EURL-MP-background doc_001 (version 1)
Performance assessment in proficiency tests organised by the EURL mycotoxins & plant toxins in food and feed

Modifications compared to previous version:
Not applicable.

Background document drafted by:
EU Reference Laboratory for mycotoxins & plant toxins in food and feed (EURL-MP)
RIKILT Wageningen University & Research
Akkermaalsbos 2, 6708 WB, Wageningen, The Netherlands
eurl.mycotoxins-planttoxins@wur.nl

Notices:
This document has been drafted for EU National Reference Laboratories as background document on proficiency tests organised by the EURL-MP. It has been produced with the utmost care. However, RIKILT does not accept liability for any claims based on the contents of this document.

Suggested Citation: Performance assessment in proficiency tests organised by the EURL mycotoxins & plant toxins in food and feed v1, 2019. EURL mycotoxins & plant toxins, RIKILT Wageningen University & Research.

Table of Contents

1 Introduction 3
2 Calculation of the assigned value (C) 3
3 Calculation of the uncertainty of the consensus value (u) 4
4 Standard deviation for proficiency assessment (σ_P) 4
5 Performance characteristics with regard to the accuracy 5
6 Dealing with false positives, '<LOQ', and false negatives 7
7 References 8

1 Introduction

RIKILT organises proficiency tests (PTs) in the field of food and feed safety. This document describes the general procedure for evaluation of data submitted by the participants in RIKILT PTs, and more specifically the assessment of the performance of laboratories participating in PTs organised as EURL mycotoxins & plant toxins. The statistical evaluation is carried out according to the International Harmonized Protocol for the Proficiency Testing of Analytical Laboratories [1], elaborated by ISO, IUPAC, AOAC and ISO/IEC 13528:2015 [2] in combination with the insights published by the Analytical Methods Committee [3, 4] regarding robust statistics.

In EURL-MP PTs, the participants are asked to submit quantitative results (numerical values). In case the analyte is below the limit used by the participant for reporting quantitative results, i.e. below the limit of quantification (LOQ) or below the reporting limit (RL) of the laboratory, this limit needs to be specified during submission of result (e.g. <100 µg/kg). Qualitative results such as 'detected', 'not detected' or '<LOQ' without specification of the LOQ, are invalid and the analyte is considered to be outside the scope of quantitative analysis of the participant's laboratory.

For evaluation of quantitative results submitted by the participant, z-scores are calculated based on the assigned value, its uncertainty, and the standard deviation for proficiency assessment. If not negligible, the uncertainty of the assigned value and, if applicable, instability of analytes in the PT material, are taken into account in the determination of the z-scores.

In case the participant reports '<[value]', proxy-z-scores are calculated as a way to assess possible false negatives and to benchmark the LOQ or RL relative to the assigned value and the LOQ/RL of the other participants.

2 Calculation of the assigned value (C)

By default, the consensus value based on the participants' results is used as assigned value. The consensus value is determined using robust statistics [1-4]. The advantage of robust statistics is that all values are taken into account: outlying observations are retained, but given less weight. When using robust statistics, the data do not have to be normally distributed in contrast to conventional outlier elimination methods.

The robust mean of the reported results of all participants, calculated from an iterative process that starts at the median of the reported results using a cut-off value depending on the number of results, is used as the consensus value [1, 3]. For determination of the consensus value, the number of results received for an analyte in a PT material needs to be at least seven. Below seven no proper evaluation of the participants' results can be performed. In this case, consensus values (and z-scores) may be provided, but are for information only and not suited for evaluation or classification of the participants performance.

In certain cases, the EURL-MP may decide to use alternative options for establishment of the assigned value. Such cases will be justified and described in the PT report.

3 Calculation of the uncertainty of the consensus value (u)

The uncertainty of the consensus value is calculated to determine the influence of this uncertainty on the evaluation of the participants' results. A high uncertainty of the consensus value will lead to a high uncertainty of the calculated participants z-scores. If the uncertainty of the consensus value and thus the uncertainty of the z-score is high, the evaluation could indicate unsatisfactory method performance without any cause within the laboratory. In other words, illegitimate conclusions could be drawn regarding the performance of the participating laboratories from the calculated z-scores if the uncertainty of the consensus value is not taken into account.

The uncertainty of the consensus value is calculated from the estimation of the standard deviation of the consensus value and the number of values used for the calculation of the consensus value [2]:

$$u = 1.25 * \frac{\hat{\sigma}}{\sqrt{n}}$$ Equation 1

where:
u = uncertainty of the consensus value;
n = number of values used to calculate the consensus value;
$\hat{\sigma}$ = estimate of the standard deviation of the consensus value resulting from robust statistics.

