Taking a closer look.

lach bruns

Lach & Bruns Partnerschaft
Consulting Chemists

Tempowerkring 1
21079 Hamburg
Germany

Phone +49 (0) 40/7 90 12-232
Phone +49 (0) 40/7 90 12-233

Office Berlin:
Elsterstraße 77
14612 Falkensee
Germany

Phone +49 (0) 3322/42 97 353

info@lach-bruns.com
www.lach-bruns.com
www.relana-online.com

Value and evaluation of Ring Tests

Guideline for appropriate interpretation of ring test results

Hamburg, 1 July 2019

Abstract

The accreditation norm ISO 17025 requires the participation in ring tests. Therefore, these participations are part of the quality assurance system of each testing laboratory.

Ring test results can be used for several purposes such as for looking of possibilities for improvements, or for delivery of evidences of the analytical quality also with regard to laboratories' clients. Within some private lab approval systems, the successful participation in ring tests is required, while putting additional commercial pressure towards the analytical approaches of the laboratories.

As discussed in this guideline, ring test providers and private lab approval systems use differing ring test designs as well as differing statistical evaluation models.

The treatment of *announced ring test* samples (homogenates) is significantly different compared to the analyses of routine samples. When being aware of a test situation, the analytical efforts deviate of course from those efforts usually applied on a routine level. Additionally, the important steps of sample pre-preparation and homogenisation are not covered by such common ring tests. As a consequence, the value of "standard" ring tests is poor related to the evaluation of the routine performances of laboratories.

Unannounced tests are closer to lab routine than announced ring tests. In unannounced tests, the test samples arrive without any pre-announcement. Furthermore, the turnaround time (= time for analyses) is similar to turnaround times demanded by clients for day-to-day samples.

In order to assess the analytical quality of laboratories' routine performances, it is preferable to use *undercover tests*. Such undercover ring tests are challenging to organise. If undercover samples remain undiscovered, they are analysed in the same way like routine samples, providing information on the real quality of the day-to-day work.

Concerning the evaluation of laboratories' performances in ring tests, the most often applied criterion "comparability" (comparison with the average analytical performance and application of the z-score model) just compares the individual laboratory result with a statistically calculated "average" performance of all participants. The laboratory performs well if it performs (as good or as bad) like the average. Additionally, the common z-score

model suffers from its broad range of accepted results, therefore a distinction regarding the quality of the participating labs is limited.

The criterion “trueness” is more appropriate for the evaluation of laboratories’ performances. This is because the laboratory should find the levels, which are actually present in the test samples - independent of any statistical average performances. Meeting the true value is of most importance in particular when taking into consideration the clients’ perspectives. In order to apply the “trueness” criterion, it is recommended to make use of the recovery criterion for validations as defined in the document SANTE/11813/2017, accepting results in the range of 70 – 120 % of the actually spiked (= actually added) levels of analytes. The target recovery of 70 – 120 % should be applied for all types of ring tests (announced, unannounced, undercover).

If a poor analytical performance is not named, no improvements are initiated.

The application of established evaluation systems only (z-score, orientation on the average performance) does not necessarily highlight analytical shortcomings. As a consequence, improvements in the analytical performance cannot be achieved. In this respect, concrete and tangible limits must be named, which are to be used as orientation for the assessment of analytical competence.

The consideration of the laboratories’ day-to-day routines as well as the application of the trueness criterion (recovery of the actually spiked level) allows the identification of dissatisfying results and thus areas of improvements.

1. Wording

Ring test, competence scheme, proficiency testing, ring trial, round robin test – this is just a selection of wordings used for similar issues, and their usage in practice is not always stringent.

The most important terms are as follows:

Ring test/ring trial:

The German accreditation body DAkkS defines in its document 71 SD 0 010 “ring test” very similarly to “interlaboratory comparison”:

interlaboratory comparison

“organization, performance and evaluation of measurements or tests on the same or similar items by two or more laboratories in accordance with predetermined conditions” (ISO/IEC 17043:2010, no. 3.4)

proficiency testing

“evaluation of participant performance against pre-established criteria by means of interlaboratory comparisons” (ISO/IEC 17043:2010, no. 3.7)

Following these definitions, a “proficiency testing” is therefore an interlaboratory comparison with a subsequent evaluation of the performance of the participants, applying pre-established criteria.

Terms such as “competence scheme” are used similarly to “proficiency testing”.

Basic to all these “tests/trials/testings/schemes” is, that test material is prepared, and sub-samples of this prepared test material is sent to a number of participating laboratories. The preparation of the test material shall ensure, that all participating laboratories have more or less the “same” sample – so the “same” analytes (f. ex. pesticides) at the “same” concentration levels. Conclusively, all participating laboratories should find the “same” pesticides at the “same” concentration levels. The laboratories report their individual results to the test provider organisation, which is summarising and evaluating the analytical performance of all labs and of each individual participant against defined performance criteria.

Method ring test (collaborative trial):

In method ring tests, also named “collaborative trials” (see ISO 17043), samples are analysed with prescribed methods, for example within the method validation for ISO or ASU methods (official German enforcement methods).
In such method ring tests, the performance of a method, which has usually been newly developed, is tested, not the performance of the labs. Within those ring tests, it can be proven that a method performs well in different labs, regardless of the varying conditions such as equipment, chemicals and personnel.

Within this guideline, the term “ring test” refers to proficiency tests, competence schemes and similar terms as well, following common habits.

2. Introduction

Proficiency tests and ring tests are part of every quality assurance system of analytical testing laboratories.

