DOI: 10.7795/810.20200724

CEQAT-DGHS interlaboratory tests for chemical safety: Validation of laboratory test methods by determining the measurement uncertainty and probability of incorrect classification including so-called "Shark profiles"

Peter Lüth [a], Steffen Uhlig [b], Kirstin Frost [c], Marcus Malow [a], Heike Michael-Schulz [a], Martin Schmidt [a] & Sabine Zakel [d]

[a] Bundesanstalt für Materialforschung und -prüfung (BAM), Berlin, Germany
[b] QuoData GmbH, Berlin, Germany
[c] QuoData GmbH, Dresden, Germany
[d] Physikalisch-Technische Bundesanstalt (PTB), Braunschweig, Germany

E-mail: peter.lueth@bam.de

Abstract

Laboratory test results are of vital importance for correctly classifying and labelling chemicals as "hazardous" as defined in the UN Globally Harmonized System (GHS) / EC CLP Regulation or as "dangerous goods" as defined in the UN Recommendations on the Transport of Dangerous Goods. Interlaboratory tests play a decisive role in assessing the reliability of laboratory test results. Interlaboratory tests performed over the last 10 years have examined different laboratory test methods. After analysing the results of these interlaboratory tests, the following conclusions can be drawn:

1. There is a need for improvement and validation for all laboratory test methods examined.
2. To avoid any discrepancy concerning the classification and labelling of chemicals, the use of validated laboratory test methods should be state of the art, with the results accompanied by the measurement uncertainty and (if applicable) the probability of incorrect classification.

This paper addresses the probability of correct/incorrect classification (for example, as dangerous goods) on the basis of the measurement deviation obtained from interlaboratory tests performed by the *Centre for quality assurance for testing of dangerous goods and hazardous substances* (CEQAT-DGHS) to validate laboratory test methods. This paper outlines typical results (e.g. so-called "Shark profiles" – the probability of incorrect classification as a function of the true value estimated from interlaboratory test data) as well as general conclusions and steps to be taken to guarantee that laboratory test results are fit for purpose and of high quality.

Keywords: *interlaboratory test, validation, measurement uncertainty, incorrect classification*

1. Introduction

Accidents such as explosions in chemical plants and fires on dangerous goods vessels can be caused in several ways. Prevention starts in the laboratory, where chemicals are tested for their hazardous properties in order to be able to assess the risks involved in their handling. For this purpose, laboratory test methods have been developed and published (see e.g. laboratory test methods published by the European Union (2008) and by the United Nations (2019)); these methods are currently applied throughout the world. Laboratory test results (amongst others) are used to correctly classify and label

 DOI: 10.7795/810.20200724

chemicals as “hazardous” as defined in the UN Globally Harmonized System (GHS) / EC CLP Regulation or as “dangerous goods” as defined in the UN Recommendations on the Transport of Dangerous Goods.

Safety experts, manufacturers, suppliers, importers, employers and consumers must be able to rely on the validity of safety-related laboratory test methods and on the accuracy of laboratory test results and assessments. Interlaboratory tests play a decisive role in assessing the reliability of laboratory test results. Participation in interlaboratory tests is a crucial element of the quality assurance of laboratories; for this reason, it is explicitly recommended in ISO/IEC 17025 (2017) (assuming such interlaboratory tests are available). Interlaboratory tests are used in laboratory test method development and validation and can be used to determine measurement uncertainties (Hässelbarth, 2004; ISO 21748:2017-04, 2017).

Over the past 10 years, the Bundesanstalt für Materialforschung und -prüfung (BAM) and the Physikalisch-Technische Bundesanstalt (PTB) in cooperation with QuoData GmbH have carried out interlaboratory comparisons to evaluate various laboratory test methods (Lueth et al., 2019). Significant differences between the results of the participating laboratories were observed in all interlaboratory tests. The deviations in the laboratory test results were caused not only by malfunctions in the laboratory equipment and laboratory faults but also by deficiencies in the different laboratory test methods (see interlaboratory test reports of the *Centre for quality assurance for testing of dangerous goods and hazardous substances* (CEQAT-DGHS), www.ceqat-dghs.bam.de).

After analysing the results of these interlaboratory tests, the following conclusions can be drawn:

- To avoid any discrepancies in the classification and labelling of chemicals, the use of validated laboratory test methods should be state of the art, with the results accompanied by the measurement uncertainty and (if applicable) the probability of incorrect classification.
- There is a need for improvement and validation for all laboratory test methods examined. Thus, interlaboratory tests should initially focus on the development, improvement and validation of the laboratory test methods (including the determination of the measurement uncertainty) and not on proficiency tests.
- Laboratory management and the practical execution of laboratory tests need to be improved in many laboratories.
- The term “experience of the examiner” must be seen critically; long-term experience involving many laboratory tests does not necessarily guarantee correct results.

However, it is not currently clear whether the measurement uncertainty is sufficiently considered in practice when a chemical is classified as a particular packing group (PG) or as a hazard class based on the measurement results combined with the measurement uncertainties. Statistical methods and graphical tools must therefore be developed which will clearly define how the measurement uncertainty should be used in practice.

This paper addresses so-called “Shark profiles”, graphical indicators of the probability of incorrect classification of dangerous goods and hazardous substances on the basis of the measurement uncertainty obtained from special interlaboratory tests performed by CEQAT-DGHS to validate laboratory test methods. “Shark profiles” were developed to characterize the quality and suitability of the laboratory test method(s) with respect to the laboratory test objective, i.e. the classification of chemicals based on specific classification criteria (Antoni et al., 2010, Antoni et al., 2011 and Kunath et al., 2011). This objective is particularly important when there are several classification levels (e.g. packing groups PG or hazard classes) and the question of the reliability of the classification arises.

