Is the \( z \) score sufficient to assess participants’ performance in proficiency testing? The hidden corrective action

Fernando Cordeiro\(^1\) · Hendrik Emons\(^1\) · Piotr Robouch\(^1\)

Received: 6 May 2021 / Accepted: 3 February 2022 / Published online: 18 March 2022
© The Author(s) 2022

Abstract
Proficiency testing providers, accreditation bodies and testing laboratories should be aware that a laboratory participating in a proficiency testing round might have reported a biased result despite a satisfactory performance indicated by an assessment using uniquely the \( z \) score. A complementary performance evaluation, based on the \( \zeta \) score and the assessment of the measurement uncertainty, is therefore highly recommended. This work presents an intuitive graphical tool (the Naji2 plot) that combines \( z \) and \( \zeta \) scores together with the reported measurement uncertainties. This tool allows a comprehensive assessment of the laboratory performance and enables to identify the need for corrective actions. The concerned laboratory should then perform a root cause analysis and investigate their bias and/or their measurement uncertainty evaluation.

Keywords
Proficiency testing · Performance assessment · Zeta score · Measurement uncertainty · Naji plot

Introduction
The international standard ISO/IEC 17025:2017 [1] requires from testing laboratories to use validated measurement procedures, certified reference materials (when available) or appropriate quality control materials and to participate in interlaboratory comparisons such as proficiency testing (PT) rounds, to ensure and/or demonstrate the validity of their results. In addition, every laboratory is expected to estimate the measurement uncertainty (MU) associated with the reported result in order to comply with the basic demand that “a measurement result is generally expressed as a single measured quantity value and a measurement uncertainty” [2]. Similarly, ISO/IEC 17043:2010 [3] is demanding from PT providers, among other requirements, to estimate the associated measurement uncertainties when determining the assigned values.

Most of the European Union official laboratories performing control activities in the Single Market, for customs or environmental surveillance, are accredited according to ISO/IEC 17025 [1]. During the accreditation audits, the technical assessors check the compliance of such testing against the requirements set in the standard. They are often relying on the laboratory performance demonstrated in the frame of regular participation in PT schemes.

While all PT providers require laboratories to report a value for each investigated measurand, only a few of them request participants to report the associated MU. Hence, the common proof of satisfactory performance relies only on the \( z \) score. This score may confirm that the method of analysis applied is fit for the intended purpose, but it will not allow concluding on the presence or absence of a potential bias of the value reported, or on the appropriateness of the estimated MU associated with that value. Hence, laboratories may miss the hidden corrective action needed.

Robouch et al. presented a simple graphical tool for the evaluation of PT results at an international workshop on “Data analysis of key-comparisons” in 2003 [4]. The original Naji plot allowed the simultaneous assessment of various performance criteria of PT participants. Two alternative visualisation tools have been subsequently developed, namely the PomPlot [5] and the Kiri plot [6].

This paper revisits the approach of the original Naji plot concept, combining several assessment criteria recommended by ISO/IEC 17043 and ISO 13528 [3, 7]. It introduces several modifications to obtain the “Naji2 plot”, a simpler and more intuitive tool. The main goal of this work is to raise the awareness of PT providers on the importance of requesting from the participants the analytical results.
together with the respective measurement uncertainties, in order to deliver a comprehensive assessment of the laboratory performance.

### Performance scoring

ISO/IEC 17043:2010 [3] and ISO 13528:2015 [7] recommend, among others, the following three performance scores to assess the reported results in the frame of a proficiency testing round.

The relative deviation ($D_{rel}$) used to estimate the deviation of the reported result ($x_i$) from the assigned value ($x_{pt}$) is calculated as (expressed as a percentage of $x_{pt}$):

$$D_{rel,i} = \frac{100(x_i - x_{pt})}{x_{pt}} \tag{1}$$

Similarly, the deviation may be normalised using the standard deviation for proficiency assessment ($\sigma_{pt}$) to calculate the $z$ score:

$$z_i = \frac{x_i - x_{pt}}{\sigma_{pt}} \tag{2}$$

Alternatively, the deviation can be normalised combining the uncertainty reported by the laboratory ($u(x_i)$) and the uncertainty associated with the assigned value ($u(x_{pt})$) to compute the zeta ($\zeta$) score:

$$\zeta_i = \frac{x_i - x_{pt}}{\sqrt{u(x_i)^2 + u(x_{pt})^2}} \tag{3}$$

According to ISO 13528 [7], a laboratory performance is considered as satisfactory (SP), questionable (QP) or unsatisfactory (UP), when the absolute value of the $z$ or $\zeta$ scores are: $|z| < 1$; $1 \leq |z| < 2$; and $|z| \geq 2$, respectively.

