Accred Qual Assur (2005) 10: 409–414
DOI 10.1007/s00769-005-0029-0

Siu Kay Wong

Evaluation of the use of consensus values in proficiency testing programmes

Received: 21 October 2004
Accepted: 23 August 2005
Published online: 23 November 2005

S. K. Wong (✉)
Government Laboratory,
88 Chung Hau Street, Homantin,
Hong Kong SAR, China
e-mail: skwong@govtlab.gov.hk

Abstract Proficiency testing (PT) is an essential tool used by laboratory accreditation bodies to assess the competency of laboratories. Because of limited resources of PT providers or for other reasons, the assigned reference value used in the calculation of *z*-score values has usually been derived from some sort of consensus value obtained by central tendency estimators such as the arithmetic mean or robust mean. However, if the assigned reference value deviates significantly from the 'true value' of the analyte in the test material, laboratories' performance will be evaluated incorrectly. This paper evaluates the use of consensus values in proficiency testing programmes using the Monte Carlo simulation technique. The results indicated that the deviation of the assigned value from the true value could be as large as 40%, depending on the parameters of the proficiency testing programmes under investigation such as sample homogeneity, number of participant laboratories, concentration level, method precision and laboratory bias. To study how these parameters affect the degree of discrepancy between the consensus value and the true value, a fractional factorial design was also applied. The findings indicate that the number of participating laboratories and the distribution of laboratory bias were the prime two factors affecting the deviation of the consensus value from the true value.

Keywords Proficiency testing · Monte Carlo simulation · Central tendency estimator · Consensus value · Experiment design

Introduction

Nowadays, it is almost a mandatory requirement that a testing laboratory has to demonstrate to its client or the accreditation body its measurement capability through participation in external quality assurance programmes such as proficiency testing (PT) programmes. Normally, a PT programme is organized in such a way that a number of participating laboratories analyze the same test material distributed by the programme provider. The performance of the individual participating laboratory would then be evaluated by the deviation of its reported results from the assigned reference value which, in most cases, is some sort of consensus value obtained from the participants' results.

Interestingly, as an external quality assessment tool, a PT programme plays a dual role of policeman and teacher [1]. Firstly, a PT programme acts in the capacity of policeman when the outcome of a PT is used by accreditation bodies to monitor laboratory competence in specific tests. Secondly, for laboratories that are not yet competent in the tests, a PT programme provides an opportunity for improvement since the PT programme reports are usually accompanied by an analysis of results and recommendations for improving performance. For many programmes, this 'teacher' function is at least as important as the monitoring component. Hence, the manner in which the performance of an individual participating laboratory is evaluated in a given PT programme becomes a very important issue.

According to international harmonized protocol [2], a performance *z*-score, as defined below, is recommended for use in proficiency testing programmes to evaluate the accuracy of the analytical results submitted by participating laboratories:

$$z - \text{score} = \frac{\text{Laboratory result} - \text{Assigned value}}{\text{Target standard derivation}}$$

This assumes that the assigned value is the best available estimate of the true value of the analyte in the test material. The assigned value for the amount of analyte of interest in the test material can be established by using the following approaches:

1. Consensus value from expert laboratories
2. Formulation
3. Direct comparison with certified reference materials
4. Consensus of participant results

In practice, the PT organizer or provider chooses the appropriate approach according to the resources that it has available and the nature of the test materials. The critical point is that the unbiased assigned value needs to be used or incorrect evaluation of participants' performance would otherwise occur, i.e. the PT programme would fail to perform its dual roles of policeman and teacher. Generally, it seemed unlikely that problems due to large deviation of the assigned value from the true value obtained using the first three approaches mentioned above would arise. However, the international protocol had highlighted a number of possible drawbacks when the consensus of participants' results was used as the assigned value. Indeed, the possibility of the consensus being biased had also been discussed in the literature [3]. In order to evaluate in detail the use of consensus values in PT programmes, the Monte Carlo method was used in this paper. With this simulation technique, the effect of the number of participating laboratories, homogeneity of the test materials, concentration level of the analyte, laboratory bias and method precision on the deviation of the consensus value from the true value were assessed.

