Accreditation and Quality Assurance (2022) 27:349–357
https://doi.org/10.1007/s00769-022-01522-x

GENERAL PAPER

A new methodology for proficiency testing scheme interpretation based on residual analysis

Luiz Henrique da Conceição Leal[1] · Fernando Gustavo Marques Violante[1] · Lucas Junqueira de Carvalho[1] · Bruno Carius Garrido[1] · Eliane Cristina Pires do Rego[1] · Gabriel Fonseca Sarmanho[1] · Werickson Fortunato de Carvalho Rocha[1]

Received: 13 August 2020 / Accepted: 29 August 2022 / Published online: 19 September 2022

Abstract
Proficiency testing schemes by interlaboratory comparisons are used to determine the performance of individual laboratories for specific tests or measurements. The international standard ISO 13528:2015 provides a description of statistical methods used for achieving this goal; one of these methods is the *z* score. The standard allows each participant to choose a measurement method and this can lead to inhomogeneity between participant variance known as heteroscedasticity. The ISO 13528:2015 standard does not mention heteroscedasticity in its statistical procedures. This paper describes a new approach, based on residuals analysis, to assess the performance of an interlaboratory comparison for determining the presence of benzoic acid in orange juice. The results indicate that the conclusions (using *z* scores) and the proposed approach are different for some laboratories. This occurs due to violation of the homoscedasticity assumption. The *z* score procedure does not consider this assumption violation while residual analysis can provide such information. Feasible generalized least squares allow one to deal with non-homoscedasticity.

Keywords Proficiency testing · Residual analysis · Heteroscedasticity · *z* score

Introduction

Proficiency testing (PT) schemes are a powerful tool for the identification of interlaboratory differences that aim to evaluate measurement results performed under similar conditions and assess the technical competence of participants to demonstrate the reliability of their measurement processes [1]. For this purpose, ISO 13528:2015 [1] presents four performance statistics, namely *z* score, *z*′ score, zeta score and E_n score [2]. These statistics also may be used in interlaboratory comparisons. The first three statistics (*z* score, *z*′ score and zeta score) have an underlying assumption of normality of the measurement results. The E_n score does not have this assumption.

In statistics, the *z* score is used for either standardization of data when they were obtained in different orders of magnitude or outlier detection procedure for univariate data set (if the data are normally distributed) [3]. Although ISO 13528:2015 notes the normality assumption inherent in the *z* score, there is no mention about methods to handle heteroscedasticity. In PT schemes, heteroscedasticity is the inhomogeneity variance between laboratory participants. The PT participating laboratories may use different analytical methods to obtain their results [1], which may lead to inhomogeneous variance between the laboratory results. The ISO 13528:2015 standard does not provide statistical procedures to deal with heteroscedasticity.

Different strategies have been proposed to deal with data sets from PT scheme instead of methodologies described in ISO 13528:2015 [4–7]. A modified *z* score was proposed in the Analytical Methods Committee document AMCTB No. 78 to evaluate participants' performance. According to ISO 13528:2015, one way to obtain the assigned value is through the consensus of the participants results. The assigned value and individual results may be correlated if there is a small number of participants (less than 15) [4]. This effect might be reduced with *z* score weighted by the

✉ Werickson Fortunato de Carvalho Rocha
wfrocha@inmetro.gov.br

[1] National Institute of Metrology, Quality and Technology (INMETRO), Av. Nossa Senhora das Graças, 50, Xerém, Duque de Caxias, RJ 25250-020, Brazil

number of laboratories. In other words, applying the finite population correction factor $1-1/n$ to z score [4].

The consensus value may be affected by the presence of outliers and provides inadequate assigned values for PT scheme. Albano et al. [5] have suggested evaluating interlaboratory comparisons based on simple robust statistics mentioned in ISO 13528:2015. The authors suggest use the participants' median results as assigned value and the normalized interquartile range as standard deviation for proficiency assessment [5]. It should be noted that, although it is mentioned in the document, ISO 13528:2015 recommends the use of more sophisticated robust estimators, such as algorithm A described in its annex C. In addition to the ISO 13528:2015 and the AMCTB document, researchers have proposed other metrics to evaluate the laboratories. Arvizu-Torres et al. [6] have proposed a relative quadratic mean error (QME_R) to evaluate the analytical competence of participating laboratories. The QME_R allows participants to evaluate the sources of declared uncertainty and makes reported results comparable to the certified value. This methodology is formulated by estimating the bias with respect to the reference value and that of the uncertainty of each laboratory's result [8]. Thompson and Wood [7] have suggested an alternative score, namely Q-scoring, which is based on the relative bias. The participants' results are evaluated by determining the percentage deviation from the assigned value [9]. The proficiency-testing provider defines this percentage in a discretionary manner.

This work proposes a new procedure to assess the performance of interlaboratory comparisons that are different from the methodologies described in ISO 13528:2015 and the researchers mentioned above. A new methodology to evaluate the PT scheme presented in this paper is derived from the concepts of analysis of variance and multiple comparison tests. Residual analysis allows one to choose the more suitable analysis of variance model.

The analysis of variance (ANOVA) proposed to evaluate PT participating performance has normality and homoscedasticity assumptions of residuals [10]. Homoscedasticity means that there is no statistically significant difference between participating variances. Both assumptions of ANOVA need to be checked by residuals analysis (Fig. 1).

