Proficiency testing for determination of pesticides residues in black tea: Comparison of three robust statistical approaches based on ISO 13528 to estimate the consensus values for small number of participants

Nurhani Aryana
Research Center for Chemistry, Indonesian Institute of Sciences, Building 452
Kawasan Puspiptek Serpong, Tangerang Selatan, Indonesia

Corresponding author email: nurh020@lipi.go.id

Abstract. In Indonesia, proficiency testing (PT) provider for determination of pesticides residues in tea is still limited. In addition, many proficiency testing schemes still use consensus value from participants’ results as the assigned value rather than reference assigned value due to some technical and economical issues. However, this mode of assigned value may lack of reliability when the number of participants are small and the test results are highly varied. In these conditions, many PT providers were not able to perform the statistical data evaluation causing the needed PT became unavailable. This study is aimed to investigate the suitable statistical approach in estimating the consensus value for small number of participants in PT for determination of pesticides residues in black tea. Three robust statistical approaches recommended by ISO 13528: variants of Algorithm A, Hampel method, and M estimator were compared to evaluate the participants’ data from the PT schemes organized by Research Center of Chemistry-LIPI in 2014 and 2015. All of the robust statistical approaches showed to be able to assess the performance of participants. The Hampel method and M estimator were the most adequate robust statistical approaches in obtaining the closest consensus values compared to the homogeneity test results, stability monitoring results, and the assigned valued in the real PT scheme, for endosulfan sulphate and bifenthrin, respectively.

1. Introduction
The amount of pesticides residues in tea commodities is routinely checked during trading to ensure that the quality of tea meets the requirements set by the food safety standards or the respective national regulations. However, an accurate measurement occasionally cannot be easily performed since some of the maximum residue limit (MRL) of the residual pesticides are at trace level. Laboratories for testing pesticides should incorporate all the management and technical requirements as specified in ISO/IEC 17025 in order to ascertain reliable analytical data in this challenging work [1]. One of the mandatory requirements for a laboratory in monitoring the validity of its results according to ISO/IEC 17025 is by participation in a proficiency testing (PT) program. PT is one of the key elements in the implementation of an appropriate quality assurance program and performance monitoring procedure for chemical analysis laboratories [2].

Metrology in chemistry laboratory-Indonesian Institute of Sciences (MiC-LIPI) has organized two PT schemes for determination of pesticides residues in black tea. Both of the PT schemes were using candidate reference materials pesticides residue in black tea, PTH14, as the test samples. The test
parameters were endosulfan sulphate and bifenthrin. The first PT scheme was conducted in 2014 and was participated by three laboratories from Indonesia. The second PT scheme was conducted in 2015 and was participated by six laboratories from Myanmar and two laboratories from India. According to ISO 13528 – Statistical methods for use in proficiency testing, where the number of participants is small, the assigned value should ideally be determined using a metrologically valid procedure and independent of the participants [3]. In these ideal conditions, the PT scheme can be rolled even with just one participant. The metrologically valid procedure can be achieved by using a certified reference material or a primary method. The assigned values in the PT schemes were derived from analytical measurements by using high sensitivity methods: gas chromatography-µ-electron capture detector (GC-µECD) method, gas chromatography-mass spectrometry (GCMS) method, and gas chromatography-isotope dilution mass spectrometry (GC-IDMS) method.

In Indonesia, many PT schemes were still operated by the consensus values derived from the participants’ results. Apart from technical issues, one of the reasons often mentioned is that PTs operated with consensus value would be cheaper than those using reference value obtained from a priori characterization measurements [4]. It has been reported that over 90 % of the proficiency testing programmes rely on consensus values [5]. However, when the PT scheme was only participated by small number of participants and the test results were highly varied, the performance evaluation by using consensus value from the participants’ results may become unreliable. In these conditions, many PT providers were not able to perform the statistical data evaluation causing the needed PT became unavailable. To overcome these issues, in this study, some data treatment processes for PT with small number of participants were investigated including outlier identification methods and three robust statistical approaches recommended by ISO 13528 for calculating consensus value from participants’ results: variants of Algorithm A, Hampel method, and M estimator. The participants’ data combined from the first and second PT schemes for determination of pesticides residue in black tea were used in this study. The effectiveness of the robust statistical approach was expressed as the percentage of participants with acceptable performance score. The results in this study were further verified by using the homogeneity test results performed by gas chromatography-µ-electron capture detector (GC-µECD) method, stability monitoring results performed by gas chromatography-isotope dilution mass spectrometry (GC-IDMS) method, and the assigned values from the real PT scheme.