With the help of ring tests, laboratories can improve their analytical quality, identify possible weakness points in the analytical work flow, monitor and compare their performance, and display their competence f. ex. to customers.

The participation in proficiency tests is not voluntary, but required by the ISO 17025 norm and verified by accreditation bodies. The selection of proficiency tests and the frequency of participation may be prescribed by the responsible accreditation body (see f.ex. the German DAkkS document 71 SD 0 010). In certain areas, the successful participation in ring trials is demanded by legal regulations, such for the analysis of water for human consumption in Germany [Trinkwasserverordnung – “potable water act”].

Furthermore, some lab approval systems require the successful participation in competence schemes as well, adding a strong commercial pressure on the performance of the labs.

Due to the importance of proficiency testing schemes, a closer look at ring tests shall be taken within this guideline, with a focus on differences in designs and possible ways of evaluating the performance of analytical laboratories.

3. Common designs of ring tests

3.1. "Common" ring tests

Common or open ring trials are offered by private or public ring test providers. The common procedure is as follows:

- The ring test provider publishes a ring test programme combining analytes and matrices (product types like oranges, mushrooms, cereals, etc), which might be relevant for the clients of the laboratory. The ring test sample may contain one or more analytes of certain analytical groups such as mycotoxins, pesticides, nutrients, or microorganisms.
- Laboratories order their participation in the ring tests of their particular interest.
- The provider prepares the ring test material by spiking the relevant analytes or by applying material with incurred analytes.
- Subsequently, the provider assures the homogeneity of the test material as well as the stability of the analytes by conducting homogeneity resp. stability tests.
- At the announced starting date, one or more ring test samples are sent out to the participating labs in an appropriate way (frozen, cooled, etc.). Commonly, the material is a homogenate.
- The participating lab analyses the ring test sample(s) and reports the results according to the announced deadline, usually by electronic transfer.
- The ring test provider evaluates the received results (of all participating laboratories) statistically and prepares a report. The applied statistics refer f. ex. to checking for outliers and for calculating an assigned value (statistical average performance of all participants – if not identified as outlier). Depending on the design of the ring trial, the recovery of spiked analytes may be determined additionally.
- Within a competence scheme, the performances of the participating labs have to be evaluated as well, using criteria such a comparability and trueness, see chapter 6.

3.2. Unannounced ring test

Unannounced ring trials are handled similarly to common ring tests, with the main difference that the lab does not know the exact arrival date of the test sample. Sometimes, a time window (such as "in year ….") is known for the reception of the test sample.

The approach is followed f.ex. by the lab approval system of the BNN e.V. (German association "Bundesverband Naturkost Naturwaren"). Some commercial ring test providers adopt aspects of this type of ring test, such as short turnaround times.

3.3. Undercover ring test

Undercover ring tests are designed in such a way, that they are (preferably) not recognised as ring test samples at arrival in the participating lab.

In order to achieve this “undercover” character, common real samples are spiked by the organiser but are sent to the laboratories by “real” customers, like for example companies, which are or may be clients of the participating labs.

If the lab does not discover the ring test character of the sample, it is going to treat it as a routine sample.

4. Advantages and shortcomings of common ring test designs

4.1. “Common” ring tests

The handling of common ring trial samples varies in several points from that of routine samples, as table 1 shows:

Table 1. Comparison of classic (“common”) ring trial and routine samples

Aspect	**Ring trial sample**	**Routine sample**
Announcement of arrival	Yes	No (usually)
Analytes	Number and kind of analytes often known (selection of pesticides, selection of mycotoxins etc.)	Number and kind of positive analytes usually unknown (which pesticide(s), which contaminants, …)
Scope of analyses	Selected by ring test provider	Ordered by client
Sample material	Homogenate	Real sample, not homogenised (like fruits, grains, etc.)
Blank material	Often available	Not available
Starting point	Sample (homogenate) weighing	Sample preparation (selection of parts, peeling, cutting etc.)
Turnaround time	Up to several weeks	Usually some days, sometimes only hours
Staffs’ extra attention	Possible	Not possible
Attention paid to sample	High	“common”
Analytical runs	Several	One; additional runs in case of confirmation and quantification analyses
Quantification method	Several may be applied (external calibration, matrix calibration, standard addition etc.)	The one used in routine
Exchange of results between laboratories	Possible	Usually not possible
Interpretation according to legal and/or private standards	Often not required	Typically, yes

In consequence of the characteristics of common ring trials, the performance of a lab in such ring test does not necessarily reflect the performance in day-to-day routines!

As a sample of high homogeneity is distributed, common ring trials have the limitation that the measurement of the analytical performance is starting from the initial weighing of the analytical sample portion. All other steps before weighing (preparation of the sample such as correct chopping, taking the correct sample pieces for analyses, degree of homogenisation achieved) are excluded.

Crucial point: The ring test provider shall check the stability of the substances both after homogenisation (= before shipment) and at the deadline (delivery of results), in order to make sure that the analytes are stable across the analysing period.

4.2. Unannounced ring tests

Unannounced ring trials share some drawbacks with common ring trials, such as the special attention paid to it, usually using several analytical runs, applying several quantification methods (matrix calibration, standard addition etc.).

Nevertheless, most unannounced ring tests are designed to be closer to routine conditions:

- unannounced arrival,
- short turnaround times (usually some days),
- typical matrix (product) with typical analytes at common levels.
- Some ring trials also demand an **interpretation** which is assessed afterwards as well, thereby approaching real samples even more.

Conclusively, unannounced ring tests reflect routine conditions in a better way. Nevertheless, these ring test samples are treated with special care as well, especially when a successful participation is required for economic reasons. Therefore, the outcome of unannounced ring tests cannot be transferred one-to-one to the general analytical performance in the laboratories' routines.