Residuals analysis

Residuals analysis is a statistical tool to investigate normality and homoscedasticity assumptions of linear models such as ANOVA [10]. In the PT scheme, residuals are fitted by subtracting the mean from each reported value for each participating laboratory. Figure 1 presents the steps that need to be followed in residual analysis. The first step consists of verifying residual normality by relating the standardized residuals to the fitted values [10] and Shapiro–Wilk test (Fig. 1a). The Shapiro–Wilk test has a low probability of residuals misclassification [11]. The homoscedasticity assumption is the second step, which is checked by the Koenker–Bassett test (Fig. 1b). This method is applicable even if residuals are not normal [12].

The first two stages of the residual analysis (Fig. 1a, b) allow choosing a suitable model to evaluate PT participating performance. ANOVA should be adopted in cases where the residuals are normal and homoscedastic parametric. The F-test from ANOVA (Fig. 1c) is an omnibus (overall) test which indicates the existence of any difference between participating measurements [10]. If the overall test from ANOVA indicates that there are no differences, then all participating laboratories are classified as acceptable. On the other hand, if the F-test indicates that there are statistically significant differences between laboratory results, it is necessary to use post hoc multiple comparison tests, such as Dunnett's and Fisher's least significant difference (LSD), to identify which results are different.

Following the above-mentioned flowchart, if residuals are normal (Fig. 1a), homoscedastic (Fig. 1b) and F-test from ANOVA (Fig. 1c) indicates differences between reported results the next step is to check the availability of the assigned value.

When the assigned value is available, in this methodology, the participating results are evaluated by analysis of variance with one control where there is a "control treatment" (reference laboratory) and the PT provider is interested in comparing each "treatment" (participating laboratory) with this "control". It is suggested to use the Dunnett's test [10] (Fig. 1d) to verify which participating results differ from the assigned value. If there are differences between the participating results and assigned value, the first result is classified as unacceptable, and the remaining results are considered acceptable.

If the assigned value is unavailable, all laboratories' results are compared two by two to determine which ones are different, using the Fisher's Least Significant Difference (LSD) [10] (Fig. 1e). A specific laboratory is classified as unacceptable if its results differ statistically from all the others, otherwise it is classified as acceptable.

Some PT schemes may not have normal residuals (Fig. 1a) but homoscedastic (Fig. 1b). In this situation, nonparametric methods are used such as Kruskal–Wallis [13] (Fig. 1f). This test provides an overall evaluation of the differences among participating laboratories. All results are classified as acceptable if Kruskal–Wallis test does not identify differences between the laboratory results. When there are differences, one should use the post hoc multiple comparison Dunn's test [14]. Checking the availability of the assigned value according to the flowchart in Fig. 1, the Dunn's test with one control (Fig. 1g) provides conclusions for this value. Laboratory results which differ from the assigned value are