2. Experimental section

2.1. Preparation of the test samples
The test samples were prepared as candidate certified reference materials in accordance with ISO/IEC 17034 [6] and ISO Guide 35 [7]. The procedure for the preparation of the test samples were described elsewhere [8]. The homogeneity and stability of the test samples were checked in accordance with ISO 13528 [3].

2.2. Analysis by the participants
The PT schemes were organized in accordance with ISO/IEC 17043 [9] and were subscribed by total of 11 participants from Indonesia, Myanmar, and India. Each participant received one bottle of the test sample together with the technical instruction and the result report form. The participants were asked to determine preferable the mass fraction of the test parameters endosulfan sulphate and bifenthrin (in three significant figures, in µg/kg) by using the participants’ routine methods.

2.3. Statistical approaches
The following procedures describe the data treatment processes for PT with small number of participants, including outlier identification methods, calculation of consensus value from participants’ results, and performance evaluation of participants’ results. All of the statistical data evaluation were done in accordance with ISO/IEC 17043 [9] and ISO 13528 [3] by using the Microsoft Excel.
2.3.1. Outlier identification. Three identification methods were used in this study to detect any outlier in the data set: Grubb’s test, Hampel’s test, and trimmed average. The most commonly used and recommended test by ISO for outliers is the Grubb’s test since it is easy to apply and operate using the mean and standard deviation of the data [10]. The Grubb’s test was performed to the highest and lowest values in the set of test results. Grubb’s statistics for the highest value and the lowest value are given as equation (1) and equation (2), respectively.

\[
G_p = \frac{x_p - \bar{X}}{s}
\]

\[
G_p = \frac{\bar{X} - x_l}{s}
\]

where \(G_p\) indicates the Grubb’s statistics, \(x_p\) indicates the highest value, \(x_l\) indicates the lowest value, \(\bar{X}\) indicates the mean value, and \(s\) indicates the standard deviation. \(G_p\) value was compared with the critical value for the \(p\) number of test results.

The Hampel’s test [3] was performed by equation (3).

\[
d_i = |x_i - \text{med}(x_i)|
\]

where \(d_i\) indicates the absolute differences, \(x_i\) is the \(i^{th}\) data of the test results (for \(i = 1\) to \(p\) number of test results), and \(\text{med}(x_i)\) indicates the median of the test results. The outlier value from the data set can be obtained by comparing the \(d_i\) value from the condition given in equation (4).

\[
|d_i| \leq 4.5 \text{ med} (d_i)
\]

where \(\text{med}(d_i)\) is the median of the absolute differences.

The trimmed average method [11] was performed by calculating the mean \(\bar{X}\) and standard deviation \(s\) from all data. The values outside of the likelihood interval (\(\bar{X} - s\) or \(\bar{X} + s\)) then were discarded.

2.3.2. Estimation of the consensus values from participants’ results. The first robust statistical approach used to estimate the consensus values from the participants’ results was variants of Algorithm A [3].

The algorithm A was performed with no iteration of location, using the median (equation 5) as a location estimate.

\[
x^* = \text{med}(x_i)
\]

The robust standard deviation of participants’ results (\(s^*\)) was calculated with Q estimator [3] and the \(s^*\) was not updated during iteration. The \(Q_n\) for a data set was calculated with \(p\) reported results. Firstly, the absolute differences (\(p(p-1)/2\)) were calculated as equation (6).

\[
d_{ij} = |x_i - x_j| \text{ for } i = 1, 2, \ldots p - 1 \text{ and } j = i + 1, i + 2, \ldots p
\]

The ordered differences (\(d_{ij}\)) were denoted by \(d_{11}, d_{22}, \ldots d_{p(p-1)/2}\). Then, the number distinct pairs (\(k\)) chosen from \(h\) objects was calculated by using equation (7),

\[
k = \frac{h(h-1)}{2}
\]

where

\[
h = \left\{ \begin{array}{ll} p/2 & \text{ } p \text{ even} \\ \frac{p-1}{2} & \text{ } p \text{ odd} \end{array} \right.
\]

The \(Q_n\) was calculated as equation (9),

\[
Q_n = 2.2219 d_{[k]} b_p
\]

where \(b_p\) was selected from Table C.2 in ISO 13528 for a particular number \(p\) of data points.