4.3. Undercover ring tests

The organisation of undercover ring tests is more demanding compared to of the other types of ring tests:

- A matrix (product) suitable for this type of ring trial must be identified: fruits, which can be spiked with injections, grains which can absorb pesticides, etc.
- A certain number of cooperating companies have to be identified by making use of them as official clients sending the samples to the participating labs. These companies should be known as common customers to the participating labs. The official clients must be able to answer sample related questions on the phone or by mail. Furthermore, the cooperation partner has to receive the analytical reports and forward them to the organiser.

- As the lab does not know that a ring test is going to arrive, the common procedure of ring test participation cannot be followed – the lab cannot order its participation for a certain trial at a known date.

This type of undercover ring trial has some important advantages:

- The sample is treated as a routine sample (as long as the lab does not discover its real nature). Consequently, the results of the undercover ring test reflect the performance in routine analysis, at least at the time when the analysis was carried out (spot check).
- As real samples are sent in, all steps (for which the lab is responsible) are tested, including sample registration, the selection of sample material, cutting and homogenisation. These steps of the entire analytical chain are not covered by common ring tests.

If no undercover ring tests are available, such tests can be carried out even without an external organiser: The QS department of the laboratory may use samples with a known content from former analysis or ring tests and let them re-analysed again as a newly arrived sample.

5. Possible ways of evaluation

Depending on the aim and the design of the ring test, several ways of evaluation of ring tests results are applied.

In general, it must be distinguished between the following criteria:

- trueness and

- comparability.

Figure 1 (next page) visualises the difference: In an ideal world (down right), analytical results are true (correct, close to the real value, represented by the bull´s eye) and comparable (all participants´ results are distributed close to the bull´s eye; the median (or average) value is virtually equal to the correct value).

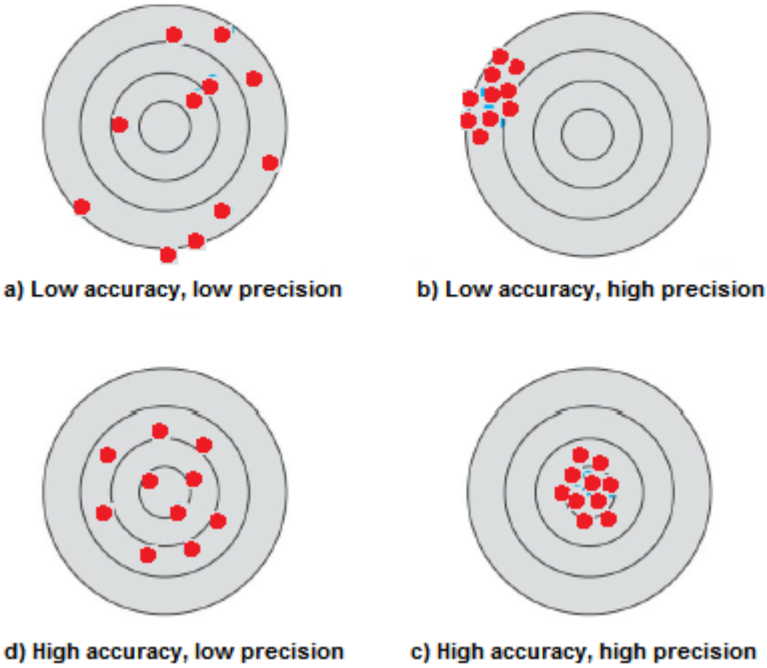
a) Low accuracy, low precision
b) Low accuracy, high precision
d) High accuracy, low precision
c) High accuracy, high precision

Figure 1. Comparability (precision) vs. accuracy (trueness / correctness) [Agbachi]

In reality, results may be accurate on average, but show a high variety (high trueness, low comparability; down left). Or results may be comparable, but not accurate (top right). As a worst case, results are both not accurate (false) and with low comparability (top left).

The most common ways to evaluate ring tests are discussed in the following subchapters.

5.1. Comparability (z-score)

When assessing the quality of the performance of a lab by its comparability, the z-score evaluation is the most commonly applied model.

The assigned value is set as the (robust) mean, median or average value of the participants.

In order to improve the quality of the assigned value, several statistical methods can be used:

5.1.1. Identification and treatment of outliers

Outliers might have a large influence on the evaluation of ring tests. Therefore, a careful approach should be followed when identifying and treating outliers. It should be kept in mind that it is not easy to determine, which values are actual “outliers”. This depends on the basis of analytical values and the applied statistical methods.

Exclusion of outliers: Several tests for outliers can be applied, such as the Cochran test and Grubbs test, see also [Bruns]. These values can be excluded from the statistical evaluation such as the calculation of the assigned value and the standard deviation.

It should be noted that the statistical basis is diminished, if too many outliers are excluded. Therefore, this step should be done carefully. Preferably, other techniques should be used to minimise the influence of extreme values, see the following paragraph.

5.1.2. Calculation of (weighed) average/mean/assigned value

For the calculation of the so-called “assigned value”, several techniques can be used, and some of them help to diminish the influence of outliers.

(Arithmetic) average/mean: The average or mean value is the statistical average value of all analytical results. Single outliers have a large influence on this value.

Median: The median is the median value of all analytical results, sorted by magnitude. Thereby, the influence of single outliers is diminished.

Robust mean: With the help of winsorisation techniques (see for example [Analytical Methods Committee]), the influence of outliers is minimised. One calculation method applying this technique is the Huber algorithm, as described in ISO 13528 and [Huber].

