proficiency testing is used to demonstrate the laboratory's competence and procedures followed in the event of unsatisfactory performance"

As a consequence of the increased use of PTs, it is not surprising that more discussions are going on regarding the interpretation of PT results, i.e. when is a result good or bad and which corrective actions are effective and which are not?

To be able to answer such questions, one should have some knowledge of the (statistical) background of the evaluation of PT results and of the goals of PTs in general. Both will be discussed in the first chapters of this article. In the remaining chapters, some possible answers regarding the questions are given.

Goals of a proficiency test

As described in the previous chapter, one of the goals of a PT is to enable objective evaluation of laboratory performance. This is the reason that in some countries the national accreditation organisation organises PTs (NATA in Australia is a well-known example, see www.nata.asn.au/).

But from the laboratory's point of view, this is not the only goal. Other relevant goals are in brief:

Comparison of own results with the results of other laboratories by analysis of identical samples

Evaluation of trueness in cases where no certified reference material is available

Quality improvement via corrective actions

Facilitation of training of personnel with sample material with known values.

Statistical background

All the above mentioned goals have one thing in common: the assigned value, being the best estimate of the unknown "true value", is of crucial importance, while the spread of the data is of minor importance. The spread is in fact merely a measure for the uncertainty of the assigned value.

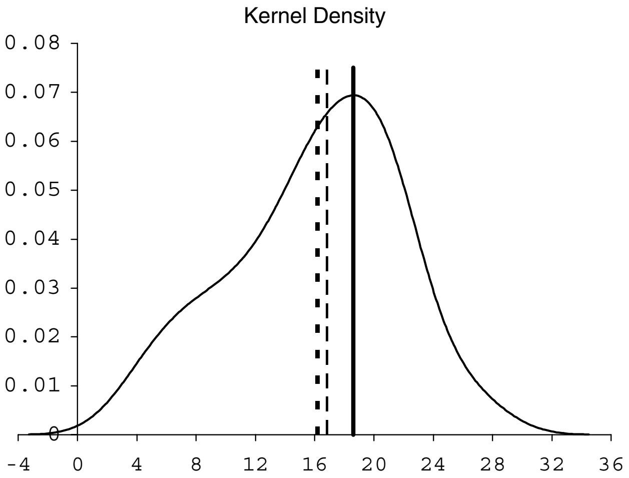

Fig. 1 Effect on assigned value of calculation method used; NaCl in Crude Oil. Data from Ref. [12]

Consequently, the task of the PT organiser can be reduced to two basic tasks including (1) prepare and provide suitable, stable and homogeneous samples and (2) provide reliable assigned values of the parameters that were evaluated in the PT.

How to produce suitable homogeneous samples is a theme that will not be discussed in this article; this has already been extensively described in the literature, e.g. ISO Guide 43 [5] and ILAC G13 [6]. Also, there are numerous publications that describe statistically sound ways to determine an assigned value. Additionally, in the last decade, several standardised methods were developed [7–11].

The large variety of calculation methods can be divided in four groups:

1. Calculation of the arithmetic mean after assumption of normal distribution and removal of outliers
2. Use of robust statistical calculations
3. Calculation of some kind of trimmed mean
4. Remaining calculation methods

The first two calculation methods are most commonly used, the first being most accepted by the laboratories due to the fact that the necessary calculations can be understood easily and the second being most appreciated by the PT organisers due to the low influence of outlying values on the assigned value. In practice, both methods have disadvantages. The biggest disadvantage of method 1 is the fact that normal (Gaussian) distributions of PT results are rarely observed. The detection of outliers is often problematic due to the so-called masking effect. Additionally, the assigned value also depends on the outlier test that is selected to be used and consequently on the number of outliers that is observed and excluded from the calculations.

The only, but not insignificant, disadvantage of method 2 is the fact that the robust calculations are hard to understand for technicians that have only learned to use normal (Gaussian) statistics.

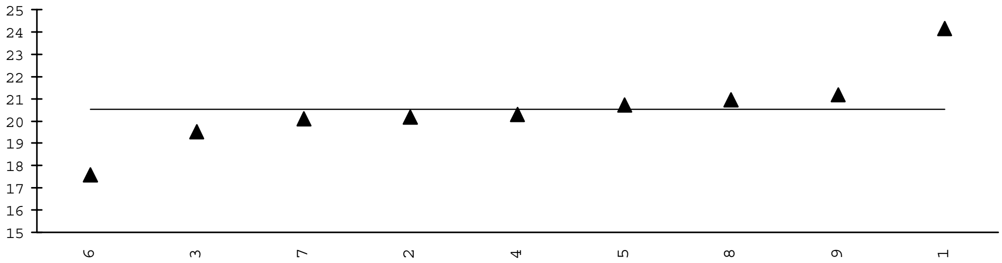

Fig. 2 Effect of the selection of the outlier test on the assigned value. Data from Ref. [13]