In addition, ISO 13528 [7] suggests to assess the reported measurement uncertainties (MU). The systematic approach implemented by the European Union Reference Laboratories (EURLs) managed by the Joint Research Centre is described in the PT report to participants, see for instance [8, 9]. This was done by comparing the reported relative MU ($u_{rel}(x_{pt}) = u(x_i)/x_{pt}$) to the relative uncertainty of the assigned value ($u_{rel}(x_{pt}) = u(x_{pt})/x_{pt}$) and to the relative standard deviation for proficiency assessment ($\sigma_{pt,rel} = \sigma_{pt}/x_{pt}$). Hence, MU is considered as realistic (case “a”), probably underestimated (case “b”) or likely to be overestimated (case “c”) when: “a” $u_{rel}(x_{pt}) \leq u_{rel}(x_i) \leq \sigma_{pt,rel}$; “b” $u_{rel}(x_i) < u_{rel}(x_{pt})$; and “c” $u_{rel}(x_i) > \sigma_{pt,rel}$, respectively.

Knowing that there are three performance evaluation categories (satisfactory, questionable and unsatisfactory) and three uncertainty evaluation cases (underestimated, realistic and overestimated) for each of the three assessments mentioned above ($z$ score, $\zeta$ score and MU), a total of 27 theoretical possibilities could be identified as shown in Table 1. However, a few possibilities marked by an asterisk in Table 1 may be unrealistic. For example, it is not possible to have simultaneously a satisfactory performance according to $z$, an unsatisfactory performance according to $\zeta$ and an overestimated MU (see possibility “I” in Table 1).

### The Naji2 Plot: a graphical solution

Assuming that the $\zeta$ score is smaller than a certain performance limit $P_L$ (equal to 2 or 3, as set in [3, 7]), and combining Eqs. 2 and 3 one can derive the following relations:

$$P_L \geq \zeta = \frac{\sigma_{pt} \cdot z}{\sqrt{u(x_i)^2 + u(x_{pt})^2}} \tag{4}$$

Equivalent to:

---

**Table 1** The 27 theoretical possibilities related to $z$ score, $\zeta$ score and the measurement uncertainty evaluation (MU)

| $z$ score | $\zeta$ score | MU | Possibility or Naji2 plot areas |
|-----------|---------------|----|---------------------------------|
| Satisfactory | Satisfactory | Underestimated | A |
| Satisfactory | Satisfactory | Realistic | B |
| Questionable | Underestimated | C |
| Questionable | Underestimated | E |
| Unsatisfactory | Underestimated | F |
| Questionable | Satisfactory | G |
| Questionable | Satisfactory | H |
| Unsatisfactory | Underestimated | I* |
| Questionable | Underestimated | J* |
| Questionable | Underestimated | K |
| Questionable | Underestimated | L |
| Unsatisfactory | Underestimated | M* |
| Questionable | Underestimated | N |
| Questionable | Underestimated | O |
| Questionable | Underestimated | P |
| Questionable | Underestimated | Q |
| Unsatisfactory | Underestimated | R |
| Unsatisfactory | Underestimated | S* |
| Questionable | Underestimated | T* |
| Questionable | Underestimated | U |
| Questionable | Underestimated | V* |
| Questionable | Underestimated | W |
| Unsatisfactory | Underestimated | X |
| Unsatisfactory | Underestimated | Y |
| Unsatisfactory | Underestimated | Z |
| Unsatisfactory | Underestimated | AA |

The letters in the last column represent various possibilities, later denoted as Naji2 plot areas. The asterisk marks unrealistic possibilities.
\[ P_t^2 \left[ u(x_i) - u(x_{pt}) \right]^2 \geq \sigma_{pt}^2 \]  

Further transformed into:

\[ \frac{u(x_i)}{\sigma_{pt}}^2 \geq \frac{z^2}{P_t^2} - \frac{u(x_{pt})}{\sigma_{pt}}^2 \]  

Equation 6 describes the original Naji plot parabolas presented by Robouch et al. [4], when plotting \( u(x_i)/\sigma_{pt} \) (y-axis) as a function of the \( z \) score (x-axis). These graphs were used in several publications related to proficiency testing rounds organised in a broad variety of fields [10–17]. Furthermore, this graphical representation is currently implemented in the commercial software “PROLab™” developed by Quodata GmbH (Dresden, Germany) for the interpretation of results reported in the frame of interlaboratory comparisons [18].