The principle of the calculation of “Shark profiles” for interlaboratory tests is explained by means of quantitative laboratory test results. This paper outlines typical results as well as general conclusions and steps to be taken to guarantee that laboratory test results are fit for purpose and of high quality.

DOI: 10.7795/810.20200724

2. Experiments

2.1 CEQAT-DGHS interlaboratory tests and minimum number of participating laboratories

As part of the CEQAT-DGHS programme, interlaboratory tests were carried out on various laboratory test methods and either qualitative or quantitative test results were obtained and evaluated.

An overview of the laboratory test methods currently listed in the CEQAT-DGHS interlaboratory test programme is shown in Fig. 1. This figure also indicates interlaboratory tests already performed by CEQAT-DGHS as well as the number of laboratories interested in taking part in these interlaboratory tests.

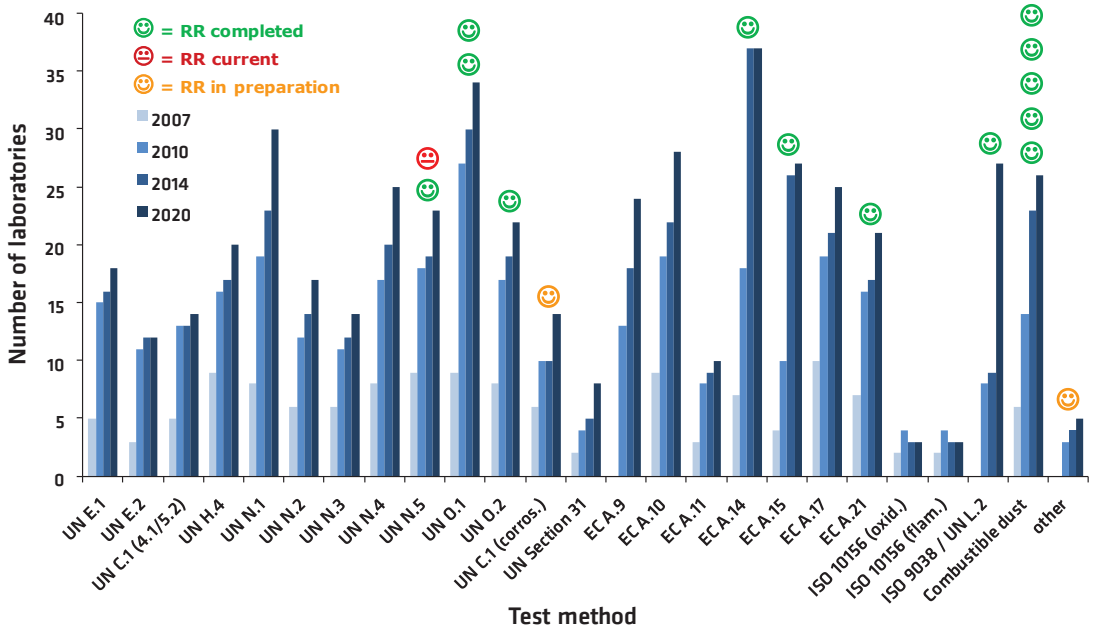

Fig. 1. *Laboratory test methods listed in the CEQAT-DGHS interlaboratory test programme, number of laboratories with interest in participation in CEQAT-DGHS interlaboratory tests and interlaboratory tests performed by CEQAT-DGHS since 2007 (RR = interlaboratory test)*

Since CEQAT-DGHS was founded in 2007, the number of laboratories interested in taking part in interlaboratory tests has steadily increased to about 98. The minimum number of participating laboratories required for statistically meaningful interlaboratory tests (i.e. for statistical reliability) has been met for almost all laboratory test methods. Hence, interlaboratory tests can now be carried out for almost all laboratory test methods listed in in Fig. 1.

2.1.1 Data verification (inspection upon receipt) in interlaboratory tests for method validation

A special feature of the CEQAT-DGHS interlaboratory comparison programme for method validation is that the data supplied by the laboratories are subjected to a strict verification before being evaluated. This is necessary to ensure data quality.

The specific review of the data submitted by the laboratories includes the following points (Antoni et al., 2010, Antoni et al., 2011, Kunath et al., 2011, Frost et al., 2016 and Lueth et al., 2019):

- Completeness of the data – check for missing data
- Conformity – check for irregular deviations from the laboratory testing method and/or from the interlaboratory test instructions
- Plausibility – check for obvious incorrectness of the data submitted (e.g. distorted data)
- Consistency – check the correctness of the values in the data input form submitted (e.g. by comparing them to raw data)

 DOI: 10.7795/810.20200724

The data were verified/validated by different experts independently directly after receiving the data from the laboratory and before starting the statistical analysis. If necessary and possible, faulty data were corrected after consulting the respective laboratory; laboratories were also asked to submit any missing data. A statistical evaluation including the determination of the measurement uncertainty and the probability of incorrect classification with corresponding "Shark profiles" was carried out on this specially validated data. The test results constituted an adequate basis for the statistical evaluation and for reliable conclusions.