Consensus of participants

Because of limited resources, PT programme providers might be unable to afford the use of certified reference materials for the test materials or arrange for the determination of the assigned value by expert/reference laboratories through a collaborative study. Formulated or synthetic test materials are seldom used in PT programmes because the analyte is likely to be in a different chemical form from the incurred analyte [4]. Hence, in most cases, PT programme providers have few alternatives but to use the consensus of participant results as the assigned value. Taking the author's laboratory as an example, among the 75 proficiency testing programs that the laboratory participated in last year, over 90% of the programmes used this approach to obtain the assigned value for the evaluation of participant's performance.

By definition, consensus is taken to be the central tendency of a data pool. Central tendency estimators such as median, arithmetic mean and robust mean are commonly used in PT programmes to determine the consensus of participants' results. However, would there be any significant differences in the value of the consensus if different central tendency estimators were used? Or, would the number of participating laboratories, sample homogeneity, concentration level, method precision or laboratory bias have significant effects on the discrepancy between the consensus value and the true value? Firstly, it is not feasible to conduct a thousand rounds of PT programmes to find out the answers. Moreover, the true values of the analyte of the PT test materials are usually not available for us for evaluating how well the consensus values serve as the assigned values. Hence, this paper proposed using a simulation technique to find out the answers. In the simulation, different central tendency estimators were used to obtain the consensus values. By varying the PT programme parameters such as the level of analyte, sample homogeneity, number of participants, laboratory bias and method precision, their influence on the discrepancy were evaluated.

Monte Carlo simulation

A Monte Carlo simulation refers to approaches that apply to any use of random numbers. It is based on the principle that any complex process could be broken down into a series of simpler independent events, each represented by a probability distribution [5]. Recently, this simulation technique has a variety of applications in chemical metrology, especially in the estimation of measurement uncertainty [5–8].

To start with, it is necessary to establish a model to delineate a laboratory testing result. According to ISO 5725 [9], a laboratory testing result, x, could be expressed by the following statistical model:

$$x = M + b + e \tag{1}$$

where M is the gross average of the sample results; b is the effect due to laboratory bias and e is the effect due to random error made on x.

Applying this model to the participating laboratory's result in a PT programme, the value M should be related to the true value of the analyte in the sample and the sample homogeneity. Also, b and e would be determined by the laboratory bias and method precision, respectively, whose contributions are usually expressed in terms of respective relative standard deviations and the concentration of the analyte. In terms of the true value T_V, sample homogeneity S_H, laboratory bias b_i, and method precision S_{Pi} (Note: S_H, b_i, S_{Pi} are all expressed as relative standard deviation), the simulation model for the result of the ith participating laboratory is proposed as follows:

$$x_i = m_i + m_i b_i + m_i k_{2i} S_{\mathrm{Pi}} \tag{2}$$

where $m_i = T_v(1 + k_{1i} S_{\mathrm{H}})$, the value of the analyte in the sample as received by the ith participating laboratory.

In the above model, k_{1i} and k_{2i} are random numbers for that particular participating laboratory in the simulation

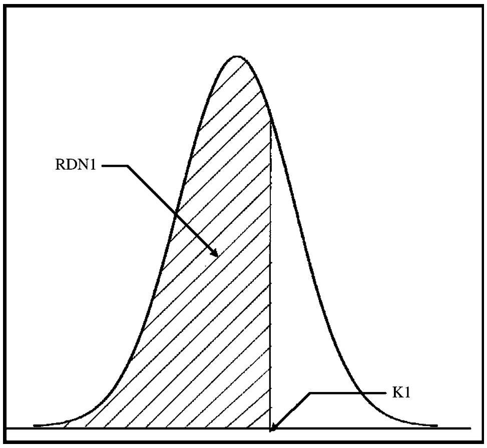

Fig. 1 Transform a random number RDN1 (0–1) to K1 (−3.5–3.5) using the probability density function of the normal distribution

whose values fell within −3.5 to 3.5, assuming normal distribution for the sample homogeneity and method precision. However, as random numbers generated by computer programmes normally range from 0 to 1, the probability density function of the normal distribution was thus used for the transformation as shown in Fig. 1 where RDN represents a random number generated from a computer programme. For easy reference, the algorithm of the simulation process is flowcharted in Fig. 2.