The second robust statistical approach used to estimate the consensus values from participants’ results was the Hampel estimator. The Hampel estimate of location was provided by an iterative
reweighting scheme [3], started with calculating the initial robust mean \( \bar{x}^* \) as calculated in equation (5). The robust standard deviation of participants’ results \( s^* \) was calculated as the scaled median absolute deviation \( \text{MAD}_e \) by using equation (10),

\[
\text{MAD}_e = 1.483 \text{med} |x_i - x^*| \quad (10)
\]

where \( i = 1, 2, \ldots p, \) with \( p \) being the number of test results.

For each data point \( x_i, q_i \) was calculated as equation (11). Then, for each \( q_i, \) the weight \( w_i \) was calculated as equation (12).

\[
q_i = \frac{|x_i - x^*|}{s^*} \quad (11)
\]

\[
w_i = \begin{cases} 
0 & |q| > 4.5 \\
\left(4.5 - \frac{q}{q^*}\right) & 3 \leq |q| \leq 4.5 \\
\frac{1.5}{q} & 1.5 \leq |q| < 3.0 \\
q & |q| \leq 1.5
\end{cases} \quad (12)
\]

The \( x^* \) was then recalculated as equation (13),

\[
x^* = \frac{\sum_{i=1}^{p} w_i x_i}{\sum_{i=1}^{p} w_i} \quad (13)
\]

The steps in equation (11) to equation (13) were repeated until \( x^* \) converges. Convergence can be assumed when the change in \( x^* \) from one iteration to the next was less than \( 0.01s^*/p^{1/2} \), corresponding to approximately 1 % of the standard error in \( x^* \).

The third robust statistical approach used to estimate the consensus values from participants’ results was the M estimator. According to Rousseeuw et al. [12], the robust mean \( (\bar{x}^*) \) and robust standard deviation \( (s^*) \) are estimated by the M estimator as equation (14) and equation (15),

\[
\frac{1}{n} \sum_{i=1}^{p} \psi \left( \frac{x_i - T_n}{S_n} \right) = 0 \quad (14)
\]

\[
\psi(x) = \frac{e^x - 1}{e^x + 1} \quad (15)
\]

where \( T_n \) is the location \( x^* \) estimator and \( S_n \) is the scale \( s^* \) estimator.

The initial robust mean \( (\bar{x}^*) \) was calculated as equation (5). Next, \( T_n \) was iteratively calculated using equation (14) computed using a Newton-Raphson algorithm. The robust standard deviation \( (s^*) \) was calculated as \( Q_n \), following equation (6)-(9).

2.3.3. Performance evaluation. Before selecting the criterion for performance evaluation, the standard uncertainty of the consensus value \( u(x_{pt}) \) was compared to the standard deviation of proficiency assessment \( \sigma_{pt} \). The standard uncertainty of the consensus value may be considered to be negligible and need not be included in the interpretation of the results of the round of proficiency testing when the criterion in equation (16) was met [3].

\[
u(x_{pt}) < 0.3\sigma_{pt} \quad (16)
\]

When this criterion was met, laboratory performance was assessed by using z-score as shown in equation (17). When the criterion in equation (16) was not met, the standard uncertainty of the consensus value need to be included in the interpretation of the results of proficiency testing and the laboratory performances was assessed by using z'-score as shown in equation (18).

\[
z_i = \frac{x_i - x_{pt}}{\sigma_{pt}} \quad (17)
\]

\[
z'_i = \frac{x_i - x_{pt}}{\left((\sigma_{pt}^2 + u^2(x_{pt}))^{1/2}\right)} \quad (18)
\]
where $x_i$ is the submitted participants result.