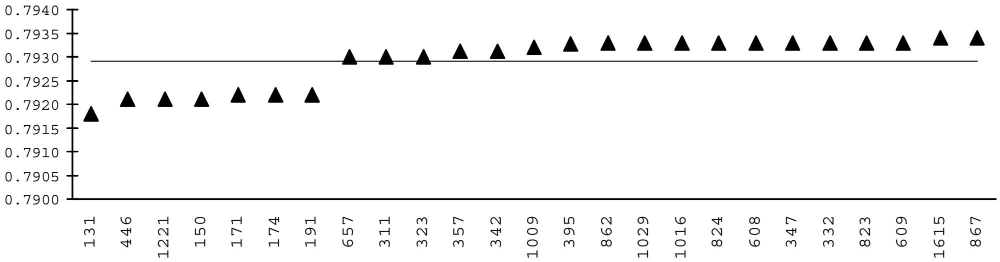

Fig. 3 Effect of data distribution; Apparent Specific Gravity of Methanol. Data from Ref. [14]

Effect of the selected statistical method on the assigned value

Both the assigned value and the standard deviation are dependent on the choices made by the PT-provider. The effect on the assigned value of this choice is visualised in Figs. 1–3 where examples with real world data are used.

Example 1: The data taken from a PT on NaCl in crude oil [12] visually show a nearly Gaussian distribution. The Lilliefors normality test confirms this assumption. No statistical outliers are present in the data set according to the Dixon and the Grubbs outlier's tests [8]. The arithmetic mean of all 34 reported analytical results is 16.1 (the bold dotted line in Fig. 1). But can this figure reliably be used for an assigned value? The median is 16.8 (the thin striped line in Fig. 1) and the maximum of the Kernel Density Plot [15] is 18.6!

Example 2: The effect of the selection of an outlier test is nicely demonstrated with a data set from ISO5725 [13], to which the Dixon and Grubbs tests give different outcomes. The data distribution (n=9) is Gaussian according to the Lilliefors normality test. No statistical outliers are present in the data set according to the Grubbs outlier test. The arithmetic mean is 20.5. However, the Dixon outlier test does recognize an outlier with 95% probability (the result from laboratory 1). When this value from laboratory 1 is excluded from the calculations, the mean and standard deviation change significantly and the new mean is 20.1.

Example 3: The raw data, taken from a PT on the apparent specific gravity of methanol [14], show an odd distribution that is clearly not Gaussian. From the 25 figures, 7 are much lower than the rest. The Arithmetic mean is 0.7929, but it is clear that this cannot be used as an assigned value.

Upon investigation, it appeared that the seven laboratories that had reported the lower values all had used the same type of apparatus to determine the density. In the manual of the apparatus, an error was present in the formula to calculate the apparent specific gravity from the measured density. When the correct formula was used, all seven new results fitted nicely in the normal distributed data of the remaining 18 laboratories. The assigned value after this correction was significantly higher than before. Mere application of statistics would not have solved this problem.

From the above three examples it is clear that each PT-provider may end up with a different assigned value for each data set, depending on the protocol that was followed.

Calculation of *z*-scores

To facilitate the performance evaluation by the participating laboratories, many PT providers calculate a z-score for each reported laboratory result.

The z-scores, z_i, are calculated from the laboratory results, the assigned value and a standard deviation in accordance to:

$$z_i = (x_i - X)/\sigma$$

where x_i =laboratory result, X = assigned value and σ = standard deviation.

The problems with the determination of the assigned value were discussed in the previous chapter. Another problem is the selection of a relevant and fit-for-use standard deviation. Often the standard deviation of the group of participating laboratories is used. Although commonly used, this choice does have a major disadvantage! By using the standard deviation of the group, about 95% of the results will get a z-score between -2 and $+2$, regardless of whether the accuracy is appropriate. Consequently these z-scores do not say anything about the fitness for purpose of the laboratory results [16]. When another standard deviation is selected to be used, this problem can be overcome.

In Table 1, an overview is given of the choices made by some PT-providers. In practice every PT provider may use a different procedure to determine the assigned value. Also every PT provider may use a different standard deviation and consequently the z-scores calculated from one data set by one PT provider can differ significantly from the z-scores calculated by another PT provider. Obviously, the rule that 'a z-score is equal to or larger than 3 or equal to or smaller than –3, represents a bad result' is not an absolute truth.