An alternative transformation of Eq. 5 is further investigated:

\[ u(x_i) \geq \left( \frac{\sigma_{pt} \cdot z}{P_t} \right)^2 - u(x_{pt})^2 \]  

Equivalent to:

\[ u(x_i) \geq \sqrt{\left( \frac{\sigma_{pt} z - u(x_{pt})}{P_t} \right)^2 + u(x_{pt})^2} \]  

Equation 8 describes two sets of hyperbolas obtained when plotting \( u(x_i) \) versus the \( z \) score, for the two acceptance criteria \( (P_t = 2 \) or \( 3) \). This graphical representation constitutes the “Naji2 plot”. Each hyperbola presents the following characteristics: (i) it is symmetric around the y-axis; (ii) it is always positive or equal to zero when \( z = \pm P_t u(x_{pt})/\sigma_{pt} \); and (iii) it becomes steeper with increasing \( |z| \) to reach the linear asymptote \( (y = \pm \sigma_{pt} z/P_t) \). However, Eq. 8 is not valid for \( z \) values ranging between \( -P_t u(x_{pt})/\sigma_{pt} \) and \( +P_t u(x_{pt})/\sigma_{pt} \), because they would lead to a negative value under the square root.

### Assessment of measurement uncertainties

In order to include the assessment of measurement uncertainties (cases “a”, “b”, “c”) in the Naji2 plots, the relation between the relative standard deviation \( u_{rel}(x_i) \) and the \( z \) score is described hereafter:

\[ u(x_i) = u_{rel}(x_i) x_i = u_{rel}(x_i) (\sigma_{pt} z + x_{pt}) \]  

Equation 9 is a linear relation of \( u(x_i) \) as a function of the \( z \) score. Two specific lines delimit the range of “realistic” measurement uncertainties (case “a”), when \( u_{rel}(x_i) \) is equal to \( u_{rel}(x_{pt}) \) or to \( \sigma_{pt,rel} \). While each line crosses the y-axis (\( z = 0 \)) at \( u(x_i) = u(x_{pt}) \) or at \( u(x_i) = \sigma_{pt} \), they both cross the x-axis \( (u(x_i) = 0) \) at \( z = -x_{pt}/\sigma_{pt} \).

#### The bias boundaries

No bias can be identified (null hypothesis \( H_0 \)) when the distribution of a measurement result and the interval of the assigned value and its expanded uncertainty (both assumed to follow a normal or approximately normal distributions) overlap with a probability (\( \alpha \)) of rejecting a true \( H_0 \) of 5%, and a probability (\( \beta \)) of accepting a false \( H_0 \) of 5% [19]. Consequently, a value \( (x_i) \) lower than the assigned value \( (x_{pt}) \) would be negatively biased \( (z < 0) \):

\[ x_i + 1.64u(x_i) < x_{pt} - 1.64u(x_{pt}) \]  

where 1.64 is the inverse of the standard normal cumulative distribution for a probability of 95% (using the built-in spreadsheet function “NORM.S.INV(0.95)”).

Combining Eqs. 2 and 10, one derives the following linear relation between \( u(x_i) \) and \( z \):

\[ u(x_i) < (-\sigma_{pt} z/1.64) - u(x_{pt}) \]  

A similar linear relation is obtained in the case of “positive range” \( (z > 0) \):

\[ u(x_i) < (\sigma_{pt} z/1.64) - u(x_{pt}) \]  

The two lines defined by Eqs. 11 and 12 delimit the “upper range” of the bias boundaries. The ranges delimited by these lines are similar to those defined by the two green hyperbolas \( (|z| > 2) \) and comply with the criteria for bias set by ISO Guide 33 [20]. Such large \( z \) scores may be caused either by a too large numerator, indicating a significant deviation (bias) of the reported value from the assigned value \( (x_i - x_{pt}) \); or by a too small denominator \( \sqrt{u(x_i)^2 + u(x_{pt})^2} \) caused by an underestimated reported MU, assuming that the uncertainty associated with the assigned value is realistic. The concerned laboratory should then perform a root cause analysis and investigate their bias and/or their measurement uncertainty evaluation, in order to resolve the identified poor performance expressed as a \( \zeta \) score.

Despite the fact that many PT results are indeed corresponding to a satisfactory performance when expressed uniquely by the \( z \) score \( (|z| \leq 2) \), some of them are located below the bias boundaries \( (|z| > 2) \) and may require corrective actions by the laboratory.
A hypothetical case study

In order to demonstrate the Naji2 plot concept, a hypothetical proficiency test round attended by 40 participants was constructed. A PT provider processed a commercially available food commodity to produce a test material with adequate homogeneity and stability, according to the recommendations of ISO/IEC 17043 [3]. The PT provider also determined the total mass fraction of a specific analyte \((x_{pt})\) of 100 mg kg\(^{-1}\) and the associated standard uncertainty \((u(x_{pt}))\) of 3 mg kg\(^{-1}\). In addition, a \(\sigma_{pt}\) of 10 mg kg\(^{-1}\) was set to assess the measurement capabilities of the laboratories to comply with some legal requirements. Since \(u(x_{pt}) \leq 0.3 \sigma_{pt}\), the test item was considered fit-for-purpose and the \(z\) score could be applied for performance assessment.