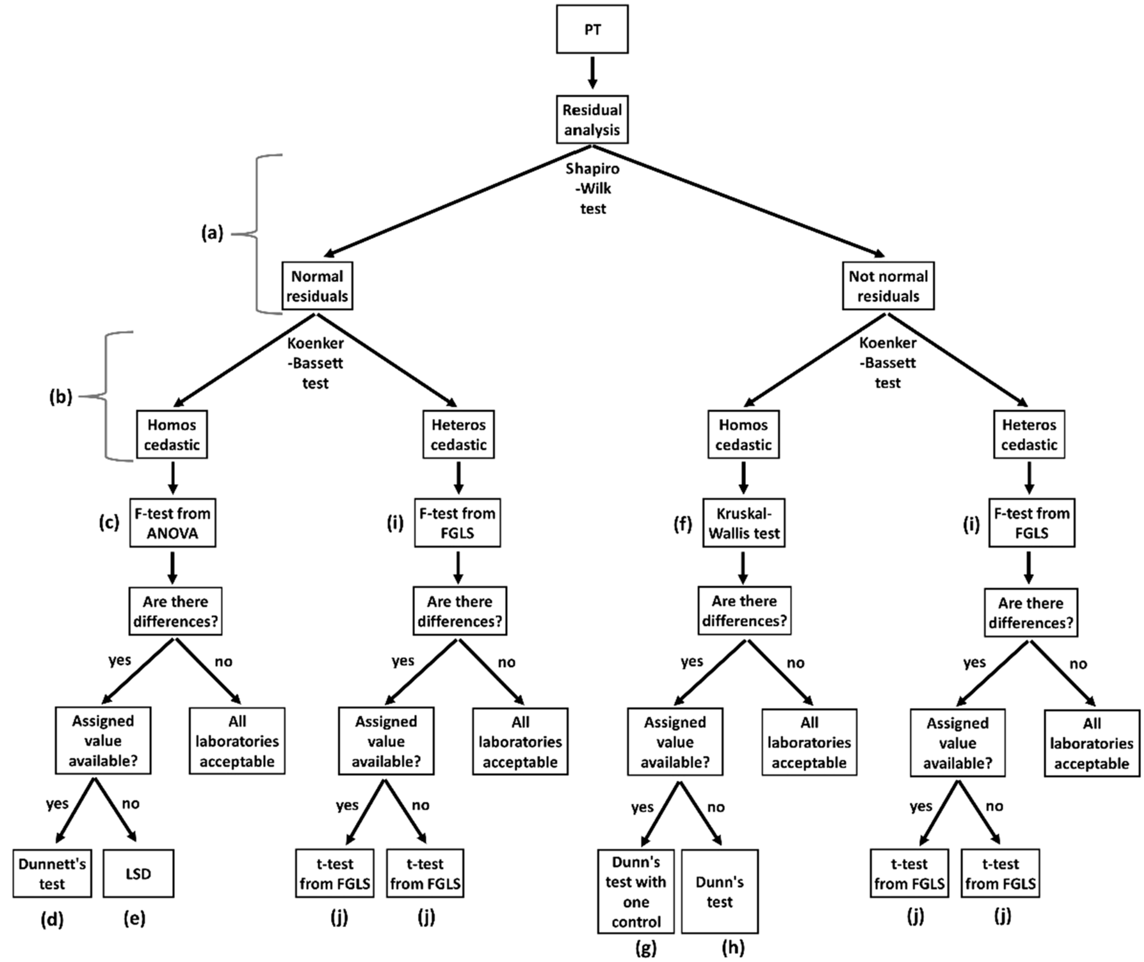

Fig. 1 Flowchart of the proposed residuals-analysis procedure

classified as unacceptable (they are acceptable if they do not differ). In the case where the assigned value is unavailable, the Dunn's test from Fig. 1h provides a two-by-two comparison for all participants. If the reported results by the participants differ from one another, then this laboratory is classified as unacceptable. If its results do not differ, the laboratory is classified as acceptable.

Lastly, regardless of normality assumption (Fig. 1a), if residuals are heteroscedastic (Fig. 1b), then the *F*-test from the feasible generalized least squares (FGLS) model in Fig. 1i is used to evaluate the measurement results. The *F*-test from the FGLS model (Fig. 1i) is an overall test that assesses the performance of the interlaboratory comparison. In the proposed methodology, all participants will have their results considered acceptable if the FGLS *F*-test does not identify differences between the reported values. Nonetheless, if the *F*-test indicates differences between results it becomes necessary to identify which of the laboratory results differ by carrying out multiple comparison tests. The *t*-test from the FGLS model (Fig. 1j) may be used as a multiple-comparison test with one control if an assigned value is available. Laboratory results are considered unacceptable if the *t*-test indicates a difference between the reported and assigned values; however, if there are no differences, then the results are considered acceptable. Moreover, the FGLS *t*-test (Fig. 1j) may be used for pairwise (two by two) comparison. If a specific laboratory result differs from all other participants, then these results are classified as unacceptable, and if there are no differences, then the results are acceptable.

FGLS model

Heteroscedasticity is a common problem present in results reported by PT participating laboratories. The standard ISO 13528:2015 allows each laboratory to choose its own measurement method [1]. This can lead to inhomogeneity in the variability between laboratory results and lead to wrong conclusions regarding which test (*F*-test from ANOVA in Fig. 1c and Kruskal–Wallis test in Fig. 1f) should be used in the proposed methodology. The FGLS model assumes that normality and homoscedasticity criteria are met for the analysis-of-variance model.

The FGLS model provides an approach to estimate the unknown parameters in a linear regression model when either the error variance is not constant (heteroscedastic) or there is a certain degree of correlation between the residuals [15, 16]. The FGLS method estimates are obtained from the matrix equation

$$b = \left(X\hat{\Omega}^{-1}X\right)^{-1}X'\hat{\Omega}^{-1}y$$

where $\hat{\Omega}$ is the diagonal matrix where each matrix element is a suitable weight [15, 16]. The F-test from the FGLS model (Fig. 1i) provides an overall test and, if there are differences, the t-test (Fig. 1j) identifies (two by two) which of them are statistically significant.

When an assigned value is available, the FGLS t-test (Fig. 1j) may be used as a post-hoc multiple-comparisons test with one control to verify the null hypothesis. The laboratory results x_i do not differ from the assigned value x_{PT} ($H_0 : x_i = x_{\mathrm{PT}}$) against the alternative hypothesis that they differ ($H_1 : x_i \neq x_{\mathrm{PT}}$). Failure to reject the null hypothesis (FTR H_0) means that there are no differences between the reported and assigned values, thus the laboratory is classified as acceptable. On the other hand, rejection of the null hypothesis means that laboratory result is classified as unacceptable.

If the assigned value is not available, the FGLS t-test (Fig. 1j) may be used as a pairwise (two by two) comparison to check the null hypothesis, namely the reported results by participant i do not differ from those of participant j ($H_0 : x_i = x_j$) against the alternative hypothesis that they differ ($H_1 : x_i \neq x_j$). Considering m participating laboratories, if the results of a specific laboratory i differ from the $m - 1$ remaining laboratories, then the null hypothesis is rejected for all $m - 1$ comparisons and its results are considered unacceptable, while the results will be acceptable if there is no difference.

For both cases mentioned above, the null hypothesis (H_0) is rejected if $\left|t_{\mathrm{obs}}\right| > t_{n-2}$ where $t_{\mathrm{obs}} = b/s(b)$ and t_{n-2} is the quantile from the t-distribution with $n - 2$ degrees of freedom. The variance–covariance matrix of the parameters $s^2(b)$ is estimated by

$$s^2(b) = \frac{SQG}{n-2}\cdot\left(X^T\hat{\Omega}^{-1}X\right)^{-1}$$

where n is the number of observations and SQG is the generalized sum of squares [15, 16] obtained by the matrix equation $SQG = (Y - Xb)^T\hat{\Omega}^{-1}(Y - Xb)$.