According to ISO/IEC 13528:2015 [2] the uncertainty of the consensus value (u) is negligible and therefore does not have to be included in the statistical evaluation if:

$$u \leq 0.3\sigma_P$$ Equation 2

where:
u = the uncertainty of the consensus value;
σ_P = standard deviation for proficiency assessment (see chapter 4).

In case the uncertainty of the consensus value does not comply with this criterion, the uncertainty of the consensus value should be taken into account when evaluating the performance of the participants' results (see chapter 5). In case the uncertainty is > $0.7\sigma_P$, it is too high to determine meaningful consensus values and z-scores, and no evaluation of laboratories' performance is possible.

4 Standard deviation for proficiency assessment (σ_P)

The standard deviation for proficiency (target standard deviation) determines the performance boundaries in a PT. For mycotoxin and plant toxin determination in food and feed, a fit-for-purpose relative target standard deviation of 25% is used, irrespective the analyte, matrix or concentration. This value has been primarily based on what is currently analytically feasible, while also keeping in mind what is desirable from a user perspective (enforcement, risk assessment). The analytical feasibility has been assessed by an inventory of the robust relative standard deviations (RSD_R, here $\hat{\sigma}$) from a large number of mycotoxin PTs EURL and Fapas in the period 2013-2018) covering a wide variety of matrices and concentrations. No clear dependencies of the RSD_R on the toxin (mostly mycotoxins), the matrix, or the concentration was observed. The median and 75 percentile of the RSD_R's were 22% and 26%, respectively (N>750). Based on this, a realistic target RSD would

be around 22%. A slightly more tolerant 25% was chosen here because of i) the trend towards increased use of LC-MS-based multi-toxin methods, ii) more challenging PTs on 'new' mycotoxins and plant toxins, and iii) to align with the value used by the EURLs on pesticides [5]. Based on the above, the standard deviation for proficiency assessment σ_P is given by:

$$\sigma_P = 0.25*C \qquad \text{Equation 3}$$

where C = assigned value.

In certain cases, the EURL-MP may decide to use alternative values for the target standard deviation. Such cases will be justified and described in the PT report.

5 Performance characteristics with regard to the accuracy

For indicating the performance of the participating laboratories with regard to the accuracy a z-score is calculated. For the evaluation of the performance of the laboratories, ISO/IEC 13528:2015 [2] is applied. According to these guidelines z-scores are classified as presented in Table 1.

Table 1 *Classification of z-scores.*

$\|z\| \leq 2$	Satisfactory
$2 < \|z\| < 3$	Questionable
$\|z\| \geq 3$	Unsatisfactory

Depending on the uncertainty of the consensus value and, if applicable, any instability of the analytes in the PT material, z-scores can be are calculated in four different ways. These are described below.

I) Uncertainty of the consensus value is negligible/no instability
Here the calculated uncertainty of the consensus value complies with the criterion mentioned in chapter 3, ($u \leq 0.3\sigma_p$), and no significant instability of the analytes in the PT material is observed. In this case the z-score is calculated from:

$$z = \frac{x-C}{\sigma_p} \qquad \text{Equation 4}$$

where:
z = z-score;
x = the result of the laboratory;
C = consensus value;
σ_P = standard deviation for proficiency assessment.

II) Uncertainty of the consensus value is not negligible/no instability
When: $0.3\sigma_p < u \leq 0.7\sigma_p$ the uncertainty of the consensus value is significant but a z-score is nevertheless calculated. Here the uncertainty could influence the evaluation of the laboratories. Although, according to ISO/IEC 13528:2015 in this case no z-scores should be calculated, we feel that evaluation of the participating laboratories is of main importance justifying the participating

laboratories' effort. Therefore, in this case, the uncertainty is taken into account by calculating the z'-score [2]:

$$z' = \frac{x-C}{\sqrt{\sigma_p^2 + u^2}} \qquad \text{Equation 5}$$

where:
z' = z-score taking into account the uncertainty of the consensus value;
x = the result of the laboratory;
C = consensus value;
σ_P = standard deviation for proficiency assessment;
u = uncertainty of the consensus value.

III) Uncertainty of the consensus value is negligible/instability is not negligible

When during assessment of the stability of an analyte in the PT material a decrease in concentration is observed that might influence the evaluation of the laboratory performance, this consequential instability is taken into account when calculating z-scores. Because instability only regards one side of the confidence interval (a decrease of the concentration) this correction only applies to negative z-scores and results in an asymmetrical confidence interval.