Figure 1 presents the Naji2 plot generated based on the above-mentioned predefined criteria \((x_{pt}, u(x_{pt}),\) and \(\sigma_{pt}\)). This plot consists of:

(i) \(u(x_i)\) (y-axis) as a function of the \(z\) score (x-axis) having:
(ii) four vertical lines at \(|z| = 2\) and \(|z| = 3\);
(iii) two sets of hyperbolas for \(|z| = 2\) and \(|z| = 3\) (Eq. 8);
(iv) two lines representing \(u_{ref}(x_i) = u_{ref}(x_{pt})\) and \(u_{ref}(x_i) = \sigma_{pt,rel}\) (Eq. 9, see dotted and dashed lines in Fig. 1); and
(v) two bias boundaries (Eqs. 11 and 12, see double blue lines in Fig. 1).

The four vertical lines \((|z| = 2\) and \(|z| = 3\)), the four hyperbolas \((|z| = 2\) and \(|z| = 3\)) and the two straight lines \((u_{ref}(x_i) = u_{ref}(x_{pt})\) or \(u_{ref}(x_i) = \sigma_{pt,rel}\)) delimit all the possible Naji2 plot areas identified by the letters “A” to “AA” in Table 1 and Fig. 1. The areas denoted by the letters with an asterisk in Table 1 could not be represented in Fig. 1, since they represent unrealistic possibilities.

The spreadsheet function “NORMINV(RAND(),m,u)” was used to generate two sets of 40 values, where \(m\) is the mean value and \(u\) the standard uncertainty. The first set of data simulated the reported results \((x_i)\) normally distributed around the assigned value \((m = 100\) mg kg\(^{-1}\)) with \(u = 25\) mg kg\(^{-1}\) (see Table 2, 2nd column). The second set simulated a broad range of reported standard measurement uncertainties \((u(x_i))\), with \(m = 6\) mg kg\(^{-1}\) and \(u = 4\) mg kg\(^{-1}\) (see Table 2, 4th column). The standard uncertainties were multiplied by a coverage factor \((k = 2)\) to derive the expanded uncertainties shown in Fig. 2.

Table 2 summarises the synthetic data set, the relative standard uncertainty \(u_{rel}(x_i)\) (to be compared to \(u_{rel}(x_{pt})\) and to \(\sigma_{pt,rel}\)) and the outcome of the four assessments \((D_{\%},z\) score, \(\zeta\) score, and MU). Similarly, Fig. 2 presents the data set in a particular order, sorted by the reported result in increasing order first, then by the performance categories (SP, QP and UP), expressed as \(z\) scores and \(\zeta\) scores. This allows the identification of 11 laboratories (from L25 to L38, see x-axis of Fig. 2) having “satisfactory” performances expressed as \(z\) scores but “questionable” or even “unsatisfactory” performances when expressed as \(\zeta\) scores.

Figure 3 presents the Naji2 plot applied to the synthetic data set (Table 2). This graphical presentation allows an intuitive and direct assessment of every data point, related to all the assessment criteria under investigation, namely the \(z\) score, the \(\zeta\) score or the bias boundaries, and the appropriateness of the reported MU. However, this graphical presentation is only accessible to PT providers collecting measurand values including measurement uncertainties from their participants.

One can easily notice that: (i) most of the results are in the central part of the plot, representing satisfactory \(z\) and \(\zeta\) scores; (ii) seven “overestimated” (case “c”) and (iii) seven seemingly “underestimated” (case “b”) MU were reported, while four laboratories did not report their MU \((u(x_i) = 0\) for L02, L04, L13 and L27). Seventeen results are below the “bias boundaries”, of which six are significantly biased with \(|z| > 2\) and \(|\zeta| > 3\) (L03, L04, L09, L14, L27, L33). The remaining eleven laboratories have “satisfactory” \(z\) scores, and “questionable” or “unsatisfactory” \(\zeta\) scores (see empty circles in Fig. 3). These laboratories may consider two types of improvement actions: either an increase in their MU
Table 2: Synthetic data set of 40 laboratories having reported measurement results ($x_i$) and expanded uncertainties ($U_i$). The standard MU ($u(x_i)$) and the relative standard uncertainty ($u_{rel}(x_i)$) are calculated accordingly. The performance scores ($D\%_i$, $\zeta$ and $\zeta$ scores) and the uncertainty assessment are based on the criteria set by the PT provider ($x_{pt}=100 \text{ mg kg}^{-1}$; $u(x_{pt})=3 \text{ mg kg}^{-1}$ and $\sigma_{pt}=10 \text{ mg kg}^{-1}$). The grey and black cells represent “questionable” and “unsatisfactory” performance scores, respectively. The last column refers to the MU assessment (cf. Case “a” realistic, “b” underestimated or “c” overestimated).