Statistical software

All statistical analyses were performed using the R statistical software, an open-source, free environment program for statistical computing [17]. The FGLS model was built using the package *nlme*. The confidence level of all tests was 95 %.

Materials and methods

Chemicals and reagents

The chemicals and reagents that were used were: pure benzoic acid certified reference material from the National Institute of Standards and Technology (NIST, Gaithersburg, MD, USA), benzoic acid-D5 from Cambridge Isotope Laboratories (Tewksbury, MA, USA), high-purity liquid-chromatography grade methanol and sulfuric acid from Tedia (Farfield, OH, USA), certified reference material of benzoic acid in orange juice from HSA, with a reference value of 766 mg kg^{-1} ± 52 mg kg^{-1} (Outram Road, Singapore) [18].

For the flow-injection analysis with mass-spectrometry-coupled (FIA-MS) method, stock solutions of the analyte (benzoic acid) were prepared in methanol at a concentration of approximately 2500 mg kg^{-1} and stock solutions of the internal standard deuterated benzoic acid were prepared in methanol at a concentration of approximately 1500 mg kg^{-1} [18].

Preparation of the calibration standards

The FIA-MS analyses were performed in a Waters Acquity ultra-pure liquid chromatography I-Class system coupled to a Xevo TQ mass spectrometer. The system was adapted to the FIA by connecting the exit of the injection valve directly to the electrospray probe with a 50 cm PEEK tube with 1.59 mm of external diameter and 1.27 mm of internal diameter [18].

The carrier stream was composed of a 60/40 methanol/water mixture flowing at a flow rate of 0.5 mL min^{-1}. The MS parameters were: electrospray in negative mode, capillary voltage: 2.8 kV; cone voltage: 27 V; desolvation temperature: 300 °C; desolvation gas flow: 650 L h^{-1}. The mass spectrometric analysis was performed in multiple reaction monitoring (MRM) mode for the transition 121 > 77 for the analyte and transition 126 > 82 for the internal standard, both with a collision energy of 10 eV. The injection volume was 1.0 mL and total run time was 0.5 min [18].

PT preparation

For the participating laboratories, the concentration of benzoic acid in the test item was introduced at intervals between

(100 and 1000) mg L^{-1}. The test items were prepared with 3.1 kg of orange juice, which were transferred to a 5 L bottle and a known amount of benzoic acid was weighted and added in this bottle. This solution was stirred at 50 °C for 18 h. Afterwards, the solution was dispensed into a 100 mL amber crimp flask, so that each vial was filled with 30 mL of mixture. The vials were sealed and stored at (-20 ± 3) °C. The refrigerator was filled 200 vials [18].

After packaging, the sample homogeneity was evaluated from 14 randomly selected bottles and analyzed by the FIA-MS. The results were obtained in mg kg^{-1}; however, the laboratories participating in the PT measured their samples in mg L^{-1}, so the sample density had to be determined to convert the results [18].

Sample preparation

An aliquot of approximately 1 g of sample or calibration standard was gravimetrically blended approximately 1:1 with D5-benzoic acid solution. An aliquot of 50 mL was then diluted with 1950 mL of methanol, filtered to 0.22 µm in a syringe filter and transferred to a 2.0 mL vial, which was analyzed in three injections per sub-sample. This technique was validated using the Health Sciences Authority (HSA) certified reference material (CRM) benzoic acid in orange juice [18].

Results and discussion

Laboratory proficiency testing

Proficiency testing with the participation of 13 food and beverage laboratories was carried out to verify the concentration of benzoic acid in orange juice. Table 1 presents the measurement results for all the laboratories.

The homogeneity of the samples was evaluated from 14 randomly selected bottles and analyzed using FIA-MS. From each bottle, a sub-sample of approximately 1 mL (considered to be sufficiently homogeneous) and homogeneity measurements carried out by a reference laboratory (Table 1).

Figure 2 shows the distribution of the data mentioned in Table 1. The data set consists of 39 observations from the 3 replicates provided by the 13 participants plus the 14 observations from the homogeneity study, making a total of 53 observations (Fig. 2). The kernel density obtained from the information mentioned in Table 1 indicates that the data present a left skewed distribution (Fig. 2). The procedure to assess the performance of interlaboratory comparisons is based on the residual data analysis a not in the distribution of data. Thus, the data distribution, shown in Fig. 2, does not interfere with the methodology proposed in this paper to check the performance of laboratories.