In the case of a consequential instability, the z-score for the laboratories that reported an amount below the consensus value is corrected for this instability by:

$$z_i = \frac{x-C}{\sqrt{\sigma_p^2+\Delta^2}} \qquad \text{Equation 6}$$

where:
z_i = z-score taking into account the instability of the consensus value;
x = the result of the laboratory;
C = consensus value;
σ_P = standard deviation for proficiency assessment;
Δ = difference between average concentration of compound stored at different storage conditions.

IV Uncertainty of the consensus value is not negligible/instability is not negligible

When both the uncertainty of the consensus value and the instability of an analyte in the PT material are not negligible, then both are taken into account for the laboratories that reported a concentration below the consensus value (for concentrations above the consensus value Equation 7 is identical to 5). Here a z'_i score is calculated:

$$z'_i = \frac{x-C}{\sqrt{\sigma_p^2+\Delta^2+u^2}} \qquad \text{Equation 7}$$

where:
z'_i = z-score taking into account the uncertainty and instability of the consensus value;
x = the result of the laboratory;
C = consensus value;
σ_P = standard deviation for proficiency assessment;

Δ = difference between average concentration of compound stored at different storage conditions;
u = uncertainty of the consensus value.

In the PT report it will be indicated which z-score, z, z', z_i, or z'_i was used for each analyte.

6 Dealing with false positives, '<LOQ', and false negatives

Besides calculation of z-scores, the data set is checked for false positives (FP) and false negatives (FN), and LOQs/RLs are benchmarked against what is analytically feasible within the NRL network.

False positives
A false positive is a quantitative result reported by the participant while the toxin is:
i) not detected in the PT material by the organiser, and/or
ii) not detected by the majority of the other participants.
A threshold may apply, below which results are not considered false positives, i.e. when the analyte concentration is below the LOQ of the organiser and/or the majority of the participants. This will be decided on a case-to-case basis. Since there is no assigned value, no z-score can be calculated. False positives will be indicated in the report as 'FP'. False positives should be interpreted as unsatisfactory performance.

Results below LOQ or RL (< x µg/kg)
Participants that analyse the PT material for a certain analyte, either report a quantitative result (numerical value) or, when the toxin was not detected or below the level the laboratory uses for reporting quantitative data, report as below the LOQ or RL, i.e. '<x µg/kg' (with specification of the value). In this case, 'proxy-z-scores' are calculated as a way to assess possible false negatives and to benchmark the LOQ (or RL) relative to the consensus value and to what is analytically feasible within the NRL network. Proxy-z-scores are calculated using:

$$proxy - z = \frac{x - C}{\sigma_p}$$

Equation 8

where:
proxy-z = value to classify <LOQ results
x = the LOQ or RL of the laboratory;
C = consensus value (see Ch2);
σ_P = standard deviation for proficiency assessment;

Proxy-z-scores are indicated in the PT report as a value between brackets and are for information only. They are not included in the graphical representations of z-scores of the participants. The interpretation is as follows:

proxy-z ≤ -3	based on the LOQ provided, the laboratory should have been able to detect and quantify the analyte. The result is classified as a false negative (FN). A false negative is interpreted as 'unsatisfactory' performance.

-3 < proxy-z < -2	based on the LOQ provided, it is highly likely that the laboratory should have been able to detect and quantify the analyte. The result is classified as a false negative (FN) and should be interpreted as 'questionable'.
-2 ≤ proxy-z ≤ 2	-2 to 0: based on the assigned value and the LOQ provided, the result cannot be classified as false negative. 0 to +2: benchmark: the LOQ is in the range of what is analytically feasible*.
2 < proxy-z < 3	benchmark: the LOQ is high compared to what is analytically feasible* The laboratory should consider to lower their LOQ/RL.
proxy-z ≥3	benchmark: the LOQ is too high compared to what is analytically feasible* The laboratory should consider to lower their LOQ/RL.

* the analytical feasibility is derived from all the participant results. When a consensus value can be determined, this means that at least seven laboratories reported quantitative results and the uncertainty of the assigned value was within acceptable limits (see chapter 3).

7 References

[1] Thompson M., Ellison R. and Wood, R., 2006, The International Harmonized Protocol for the Proficiency Testing of Analytical Chemistry Laboratories, Pure Appl. Chem, 78(1), 145-196.

[2] ISO/IEC 13528:2015, Statistical methods for use in proficiency testing by interlaboratory comparison.

[3] Analytical Methods Committee, 1989, Robust statistics - How not to reject outliers Part 1. Basic concepts, Analyst, 114, 1693-1697.

[4] Analytical Methods Committee, 1989, Robust statistics - How not to reject outliers Part 2. Interlaboratory trials, Analyst, 114, 1699-1702.

[5] Pesticides EURLs, EUPT General Protocol. http://www.eurl-pesticides.eu/docs/public/tmplt_article.asp?CntID=821&LabID=100&Lang=EN