Bootstrapping: In case the results are not distributed around one peak as expected according to Gauss but show two or more peaks, methods such as bootstrapping [Thompson 2002] can be applied.

5.1.3. Evaluation according to comparability: z-score model

The z-score describes the deviation of the result of a lab from the assigned value, in dependence of the standard deviation ([Albrecht et al.], cited in [Bruns]):

$z = (x – xa) / \hat{\boldsymbol{\sigma}}$

with
x = analytical result as reported by the participant
xa = assigned value
$\hat{\boldsymbol{\sigma}}$ = target standard deviation („fit-for-purpose standard deviation”, see below)

In order to evaluate the performance of a lab with respect to comparability, the pattern shown in table 2 is commonly applied, see also [Thompson et al. 2006]:

Table 2: Evaluation of ring tests with z-score-model

z-score	Evaluation	Probability according to Gauss
$\lvert z \rvert \le 2$	satisfying	95.4 %
$2 < \lvert z \rvert \le 3$	questionable/suspicious	4.3 %
$\lvert z \rvert > 3$	dissatisfying	0.3 %

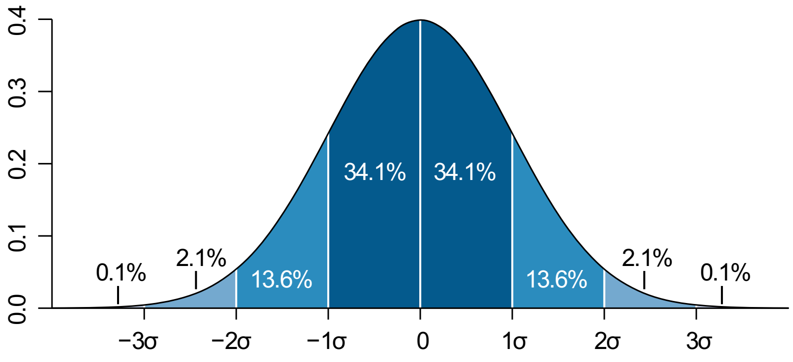

Figure 2. Gauss distribution of analytical results (σ: standard deviation) [Toews]

Usually a z-score between -2 and 2 is accepted. The value thereby is within the double standard deviation. In an ideal Gaussian distribution (curve, see figure 2), approx. 95 % of the result are within this range.

Concerning the determination of the standard deviation, several models can be applied:

- Use of experimental standard deviation as gathered in the ring trial: This model displays the analytical challenges of the single ring trial and the analytical quality of the participating labs.
- Use of standard deviation calculated according to Horwitz. The mathematician Horwitz described the standard deviation as a function of the concentration of an analyte – independent from the nature of the sample (fruit, oil, herbs etc.) and of the analyte (pesticides, heavy metals, food additives etc.):

$$RSD_r = 2^{(1 - 0,5 \log c)}$$

with:

RSD$_R$ relative standard deviation under reproducibility conditions

log common logarithm

c concentration, expressed as dimensionless mass fraction

In the concentration range for pesticides and contaminants, the Horwitz equation gives standard deviations of 18 – 22 %. Alternatively, the Fit-For-Purpose Relative Standard Deviation (FFP RSD) may be used, allowing a default standard deviation of ± 25 %.

For more aspects on this topic, see also relana® Position Paper No. 19-03 “*Differing results of high-quality labs: reasons and what is “normal?”.*

As a consequence of applying this evaluation model, a lab performs better, the better it meets the assigned value – the more “mediocre” it works, so the better it meets the average performance of all participants.

The z-score model has several advantages:

- It can be used with incurred analytes, where the true value (= spiked level) is unknown.

- It can be applied for fast degradable analytes such as dithiocarbamates, which may have partly metabolised in the time period between spiking and analysis.

The main disadvantage is that the trueness is not considered. Especially for challenging analytes, the majority of labs may deliver comparable results, which are not correct in terms of recovery, as the examples presented in chapter 5.3. and 5.5. highlight.

To cut a long story short: The majority might be right – but not necessarily in every case.

A comparison of the approaches "comparability" and "trueness" is discussed in chapters 5.3. and 5.5. also.

5.2. Trueness & recovery

In order to assess how good a laboratory can measure an analyte, the trueness criterion is preferably used, as the clients of the lab usually want to know the true (and not the comparable) levels of analytes.

Derived from SANTE document 11813/2017, the following ranges for recovery can be used as an acceptance criterion:

70 -120 %:

The a.m. SANTE document gives this range in chapter G6 as acceptable mean recoveries for method performance acceptability criteria.

60 -140 %

This range is given in chapter C44 as "a practical default range" for acceptance criteria for routine recoveries.

5.3. Comparison of trueness (recovery) versus comparability (assigned value / z-score)

What does a satisfying *z*-score of |*z*| ≤ 2 mean? And how is the result to be understood? The following evaluation scheme based on the z-score model is commonly applied (see also chapter 6.1.3):

<u>z-score interpretation</u>

|z| ≤ 2	Satisfactory
2 < |z| < 3	Questionable result
|z| ≥ 3	Unsatisfactory result

A satisfactory result |*z*| ≤ 2 is achieved, if the reported result deviates from the assigned value (statistical average) by less than twice the target standard deviation (σ) (Horrwitz / Thompson), which can be at ± 25 % (Fit-For-Purpose Relative Standard Deviation (FFP RSD).