Table 1 Results reported by the laboratory participants and summarized reference laboratory result

Lab	Concentration measurements (mg/L)		
	Replicate 1	Replicate 2	Replicate 3
04	125.7	125.7	125.7
27	722.8	721.1	721.4
39	830	806	782
41	529.7	527.8	529.8
44	605	599	602.6
59	593	593.5	592.7
61	802.5	798.9	800.1
63	676.3	675.5	680.2
69	733.2	711.4	711.5
77	632.1	645.7	654.4
83	723.2	722.2	717.3
88	715.9	751.8	761.1
98	711.8	714.6	712.9
Ref lab[a]	Minimum[b]	Mean[c]	Maximum[b]
	706.63	721	731.43

[a] Summarized reference laboratory measurement (homogeneity study)

[b] Minimum and maximum of 14 (bottles) homogeneity measures

[c] Mean of 14 homogeneity measures (assigned value)

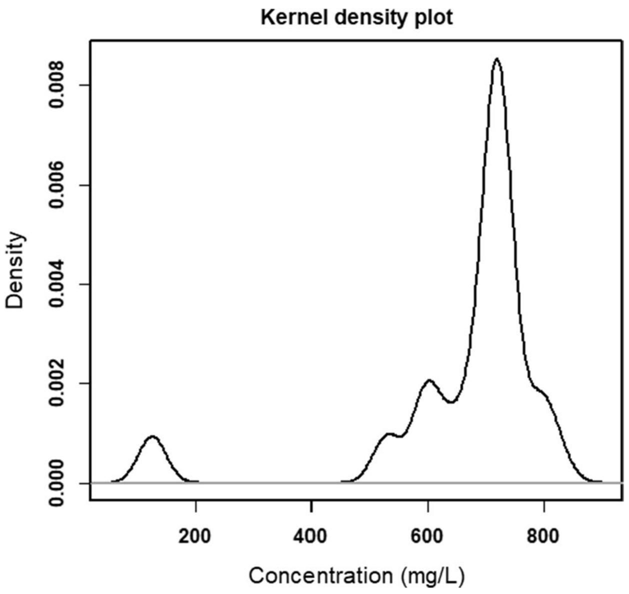

Fig. 2 Gaussian kernel density plot obtained from data set mentioned in Table 1

The next step in the proficiency test was to define the assigned value. The assigned value $x_{PT} = 721$ mg/L was the average value of the results obtained in the homogeneity study (reference laboratory). The uncertainty of the assigned value was calculated from the homogeneity test data set, while also considering the sources of uncertainty from the analysis method.

The standard deviation for the proficiency assessment was $\sigma_{\mathrm{PT}} = 43.1$ mg/L, which was calculated by a general model according to ISO 13528:2015. The model adopted for this purpose was based on the Horwitz curve which expresses interlaboratory precision in terms of a standard deviation of reproducibility [1]. For this PT scheme, the Horwitz curve equation was

$$\sigma_{\mathrm{PT}} = \left(\left(0.02\left(k \cdot 10^{-6}\right)^{0.8495}\right)10^{6}\right)d$$

where k is the average value from homogeneity study in mg/kg and d is the density ($d = 1.04$ g/cm^3).

Residuals analysis

For the proposed methodology, the PT scheme is considered as a one-way fixed effects analysis-of-variance problem where the participating and reference laboratories represent levels of a single factor. Table 1 values represent the information taken under factor level. From these values, the appropriate model may be chosen following the flowchart scheme described in Fig. 1.

According to the proposed methodology, the first model assumption which is checked is the residuals normality (Fig. 1a). This assumption is rejected by the Shapiro–Wilk test (p value of 5.2×10^{-5}) considering a 5 % significance level. The second assumption is the homoscedasticity (Fig. 1b). The residual plot in Fig. 3a shows evidence of heteroscedasticity ("funnel shaped"). It is expected that the residuals plot will not have any kind of pattern [10, 12, 13]. The assumption of normality homoscedasticity is rejected by the Koenker–Bassett test at 5 % significance level (p value of 0.03).

Both assumptions of residual normality and homoscedasticity are not met from the data showed in Table 1. The proposed flowchart in Fig. 1 indicates the need to fit the FGLS model to the data from Table 1 (Fig. 1i). Figure 3b shows that the FGLS model provided a better residual pattern, which is not funnel shaped. The residuals for the FGLS model are normal (p value of 0.93) and homoscedastic (p value of 0.71) considering a 5 % significance level. Thus, FGLS is a more appropriate model to assess the laboratory performance.

FGLS model versus *z* score

The assigned value obtained from the homogeneity study and the standard deviation for proficiency assessment calculated by the Horwitz curve allows estimation of z score z_i which is interpreted as follows: $\left|z_i\right| \leq 2$ the laboratory result is considered acceptable, $\left|z_i\right| \geq 3$ the result is considered unacceptable (or action signal) and $2 < \left|z_i\right| < 3$ is considered a 'warning signal' [1]. The z score and its interpretation are provided in Table 2. Nine laboratories were classified as acceptable, two as warning signals, and two as unacceptable (Table 2).