2.2 Probability of incorrect classification and corresponding "Shark profiles"

The probability of incorrect classification of a tested substance, as defined by the European Union (2008) and by the United Nations (2019), depends on the measurement uncertainty of the test result(s) and the difference between the test result and the classification criterion. However, the percentage of laboratories which have incorrectly classified the test samples may be only a very rough estimate of the actual probability of the incorrect classification. Therefore, statistical methods and graphical indicators must be developed which can provide the probability of incorrect classification regardless of the substance and characterize the quality and the suitability of the laboratory test method(s) with regard to the test objective (i.e. the classification of chemicals based on special classification criteria). For this reason, the concept of the probability of incorrect classification and the visual shark profiles explained below were developed. The methods and development relevant to this topic are demonstrated below using several examples.

During the development of the "Shark profiles" the focus was on laboratory test methods whose quantitative test results are used to classify substances based on classification criteria defined by the European Union (2008) and the United Nations (2019). The interlaboratory tests evaluated the laboratory test methods listed in Table 1:

Table 1: *Quantitative laboratory test methods of the CEQAT-DGHS interlaboratory tests, year in which each interlaboratory test was performed and name of report published*

Laboratory test method	**Year of the interlaboratory test**	**Report of the interlaboratory test**
UN Test O.1 "Test for oxidizing solids"	2005-2006	Antoni et al., (2010)
UN Test N.5 "Substances which in contact with water emit flammable gases"/EC A.12 "Flammability (contact with water)"	2007	Kunath et al., (2011)
UN Test O.2 "Test for oxidizing liquids"/ EC A.21 "Oxidizing Properties (Liquids)"	2009-2010	Antoni et al., (2011)

The laboratory test methods shown in Table 1 were examined in interlaboratory tests for the improvement and validation of laboratory test methods (i.e. there were no proficiency tests). For reasons of simplification and better readability, an alternative packing group designation – PG 1, PG 2 and PG 3 – is used in the following instead of the legally correct designation – PG I, PG II and PG III.

2.2.1 Measurement uncertainty of laboratory test methods

The measurement uncertainty of the laboratory test methods can be determined efficiently via interlaboratory tests for method validation (Hässelbarth, 2004, ISO, 2017-04), and can be expressed as shown in eq (1).

'Laboratory result' = 'Measurement result' $\pm$ U,

where U denotes the expanded measurement uncertainty $U = k * u$ with $u = s_R$ (1)

 DOI: 10.7795/810.20200724

The factor k corresponds to the coverage factor k defined in GUM (Guide to Uncertainty in Measurement, Joint Committee for Guides in Metrology, 2008) and the factor s_R denotes the reproducibility standard deviation obtained in the interlaboratory tests for method validation.

The reproducibility standard deviation s_R is determined by means of standardized procedures and statistical methods (Antoni et al., 2011). These standardized statistical calculations are commonly used by statistical experts in proficiency testing and therefore do not require any further explanation at this point. However, detailed information on the calculation methods can be found in the respective interlaboratory test reports (Antoni et al., 2010, Antoni et al., 2011, Kunath et al., 2011).

2.2.2 Calculation of the probability of incorrect classification and "Shark profiles"

The procedure used to calculate the probability of incorrect classification and the "Shark profiles" is similar to statistical testing with a null hypothesis (e.g. the test sample belongs to PG 1) and with an alternative hypothesis (e.g. the test sample belongs to PG 2). In general, the objective of a statistical test is to keep the probability of a false positive result (rejection of null hypothesis and acceptance of alternative hypothesis) below a certain limit (statistical significance level), whereas the probability of the false negative error (acceptance of null hypothesis) can only be controlled with great difficulty. These probabilities always depend on the "true" value of the measurand (e.g. combustion time) for the specific test sample.

Using interlaboratory test data, the percentage of laboratories which incorrectly classified test samples provides only a very rough estimate of the actual probability of incorrect classification. A more reliable estimation of this probability can be obtained using the quantitative properties of the data. Here, the probability of incorrect classification of the test samples is derived by evaluating the ratios between the laboratory results of the test sample and the classification criterion (e.g. a firmly defined threshold value taken from a legal regulation or a laboratory test value obtained by testing a reference substance which characterizes the classification criterion (threshold value)). The procedure is as follows (and is initially carried out for each PG separately):

Step 1: Calculation of the characteristic statistical values and determination of the true value

Based on interlaboratory test data, the mean value and the standard deviation of the laboratory-specific ratios between the test results of the test sample and those of the reference sample are determined using the robust Q/Hampel method (described e.g. in ISO 13528:2015). Thus, with the interlaboratory test data, a robust estimate of the "true" ratio value (which is typically not exactly known) will be obtained as well as the respective reproducibility standard deviation regarding the ratios. The advantage of using robust methods is that no special outlier tests are required. Especially the Hampel estimator is a weighted arithmetic mean, with lower weights for the outlying values. Obvious "outliers" have a weight of zero and no influence on the mean value.

Step 2: Calculation of the probability of incorrect classification

Based on the "true" ratio value (i.e. the robust mean value), and the corresponding reproducibility standard deviation obtained in Step 1 above), the expected reproducibility standard deviation for the ratio between the test results of an arbitrary test sample and those of the reference sample (which characterizes the classification limit) can be estimated. Based on this estimation, the probability of the test sample being classified in a PG for more dangerous substances can be calculated as a function of the ratio of the "true" value of the measurand (see Fig. 2). Here, the "decision direction" (i.e. from a PG for less dangerous substances to a PG for more dangerous substances) must be taken into account. For the packaging groups, the order from PG 3 to PG 2 to PG 1 means an increase in the dangerousness of the substance. Once a certain classification threshold value (criterion for classification) has been reached, the substance is to be classified as a defined PG and then remains in this PG until the next higher classification criterion (i.e. for more dangerous substances) has been reached. Once the next higher criterion value (for the next PG for more dangerous substances) has

DOI: 10.7795/810.20200724

been reached, the substance is classified in the next PG for more dangerous substances. Below, Fig. 2 shows the probability of incorrect classification of the measured value for which the next-higher classification (i.e. for more dangerous substances, ratio < 1) is required.

