

QAWG/03/06

COMMERCIAL-IN-CONFIDENCE

EURACHEM/CITAC Guide: The Expression of Uncertainty in Qualitative Testing

Committee Draft September 2003

LGC/VAM/2003/048



*Setting standards
in analytical science*

Expression of Uncertainty in Qualitative Testing

Contents

Introduction	1
Part 1: Uncertainties in qualitative testing and analysis	2
1 Introduction	2
2 Importance of uncertainty in qualitative testing	2
3 Forms of uncertainty information in qualitative testing	2
4 Nomenclature relating to qualitative testing uncertainties	3
5 The reliability of probabilistic information used to characterise uncertainties in qualitative testing	3
6 Reporting uncertainties in qualitative testing	3
7 Methods of evaluating uncertainties in qualitative testing	4
.7.1 False response rates from experiment.	4
.7.2 Predicted false response rates	4
8 The relevance of measurement uncertainty	5
9 The relevance of traceability	5
10 The current state of the art	5
11 Future developments	6
12 Implications	6
Part 2: Expression of uncertainty in identification	7
13 Summary	7
14 Introduction	7
15 Obtaining False Positive/Negative Rates	9
16 Additional Costs	12
17 The Limit of Detection (LOD)	12
18 Selectivity	12
19 Relevance	12
20 Terminology and definitions	13
21 Summary	13
Part 3: Examples	15
Example 1: A Reported Study on Identification Certainty in Mass Spectrometry Using Database Searching	15
Example 2: Chance Matches When Using an IR Database	18
Example 3: Sample databases for assessing drug identification performance	20
References	22

Introduction

The problem of the reliability associated with qualitative testing has received relatively little coverage in the literature compared to that afforded to measurement uncertainty. While a few authors have addressed this area,¹⁻³ much still remains to be done.

Uncertainty in relation to qualitative analysis is a topic of current interest to the EURACHEM Measurement Uncertainty Working Group. This report is intended to stimulate debate within the WG and to aid the development of policy in this area. The report is presented in two parts. The first part consists of a general overview of the main issues while the second part explores the use of several measures of reliability in more detail with an emphasis on the use of false response rates.

Part 1 of this paper comprises a Eurachem discussion paper first published in Accreditation and Quality Assurance. It aims to describe the present state of the art and to give an indication of what may reasonably be expected from laboratories, for example by accreditation bodies.

Part 2 sets out a range of existing measures of uncertainty in identification. In combination, these two parts could form the basis for formal Eurachem guidance on the topic. Comment is accordingly invited.

Part 1: Uncertainties in qualitative testing and analysis

1 Introduction

Uncertainties associated with quantitative measurement results have been the subject of considerable activity since the publication of the ISO Guide on the topic⁴. By comparison, the issue of uncertainties in qualitative testing and analysis (referred to elsewhere as “identification certainty”³) has received less attention. With the publication of ISO 17025:1999, however, interest in uncertainties in testing operations has increased. The problems of establishing uncertainty associated with qualitative tests, such as ‘pass/fail’, identity and comparative identity tests have accordingly become more important.

This paper sets out some of the main issues arising for analysts in testing laboratories and accreditation bodies interested in the assessment of uncertainty in qualitative testing. While it does not provide detailed statistical methods for the characterisation of uncertainties in qualitative testing, it does provide general guidance on the main issues.

2 Importance of uncertainty in qualitative testing

Broadly, qualitative testing provides a simple statement or categorisation of a test item or material. Decisions are invariably taken as a result; for example, whether or not to issue a batch of fertiliser, whether water is fit to drink, whether a person is in possession of controlled substance or not, or whether a newly synthesised material has the desired structure. Clearly, incorrect classifications – such as ‘passing’ a product when in fact it is unfit for use – carry risks to all parties. To control those risks, professionals involved in testing take pains to ensure that their methods lead to acceptably low risks of incorrect classification.