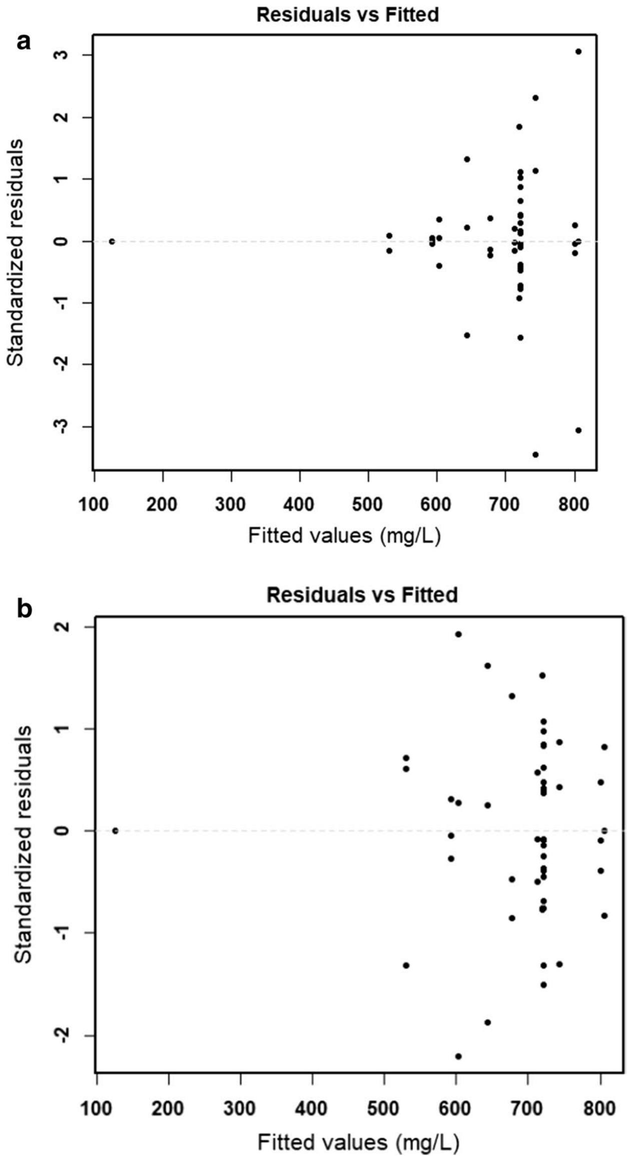

Fig. 3 Variation of the **a** standardized residuals with respect to the fitted values for the linear model and **b** FGLS model

The residual analysis indicates that FGLS is a suitable model for dealing with violations of analysis of variance. The F-test from the FGLS (Fig. 1i) model indicates that there are differences between the laboratory concentrations (p value $< 10^{-4}$) at 5 % significance level. Thus, t-test from FGLS model (Fig. 1j) should be used to check which laboratory results differ statistically from the assigned value.

From the FGLS model t-test (Fig. 1j), a p value greater than or equal to the pre-established significance level (5 %

Table 2 Comparison of results using the FGLS model and *z* score procedure

Lab	*z* score procedure		FGLS model		
	z score	Interpretation	*t*-value	*p* value[a]	Interpretation
04	-13.81	Unacceptable	-231.36	3.4×10^{-61}	Unacceptable
41	-4.45	Unacceptable	-72.87	5.5×10^{-42}	Unacceptable
59	-2.97	Warning signal	-47.53	5.3×10^{-35}	Unacceptable
44	-2.76	Warning signal	-43.94	8.1×10^{-34}	Unacceptable
77	-1.79	Acceptable	-17.13	6.6×10^{-19}	Unacceptable
63	-1.01	Acceptable	-15.3	2.7×10^{-17}	Unacceptable
98	-0.18	Acceptable	-2.68	0.05	Acceptable
69	-0.05	Acceptable	-0.4	1	Acceptable
83	-0.002	Acceptable	-0.07	1	Acceptable
27	0.02	Acceptable	0.22	1	Acceptable
88	0.51	Acceptable	1.78	0.34	Acceptable
61	1.85	Acceptable	22.54	4.7×10^{-23}	Unacceptable
39	1.97	Acceptable	5	6.5×10^{-05}	Unacceptable

[a] *p* values were adjusted by the Benjamini–Yekutieli procedure

for example) means that the laboratory measures do not differ statistically from the assigned value and their results are classified as acceptable. Table 2 shows that five laboratories were classified as acceptable (for a 5 % significance level). The remaining laboratories had a *p* value of less than 5 %, so these results were considered unacceptable.

One difficult issue to be addressed in multiple comparison tests is the potential accumulation of decision errors. As the number of tests increases, the probability of making at least one type of classification error (i.e., where a laboratory is classified as unacceptable when, indeed, it is acceptable) also increases. Benjamini and Hochberg [19] and Benjamini–Yekutieli [20] have proposed a procedure to deal with this issue. There are no restrictions for Benjamini–Yekutieli procedure [20] so *p* values from the FGLS model are adjusted by this method (Table 2).

Table 2 shows that the FGLS model and *z* score provide the same conclusion for seven participating laboratories. For the rest of the participants, the conclusions were different. Laboratories 77, 63, 61, and 39 were classified as acceptable by *z* score and unacceptable by the FGLS model. Laboratories 59 and 44 were classified as a warning signal by the *z* score but the FGLS model considered the results unacceptable.

Methodology validation

The dataset extracted from Banzatto and Kronka [21] was used to validate the proposed methodology. The data have heteroscedastic residuals [21] and were adapted to be considered from a hypothetical proficiency test (Table 3).

The data show right skewed distribution. Through the Anderson–Darling test, it was concluded that the data are compatible with the gamma probability distribution with parameters shape 0.87 and rate 10^{-3} (*p* value 0.47). This feature does not influence the methodology described in Fig. 1.

In the example shown in Table 3, a proficiency testing was considered in which each of the 4 participants provided 6 measurements of a given analyte. Additionally, it was considered that the proficiency testing provider performed a homogeneity study on 6 samples (Ref lab in Table 3).