The following example should help to visualise what "satisfying" in terms of the z-score model means:

Following assumptions are provided:

- assigned value xa = 100 µg / kg (thus the calculated statistical average)
- Target standard deviation $\hat{\sigma}$ = 25 µg/kg (thus 25%)
- Results = 75 µg/kg (Lab A) / 150 µg/kg (Lab B) / 50 µg/kg (Lab C) / 40 µg/kg (Lab D) / 180 µg/kg (Lab E)
- Formula for calculation of the z-score: $z = (x - x_a) / \hat{\sigma}$

Laboratory A: z-score = (75 µg / kg - 100 µg / kg) / 25 µg / kg = -1.0. Amount | -1,0 | = 1,0
Laboratory B: z-score = (150 µg / kg - 100 µg / kg) / 25 µg / kg = 2.0. Amount | 2.0 | = 2,0
Laboratory C: z-score = (50 µg / kg - 100 µg / kg) / 25 µg / kg = -2.0. Amount | -2,0 | = 2,0
Laboratory D: z-score = (40 µg / kg - 100 µg / kg) / 25 µg / kg = -2.4. Amount | -2,4 | = 2,4
Laboratory E: z-score = (180 µg / kg - 100 µg / kg) / 25 µg / kg = 3.2. Amount | 3,2 | = 3,2

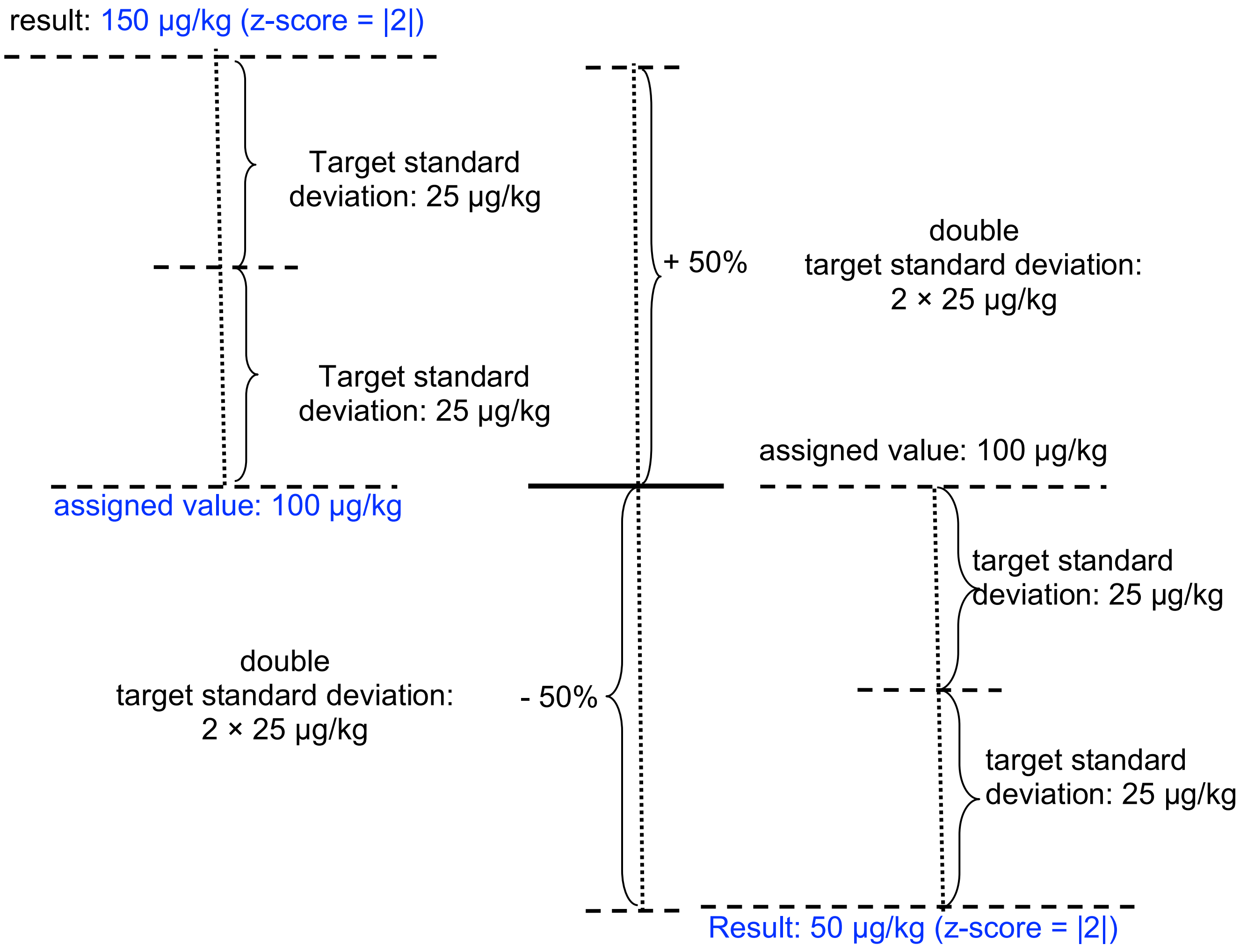
result: 150 µg/kg (z-score = |2|)
Target standard deviation: 25 µg/kg
Target standard deviation: 25 µg/kg
assigned value: 100 µg/kg
+ 50%
double target standard deviation: 2 × 25 µg/kg
assigned value: 100 µg/kg
double target standard deviation: 2 × 25 µg/kg
- 50%
target standard deviation: 25 µg/kg
target standard deviation: 25 µg/kg
Result: 50 µg/kg (z-score = |2|)

Figure 3: A laboratory reporting 50 µg/kg is just as comparable to the reference value of 100 µg / kg (assigned value) as a laboratory with a result of 150 µg/kg.