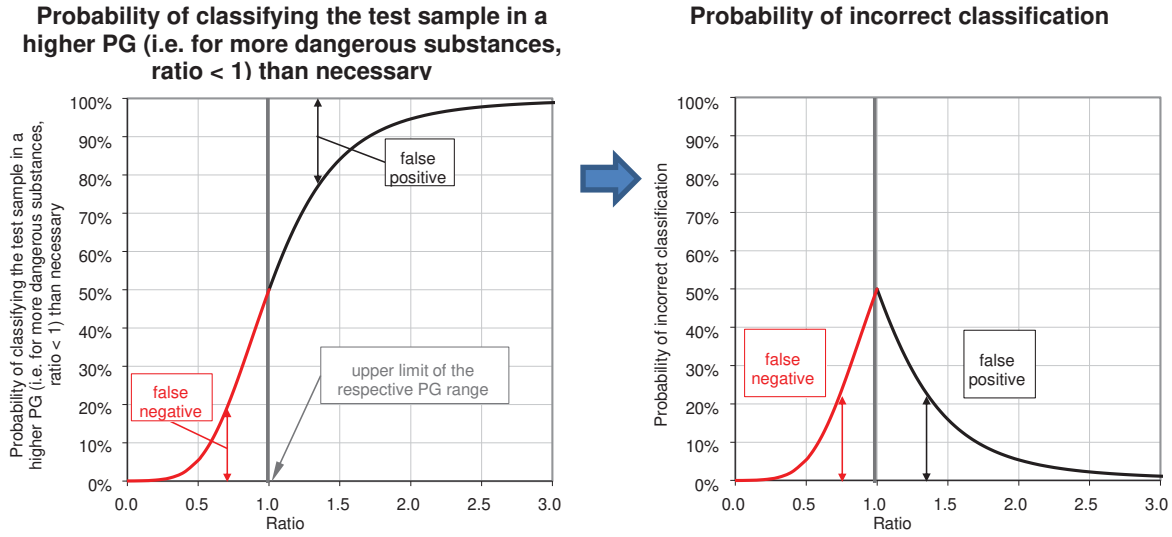

Fig. 2 *Probability of classifying the test sample in a higher PG (i.e. for more dangerous substances, ratio < 1) than necessary as well as the resulting false negative and false positive error (left diagram) and the resulting probability of incorrect classification (right diagram), (ratio = ratio between the laboratory results of the test sample and the reference substance/classification criterion)*

As given in Fig. 2 (left diagram), for a ratio less than 1, the resulting probability indicates the probability of a false negative classification, while for a ratio greater than 1, the probability of the false positive classification can be calculated by subtracting the given probability from 100 % (see Fig. 2, right diagram). For a ratio of exactly 1, the probability of both the false negative and the false positive classifications are equal to 50 %.

Step 3: Transformation of the ratio-based diagram into a “Shark profile” diagram

For the resulting “Shark profile”, the y-axis indicates the probability of false negative classification for arbitrary test samples with a value less than that of the reference sample as well as the false positive classification for arbitrary test samples with a value greater than that of the reference sample, which creates a shape like a shark’s dorsal fin. Instead of the ratios, the x-axis shows the value of the measurand as a continuous scale. The robust mean value of the reference sample (depending on the PG) is used for the transformation. It is therefore possible to depict the results in one figure for all PGs considered (see Fig. 3).

A laboratory can generally assess the test result of an arbitrary sample using the “Shark profile” as follows: Assume that the sample has been characterized as PG 2 in the measurement (see Fig. 3). Then, the corresponding right arm of the curve of PG 1 (red curve) can be interpreted as the probability that the sample was incorrectly classified as PG 2, although it belongs to PG 1. This probability naturally depends on the “true” value of the test sample (e.g. combustion time). If the probability is less than 5 %, it can be argued that the classification based on the test result is statistically significant (i.e. the assumption that the sample belongs to PG 1 would be rejected at a statistical significance level of 5 %). The level at which the 5% limit is attained can be considered the lower limit of the measurement uncertainty (k = 2) when the test result belongs to PG 2.

Conversely, in Fig. 3, the corresponding left arm of the curve from PG 2 (blue curve) can be interpreted as the probability that the sample is incorrectly assigned to PG 2, although it belongs to

 DOI: 10.7795/810.20200724

PG 3. The statistical significance level at which the 5 % limit is attained can be considered the upper limit of measurement uncertainty (k = 2) when the test result belongs to PG 2.

The probability of incorrect classification and the corresponding "Shark profiles" were calculated in three different CEQAT-DGHS interlaboratory tests. The interlaboratory tests on the methods given in Table 1 were used. The procedure for and calculation of the probability and the corresponding "Shark profiles" have been explained in greater detail in the respective interlaboratory test reports (Antoni et al., 2010, Antoni et al., 2011, Kunath et al., 2011). Examples of the “Shark profiles” are explained in the next chapter.

3. Results and discussion

In the following, it is assumed that a false positive classification indicates that the test sample mixture has been classified by a laboratory in a PG for more dangerous substances, although a PG for less dangerous substances would have been correct because this is the “true” packing group. A false negative classification, however, indicates that the test sample mixture is classified by a laboratory in a PG for less dangerous substances, although a PG for more dangerous substances would have been correct.