It follows that, at some point in the development of any such test method, an evaluation must be made as to the risk of incorrect classification. For most such methods, therefore, it is reasonable to expect a laboratory to establish, or have access to, information on the risks of incorrect results.

An important exception is the use of standard test methods, established by groups outside the particular laboratory as fit for the purpose in question. The laboratory may well have limited, or even no access to the risk information leading to that decision. However, such methods invariably specify a test procedure in some detail, and the laboratory will generally be expected to show that those factors which are within its control do indeed meet the requirements of the test method. That, in turn, may involve demonstrating that the uncertainty of reference values, calibration operations or intermediate measurements leading to a decision is sufficiently small.

3 Forms of uncertainty information in qualitative testing

Qualitative testing generally relates to categorical statements, such as ‘present/absent’, ‘pass/fail’, chemical species, or perhaps membership of a class of compounds. Such classification statements are not usually associated with a range of expression; one does not, in general reporting, generally speak of an artefact or material being a 90% pass, or 99% present*. The typical form of uncertainty

* Partial class membership is used extensively in “fuzzy logic” systems, but the relevant terminology and treatment is very rare in ordinary testing activities.

information is, as a result, typically probabilistic in nature. That is, one gives an indication of the probability of a given classification being correct.

The most familiar and widely used form of such information is, at present, the use of false response rates, particularly “false positive rates” and “false negative rates”.

Probably the most important alternative to simple statements of false response rates is the use of values derived from Bayes’ theorem (a summary of Bayes’ theorem is given in reference 2). Examples include likelihood ratio (an indication of the additional information provided by a test result) and posterior probability, an indication of the probability of an object fitting a given category given a test result. Bayesian estimates are particularly widely used in evaluating forensic evidence, for example DNA matching or blood group matching. Further details can be found elsewhere [ref. 2 and references cited therein]. Bayesian estimates can be calculated by appropriate combination of false positive and false negative rates.

4 Nomenclature relating to qualitative testing uncertainties

The nomenclature for qualitative testing is not fully developed. An example will illustrate a current problem. The term ‘false negative rate’ can, in principle, have two quite different interpretations.

- i) The chance, or frequency, of negative responses given that the response should be positive. Broadly, this is the fraction of ‘true positive’ test items that return negative responses.
- ii) The frequency of incorrect negative responses in a series of tests, that is, the fraction of the testing population which returns false negatives.

The difference appears subtle, but is important. In case i), the fraction is not expected to change with the number of ‘true positives’ in the population. This fraction could be established by appropriate method performance studies with known ‘true positive’ samples. But in the second case, a very small fraction of ‘true positives’ in the population leads to a very low fraction of ‘false negatives’ irrespective of the performance of the method. Current nomenclature in analytical chemistry does not distinguish these terms. It follows that there is a strong risk of confusion in using even familiar terminology at present.

5 The reliability of probabilistic information used to characterise uncertainties in qualitative testing

Because false response rates are, in general, low for effective methods, it often takes an extremely large number of experiments to obtain even indicative values. Further, observed false response rates are influenced very considerably by the characteristics of the test population. For example, false response rates are much higher when the typical level of a material falls close to the response threshold of a simple spot test. Thus, it is unrealistic to expect great reliability in false response rates obtained within a laboratory; it is often difficult to obtain false response rate figures accurate to within an order of magnitude.

Quantitative expression and reporting of qualitative testing uncertainties is accordingly unlikely to give indicative, but not very accurate information.

6 Reporting uncertainties in qualitative testing

Three main factors bear on the reporting of uncertainty information in qualitative testing. First, probabilistic statements are frequently misinterpreted by non-statisticians. Second, reliable figures are difficult to obtain by observation. Third, some indicators in common use are subject to

misinterpretation even by professionals. For these reasons, quantitative reporting of qualitative uncertainty information is very much the exception, rather than the rule. To underline the point: in one recent UK court case, the Judge ruled that the quantitative probability evidence presented by a leading forensic statistician and the accompanying information on its interpretation in relation to other evidence so confused the jury that the case was dismissed.