With the preset values for T_V and S_H, m_i could be simulated with an input of a random number which was first transformed to k_{1i} through the probability density function of a normal distribution as described above. For the laboratory bias, it was presumed that the distribution could be a normal one or skewed to either side depending on the analytical methods the participating laboratories used. (Of course, if all participating laboratories had their results bias-corrected using appropriate certified reference materials, there should not be any problem of laboratory bias.) Similarly, with an input of a random number and the probability density function of a pre-set distribution, the b_i could be simulated (Fig. 3). However, a similar approach could not be applied to the simulation of S_{Pi} since it was envisioned that the participating laboratories might be divided into groups having good, normal or poor method precision. In this study, good or poor precision are defined using multiples of the p value calculated from the well-recognized Horwitz equation [10] as shown below for analyte with concentration C in the test sample.

$$p = 2^{(1-0.5\ \log_{10} C)} \tag{3}$$

For inter-laboratory collaboration, the p value is widely used to determine the acceptable reproducibility (between laboratory) precision for the analyte concentration C. Also, it was well recognized that the ratio between reproducibility (between laboratory) precision and repeatability (within laboratory) precision should not be greater than 2 [11]. In view of this, ranges for good, normal and poor precision were defined as $0.1p$—$0.5p$, $0.5p$—$2p$, and $2p$—$5p$ respectively. Then with a pre-set distribution pattern and an input of a random number, the S_{Pi} for the ith participating laboratory could be simulated (Fig. 4). With the simulated values for m$_i$, b_i, S_{Pi} and one more random number for k_{2i}, the result x_i of the ith participating laboratory could be simulated. By repeating the above procedure n times for n participating laboratories, the results for one round of

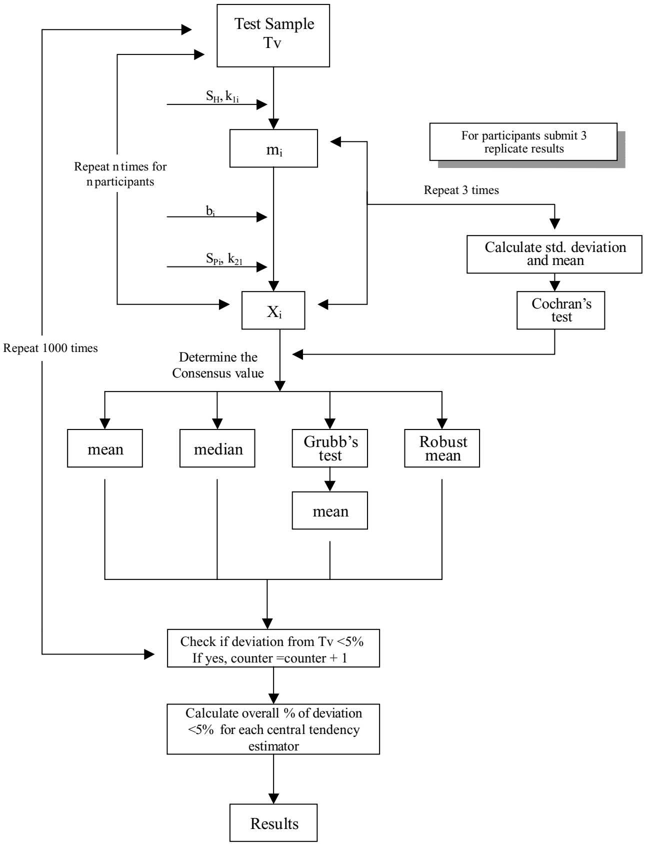
Test Sample
Tv
S$_H$, k$_{1i}$
m$_i$
For participants submit 3 replicate results
Repeat n times for n participants
Repeat 3 times
b$_i$
Calculate std. deviation and mean
S$_{Pi}$, k$_{21}$
X$_i$
Cochran's test
Repeat 1000 times
Determine the Consensus value
mean
median
Grubb's test
Robust mean
mean
Check if deviation from Tv <5%
If yes, counter =counter + 1
Calculate overall % of deviation <5% for each central tendency estimator
Results