3.1 Example 1: Probability of incorrect classification and “Shark profiles” of the UN Test O.1 interlaboratory test for an arbitrary test sample mixture

The probability of incorrect classification can be calculated for the test sample mixtures used in the interlaboratory test as well as for an arbitrary test sample mixture. For example, in Fig. 3 below, the “Shark profiles” calculated on the basis of the UN Test O.1 interlaboratory test, “Test for oxidizing solids” (Antoni et al., 2010), are shown. Due to its curves, the profile in Fig. 3 is also referred to as a “Shark profile”. “Shark profiles” of incorrect classification as packing groups are to be interpreted in accordance with Table 2. The colours of the different curves represent the probability of incorrect classification according to the respective classification criteria of PG 1 (red curve), PG 2 (blue curve) and PG 3 (green curve). The probability of incorrect classification depends on the measurement uncertainty.

Table 2: *Interpretation of the curves of the probability of incorrect classification (“Shark profiles”) of the CEQAT-DGHS interlaboratory test of the UN Test O.1 method (Antoni et al., 2010)*

Arm of the “Shark profile” *	**Classification**	**Classification error**
Left arm of PG 1 (red)	Probability that a laboratory has classified the test sample as PG 2 although PG 1 would have been correct (i.e. the true classification is PG 1 for more dangerous substances)	false negative classification
Right arm of PG 1 (red)	Probability that a laboratory has classified the test sample as PG 1 although PG 2 would have been correct (i.e. the true classification is PG 2 for less dangerous substances)	false positive classification
Left arm of PG 2 (blue)	Probability that a laboratory has classified the test sample as PG 3 although PG 2 would have been necessary (i.e. the true classification is PG 2 for more dangerous substances)	false negative classification
Right arm of PG 2 (blue)	Probability that a laboratory has classified the test sample as PG 2 although PG 3 would have been correct (i.e. the true classification is PG 3 for less dangerous substances)	false positive classification
Left arm of PG 3 (green)	Probability that a laboratory has not classified the test sample although PG 3 would have been necessary (i.e. the true classification is PG 3 for more dangerous substances)	false negative classification
Right arm of PG 3 (green)	Probability that a laboratory has classified the test sample as PG 3 although no classification for less (no) dangerous substances would have been correct (i.e. the true classification is “no” PG)	false positive classification

*…Curve of the probability of incorrect classification according to the respective PG criterion

DOI: 10.7795/810.20200724

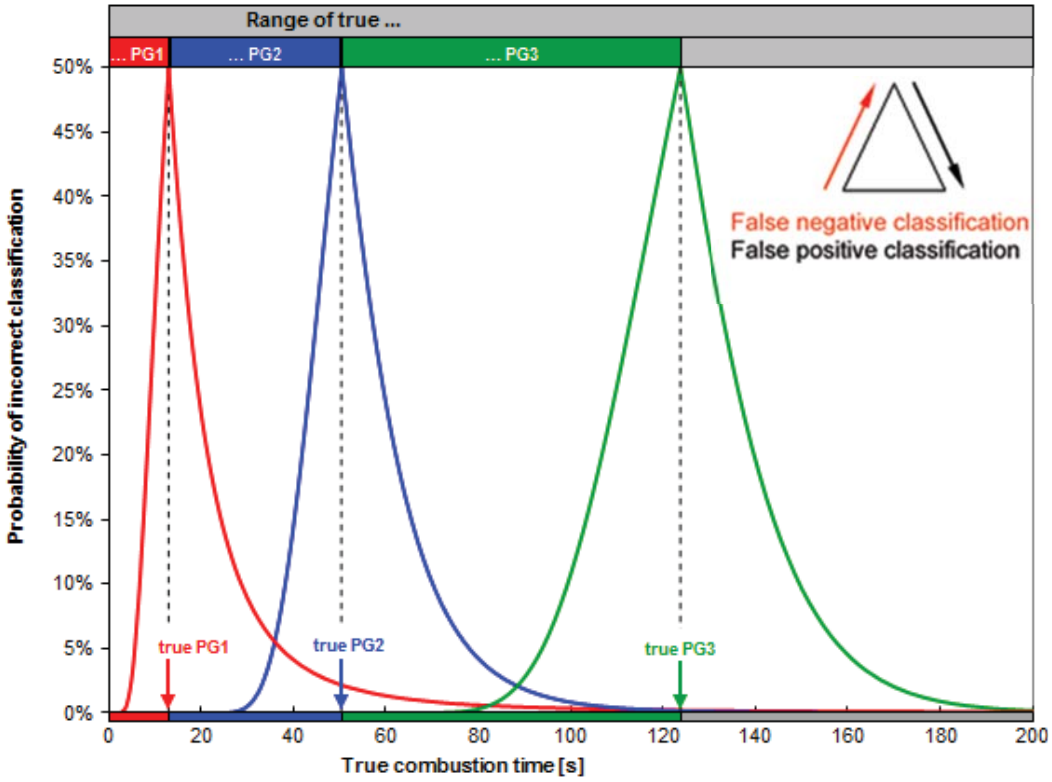

Fig. 3 *"Shark profiles" calculated for the UN Test O.1 interlaboratory test – probability of incorrect classification as PG 1, PG 2 or PG 3 or no classification as a function of the "true" combustion times of an arbitrary sample mixture (PG=packing group) (Antoni et al., 2010)*

If the curves in Fig. 3 do not overlap, only one kind of incorrect classification can occur. Therefore, in a range where two or more curves overlap, more than one type of error must be taken into account.