Both laboratories report comparable results of 50% relative to the assigned value (100 µg/kg). Both results are still satisfying taking into consideration the z-score model.

Making use of the recovery (70 – 120 %) of the spiked level for the evaluation of the results, only laboratories with results of 70 µg/kg up to 120 µg/kg would be considered satisfying (assuming the spiked level is at 100 µg/kg).

Another example is based on an undercover test (pesticides in grapes):

The results for the pesticide Fenhexamid of 19 undercover tested laboratories were evaluated according to 3 different models:

- Making use of the median as a reference of the average performance and calculation of the corresponding z-scores making use of a target standard deviation of 25%.
- Making use of the robust mean (assigned value) as a reference of the average performance and calculation of the corresponding z-scores making use of a target standard deviation of 25%.
- Making use of the trueness criterion thus taking into consideration a target recovery of 70 – 120 %.

Table 3: Evaluation of the laboratories' performances by making use of 3 different models

Fenhexamid Spiked level: 355 µg/kg				
Lab code	reported result (µg/kg)	z-score* with Median as reference (=292 µg/kg)	z-score* with assigned value (robust statistic) as reference (= 309 µg/kg)	Recovery of the spiked level (%)
1	246	-0,6	-0,9	**69**
2	170	-1,7	-1,9	**48**
3	420	1,8	1,5	118
4	522	**3,2**	**2,9**	**147**
5	283	-0,1	-0,4	80
6	380	1,2	1,0	107
7	350	0,8	0,6	99
8	328	0,5	0,3	92
9	210	-1,1	-1,4	**59**
10	260	-0,4	-0,7	73
11	300	0,1	-0,1	85
12	282	-0,1	-0,4	79
13	420	1,8	1,5	118
14	1140	**11,6**	**11,4**	**321**
15	360	0,9	0,7	101
16	270	-0,3	-0,5	76
17	202	-1,2	-1,5	**57**
18	20	**-3,7**	**-4,0**	**6**
19	320	0,4	0,2	90

* Target standard deviation: 25%

Whereas the first two criteria (assessments against the median and against the robust mean) show satisfying results (z-score) for 16 out of 19 participants, the trueness criterion shows dissatisfying results for 7 participants. A recovery of 48 %, 57 % and 59 % is considered satisfying according to the first two assessments. However, it is more than questionable in the client's perspective whereas a result of these low recoveries is a helpful tool to decide about the real quality of analysed goods.

Recommendation:

Both common ring tests as well as undercover samples aim checking the quality of the analytical performance of the lab. In order to identify areas of improvement, the trueness criterion (**70 – 120 % recovery**) could be applied to all sorts of proficiency tests (common, unannounced, undercover) when assessing trueness, provided that the analytes are stable in the test samples. The identification of possible shortcomings will be easier and straightforward.

A further aspect concerning trueness: both **false-positive and false-negative results** lead to the evaluation "proficiency test failed", as both incidents must be considered "false" and can have a strong commercial impact on lab clients.

A general comparison of both approaches "comparability" and "trueness" are discussed in chapter 5.3. and 5.5.

5.4. Consensus method

The consensus method can be applied in qualitative ring trails when the correct value (positive/negative) is not known as the sample has not been spiked.

This method is used in the area of analysis for genetically modified organisms (GMO), for instance, when the sample has not been spiked and a positive result cannot be excluded due to ubiquitous contaminations with GMO.

The consensus shows, whether a majority of the participating labs measured the same result (positive/negative, presence/absence). The criteria for the consensus (required percentage of labs which gained the same result, etc.) is fixed prior to the ring trial.

The drawback of this method is the same as for the comparability criterion (6.1): The majority might be right – but not necessarily in every case!

5.5. Comparability vs. trueness criterion

As explained in the introduction of this chapter, trueness and comparability are the most important criteria when evaluating ring test results. Ideally, the results are true AND comparable.

Figures 4 and 5 show the example "pymetrozine in lettuce" taken from a Lach & Bruns ring trial visualising the differences between the evaluations:

As shown in figure 4, the majority of participants delivered satisfying results in terms of comparability: 15 out of 19 (79 %) labs achieved a z-score between – 2 and +2.