The performance scores calculated by the standard deviation for proficiency assessment ($\sigma_{pt}$) derived from the robust standard deviation of participants’ results ($s^*$) were compared with the standard deviation from Horwitz Equation as calculated by equation (19),

$$\sigma_{pt} = 0.22c$$

(19)

3. Results and discussion

From the total of 11 registered participants, there were 8 and 10 laboratories submitted the test results for endosulfan sulphate and bifenthrin, respectively. In order to see the scattering of the data, the participants’ test results are presented in figure 1 and 2, sorted by increasing value with the red line represents the median value. Figure 1 and 2 show wide range of test results for both of endosulfan sulphate and bifenthrin. In the PT with small number of participants, it is important to treat extreme results correctly if they are not caused by a known gross error or miscalculation [13]. Information about the distribution patterns of the data (such as normality) is needed before applying any statistical analysis tools since typical statistical tests incorporate assumptions about the underlying distribution of data [9]. The histograms were drawn to assess the distribution of the data set and detect visually any outliers. The histograms of the test results of endosulfan sulphate and bifenthrin show tailed distributions and the presence of some test results which can be considered as outliers.

![Figure 1. The participants’ test results and the histograms of the data in the PT for determination of endosulfan sulphate in black tea.](image)

When the number of participants is small, extreme results cannot usually be identified as outliers by known statistical tests because of low power of these tests [13]. Outlier tests are seldom completely satisfactory, especially when many outliers may be present [14]. In this study, three outlier identification methods were employed to identify the outliers: Grubb’s test, Hampel’s test, and trimmed average. Outlier identification of the test results are summarized in table 1 and 2 for endosulfan sulphate and bifenthrin, respectively. The Grubb’s and Hampel’s tests compare the difference between an extreme value and the centre of the data with the variability of the data and identify the extreme value as an outlier if the ratio is too large [15]. For endosulfan sulphate, the $G_p$ value for the lowest test result was 1.12 and the $G_p$ value for the highest test result was 1.60. These $G_p$ values were smaller than the critical value of Grubb’s statistic ($G_{pcrit} = 2.03$, $n=8$, at confidence level 95 %), therefore, there was no outlier can be detected by the Grubb’s test. For bifenthrin, the $G_p$ value for the lowest test result was 1.26 and the $G_p$ value for the highest test result was 1.70. These $G_p$ values were also smaller than the critical value.
of Grubb’s statistic \( (G_{pr} = 2.18, n=10, \text{ at confidence level } 95\%) \), therefore, there was no outlier can be detected by the Grubb’s test. According to ISO 13528, the Grubb’s test is most useful for \( p>10 \) and can fail when there are multiple outliers \([3]\). There was also no outlier detected by the Hampel’s test since the test results for endosulfan sulphate and bifenthrin have not met the criterion in equation (4).

![Figure 2. The participants’ test results and the histograms of the data in the PT for determination of bifenthrin in black tea.](image)

| Lab Code | Test result (µg/kg) | Grubb’s test | \( |d| \)  |
|----------|---------------------|-------------|----------|
| LC0001   | 919                 | OK          |         |
| LC0003   | 71                  | OK          |         |
| LC0004   | 1082                | OK          |         |
| LC0005   | 121                 | OK          |         |
| LC0007   | 552                 | OK          |         |
| LC0009   | 235                 | OK          |         |
| LC0010   | 300                 | OK          |         |
| LC0011   | 620                 | OK          |         |

OK indicates inlier test result  
$$ $$ indicates outlier test result

The values of 4.5med\((d)\) were 1123 µg/kg and 740 µg/kg, for endosulfan sulphate and bifenthrin, respectively. Outliers finally can be detected by the trimmed average method as shown in table 1 and 2. For endosulfan sulphate, the outlier-corrected average value was change from 488 µg/kg to 366 µg/kg when the outliers detected by trimmed average method were excluded. For bifenthrin, the outlier-corrected average value was change from 660 µg/kg to 596 µg/kg when the outliers detected by trimmed average method were excluded. The trimmed average method was the most suitable method to detect the outliers in this study.
Table 2. Identification of outliers using different methods in the PT for determination of bifenthrin in black tea.