If, for example, the "true" combustion time of an arbitrary test sample mixture equals 40 s, the probability of a false positive classification with regard to PG 1 (i.e. the probability of the test sample mixture being classified as PG 1 instead of the "true" PG 2 due to the measurement uncertainty) equals about 4 % (value of red curve for 40 s). At the same time, for a mean combustion time of 40 s, there is the possibility of a false negative error as well: the probability of a false negative classification (i.e. the probability of the test sample mixture being classified as PG 3 instead of the "true" PG 2) equals about 15 % (value of blue curve for 40 s).

3.2 Example 2: Probability of incorrect classification and "Shark profiles" of the UN Test O.2/EC A.21 interlaboratory test for the "aqueous solution of sodium nitrate" test sample and for an arbitrary test sample mixture

The significance and advantage of the "Shark profiles" can be seen particularly well in UN Test O.2 interlaboratory test because, in this interlaboratory test for method improvement, two different series of tests (see Table 3) were carried out with the same interlaboratory test sample, "aqueous solution of sodium nitrate" (mass fraction 44.6 % g/g (checked by gravimetry)/44.7 % g/g to 45.0 % g/g (checked by means of ion chromatography)). Sodium nitrate (45 % aqueous solution) is listed in Table 34.4.2.5, "Examples of result", of UN Test O.2 as an example of an oxidizing liquid which must be classified as Division 5.1, PG 3 according to the UN Manual of Tests and Criteria (United Nations, 2019).

 DOI: 10.7795/810.20200724

Table 3: *Test series in UN Test O.2/EC A.21 (Antoni et al., 2011)*

Test series	**Test method**	**Instruction for the mixing procedure (short description)**
1	UN O.2/EC A.21 (without modification)	No special recommendation* ("should be applied as usual in your laboratory")
2 (optional)	Modification of UN Test O.2/EC A.21	Modified mixing procedure: 1. Use a ceramic mortar and a pestle 2. Homogenize with only very little force 3. Homogenize within exactly 2 minutes

The current application of UN Test O.2 in different laboratories and the classification error were assessed in the first part of the interlaboratory test (test series 1). The influence of the mixing procedure on the test results/classification was evaluated using a second, optional test (test series 2) the participating laboratories were requested to perform. The mixing procedure in this test series was standardized with regard to the mixing equipment and the mixing duration. The modification of the mixing procedure in test series 2 was based on a proposal by BAM.

The probabilities of a false positive classification of the “aqueous solution of sodium nitrate” interlaboratory test sample are given in Table 4 with regard to all classifications/packing groups in both test series.

Table 4: *Probability of a false positive classification as PG 1, PG 2 or PG 3 of the “aqueous solution of sodium nitrate” interlaboratory test sample depending on the test series (see Table 3) of UN Test O.2/EC A.21 (PG=packing group) (Antoni et al., 2011)*

Test series	**Packing group**	**Probability of incorrect positive classification**
1	PG 1	7 %
1	PG 2	19 %
1	PG 3	27 %
2	PG 2	10 %
2	PG 3	5 %

Additionally, for test series 1 and test series 2, the probability of incorrect classification can also be predicted for an arbitrary test sample mixture (see Fig. 4 and Fig. 5). The “Shark profile” curves for test series 1 in Fig. 4 should be read as follows:

If, for example, in test series 1, the "true" pressure rise time of an arbitrary test sample mixture equals 500 ms, the probability of a false positive classification with regard to PG 1 (i.e. the probability of the test sample mixture being classified as PG 1 instead of as the “true” PG 2) is about 30 % (value of red curve for 500 ms). At the same time, there is the possibility of a false negative error of classification for a mean pressure rise time of 500 ms as well: the probability of a false negative classification (i.e. the probability of the test sample mixture being classified as PG 3 instead of the “true” PG 2) is about 15 % (value of blue curve for 500 ms).