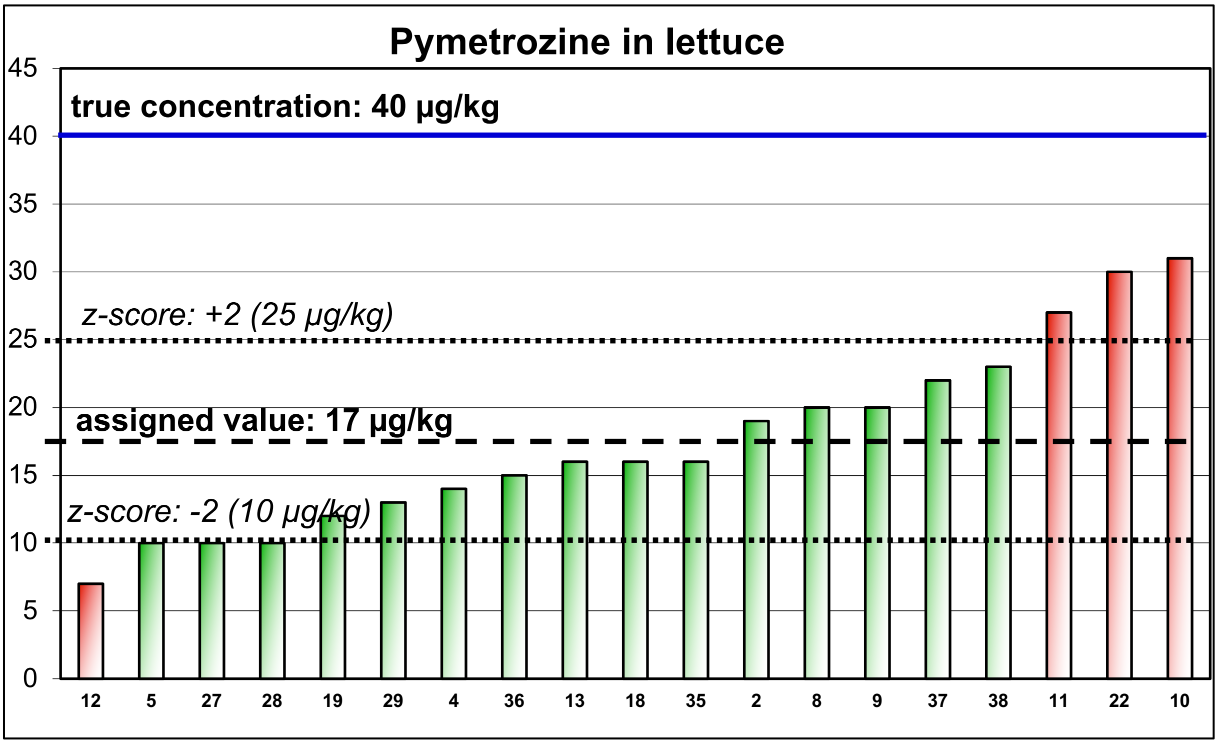

Figure 4. Ring trial "pymetrozine in lettuce": z-score (comparability) evaluation

When talking about the criterion "trueness", i.e. recovery of the spiked analyte, only 2 (!) of the participating labs in the ring trial showed good results, see figure 5. Only these 2 labs achieved recoveries in the accepted range between 70 and 120 % (see also 6.2.) - but both labs did not fulfil the criterion "comparability", as their z-scores were higher than 2!

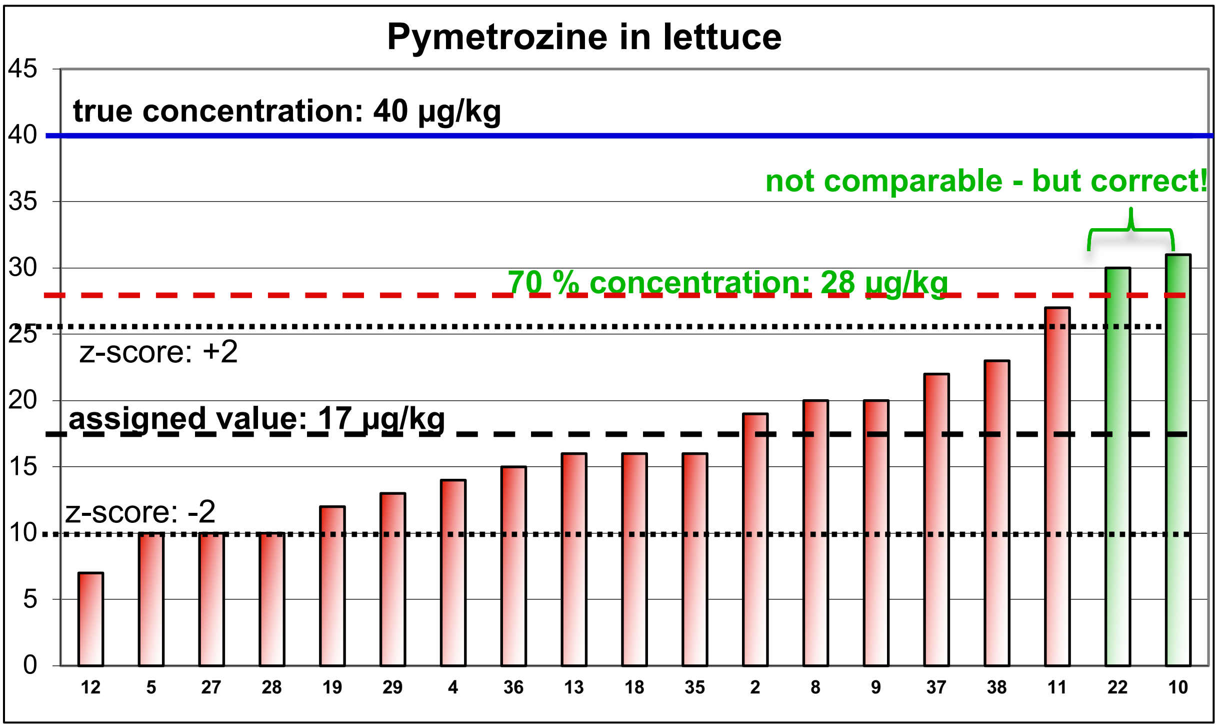

Figure 5: Ring trial "pymetrozine in lettuce": trueness (recovery) evaluation

As a consequence, both labs, which achieved correct results would have failed if the ring test had been evaluated according to the criterion “comparability” only!

The main disadvantage of ring tests applying the comparability criterion only is highlighted by this example: The trueness is not considered. This can end up in dissatisfying situations as described above.

Background: Pymetrozine is a pesticide which is unstable under the conditions during common sample preparation for pesticide analysis by the QuEChERS method [EN15662], as explained in the analytical German pre-norm ASU L 00.00-74 (V).

Especially in the case of “tricky” analytes, which need a special treatment, the application of the comparability criterion can lead to the situation that one lab shows a result close to the true (spiked) value but falling outside the accepted z-score range. Or in other words: The lab was too good!