| Lab Code | Test result (µg/kg) | Grubb’s test | Hampel’s test $|d_i|$ | Outlier test | Trimmed average |
|----------|--------------------|--------------|----------------|--------------|----------------|
| LC0001   | 1307               | OK           | 684            | OK           | $$             |
| LC0002   | 377                | OK           | 246            | OK           | OK             |
| LC0003   | 162                | OK           | 461            | OK           | $$             |
| LC0004   | 1333               | OK           | 710            | OK           | $$             |
| LC0005   | 221                | OK           | 402            | OK           | $$             |
| LC0007   | 671                | OK           | 48             | OK           | OK             |
| LC0009   | 646                | OK           | 23             | OK           | OK             |
| LC0010   | 600                | OK           | 23             | OK           | OK             |
| LC0011   | 706                | OK           | 83             | OK           | OK             |
| LC0100   | 573                | OK           | 50             | OK           | OK             |

OK indicates inlier test result
$$ indicates outlier test result

According to IUPAC/CITAC Guide – Selection and use of proficiency testing schemes for a limited number of participants [13], the assigned value of the PT items cannot be calculated safely from the measurement results obtained by the participants as a consensus value. The consensus value and the robust standard deviation from participants’ results, in principle, cannot be used for a reliable assessment of an individual laboratory performance when the number of participants is small. Methods of obtaining traceable assigned values are to be used wherever possible to provide comparable PT results. In this study, the consensus values are used only for comparison of PT results as a statistical sample.

In general, the consensus value from participants’ results can be obtained by central tendency estimators such as the arithmetic mean or robust mean [5]. The arithmetic mean is the minimum variance estimator, but it is sensitive to the influence of outliers. Meanwhile, robust statistics is a modern approach to the outlier problem in which the influence of outliers and heavy tails is downweighted. Robust statistics have been shown to be applicable to analytical data in number of situations [16]. However, when the number of participants is small ($p<20$), the choice of statistical approaches should be taken carefully since most of them become increasingly unreliable and deviations from a normal distribution are harder to identify. Among many types of robust statistics, the median, MADs, and Q/Hampel methods were selected in this study since these methods can tolerate a very large proportion of outliers.

When the data were calculated by classical mean and standard deviation without exclusion of outliers, the obtained consensus values and standard deviation of participants’ results were 488 µg/kg and 370 µg/kg, for endosulfan sulphate, and 660 µg/kg and 244 µg/kg, for bifenthrin. These standard deviation were very large to be used in the calculation of performance scores. In this condition, limit on the largest standard deviation that will be used should be placed so that results that are not fit for purpose will receive an action signal. According to ISO 13528, there are two specific recommendations for robust estimates of dispersion in very small data where $p>2$: M-estimate of standard deviation based on a logarithmic weighting function as recommended by Rousseeuw [12], and algorithm A with no iteration of location, using the median as a location estimate. Based on the outlier identification results, the proportion of outliers were expected to be more than 20 %, therefore, $s^*$ was replaced with the Q estimator and was not updated during iteration. In this study the Hampel method was also employed in estimating the consensus values since this method has high breakdown point and high efficiency.
Table 3. The obtained consensus value of participants’ results by using three robust statistical approaches for endosulfan sulphate.

| Statistical approach          | $x^*$ (µg/kg) | $u$ (µg/kg) | $U$ (µg/kg, $k=2$) | $U_{rel}$ (%) | $s^*$ (µg/kg) | $\sigma_{pt}$ (µg/kg) | 0.3 $\sigma_{pt}$ (µg/kg) |
|------------------------------|---------------|-------------|---------------------|---------------|---------------|------------------------|---------------------------|
| Variants of Algorithm A      | 426           | 107         | 214                 | 50            | 243           | 94                     | 28                        |
| Hampel method                | 482           | 164         | 327                 | 68            | 370           | 106                    | 32                        |
| M estimator                  | 459           | 107         | 214                 | 47            | 243           | 101                    | 30                        |

Table 4. The obtained consensus value of participants’ results by using three robust statistical approaches for bifenthrin.

| Statistical approach          | $x^*$ (µg/kg) | $u$ (µg/kg) | $U$ (µg/kg, $k=2$) | $U_{rel}$ (%) | $s^*$ (µg/kg) | $\sigma_{pt}$ (µg/kg) | 0.3 $\sigma_{pt}$ (µg/kg) |
|------------------------------|---------------|-------------|---------------------|---------------|---------------|------------------------|---------------------------|
| Variants of Algorithm A      | 623           | 124         | 248                 | 40            | 314           | 137                    | 41                        |
| Hampel method                | 596           | 96          | 193                 | 32            | 244           | 131                    | 39                        |
| M estimator                  | 690           | 124         | 248                 | 36            | 314           | 152                    | 46                        |

Table 3 and 4 present the obtained consensus values from participants’ results, for endosulfan sulphate and bifenthrin, respectively. All of the test results of endosulfan sulphate and bifenthrin were included in the calculation of the consensus values by variants of algorithm A and Hampel method. Meanwhile, the outliers detected by the trimmed average method were excluded in the calculation of the consensus values by M estimator.