DOI: 10.7795/810.20200724

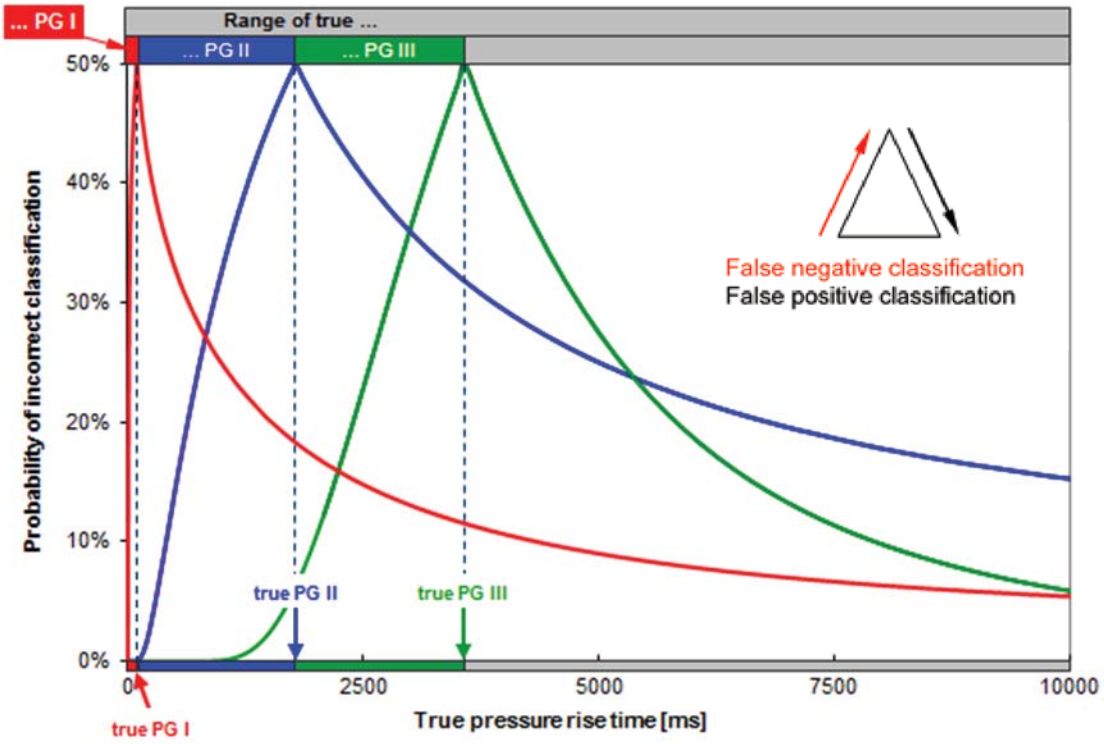

Fig. 4 *"Shark profiles" of test series 1 of the UN Test O.2/EC A.21 interlaboratory test – probability of incorrect classification as PG 1, PG 2 or PG 3 or no classification as a function of the "true" pressure rise time of an arbitrary sample mixture (Antoni et al., 2011)*

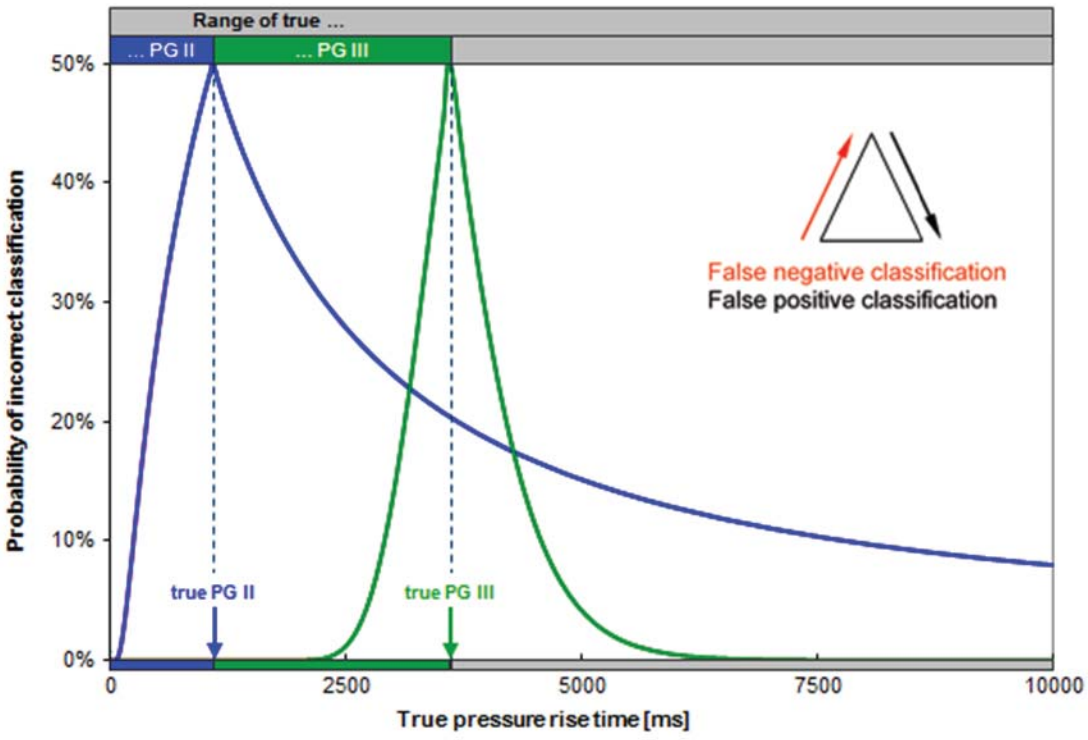

Fig. 5 *"Shark profiles" of test series 2 of the UN Test O.2/EC A.21 interlaboratory test – probability of incorrect classification as PG 2 or PG 3 or no classification as a function of the "true" pressure rise time of an arbitrary sample mixture (Antoni et al., 2011)*

 DOI: 10.7795/810.20200724

As can be seen in Fig. 4, the “Shark profiles” of the UN Test O.2 /EC A.21 interlaboratory test in test series 1 are rather blunt; thus, the classification error (i.e. the probability of incorrect classification) is relatively high and a clear distinction between the packing groups is not possible with acceptable certainty.

However, as shown in Fig. 5, in test series 2 of the interlaboratory test, the “Shark profile” curve of the probability of incorrect classification as PG 3 has become significantly narrower and more pointed, which supports the idea of better selectivity between the packing groups. This means that, even if the given laboratory test method is not yet optimal, the modifications in test series 2 (see Table 3) have already led to a significant improvement of the laboratory test method: “Shark profiles” are a useful graphic tool for clarifying the safety level of classification as a particular packing group or for deciding whether classification of a chemical substance in a particular packing group is safe.

4. Conclusion

A sound database is a prerequisite for the reliable evaluation of interlaboratory tests and for the correct validation of laboratory test methods. Within the scope of interlaboratory tests, the laboratory data submitted must be verified in order to validate laboratory test methods. Although it is very labour-intensive and time-consuming, this validation must not be neglected, as doing so will minimize the impact of deficiencies in the data on test conclusions. The comprehensive database compiled within the scope of the interlaboratory test, together with the verification of the data, created a clear overview of the probability of incorrect classification to be obtained and the corresponding “Shark profiles” to be determined.