Where an indication of the test result's reliability is required, it may be most useful to adopt a semiquantitative reporting system. For example, forensic scientists in the UK have recommended a 'weight of evidence' scale along the lines of "indication"/ "strong indication"/"very strong indication", with each expression related to (overlapping) ranges of a Bayesian likelihood ratio.

7 Methods of evaluating uncertainties in qualitative testing

Broadly, there are two general methods of evaluating false response probabilities. The first relies on observation of false probabilities in a series of controlled tests. The second relies on prediction from known population characteristics, including statistics of quantitative measurements and known distributions of test sample characteristics in a population. The latter might include, for example, the observed peak incidence rate at different positions in an IR spectrum)

.7.1 False response rates from experiment.

False response rates are hard to observe in a realistic number of experiments unless the rate is high (near 50%). The most practical experiments thus concentrate on regions where false responses are likely. Typical approaches include

i) False positive rates in the presence of high levels of known cross-reacting interferences. In these experiments, the choice of interferent is critical; the experimenter must at a minimum observe false response rates in the presence of the worst case interferent at levels significantly above those found for the interferent in the normal test population.

ii) False negative rates at very low levels of analyte.

A related experiment involves chance mismatch studies in reference databases. In some cases, this allows the equivalent of many thousands of experiments. However, though informative and powerful, a current limitation is that such databases are often quite unrepresentative of the testing population; for example, while the prevalence of different materials in general use varies widely, a typical reference database will only contain one of each. This may lead to significantly biased probability estimates; again, the values obtained are unlikely to be better than order-of-magnitude estimates.

.7.2 Predicted false response rates

Examples include:

i) Prediction of chance spectroscopic peak matching from uncertainties in peak position, for example using binomial or hypergeometric statistics

Note: In estimating the probabilities of multiple events (e.g. a six-peak match in a spectrum), predicted probabilities are often extremely sensitive to choice of the probabilities assigned to individual events; predictions are therefore unlikely to be very accurate.

ii) Prediction of chance threshold exceedence from the known or estimated dispersion of measurement results or from the measurement uncertainty of the results.

Note: If normal distributions are assumed, probabilities fall off very sharply with increasing distance between threshold levels and typical levels of response. However, very considerable caution is advisable in extrapolating much beyond 95% confidence bounds. Due to such factors as human error, it is generally observed in routine measurements that the probabilities of very extreme results, though still quite low, are nonetheless many orders of magnitude higher than would be expected on the basis of the normal distribution.

8 The relevance of measurement uncertainty

Measurement uncertainty as described in the GUM⁴ impacts qualitative measurements in two ways.

- i) Control of uncertainties in test parameters, such as times, temperatures, lengths etc, is vital for reliable qualitative testing. Typically, a laboratory is expected to control factors affecting the test result to within well established tolerances, or to show that the uncertainty is sufficiently small to have no significant influence on the outcome of the test.

Note: Because false response rates are hard to measure, good data on the sensitivity of the test result to variation in input factors is subject to the same practical limitations as the determination of false response rates. Typical experiments would accordingly aim to demonstrate that substantial change – say, larger than 3-5 times the uncertainty – in a particular parameter had limited effect. It is unrealistic to expect quantitative sensitivity analysis in routine testing.

- ii) Measurement uncertainty related to intermediate measurements may inform predictions of false response rates (see above).

In either case, the laboratory will typically be expected to provide uncertainty estimates based on established principles (i.e. the GUM⁴). Of course, where equipment is calibrated by a third party, the relevant uncertainty values will usually be provided to the laboratory on calibration certificates etc.

9 The relevance of traceability

Traceability of measurement results, reference values and calibration values is essential in qualitative testing. It is particularly critical where the qualitative test relies on comparison with reference values (such as in comparing wavelength data in an IR spectrum with a reference database, or comparing melting point data with literature values). This follows from the observation that reference data is often collated by organisations and at times remote from the testing laboratory. Realistic comparison is only feasible if both the reference data supplier and the test laboratory are using measurements traceable to common references with acceptably small uncertainties.