Figure 3. Comparison between the obtained consensus values of participants’ results with the homogeneity test results derived from GC-µECD measurements, stability monitoring results derived from GC-IDMS measurements, and the assigned values in the real PT scheme of (a) endosulfan sulphate and (b) bifenthrin. The solid red line indicates the assigned value in the real PT scheme. The dotted blue lines indicate range of tolerance ($|z^*\text{-score}|\leq 2.00$).
Figure 3 shows the comparison between the obtained consensus values of participants’ results with the homogeneity test results derived from GC-FTIR measurements, stability monitoring results derived from GC-IDMS measurements, and the assigned values in the real PT scheme. For endosulfan sulphate, the consensus values of participants’ results obtained from three robust statistical approaches were inside the range of tolerances and close to the homogeneity and stability test results and the assigned value in the real PT scheme. Meanwhile, for bifenthrin, the obtained consensus values of participants’ results were close to the low limit of the range of tolerance. This may be caused by some technical issues in the sample preparation and analysis of the target analyte.

As shown in Table 3 and 4, the standard uncertainties of the consensus values obtained from the calculation of three robust statistical approaches have not met the criterion set in equation (16), for both of test parameters. Therefore, the performance scores of participants’ results were evaluated by z’-score. The standard deviation for proficiency assessment ($\sigma_p$) was taken from the Horwitz equation since performance score evaluation by using $\sigma_p$ from $s$ cannot detect the outliers as action signal. Table 5 and 6 presents the performance scores of participants in terms of z’-score, for endosulfan sulphate and bifenthrin, respectively. Participants’ performance was ranked into three categories based on ISO/IEC 17043 [9], where $|z'| \leq 2.00$ is considered to be satisfactory (OK), $2.00 < |z'| < 3.00$ is considered to be questionable ($\$$), and $|z'| \geq 3.00$ is considered to be unsatisfactory ($$$).

### Table 5. The performance scores of participants results in terms of z’-score for endosulfan sulphate calculated by variants of Algorithm A, Hampel method, and M estimator.

| Lab Code | Test result (µg/kg) | Variants of Alg A | Hampel method | M estimator |
|----------|--------------------|-------------------|---------------|-------------|
| LC0001   | 919                | 3.46              | $\$$          | 2.24        | $\$$          |
| LC0003   | 71                 | -2.49             | $              | -2.11       | $              |
| LC0004   | 1082               | 4.61              | $\$$          | 3.08        | $\$$          |
| LC0005   | 121                | -2.14             | $              | -1.85       | OK             |
| LC0007   | 552                | 0.88              | OK            | 0.36        | OK             |
| LC0009   | 235                | -1.34             | OK            | -1.27       | OK             |
| LC0100   | 300                | -0.88             | OK            | -0.93       | OK             |
| LC0011   | 620                | 1.36              | OK            | 0.71        | OK             |

| Lab Code | Test result (µg/kg) | Variants of Alg A | Hampel method | M estimator |
|----------|--------------------|-------------------|---------------|-------------|
| LC0001   | 1307               | 3.70              | $\$$          | 4.37        | $\$$          |
| LC0002   | 377                | -1.33             | OK            | -1.34       | OK             |
| LC0003   | 162                | -2.49             | $              | -2.67       | $              |
| LC0004   | 1333               | 3.84              | $\$$          | 4.53        | $\$$          |
| LC0005   | 221                | -2.18             | $              | -2.30       | $              |
| LC0007   | 671                | 0.26              | OK            | 0.46        | OK             |
| LC0009   | 646                | 0.12              | OK            | 0.31        | OK             |
| LC0100   | 300                | 0.45              | OK            | 0.68        | OK             |
| LC0100   | 573                | -0.27             | OK            | -0.14       | OK             |
Table 7. Percentage of acceptable z’score of endosulfan sulphate and bifenthrin calculated by the consensus values from participants’ results derived by three robust statistical approaches.