One striking example is the following. In proficiency tests for liquids (like gasoline, diesel, etc), density is one of the physical parameters that are evaluated. Density is an absolute parameter that is directly related to the SI units mass and length. All standardised methods to determine the density are unbiased by the presence of correction factors. Therefore, in these interlaboratory studies, no method is prescribed and a mixture of all different measurement methods is used from manual hydrometer methods to automated oscillating U-tube methods. Although the bias of these methods is indeed very small, the precisions differ very much. The reproducibility of ASTM D1298 (a hydrometer method) is 0.0015 kg/m^3 and the reproducibility of ASTM D4052 (an oscillating U-tube method) is three times smaller: 0.0005 kg/m^3. When the z-scores of a proficiency test on the density of a fuel are calculated in an equal manner for all results using the standard deviation of the study, the laboratories that used the hydrometer method may face very large z-scores, while similar performing laboratories that used the oscillating U-tube method may have very small z-scores. This is only caused by the precision of the method used and not by the performance of the laboratories.

How PT-results can be evaluated

As discussed in the previous chapters, the use of an appropriate standard deviation will allow the laboratory to evaluate its performance in a PT independently from the performances of the other participants. When this standard deviation has a fixed value, evaluation of the laboratory performance over a long time period is made possible.

This appropriate and fixed standard deviation may be determined in many different ways and may come from many sources. It may, for instance, be taken from the Standard Method that the laboratory used to require the analytical result as well from legal or customer requirements, etc.

Using the assigned value of the PT (provided that this assigned value has been determined adequately), the reported laboratory result and the selected standard deviation, the laboratory can calculate its z-score using formula 1. With this z-score the usual performance evaluation now can be performed:

$$|z| \leqslant 2 \text{ satisfactory}$$

$$2 < |z| < 3 \text{ questionable}$$

$$3 \leqslant |z| \text{ unsatisfactory}$$

How not to follow-up bad PT-results

When a laboratory manager receives a proficiency testing report, he will first look at the z-scores that are calculated by the PT-provider concerning his laboratory. When he finds a bad result from his laboratory, it is not unusual that his first reaction will be to ask the analyst to perform the analysis on the respective sample once again. In the case that the new result is in agreement with the assigned value of the PT, the usual conclusion is that the reason for the bad PT result has been an incident, possibly a human error, or cannot be identified. As a result, no further action is taken, thus rejecting an opportunity for quality improvement.

How bad PT-results can be followed-up effectively

Before rushing into things, one has to realise that in principle every action will be already (too) late. Even in the case of a PT with a fast turn-around time, the results will have already been produced a considerable time ago and, since then, many other results will have already been produced and reported to customers.

Table 1 Assigned values, outlier tests and standard deviations used by some PT-providers

PT-provider	Country	Selected assigned value	Selected outlier test(s)	Selected standard deviation
Aquacheck [17]	UK	Mean	None	Error treshold
CHEK [18]	NL	Mean	Cochran, Grubbs	Horwitz
FAPAS [19]	UK	Trimmed mean	ANOVA	Fixed target
iis [20]	NL	Mean	Dixon, Grubbs	Fixed target
KDDL [21]	NL	Mean	Cochran, Grubbs	Actual
KIWA [22]	NL	Mean	Grubbs, Veglia	Actual
QM [23]	UK	Robust	None	Fixed target
LVU [24]	DE	Robust	None	Horwitz
RIZA [25]	NL	Mean	Cochran, Grubbs	Actual
SMPCS [26]	NL	Robust	None	Fixed target
WASP [27]	UK	Trimmed mean	$z>2 =$ excluded	Actual
WEPAL [28]	NL	Trimmed mean	None	Actual

Also, before starting investigations, one must consider whether the initially reported result indeed was a bad result. The conclusion 'bad result' was made on the basis of a *z*-score that was calculated by the PT-provider from the laboratory result, the assigned value and a standard deviation. Also, as discussed before, the choices of the PT-provider may not be applicable for your test results.

After the identification of a bad result, the root cause of the problem must be searched for effective follow-up. Looking for the root cause is not always easy. One tends to stop at the immediate cause such as lack of training; whereas, the real root cause is why the person was not trained.

Therefore a laboratory will have to have a 'root cause analysis' procedure that describes how to start an efficient follow-up and that ensures effective results. In the Eurachem guide [1], an example procedure is given as an appendix. Items that may be present in such a procedure are:

Check calculation of *z*-score. Is the assigned value reliable? Is an appropriate standard deviation chosen for the *z*-score calculation?
Can the evaluation of the PT be used? Is the number of data sufficient?
Check the appropriateness of matrix, concentration and test method
Check whether the reported result is identical to the suspect result present in the PT report
Check the raw data of the suspect test result (e.g. calculation, dilution or typing errors)
Check the data transfer from raw data to report (e.g. transcription)
Check the QA/QC during the analysis period of the suspect sample (e.g. blanc, QC-sample, calibration)
Check the laboratory conditions during the analysis period of the suspect sample (e.g. temperature, humidity, apparatus status, maintenance, power failure)
Check the human conditions during the analysis period of the suspect sample (e.g. authorised chemist, illness, replacement, trainee, conflicts).