A further important property of reference data, where used, is that its origin should be well established and the conditions under which it was obtained well documented.

10 The current state of the art

- Most competent laboratories currently evaluate one (the most critical for the application) of the false response rates, typically by experiments during method validation. Best practice involves stressing the limits of the method, for example by locating ‘worst case’ scenarios well outside normal usage or by progressive departure from normal operating conditions until false responses become significant.
- Few explore the less critical response rate, either because it is unimportant to the customer or because it is impractical.

- A few sectors have started to use Bayesian probability estimates in assessing the performance of qualitative tests; the forensic sector is probably the most advanced. Even here, direct reporting of probability information is rare because of uncertainties in the various terms needed for the estimate.
- Though there are publications on qualitative test failure probabilities and risks in the specialist literature, few laboratories can be expected to have access to the wide range of journals involved. Further, such papers tend to be written for specialists, and are accordingly not easy to implement for a routine test lab. There is thus little detailed and accessible guidance available to the general laboratory population.
- There is often sectoral or more general guidance on good practice in qualitative testing, and while this may not address uncertainty directly, it typically addresses other issues associated with quality control and assurance for the type of tests involved.

11 Future developments

There is a need to standardise the nomenclature relating to false response rates. There is an additional need to provide accessible and consistent guidance on the study of qualitative test performance.

EURACHEM is pursuing both these ends through the measurement uncertainty working group, and hopes to obtain wider input from other international organisations.

12 Implications

1. It is realistic to expect that testing laboratories have qualitative test method parameters (conditions of testing) under adequate control. Evidence of that will typically involve
 - clear evidence of traceability for the values of important control parameters prescribed by the method
 - evidence that uncertainties in these parameters are sufficiently small for the purpose
2. It is important for laboratories to check at least the most critical false response rate for a qualitative test.
3. It is reasonable to expect laboratories to be following published codes of best practice in qualitative testing where they are available.
4. Quantitative (i.e. numerical) reports of uncertainties in qualitative test results should not generally be expected.

Part 2: Expression of uncertainty in identification

13 Summary

A number of different measures of reliability for methods of qualitative analysis have been investigated. It is evident that the nomenclature for these measures is confusing and that different measures tell the analyst different things. Some guidance on when to use the different reliability measures would be useful. A further point is that the large amounts of practical experimental data required will generally be expensive to obtain, while inferences drawn from smaller data samples will have limited reliability.

14 Introduction

In analytical science, the purpose of qualitative analysis is to classify materials. In order to do this the materials of interest must first be detected. The ability of an analytical method to detect a target material depends upon the amount of the material which is present in the analytical system as well as upon the performance characteristics of the analytical method. Thus, if the aim of an analysis is to determine whether or not a particular substance is present, it will be necessary to specify a minimum concentration which must be capable of detection.

Just as it is possible to make an erroneous identification of a person under poor observation conditions so too is it possible to make an erroneous identification of a material submitted for qualitative analysis. It is hence desirable to provide users of qualitative analysis results with some indication of the reliability of an identification.

The degree of confidence in the correctness of an identification can be expressed in a number of ways. For a given test method, the basic properties that need to be measured are the numbers of true positive and negative results and the numbers of false positive and negative results obtained on a range of samples. From these numbers the fundamental measures of reliability *viz.* the *false positive* and *false negative* rates can be calculated. Several other measures can also be derived from these numbers (Table 1). The false positive and negative rates can be combined into a single figure expressed by the Bayesian likelihood ratio. If the analyst is able to quantify his initial degree of belief in the outcome of a test applied to a particular sample – before the test is applied – then a further reliability measure in the form of a Bayesian posterior probability can be calculated. One other important method parameter which needs to be determined is the *limit of detection*; knowing this enables the analyst to select a method capable of satisfying the customer's requirement relating to minimum detectable amount.