Fig. 2 Flow chart of the simulation process

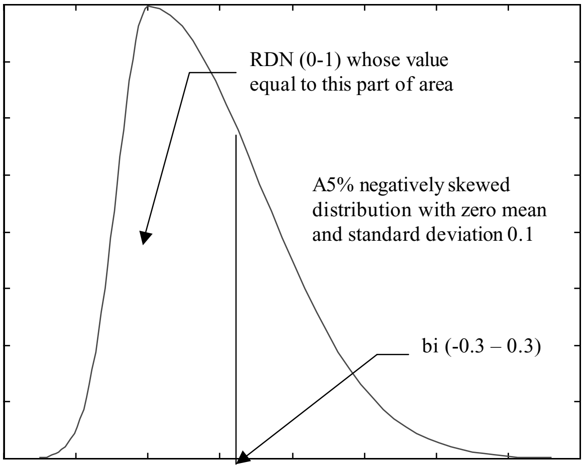

Fig. 3 Simulate the b_i value from a random number RDN using the probability density function for a skewed distribution

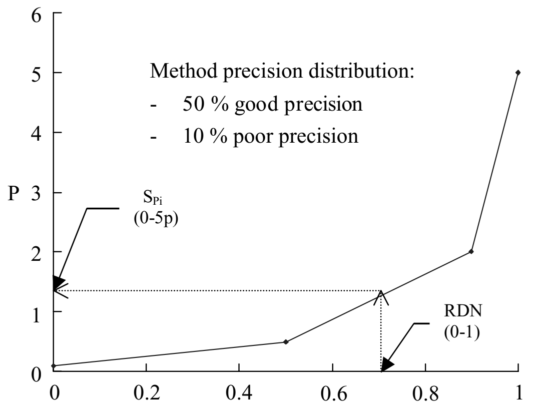

Fig. 4 Simulate S_{Pi} from a random number RDN for a particular method precision distribution

PT programme could then be simulated and the consensus value could be subsequently determined.

However, differenct practices for data handling among different PT programmes should be considered during the course of the simulation process. For instance, different central tendency estimators such as mean, median and robust mean [12] might be used for the determination of the consensus values. Also, in some PT programmes, participating laboratories were asked to provide a single result, while other programmes required that the results be reported in triplicate; the assigned values could be obtained from the participating laboratories' results with or without the removal of outliers. To take into account these variations, the simulation process was sub-branched and expanded accordingly (Fig. 2). Moreover, sub-routines were added to check if the application of outlier tests including Cochran's test [13] (for the repeatability of results from an individual participating laboratory) and Grubb's test [14] (for the data pool among participating laboratories) would improve the discrepancy between the consensus value and the true value. As it was statistically infeasible to draw any conclusion for each combination of parameter with only one round of simulation process, the simulation process was repeated 1,000 times and the probability of having an absolute deviation of less that 5% from the true values, $P(\text{deviation} < 5\%_{95\%})$ was reported for the evaluation, assuming that the acceptable deviation from the true value was 5%.

To facilitate the computation, the simulation process was programmed using the MATLAB software. Table 1 lists out the proposed ranges of each parameter discussed above for the simulation process and the evaluation study was proceeded with 15 random combinations of these parameters at different levels (Table 2).

In the study described above, the objective was to evaluate the use of consensus values in PT programmes in view of the possible variations of parameters or practices of data treatment. However, PT programme providers or participants are much more concerned about which parameters would have significant effects on the deviation from the true values. To achieve this, the experimental design technique together with the simulation process had to be used. Considering the number of parameters involved, a 3^{5-2} fractional factorial design [15] consisting of 27 different combinations were proposed for the simulation process. After that, the effects of each individual parameters on the $P(\text{deviation} < 5\%_{95\%})$ value were assessed.