Based on the results obtained to date regarding the probability of incorrect classification and the corresponding “Shark profiles”, the following conclusions can be drawn:

1. “Shark profiles” are a good graphic tool for clarifying the accuracy of classification and for deciding whether the classification of a chemical substance is accurate.
2. The probability of incorrect classification and the corresponding “Shark profiles” can be calculated via different laboratory test methods.
3. To avoid any discrepancy concerning the classification and labelling of chemicals, using validated laboratory test methods should be state of the art, with the results accompanied by the measurement uncertainty and (if applicable) the probability of incorrect classification and the corresponding “Shark profiles”.
4. By visualizing the probability of incorrect classification in “Shark profiles”, the quality of the laboratory test methods (and ultimately the quality of the classification of chemical substances as defined by the European Union (2008) and the United Nations (2019)) can be assessed in a simple manner. This statistical tool should be used in further method development tests and method validation.
5. The probability of incorrect classification and the corresponding “Shark profiles” can be used to discuss general quality requirements for classification rules in a simple manner. The probability can be used to define specific requirements for the quality of the laboratory test methods and for the corresponding classification (e.g. maximum permissible measurement uncertainties of the laboratory test method, tolerable probability, target measurement uncertainties and target probabilities of incorrect classification). The application of the measurement uncertainty of laboratory test results to the classification of chemical substances should be investigated in greater detail.

Continued efforts are necessary to further develop interlaboratory test programmes such as CEQAT-DGHS and to improve the corresponding laboratory test methods and the accuracy of the classification of chemical substances. Currently, the CEQAT-DGHS interlaboratory test programme is operated by BAM in collaboration with PTB and QuoData GmbH. Method validation by means of interlaboratory tests requires considerable effort with regard to time, money and personnel; the

 DOI: 10.7795/810.20200724

resources available at the CEQAT-DGHS competence centre limit the number of interlaboratory tests that can be offered to approximately one per year.

Interested laboratories can obtain detailed information and register for interlaboratory tests at the CEQAT-DGHS website (www.ceqat-dghs.bam.de).

Acknowledgements

The authors gratefully acknowledge the personnel and financial support provided by the Bundesanstalt für Materialforschung und -prüfung (BAM), Physikalisch-Technische Bundesanstalt (PTB), QuoData GmbH and by the laboratories participating in the interlaboratory tests.

References

Antoni S., Kunath K., Lüth P., Schlage R., Simon K., Uhlig S., Wildner W., Zimmermann C. (2010). Evaluation of the interlaboratory test on the method UN test O.1 'Test for oxidizing solids' with sodium perborate monohydrate 2005 / 06 Final report, *BAM, Berlin.* ISBN 978-3-9814281-2-4 https://opus4.kobv.de/opus4-bam/frontdoor/index/index/docId/25091.

Antoni, S., Kunath, K., Lüth, P., Simon, K., Uhlig, S. (2011). Evaluation of the interlaboratory test on the method UN O.2 / EC A.21 Test for oxidizing liquids 2009 - 2010 Final report, *BAM, Berlin,* ISBN 978-3-9814634-0-8, https://opus4.kobv.de/opus4-bam/frontdoor/index/index/docId/25090.

European Union (2008). Commission Regulation (EC) No 440/2008 of 30 May 2008 laying down test methods pursuant to Regulation (EC) No 1907/2006 of the European Parliament and of the Council on the Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH), OJ L 142, 31.5.2008, p.1.

Frost K., Lüth P., Schmidt M., Simon K., Uhlig S. (2016). Evaluation of the interlaboratory test 2015-2016 on the method DIN EN 15188:2007 “Determination of the spontaneous ignition behaviour of dust accumulations” Final report, *BAM, Berlin*. ISBN 978-3-9818270-0-2, https://opus4.kobv.de/opus4-bam/frontdoor/index/index/docId/38734.

Hässelbarth W. (2004). BAM-Leitfaden zur Ermittlung von Messunsicherheiten bei quantitativen Prüfergebnissen. *Forschungsbericht* 266. BAM, Berlin. ISBN 3-86509-212-8.

ISO 13528:2015-08 Statistical methods for use in proficiency testing by interlaboratory comparison, International Organization for Standardization, Geneva.

ISO/IEC 17025:2017-11 General requirements for the competence of testing and calibration laboratories, International Organization for Standardization, Geneva.

ISO 21748:2017-04 Guidance for the use of repeatability, reproducibility and trueness estimates in measurement uncertainty evaluation, International Organization for Standardization, Geneva.

Joint Committee for Guides in Metrology (2008) Evaluation of measurement data — Guide to the expression of uncertainty in measurement (GUM 1995 with minor corrections), BIPM <www.bipm.org/en/publications/guides/gum.html>, accessed 06.02.2020.

Kunath K., Lüth P., Uhlig S. (2011). Interlaboratory test on the method UN test N.5 / EC A.12 'Substances which, in contact with water, emit flammable gases' 2007 Short report, *BAM, Berlin*. ISBN 978-3-9814634-1-5, https://opus4.kobv.de/opus4-bam/frontdoor/index/index/docId/25094.

Lueth P., Frost K., Kurth L., Malow M., Michael-Schulz H., Schmidt M., Schulte P., Uhlig S., Zakel S. (2019). CEQAT-DGHS Interlaboratory Test Programme for Chemical Safety - Need of Test Methods Validation. *CET Chemical Engineering Transactions*, 77: 1-6. DOI: 10.3303/CET1977001.

United Nations (2019). Manual of Tests and Criteria, Seventh revised edition, *United Nations*, New York and Geneva, <www.unece.org/trans/danger/publi/manual/rev7/manrev7-files_e.html>, accessed 06.02.2020