| Lab code | $x_i$ (mg kg$^{-1}$) | $U_i(k=2)$ (mg kg$^{-1}$) | $u(x_i)$ (mg kg$^{-1}$) | $u_{rel}(x_i)$ | $D\%_i$ | $\zeta$ score | $\zeta$ score | Case |
|----------|----------------------|-----------------------------|--------------------------|---------------|----------|---------------|---------------|------|
| L01      | 93.40                | 14.50                       | 7.25                     | 0.078         | -6.6%    | -0.66         | -0.84         | a    |
| L02      | 111.00               | 0.00                        | 0.00                     | 0.000         | 0%       | 0             | 0             | a    |
| L03      | 129.90               | 10.00                       | 5.00                     | 0.038         | 29.9%    | 2.99          | 5.13          | a    |
| L04      | 76.70                | 0.00                        | 0.00                     | 0.000         | -23.3%   | -2.33         | -7.77         | b    |
| L05      | 112.90               | 9.50                        | 4.75                     | 0.042         | 12.9%    | 1.29          | 2.30          | a    |
| L06      | 118.60               | 25.90                       | 12.95                    | 0.109         | 18.6%    | 1.86          | 1.40          | c    |
| L07      | 114.10               | 22.60                       | 11.30                    | 0.099         | 14.1%    | 1.41          | 1.21          | a    |
| L08      | 98.00                | 5.00                        | 2.50                     | 0.026         | -2.0%    | -0.20         | -0.51         | b    |
| L09      | 73.90                | 13.50                       | 6.75                     | 0.091         | -26.1%   | -2.61         | -3.53         | a    |
| L10      | 125.00               | 18.00                       | 9.00                     | 0.072         | 25.0%    | 2.50          | 2.64          | a    |
| L11      | 80.10                | 5.80                        | 2.90                     | 0.036         | -19.9%   | -1.99         | -4.77         | a    |
| L12      | 108.20               | 16.80                       | 8.40                     | 0.078         | 8.2%     | 0.82          | 0.92          | a    |
| L13      | 88.00                | 0.00                        | 0.00                     | 0.000         | -12.0%   | -1.20         | -4.00         | b    |
| L14      | 62.20                | 18.00                       | 9.00                     | 0.145         | -37.8%   | -3.78         | -3.98         | c    |
| L15      | 91.40                | 17.50                       | 8.75                     | 0.096         | -8.6%    | -0.86         | -0.93         | a    |
| L16      | 89.60                | 7.30                        | 3.65                     | 0.041         | -10.4%   | -1.04         | -2.20         | a    |
| L17      | 99.20                | 18.60                       | 9.30                     | 0.094         | -0.8%    | -0.08         | -0.08         | a    |
| L18      | 102.00               | 10.00                       | 5.00                     | 0.049         | 2.0%     | 0.20          | 0.34          | a    |
| L19      | 127.60               | 23.00                       | 11.50                    | 0.090         | 27.6%    | 2.76          | 2.32          | a    |
| L20      | 83.40                | 8.60                        | 4.30                     | 0.052         | -16.6%   | -1.66         | -3.17         | a    |
| L21      | 113.00               | 13.50                       | 6.75                     | 0.060         | 13.0%    | 1.30          | 1.76          | a    |
| L22      | 96.00                | 25.60                       | 12.80                    | 0.133         | -4.0%    | -0.40         | -0.30         | c    |
| L23      | 119.20               | 13.50                       | 6.75                     | 0.057         | 19.2%    | 1.92          | 2.60          | a    |
| L24      | 102.10               | 15.70                       | 7.85                     | 0.077         | 2.1%     | 0.21          | 0.25          | a    |
| L25      | 87.00                | 9.60                        | 4.80                     | 0.055         | -13.0%   | -1.30         | -2.30         | a    |
| L26      | 95.70                | 18.40                       | 9.20                     | 0.096         | -4.3%    | -0.43         | -0.44         | a    |
| L27      | 136.80               | 0.00                        | 0.00                     | 0.000         | 36.8%    | 3.68          | 12.27         | b    |
| L28      | 78.00                | 16.20                       | 8.10                     | 0.104         | -22.0%   | -2.20         | -2.55         | c    |
| L29      | 84.10                | 16.50                       | 8.25                     | 0.098         | -15.9%   | -1.59         | -1.81         | a    |
| L30      | 87.10                | 10.40                       | 5.20                     | 0.060         | -12.9%   | -1.29         | -2.15         | a    |
| L31      | 81.60                | 25.50                       | 12.75                    | 0.156         | -18.4%   | -1.84         | -1.40         | c    |
| L32      | 93.80                | 26.70                       | 13.35                    | 0.142         | -6.2%    | -0.62         | -0.45         | c    |
| L33      | 69.00                | 4.80                        | 2.40                     | 0.035         | -31.0%   | -3.10         | -8.07         | a    |
| L34      | 108.00               | 2.50                        | 1.25                     | 0.012         | 8.0%     | 0.80          | 2.46          | b    |
| L35      | 94.40                | 9.10                        | 4.55                     | 0.048         | -5.6%    | -0.56         | -1.03         | a    |
| L36      | 103.10               | 25.80                       | 12.90                    | 0.125         | 3.1%     | 0.31          | 0.23          | c    |
| L37      | 72.00                | 19.00                       | 9.50                     | 0.132         | -28.0%   | -2.80         | -2.81         | c    |
| L38      | 117.10               | 7.00                        | 3.50                     | 0.030         | 17.1%    | 1.71          | 3.71          | a    |
| L39      | 104.00               | 0.90                        | 0.45                     | 0.004         | 4.0%     | 0.40          | 1.32          | b    |
| L40      | 105.20               | 16.40                       | 8.20                     | 0.078         | 5.2%     | 0.52          | 0.60          | a    |
(vertical translation in the Naji2 plot) or a bias correction (horizontal translation in the Naji2 plot).