Results and discussion

The built-in random-number generating function provided with MATLAB was used to generate random numbers. Despite the fact that numbers generated by computer programmes are only pseudo-random, they were found to be sufficient for the use in Monte Carlo simulation [16]. For instance, in Fig. 5, the *z*-score plot obtained from the results of a simulated PT programme does resemble a real one. Also, the repeatability of the simulation programme was checked to be satisfactory. Under a particular set of parameters, the simulation process was repeated 10 times and the relative standard deviation of the $P(\text{deviation} < 5\%_{95\%})$ values obtained was only about 2%.

In the evaluation study, the $P(\text{deviation} < 5\%_{95\%})$ values obtained varied greatly among the 15 random combinations of parameters (Table 2). Overall, only a few of them managed to obtain an $P(\text{deviation} < 5\%_{95\%})$ value close to 1. In the extreme case, a value of 0.27 was noted, which implied, under that combination of parameters, the chance of getting a consensus value which has a deviation of less than 5% from the true value is only 0.27. When we examined the spreading of the deviation for that particular combination of parameters, it was noticed that more than one third of the cases would have a deviation in the range

Table 1 Proposed different parameter level values for the evaluation study

Parameter	Level 1	Level 2	Level 3
No. of participants	12	50	100
True value, ppm	0.1	5	100
Sample homogeneity, %	0.1	1	10
Distribution of laboratory bias (in terms of R.S.D.)	5% negatively skewed normal distribution with zero mean and standard deviation of 0.1	Normal distribution with zero mean and standard deviation of 0.1	5% positive skewed normal distribution with zero mean and standard deviation of 0.1
Distribution of method precision (in terms of R.S.D.)	Distribution pattern: 50%—good precision, 10%—poor precision	Distribution pattern: 10%—good precision, 10%—poor precision	Distribution pattern: 10%—good precision, 50%—poor precision

Table 2 Results of simulation process for the evaluation of consensus values in PT programme

Level of parameters					Single result				Average of triplicate results				Average of triplicate results (removal of repeatability outliers)			
N	T_V	S_H	B	S_{Pi}	Median	Mean	Mean (outliers)	Robust mean	Median	Mean	Mean (outliers)	Robust mean	Median	Mean	Mean (outliers)	Robust mean
3	3	3	2	2	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
1	2	3	3	1	0.56	0.56	0.57	0.58	0.62	0.63	0.64	0.63	0.62	0.63	0.63	0.65
1	3	1	3	3	0.53	0.57	0.56	0.58	0.68	0.68	0.68	0.69	0.69	0.71	0.70	0.72
1	1	1	1	2	0.48	0.45	0.50	0.51	0.58	0.53	0.58	0.59	0.58	0.56	0.58	0.59
1	3	1	1	3	0.54	0.58	0.57	0.58	0.64	0.62	0.64	0.64	0.66	0.66	0.66	0.68
2	2	1	2	2	0.99	0.97	0.98	0.99	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
2	1	1	2	3	0.79	0.49	0.48	0.63	0.94	0.76	0.76	0.87	0.95	0.75	0.73	0.87
3	1	1	3	1	0.64	0.66	0.67	0.76	0.84	0.80	0.78	0.89	0.87	0.79	0.82	0.89
2	3	2	3	1	0.59	0.78	0.79	0.75	0.85	0.92	0.93	0.92	0.86	0.93	0.92	0.92
1	3	2	3	3	0.53	0.58	0.58	0.58	0.67	0.67	0.68	0.68	0.68	0.68	0.68	0.70
3	2	1	1	1	0.66	0.75	0.78	0.77	0.87	0.87	0.89	0.93	0.88	0.86	0.91	0.93
2	3	2	3	2	0.64	0.80	0.80	0.76	0.88	0.91	0.92	0.91	0.84	0.91	0.92	0.92
2	2	2	3	1	0.61	0.71	0.73	0.72	0.83	0.83	0.88	0.88	0.84	0.85	0.88	0.89
1	1	3	2	3	0.42	0.28	0.27	0.33	0.54	0.42	0.42	0.50	0.52	0.42	0.43	0.50
2	2	1	2	3	0.94	0.80	0.80	0.90	0.99	0.98	0.97	0.99	1.00	0.97	0.96	0.99

10–40% (Fig. 6). Under these circumstances, there would be a high chance of getting undesirable interpretation of participants' performance if the consensus value obtained was used as the assigned value.