If all above checks have not as yet revealed a possible cause of the suspect result, a retest of the respective sample may be performed, provided that the sample has been stored properly and that the analyte is sufficiently stable to allow successful retesting.

Corrective actions

As the number of corrective actions is in principle unlimited, it is not possible to give many details here. An effective corrective action will prevent reoccurrence of the problem that caused the bad PT-performance. Depending on the cause of the problem, the action may vary from changing an existing procedure (e.g. increase of calibration frequency), implementation of a new procedure (e.g. introduction of new quality control measures), increase of awareness of employees by additional training sessions, to the abandoning of a test method and stopping the analysis of a certain type of product.

Conclusions

Although literature is already available on this subject, in practise, many laboratories are not familiar with this literature and consequently do not sufficiently undertake thorough investigations to find the real causes of aberrant results in proficiency tests. Also, corrective actions often are only remedial and do not prevent reoccurrence of the problem. This, together with the too less critical attitude towards proficiency test reports and opinions of the accreditation bodies, causes the participation in proficiency studies to be of little effect. The laboratory is responsible for selecting an appropriate PT to participate in, for reporting test results, for evaluating its PT-performance and for implementing effective corrective actions.

With the guidelines described in this article, the laboratory should be able to improve the follow up of proficiency test results and consequently improve the effectivity of proficiency test participation.

References

1. Eurachem (2000) Selection, use and interpretation of proficiency testing (PT) schemes by laboratories Eurachem, http://www.eurachem.ul.pt/
2. ISO/IEC 17025 (2005), General requirements for the competence of testing and calibration laboratories, ISO, Geneva
3. EA-02/10, rev. 00 (2001), EA policy for participation in national and international proficiency testing activities, http://www.european-accreditation.org/
4. EA-03/04, rev. 01 (2001), Use of proficiency testing as a tool for accreditation in testing, http://www.european-accreditation.org/
5. ISO Guide 43-1 (1997), Proficiency testing by interlaboratory comparison. Part 1: Development and operation of proficiency testing schemes, ISO, Geneva
6. ILAC G13 (2000), ILAC Guidelines for the Requirements for the Competence of Providers of Proficiency Testing Schemes, ILAC, http://www.ilac.org
7. ISO 4259 (1992), Petroleum products: determination and application of precision data in relation to methods of test, ISO, Geneva
8. ISO 5725-4 (1994), Accuracy (trueness and precision) of measurement methods and results. Part 4: Basic methods for the determination of trueness of a standard measurement method, ISO, Geneva
9. ASTM E178 (2002), Standard practice for dealing with outlying observations, ASTM International, West Conshohocken, USA
10. ASTM E1301 (2003), Standard guide for proficiency testing by interlaboratory comparisons, ASTM International, West Conshohocken, USA
11. ISO/FDIS 13528 (2005), Statistical methods for use in proficiency testing by interlaboratory comparisons, ISO, Geneva
12. IIS04R01 (2004), Results of proficiency test crude oil IIS, http://www.iisnl.com
13. ISO 5725-2 (1994), Annex B3, Table B3, ISO, Geneva
14. IIS04C06 (2004) Results of proficiency test methanol IIS, http://www.iisnl.com
15. Lowthian PJ, Thompson M (2002) Analyst 127:1359–1364
16. AMC (2002) Understanding and acting on scores obtained in proficiency testing schemes, AMC Technical Brief 11, Royal Society of Chemistry, London
17. http://www.aquacheck.net/Pages/returnsnreporting.htm
18. http://www.chek-ps.nl
19. FAPAS (2002) Protocol, organisation and analysis of data, 6th edn., http://ptg.csl.gov.uk/fapasprotocol.cfm
20. IIS (2003) Protocol for the organisation, statistics and evaluation, version 3, IIS, http://www.iisnl.com
21. KDLL (1994) Report R94.012, KDLL, Zeist, NL
22. http://www.kiwa.nl/uploadedFiles/Kiwa_website/03_Water/020_Diensten/Toelichting rapportage.doc
23. QM (2002) Statistical protocol, QM, Bury, UK
24. http://www.lvus.de/html/download.htm
25. RIZA (1998) Procedure W 5003 8.301, RIZA, Lelystad, NL
26. Shell Global Solutions (2004) Shell main product correlation scheme, Shell Global Solutions, Amsterdam, NL, confidential communication
27. WASP (1996) Information book for participants, 4th edn. WASP, UK
28. van Montfort MAJ (1992) Statistical remarks on round robin data of IPE and ISE, Wageningen Agricultural University, Wageningen, NL