A further aspect [Bruns] that needs to be stressed is that the comparability criterion is generally not suitable to distinguish between high- and low-quality performances, as the accepted range of -2 < z-score < +2 is too wide.

6. Interpretation of analytical results

In some proficiency test reports you can read:

“Proficiency testing aims to provide an independent **assessment of the competence** of participating laboratories.” (accentuation by author)

Despite this statement, most ring test providers do only test the **analytical competence** of the participants, thereby not evaluating the **interpretation competence** of the labs.

The competence in interpreting analytical results is a crucial indicator for the entire quality of the lab, as the interpretation is fundamental for the clients in order to decide what to do with the analysed goods. If the interpretation is incorrect or misleading, this will devaluate even the best analytical performance.

In some ring tests and competence schemes, an evaluation considering legal and other standards is required. Furthermore, specialised competence tests focussing on the evaluation of analytical results are available.

7. Conclusion

The participation in ring tests and proficiency tests is not only demanded by the accreditation norm ISO 17025 but can be a very valuable tool for the identification of areas of improvements. Thus, ring test results contribute to the quality assurance and permanent improvement of each participating laboratory.

Nevertheless, common ring tests show a range of restrictions as the way the test samples are analysed differs from routine samples.

Unannounced ring trials, carried out on homogenates as well, show conditions which are closer to routine than common ring tests, thereby generating results closer to “normal” samples.

As discussed in this paper, the best information related the lab performance in day-to-day routines can be received by undercover ring tests, which are more complex to organise than the other types of ring tests.

Concerning the evaluation of results, the application of the “trueness criterion” is recommended, as it is of higher significance compared to the “comparability criterion”, which does not necessarily show if a lab is analysing correctly.

Exemptions should be applied in case of instable substances and incurred analytes.

In order to evaluate the “trueness”, a recovery range of 70 – 120 % (pesticides) is recommended.

Finally, if a poor analytical performance is not named, no improvements are initiated.

Only the application of established evaluation systems (z-score, orientation on the average performance) does not highlight analytical shortcomings, so that no improvements in the analysis can be achieved.

8. Literature

ISO/IEC 17025:2017: General requirements for the competence of testing and calibration laboratories

Deutsche Akkreditierungsstelle (DAkkS): Einbeziehung von Eignungsprüfungen in die Akkreditierung, 71 SD 0 010, Revision: 1.2, 14. April 2016

Trinkwasserverordnung in der Fassung der Bekanntmachung vom 10. März 2016 (BGBl. I S. 459), last changed by Artikel 1 der Verordnung vom 3. Januar 2018 (BGBl. I S. 99)

ISO/IEC 17043:2010-2: Conformity assessment - General requirements for proficiency testing

B. Albrecht, W. Luginbühl, C. Balsiger, R. Bögli, U.P. Buxtorf, F. Bühler, H. Emch, P. Roos, A. Jakob, G. Gremaud, Ph. Hübner, M. Schurter, L. Spack, P. Wenk and M. Wolfensberger (2004) Leitfaden zur Validierung chemischphysikalischer Prüfverfahren und zur Abschätzung der Messunsicherheit. Mitt. Lebensm. Hyg. 95, 199-222

W. Horwitz, L. V. Kamps, K. W. Boyer (1980): Quality assurance in the analysis of foods for trace constituents, J. Assoc. Off. Anal. Chem., 63, 6, 1344-1354

S. Bruns: Bewertung von rückstandsanalytischen Laborkompetenzen anhand experimenteller Daten, PhD thesis, Technische Universität München, 2012

M. Thompson, S.L.R. Ellison, and R. Wood (2006) The International Harmonised Protocol for the Proficiency Testing of Analytical Chemistry Laboratories. Pure Appl. Chem., 78, No. 1, 145-196

Analytical Methods Committee (1989) Robust statistics - how not to reject outliers: Part 1 Basic concepts. Analyst 114, 1693

International Organization for Standardization. ISO 13528: Statistical methods for use in proficiency testing by inter-laboratory comparisons. English version ISO 13528:2005

P.J. Huber, P.J. (1981) Robust statistics. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York etc., 1981

M. Thompson (2002) Bump-hunting for the proficiency tester: Searching for multimodality. Analyst 127,1359–1364

European Commission, DG SANTE: Guidance document on analytical quality control and method validation procedures for pesticide residues and analysis in food and feed, SANTE/11813/2017

Amtliche Sammlung von Untersuchungsverfahren nach § 64 LFGB (ASU): Untersuchung von Lebensmitteln - Multiverfahren zur Bestimmung von Pestizidrückständen mit GC und LC nach Acetonitril-Extraktion/Verteilung und Reinigung mit dispersiver SPE in pflanzlichen Lebensmitteln, Modulares QuEChERS-Verfahren, L 00.00-115, Oktober 2018

Amtliche Sammlung von Untersuchungsverfahren nach § 64 LFGB (ASU): Untersuchung von Lebensmitteln - Hochdruckflüssigchromatographische Bestimmung von Pymetrozin in pflanzlichen Lebensmitteln, L 00.00-74 (V), Dezember 2002

9. References for figures

Figure 1: Ch. Agbachi: A Metric for Qualitative Analysis: Case Study of Standard Deviation, International Journal of Engineering Trends and Technology (IJETT) – Volume 63 Number 1- September 2018, DOI: 10.14445/22315381/IJETT-V63P208

Figure 2: M. W. Toews – own work, based (in concept) on figure by Jeremy Kemp, on 2005-02-09, CC BY 2.5, https://commons.wikimedia.org/w/index.php?curid=1903871

Figures 3, 4, and 5: Lach & Bruns Partnerschaft