However, as indicated in Table 2, a robust mean should be preferred to other central tendency estimators for the determination of the consensus values. Also, removal of repeatability outliers with the Cochran Test before the determination of the consensus value helped increase the P(deviation $<5\%_{95\%}$), i.e. improve the discrepancy between the consensus value and the true value. Of course, to achieve this, PT organizers might need to request participating laboratories to provide results in triplicate or at least in duplicate.

To look for the dominant parameters that contribute significant effects to the deviation from the true values, the P(deviation $<5\%_{95\%}$) values for the 27 different combinations of parameters according to the 3^{5-2} fractional factorial design were obtained. The change in the P(deviation $<5\%_{95\%}$) values due to the changes in respective parameters are summarized in Table 3. It was observed that changing the number of participants from 12 to 100 would increase the P(deviation $<5\%_{95\%}$) value by 0.19. Also, there would be a difference of 0.21 in the P(deviation $<5\%_{95\%}$) value if the distribution of the laboratory bias changed from a normal one to a 5% negatively skewed distribution. However, increasing the proportion of laboratories having good precision did not significantly improve the P(deviation $<5\%_{95\%}$) value whilst increasing the proportion of laboratories having poor precision caused a decrease of 0.1 in the P(deviation $<5\%_{95\%}$) value. Finally, the effects of sample homogeneity and concentration levels of analyte are found to be similar and comparatively less significant. To conclude, the number of participating laboratories and the distribution of the laboratory bias among the participants were identified as the dominant parameters.

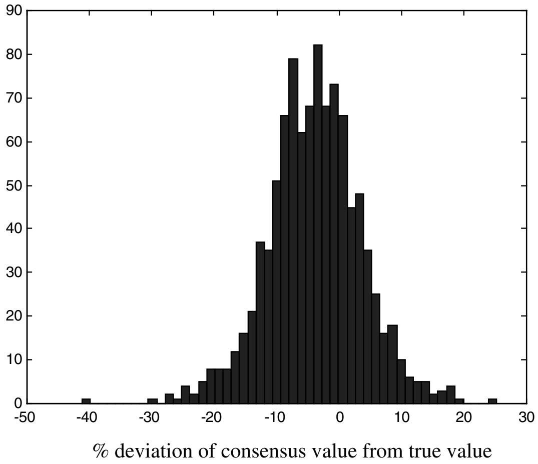

Fig. 6 Distribution of % deviation from true value of the results of a simulated PT programme

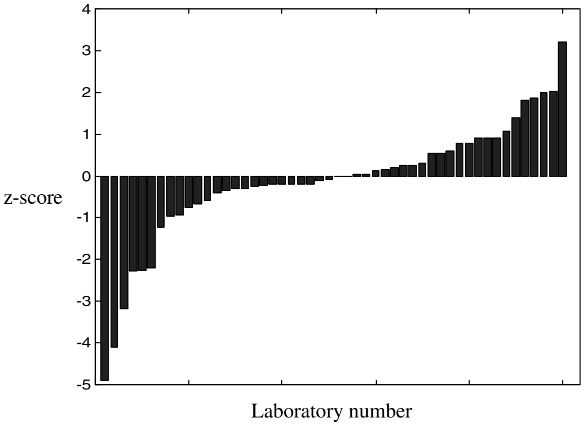

Fig. 5 z-Scores of simulated PT programme results

Table 3 Evaluation of the effects of individual parameters on P(deviation $<5\%_{95\%}$)