**Discussion**

**Measurement uncertainty evaluation**

Since 2000, the PT providers of the Joint Research Centre in Geel are assessing the participants’ performance using the $z$ and $\zeta$ scores for laboratories participating in various IMEP, REIMEP, NUSIMEP PTs or for PT schemes organised by JRC’s European Union Reference Laboratories for contaminants in food. The organisers have noticed that extremely large reported MU lead to $|\zeta| \leq 2$ (satisfactory performance), while the blue and red dashed lines represent the “assigned range” ($x_{pt} \pm 2\ u(x_{pt})$) and the “acceptance range” ($x_{pt} \pm 2\ \sigma_{pt}$), respectively. The green, yellow and red bars in the lower part of the figure indicate the “satisfactory”, “questionable” or “unsatisfactory” performance, expressed as $z$ and $\zeta$ scores.

Let us consider, for example, the results submitted by L14 and L19 ($62.2 \pm 9.0$ mg kg$^{-1}$ /0.14/; and $127.6 \pm 11.5$ mg kg$^{-1}$ /0.09/, expressed as $x_i \pm u(x_i)/u_{rel}(x_i)$, see Table 2), in the frame of the hypothetical PT round (where $x_{pt} = 100$ mg kg$^{-1}$; $u(x_{pt}) = 3$ mg kg$^{-1}$; $\sigma_{pt} = 10$ mg kg$^{-1}$; $u_{rel}(x_{pt}) = 0.03$; and $\sigma_{pt,rel} = 0.10$). According to the above-mentioned criteria, the absolute MU reported by L14 would be considered as “realistic” ($9 < 10$, in mg kg$^{-1}$), while the one of L19 would be flagged as potentially “overestimated” ($11.5 > 10$, in mg kg$^{-1}$).

However, according to Ellison and Williams [21, 22], the relative uncertainty of chemical measurements can be
taken as constant for a range of measurand values. Based on this, the alternative approach described in this paper had been implemented by several EURLs of the JRC. Instead of comparing $u(x_i)$ to $u(x_{pt})$ and to $\sigma_{pt}$, the reported relative measurement uncertainty ($u_{rel}(x_i) = u(x_i)/x_i$) is compared to the relative standard uncertainty of the assigned value ($u_{rel}(x_{pt}) = u(x_{pt})/x_{pt}$) and to the relative standard deviation for proficiency assessment ($\sigma_{pt,rel} = \sigma_{pt}/x_{pt}$) over the whole range of reported mass fractions (e.g. from 60 to 140 mg kg$^{-1}$ or $\xi \leq 4$ in Figs. 1 and 3).

Knowing that a constant relative uncertainty $u_{rel}(x_i)$ implies a proportional increase of $u(x_i)$ with $x_i$, one may have for $z > 0$ ($x_i > x_{pt}$) a realistic $u(x_i)$ larger than $u(x_{pt})$, or inversely, a realistic $u(x_i)$ smaller than $u(x_{pt})$ for $z < 0$ ($x_i < x_{pt}$). Hence, according to the new criteria and unlike to the assessment presented in ISO 13528, the MU statement of L14 is flagged as potentially “overestimated” (0.14 < 0.10), while the one of L19 is considered as “realistic” (0.09 < 0.10).

These new criteria for the evaluation of reported measurement uncertainties may be taken into consideration to replace those prescribed in the current ISO 13528 standard [7].