| Test parameter     | Variants of Algorithm A | Hampel method | M estimator |
|--------------------|--------------------------|---------------|-------------|
| Endosulfan sulphate| 50 %                     | 62.5 %        | 50 %        |
| Bifenthrin         | 60 %                     | 60 %          | 60 %        |

All of the three robust statistical approaches can be used to evaluate the PT results and were able to detect the acceptable, warning, and action signals. For endosulfan sulphate, the highest percentage of acceptable z’score was achieved by the Hampel method (table 7). The consensus value of endosulfan sulphate obtained by the Hampel method was also the closest to the homogeneity test results, stability monitoring result, and the assigned value in the real PT scheme. For bifenthrin, three robust statistical approaches give same percentage of satisfactory results (table 7). The consensus value of bifenthrin obtained by the M estimator was the closest to the homogeneity test results, stability monitoring result, and the assigned value in the real PT scheme.

4. Conclusion

The effectiveness of three robust statistical approaches recommended by ISO 13528 for small number of participants: variants of algorithm A, Hampel method, and M estimator were identified. The calculation of consensus value by variants of algorithm A and Hampel method were done without outlier exclusion. Meanwhile, for M estimator, the outlier identified by trimmed average were excluded before calculation of the consensus value. The robust standard deviation of participants’ results ($\sigma_p$) were calculated by Q estimator and the scaled median absolute deviation (MADe). However, the obtained $\sigma_p$ cannot detect the outliers as warning or action signal since the variation was too high. The standard deviation for proficiency assessment ($\sigma_{pl}$) for performance score evaluation was finally taken from the Horwitz equation. All of the obtained consensus values and $\sigma_{pl}$ showed to be able to detect the acceptable, action, and warning signals. The highest effectiveness in achieving the acceptable z’score for endosulfan sulphate was achieved by the consensus value from the Hampel method. Meanwhile, for bifenthrin, all of the robust statistical approaches gave same effectiveness in assessing the performance of participants.

References

[1] International Organization for Standardization 2017 ISO/IEC 17025 - General requirements for the competence of testing and calibration laboratories (Switzerland: ISO)
[2] Yarita T, Otake T, Aoyagi Y, Kuroiwa T, Numata M and Takatsu A 2015 Talanta 132 269–77
[3] International Organization for Standardization 2015 ISO 13528 - Statistical methods for use in proficiency testing by interlaboratory comparison (Switzerland: ISO)
[4] Baldan A, Van Der Veen A M H, Prauβ D, Recknagel A, Boley N, Evans S and Woods D 2001 Accred. Qual. Assur. 6 164–7
[5] Wong S K 2005 Accred. Qual. Assur. 10 409–14
[6] International Organization for Standardization 2016 ISO 17034 - General requirements for the competence of reference material producers (Switzerland: ISO)
[7] International Organization for Standardization 2017 ISO Guide 35 - Reference materials — Guidance for characterization and assessment of homogeneity and stability (Switzerland: ISO)
[8] Aryana N, Arisiawan Y, Putri D and Styarini D 2016 Indones. J. Chem. 16 72–9
[9] International Organization for Standardization 2010 ISO/IEC 17043 - Conformity assessment — General requirements for proficiency testing (Switzerland: ISO)
[10] Aslam M 2020 J. King Saud. Univ. Sci. 1 5
[11] Einax J 1995 Chemometrics in environmental chemistry: applications (New York: Springer)
[12] Rousseeuw P J and Verboven S 2002 Computational Statistics & Data Analysis 40 741-58
[13] Kuselman I and Fajgelj A 2010 Pure Appl. Chem. 82 1099–135
[14] Chakravarty S, Mohanty A, Ghosh B, Tarafdar M, Aggarwal S G and Gupta P K 2014 Mapan - J. Metrol. Soc. India 29 87–95
[15] Coucke W and Soumali M R 2017 Biochem. Medica 27 37–48
[16] Lawn R E, Thompson M, Thompson M and Walker R F 1997 Proficiency Testing in Analytical Chemistry (Cambridge: Royal Society of Chemistry Publishing)