Parameter	Change in setting	Change in P(deviation $<5\%_{\ 95\%}$)
No. of participating laboratories, n	12→100	0.19
Concentration level of analyte, T_V	0.1 ppm→100 ppm	0.08
Sample homogeneity, S_H	0.1%→10%	−0.08
Distribution of laboratory bias, b_i	−5%→0%	0.21
	5%→0%	0.16
Distribution of method precision, S_{Pi}	Portion of laboratories with poor precision: 50%→10%	0.10
	Portion of laboratories with good precision: 50%→10%	0.03

To enable a reasonable evaluation of participating laboratories' performance, the PT programme providers are hence encouraged to enrol as many participating laboratories as possible. Although it might not be feasible to control the distribution of laboratory bias among the participating laboratories, they could be encouraged to use certified reference materials (CRMs) to control their method bias. Or the PT programme providers could consider using only the results of those participating laboratories which had bias corrected using appropriate CRMs for the determination of the consensus value.

Further application

Similar to the case of the assigned value, the target standard deviation, i.e. the *s* term in the calculation of *z*-score, was also usually obtained from the participating laboratories results. Using the Monte Carlo simulation technique again, we could study how the *s* values obtained would affect the *z*-score values for the participating laboratories. Moreover, it is not uncommon to come across data obtained from PT programmes which were multimodal or skewed [17]. Under these circumstances, the consensus values obtained using the common central tendency estimator could deviate significantly from the true values. Using the Monte Carlo technique, we could simulate the effect due to multimodality in the participants' result and hence evaluate the use of consensus value for this special situation.

Conclusion

Using the Monte Carlo simulation technique, this paper evaluated the use of consensus value as the assigned value in PT programmes. The results revealed that the number of participating laboratories, sample homogeneity, concentration level of the analyte, distribution of laboratory bias, method precision, and the different practices of data handling would all affect the discrepancy between the consensus value and the true value. In some cases, this would render an undesirable evaluation of the participating laboratories' performance. With the use of experimental design technique, the number of participating laboratories and the distribution of the laboratory bias among the participants were identified as the dominant parameters. That means that if the PT programme providers choose to use the consensus value as the assigned values for the calculation of the *z*-scores, they should try to enrol as many participants as possible and encourage the participants to use CRMs to control their method bias. Moreover, the findings of the study indicated that a robust mean should be preferred as the central tendency estimator and it would be better to have the repeatability outliers removed before the determination of the consensus value.

References

1. Örnemark U, Boley N, Saeed, Petronella K, De Bievre P et al. (2001) Accred Qual Assur 6:140–146
2. Thompson M, Wood R (1993) International harmonized protocol for proficiency testing of (chemical) analytical laboratories. J Int AOAC 76:926–940
3. Visser RG (2001) Accred Qual Assur 6:442–443
4. Lawn RE, Thompson M, Walker RF (1997) Proficiency testing in analytical chemistry, Royal Society of Chemistry, Cambridge UK
5. Guell OA, Holcombe JA (1990) Analytical Chemistry 62:529A–542A
6. Maroto A, Boque R, Riu J, Rius FX (2003) Analyst 128:373–378
7. Hill ARC, Holst CV (2001) Analyst 126:2044–2052
8. Lepek A (2003) Accred Qual Assur 8:296–299
9. ISO 5725-1 (1994) Accuracy, trueness and precision of measurement methods and results. Part 1, General principles and definitions, ISO, Geneva, Switzerland
10. Thomspon M (2004) AMC Technical Brief No. 17, Royal Society of Chemistry, Cambridege, UK
11. Thomspon M, Lowthian PJ (1995) Analyst 120:271–272
12. Analytical Methods Committee of RSC (1989) Analyst 114:1693–1702
13. Mullins E (2003) Statistics for the quality control chemistry laboratory, Royal Society of Chemistry, UK
14. Miller JN, Miller JC (2000) Statistics and chemometrics for analytical chemistry, 4th ed, Prentice Hill Englewood Cliffs, New Jersey
15. Berthouex PM, Brown LC (1994) Statistics for environmental engineers, Lewis Publishers Boca Raton, Florida
16. Gonzalez AG, Herrador MA, Asuero AG (2005) Acced Qual Assur 10:149–151
17. Lowthian PJ, Thompson M (2002) Analyst 127:1359–1364