Finally, the assigned value $x_{\mathrm{PT}} = 527.17$ and the standard deviation for the proficiency assessment $\sigma_{\mathrm{PT}} = 227.04$ were obtained using algorithm A from ISO 13528:2015 [1].

Following the flowchart proposed in Fig. 1, both assumptions of residual normality (*p* value of 6.7×10^{-3}) and homoscedasticity (*p* value of 1.2×10^{-3}) are rejected considering a 5 % significance level. The residual plot shows evidence of heteroscedasticity due “funnel shaped” (Fig. 4a).

The FGLS model provided a better residual pattern (no funnel shaped in Fig. 4b) and the residuals are normal (*p* value of 0.11) and homoscedastic (*p* value of 0.59) considering a 5 % significance level.

According to the proposed methodology, the FGLS more is a more suitable to assess the laboratory performance. The *F*-test (Fig. 1i) shows that there are differences between the laboratory’s results (p value $< 10^{-4}$) at 5 % significance level. The *t*-test (Fig. 1j) indicates that all laboratory’s results differ statistically from the assigned value (*p* value of less than 0.05). Thus, all of them are considered unacceptable (Table 4).

To compare the proposed methodology with the performance statistics described in the ISO 13528:2015 standard,

Table 3 Hypothetical proficiency testing

Lab	Measurements (number of aphids)					
	Rep. 1	Rep. 2	Rep. 3	Rep. 4	Rep. 5	Rep. 6
A	2370	1687	2592	2283	2910	3020
B	1282	1527	871	1025	825	920
C	173	127	132	150	129	227
D	193	71	82	62	96	44
Ref lab[a]	562	321	636	317	485	842

[a] Measurements by the proficiency testing provider (homogeneity study)

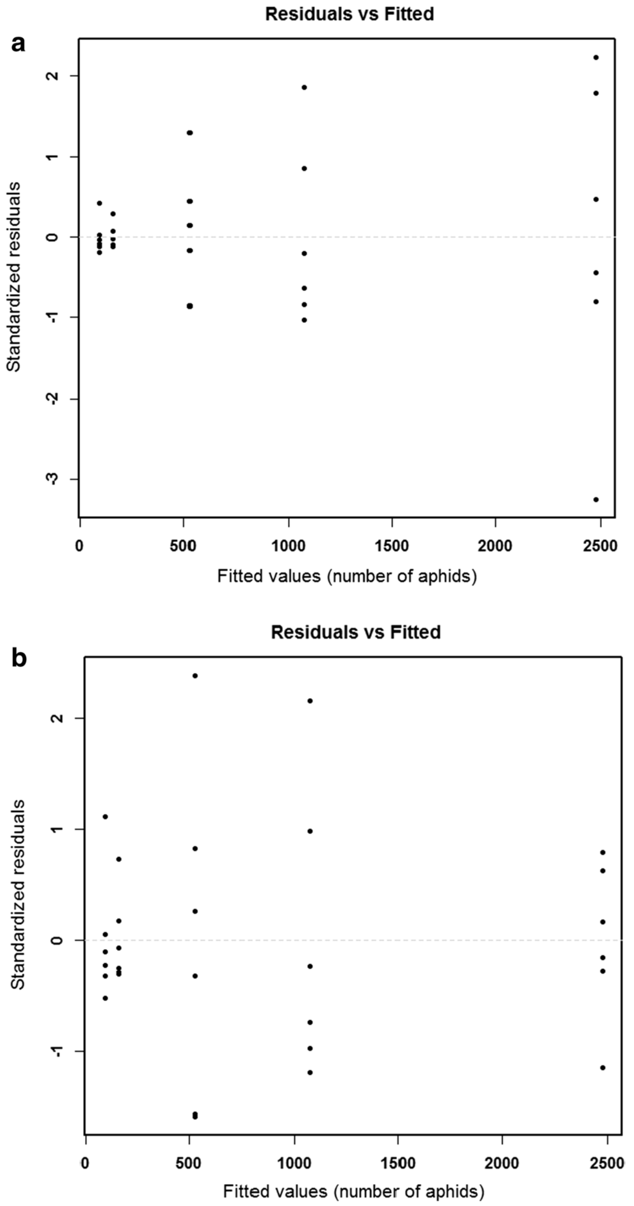

Fig. 4 Variation of the **a** standardized residuals with respect to the fitted values for the linear model and **b** FGLS model

Table 4 Comparison of results using the FGLS model and *z* score procedure

Lab	*z* score procedure		FGLS model		
	z score	Interpretation	*t*-value	*p* value[a]	Interpretation
04	8.59	Unacceptable	6.81	2.39×10^{-06}	Unacceptable
41	2.41	Warning signal	5.41	2.72×10^{-05}	Unacceptable
59	− 1.63	Acceptable	− 5.55	2.49×10^{-05}	Unacceptable
44	− 1.92	Acceptable	− 6.65	2.39×10^{-06}	Unacceptable

[a] *p* values were adjusted by the Benjamini–Yekutieli procedure

the *z* score was calculated, whose assigned value and standard deviation for proficiency assessment were obtained by algorithm A. Two laboratories were classified as acceptable, one as warning signal, and one as unacceptable (Table 4).

The preliminary results discussed in this section (real and hypothetical proficiency testing) allow us to infer that the proposed methodology is more sensitive to detecting differences among participating results and assigned value when compared to *z* score.