**Naji2 plot applied to real cases**

In 2018, the European Union Reference Laboratory for Food Contact Materials (EURL-FCM) organised a PT round (FCM-18–02, [23]) for the determination of the mass fractions of total zinc and other metals in food simulant B (acetic acid, 3% w/v). The test items were prepared gravimetrically, which resulted in an accurate assigned value ($x_{pt} = 5.024$ mg kg$^{-1}$) with a small associated standard measurement uncertainty ($u(x_{pt}) = 0.033$ mg kg$^{-1}$). Based on the expert opinion, $\sigma_{pt,rel}$ was set to 0.12. A total of 46 National Reference Laboratories (NRLs) and Official Control Laboratories (OCLs) reported results. The outcome of this PT is presented in Fig. 4a, where the Naji2 plot includes the “bias boundaries”. Most of the laboratories reported accurate values and realistic measurement uncertainty estimations (case “a”), while some of them reported overestimated MUs (case “c”). However, according to ISO Guide 33 [20] six laboratories have to investigate their significant bias (with $|\zeta| > 2$), of which three have a $|\xi| \leq 2$ (see empty circles in Fig. 4a).

In 2019, the European Union Reference Laboratory for Genetically Modified Food and Feed (EURL GMFF) organised a PT round (GMFF-19/02, [24]) for the determination of GMOs in food and feed materials to support Regulation (EU) 2017/625 on official controls [25]. One of the test items consisted of a pig feed spiked with a genetically modified “40–3-2 soybean”. The EURL characterised the mass fraction of the GM event applying the EURL validated method and used the following values for the performance assessment of the participants: $x_{pt} = 1.014$ m/m %; $u(x_{pt}) = 0.061$ m/m %, and $\sigma_{pt,rel} = 0.25$. A total of 67 NRLs and OCLs reported results. Figure 4b shows the outcome of this PT round, where the Naji2 plot presents the set of hyperbolas. Most of the laboratories reported accurate values and realistic measurement uncertainty estimations, while some of them reported overestimated MUs (case “c”). However, according to ISO Guide 33 [20] sixteen laboratories have to investigate their significant bias (with $|\zeta| > 2$), of which twelve have a $|\zeta| \leq 2$ (see empty circles in Fig. 4b). It is worth noting that, in this particular case, the realistic range of MU (defined by Eq. 9) goes to zero at $z = -4 (= 1/\sigma_{pt,rel})$.

### Conclusions

The Naji2 plot is an intuitive graphical representation that allows the simultaneous assessment of the three performance evaluations ($z$, $\zeta$ scores and the MU assessment) and the identification of potential biases. This comprehensive assessment may indicate to participants the need for an appropriate corrective action that, otherwise, would have been hidden by a satisfactory performance, if expressed uniquely as a $z$ score. Similarly to the original Naji plot, the Naji2 plot can be applied to all analytes, matrices, concentration/content levels and proficiency testing schemes. It can be used to summarise the outcome of a PT round presenting the results of all laboratories for a given measurand, or to present a historical overview of a laboratory participation in different rounds of a specific PT scheme.
This tool is accessible to the PT providers that request from laboratories who participate in their PT schemes, to report a measurement result as stipulated by ISO/IEC 17025, i.e. including the associated measurement uncertainty. Unfortunately, this is still rarely the case. The situation may improve as soon as the next edition of ISO/IEC 17043 would require PT providers to assess laboratory performances based on the measured value and the corresponding measurement uncertainty.

Furthermore, the authors consider that the evaluation of the reported measurement uncertainties described in Sect. 9.8 of ISO 13528:2015 may be reviewed to take into account the relative uncertainties as an assessment criterion.

Acknowledgements The authors wish to acknowledge Alexander Bernreuther, Ioannis Fiamegkos, Emanuel Tzokatsis (former JRC colleagues) and Marizia Mancin (from the Istituto Zooprofilattico Sperimentale delle Venezie) for their valuable comments.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Fig. 4 Two Naji2 plots presenting the outcome of two PT rounds (FCM-18–02 and GMFF-19–02). a includes the two blue double lines defining the bias boundaries. b includes the two hyperbolas defining the |ζ| ≤ 2 and the |ζ| ≤ 3 ranges, respectively. Empty circles indicate the results having |ζ| ≤ 2 and |ζ| > 2