Conclusion

A new approach is presented to assess the laboratory performance of proficiency testing by interlaboratory comparisons based on residual analysis. This methodology considers aspects such as assumptions of normality and homoscedasticity. It is worth mentioning that the latter assumption is not considered by the *z* score procedure describe in ISO 13528:2015. The proposed methodology derives concepts from the one-way, fixed-effects analysis of variance followed by post-hoc multiple comparison tests.

Proficiency testing for estimating benzoic acid concentration in orange juice was conducted by National Institute of Metrology, Quality and Technology (INMETRO). Homogeneity and stability studies were performed beforehand to ensure that the samples that were sent to participating laboratories were fit for the interlaboratory comparison. Thirteen food and beverage analysis laboratories participated in the PT study, in which their analytical competence was evaluated by *z* score, as described in ISO 13528:2015. This performance standard compares laboratory results against assigned value of x_{PT}, which was obtained from homogeneity studies. The standard deviation for proficiency assessment was computed by the Horwitz equation.

The proposed procedure to evaluate PT laboratory results followed the flowchart scheme suggested in Fig. 1. Assumptions of normality) and homoscedasticity (were checked and the FGLS model proved to be more suitable for the data analysis. The *F*-test from the FGLS model (Fig. 1i) indicated statistically significant differences between participating laboratory results and the assigned value. The *t*-test showed which of these results differed from x_{PT}.

The FGLS model and *z* score provided different conclusions for six participating. Four of them were classified as acceptable by the *z* score and unacceptable by the FGLS model. Two participating were classified as a warning signal by the *z* score and unacceptable by the FGLS model.

The *z* score procedure did not take into consideration heteroscedasticity between participating laboratories while the FGLS model provided tools to deal with laboratory heteroscedasticity. Due to *z* score limitations, the residual analysis scheme may be considered an alternative to assess the performance of the participants. For this proposed methodology, when an assigned value is available, it is recommended to use more observations for the laboratory which provides the assigned value, i.e., “control treatment” (n_p), than for the other laboratory, i.e., “treatments” (n). For this case, the ratio n_p/n should be chosen to approximately equal the square root of the total number of “treatments” [10].

The validation of the methodology was carried out through a set of data which had previously been known about heteroscedasticity. In general, it was observed that in the proposed methodology more laboratories were classified as unacceptable when compared to z score. This provides evidence that heteroscedasticity influences performance assessment in interlaboratory comparisons.

References

1. ISO 13528 (2015) Statistical methods for use in proficiency testing by interlaboratory comparison. ISO, Geneva
2. Wong SK (2007) A comparison of performance statistics for proficiency testing programmes. Accred Qual Assur 12:59–66
3. Rousseeuw PJ, Hubert M (2011) Robust statistics for outlier detection. WIREs DMKD 1:73–79. https://doi.org/10.1002/widm.2
4. AMC TB 78 (2017) Proficiency testing of sampling. Royal Society of Chemistry, AMC Technical Briefs 78
5. Albano FM, Rodrigues M, Albano JF (2007) Garantia da qualidade analítica através de programas de comparação interlaboratorial. VII SEPROSUL—Semana de Engenharia de Produção Sul-Americana, Universidad de la República, Salto
6. Arvizu-Torres R, Perez-Castorena A, Salas-Tellez JA, Mitani-Nakanishi Y (2001) Biological and environmental reference materials in CENAM. Fresenius J Anal Chem 370:156–159
7. Thompson M, Wood R (1993) The international harmonized protocol for the proficiency testing of (chemical) analytical laboratories. Pure Appl Chem 65:2123–2144
8. Mitani Y, Lara-Manzano JV, Rodrigues-Lopez A (2008) Proficiency testing scheme for the harmonization and comparability of analytical measurements. Accred Qual Assur 13:421–426
9. Ellison SLR, Barwick V, Farrant TJD (2009) Practical statistics for the analytical scientist: a bench guide. RSC Publishing, London
10. Montgomery DC (2001) Design and analysis of experiment. Wiley, Tempe
11. Yap BM, Sim H (2011) Comparisons of various types of normality tests. J Stat Comput Simul 81:2141–2155. https://doi.org/10.1080/00949655.2010.520163
12. Gujarati DN (2004) Basic econometrics. McGraw-Hill (India), Noida
13. Kutner MH, Nachtsheim CJ, Neter J, Li W (2005) Applied linear statistical models. McGraw-Hill, Atlanta
14. Dunn OJ (1964) Multiple comparisons using rank sums. Technometrics 6:241–252
15. Greene WH (2003) Econometric analysis. Prentice Hall, New Jersey
16. Florens J, Marimoutou V, Péguin-Feissolle A (2007) Econometric modeling and inference. Cambridge University Press, Cambridge
17. R Core Team (2020) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna
18. Carvalho LJ, Rego ECP, Garrido BC (2016) Quantification of benzoic acid in beverages: the evaluation and validation of direct measurement techniques using mass spectrometry. Anal Methods 8:2955–2960
19. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc 57:289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x
20. Benjamini Y, Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. Ann Stat 29:1165–1188. https://doi.org/10.1214/aos/1013699998
21. Banzatto DA, Kronka SN (2006) Experimentação Agrícola. FUNEP, São Paulo

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.