References

1. ISO 17025:2005. General requirements for the competence of testing and calibration laboratories. International Organization for Standardization, Geneva, Switzerland
2. BIPM (2012) International vocabulary of metrology—basic and general concepts and associated terms (VIM 3rd edition). JCGM 200. https://www.bipm.org/en/publications/guides/vim.html. Accessed 31 Jan 2021
3. ISO 17043:2010. Conformity assessment—general requirements for proficiency testing. International Organization for Standardization, Geneva, Switzerland
4. Robouch P, Younes N, Vermaereck P (2003) The “Naji Plot”, a simple graphical tool for the evaluation of inter-laboratory comparisons. In: Proceedings of the international workshop: data analysis of interlaboratory comparisons; data analysis of key comparisons. pp 149–160. Edited by the Federal Institute for Material (PTB), Berlin, Germany. ISBN: 3897019333
5. Pommé S (2006) An intuitive visualisation of intercomparison results applied to the KCDB. Appl Radiat Isot 64:1158–1162. https://doi.org/10.1016/j.apradiso.2006.02.017
6. Harms AV (2009) Visualisation of proficiency test exercise results in Kiri plots. Accred Qual Assur 14:307–311. https://doi.org/10.1007/s00769-009-0512-0
7. ISO 13528:2015. Statistical methods for use in proficiency testing by interlaboratory comparison. International Organization for Standardization, Geneva, Switzerland
8. Tsochatzis ED, Alberto Lopes J, Emteborg H, Robouch P, Hoekstra E (2019) Determination of PBT cyclic oligomers in and migrated from food contact materials—FCM-19/01 Proficiency Test Report. JRC 118572. https://europa.eu/!pD99nU. Accessed 30 Jan 2021
9. Tanaskovski B, Broothaerts W, Buttinger G, Corbisier P, Emteborg H, Robouch P, Emons H (2020) Determination of GM Maize MON88017 in bird feed and GM Maize GA21 in maize
flour. EURL GMFF Proficiency Testing Report GMFF-20/01. JRC122118. https://europa.eu/TV99gT. Accessed 30 Jan 2021

10. Aregbe Y, Harper C, Nørgaard J, De Smet M, Smeyers P, Van Nevel L, Taylor PDP (2004) The interlaboratory comparison “IMEP-19 trace elements in rice”—a new approach for measurement performance evaluation. Accred Qual Assur 9:323–332. https://doi.org/10.1007/s00769-003-0693-x

11. Bednarova M, Aregbe Y, Harper C, Taylor PDP (2006) Evaluation of laboratory performance in IMEP water interlaboratory comparisons. Accred Qual Assur 10:617–626. https://doi.org/10.1007/s00769-005-0031-6

12. Serapinas P (2007) Approaching target uncertainty in proficiency testing schemes: experience in the field of water measurement. Accred Qual Assur 12:569–574. https://doi.org/10.1007/s00769-007-0310-5

13. Frazzoli C, Robouch P, Caroli S (2010) Analytical accuracy for trace elements in food: a graphical approach to support uncertainty analysis in assessing dietary exposure. Toxicol Environ Chem 92:641–654. https://doi.org/10.1080/02772240903257941

14. Koster-Ammerlaan M, Bode P (2009) Improved accuracy and robustness of NAA results in a large throughput laboratory by systematic evaluation of internal quality control data. J Radioanal Nucl Chem 280:445–449. https://doi.org/10.1007/s10967-009-7475-9

15. Venciarutti C, Varga Z, Richter S, Jakopič R, Mayer K, Aregbe Y (2015) REIMEP-22 inter-laboratory comparison: “U age dating—determination of the production date of a uranium certified test sample.” Radiochim Acta 103:825–834. https://doi.org/10.1515/ract-2015-2437

16. Cordeiro F, Beldi G, Snell J, García-Ruiz S, Van Britsom G, Cizek-Stroh A, Robouch P, Hoekstra E (2018) Determination of the mass fractions of total aluminium, nickel, antimony and zinc in food simulant B. JRC report JRC113663. https://europa.eu/rup37WY. Accessed 30 Jan 2021

17. Broothaerts W, Beaz Hidalgo R, Buttinger G, Corbişier P, Cordiero F, Dimitrievska B, Emteborg H, Maretti M, Robouch P, Emons H (2019) Determination of GM Soybean 40–3–2 and MON87708 and GM Cotton GHB119 in Animal Feed and GM Soybean DAS-44406 in Soybean Flour - EURL GMFF Proficiency Testing Report GMFF-19/02. JRC report JRC11837. https://europa.eu/!uP37WY. Accessed 30 Jan 2021

18. ProLab—Software for PT programs and collaborative studies according to ISO 17043. Edited by Quodata, Dresden, Germany. https://quodata.de/content/prolab-software-interlaboratory-studies-quodata-gmbh#0. Accessed 31 Jan 2021

19. ISO 3534–1:2006. Statistics—vocabulary and symbols—Part 1: general statistical terms and terms used in probability. International Organization for Standardization, Geneva, Switzerland

20. ISO Guide 33:2015. Reference materials—good practice in using reference materials. International Organization for Standardization, Geneva, Switzerland

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