Table 1: Reliability Measures

Reliability Measure	Expression
False positive rate	$\frac{FP}{TN + FP}$
False negative rate	$\frac{FN}{TP + FN}$
Sensitivity	$\frac{TP}{TP + FN}$
Specificity	$\frac{TN}{TN + FP}$
Efficiency	$\frac{TP + TN}{TP + TN + FP + FN}$
Youden Index	Sensitivity + Specificity - 100
Likelihood Ratio	$\frac{1 - \text{False negative rate}}{\text{False positive rate}}$
Bayes posterior probability	Bayes rule

Where: TP = number of true positives; FP = number of false positives; TN = number of true negatives; FN = number of false negatives.

In Table 1 the terms *sensitivity* and *specificity* are used in the clinical chemistry sense *viz.*: sensitivity is the fraction of true positive results obtained when a test is applied to positive samples (it is the probability that a positive sample is identified as such); specificity is the fraction of true negative results obtained when a test is applied to negative samples (it is the probability that a negative sample is identified as such).

For the purposes of this study, the following definitions, based on those of AOAC, apply:

- True positive: Results obtained using the confirmatory technique and another analytical technique are both positive.
- True negative: Results obtained using the confirmatory technique and another analytical technique are both negative.
- False positive: Result obtained using the confirmatory technique is negative but that obtained using another analytical technique is positive.
- False negative: Result obtained using the confirmatory technique is positive but that obtained using another analytical technique is negative.

The false positive and false negative rates referred to in these definitions are based on those defined in the AOAC Research Institute Policies & Procedure document and commonly employed by clinical chemists *viz.*:

$$\text{False positive rate (\%)} = \frac{\text{false positives} \times 100}{\text{total known negatives}}$$

$$\text{False negative rate (\%)} = \frac{\text{false negatives} \times 100}{\text{total known positives}}$$

Since false positive/negative rates are interpreted as probabilities for the purpose of calculating the Likelihood Ratio they are expressed as fractions rather than as percentages in Table 1.

15 Obtaining False Positive/Negative Rates

For a given method, utilising a particular technique, it is necessary to be able to detect false positive and false negative results and this can be done by re-analysis using a different, confirmatory, technique. In most cases the confirmatory technique will be gc-ms but other techniques may be appropriate depending upon the nature of the analytical problem.

There are basically two ways of obtaining false response rates for a given analyte/technique combination. The first involves a review of the literature for the particular analyte and technique to see if studies on the false response rates have already been carried out and recorded. For analytes of general interest and commonly used techniques, this information might be expected to be in the public domain. For in-house methods the information should have been generated during method validation studies. Published false response rates should be used with caution; they will have been obtained using particular equipment, reagents and personnel and will refer to particular sample matrices and analyte levels so the analyst must consider whether his situation is likely to be comparable.

If information on the false response rates for a particular analyte/technique is not available it will have to be generated by an experimental study. The essential study parameters are:

- the analyte;
- the matrix;
- the analyte level;
- the detection techniques;
- the number of samples to test.

Two mechanisms can contribute towards the production of false responses. In the first of these, false responses are caused by sample matrix effects. One or more components of a matrix containing none of a target analyte can interact with the detection system to produce a false positive response. Similarly, one or more components of a matrix, other than the target analyte, can interact with the detection system to inhibit the production of a genuine positive response thereby leading to a false negative response.

A second mechanism can operate near the cut-off region of a test. Here, the number of false positives depends upon the distribution of values obtained on blanks. A cut-off value is selected - typically at a level of 3 standard deviations of the blank - below which values are regarded as negative and above which they are regarded as positive. Thus (for a 3 standard deviations cut-off) there is an *a priori* probability of obtaining 1 or 2 false positive results in every 1000 tests on genuinely negative samples.

As well as depending on the cut-off value, the number of false negatives is also influenced by the level of the analyte and the distribution of values that could be obtained at a given level. For high levels of analyte the likelihood of false negatives will be very low and for low levels of analyte it will be relatively higher. The false negative rate therefore depends upon the distribution of analyte values in the population being sampled.

Table 2: Effect of cut-off level on false response rates at low levels of analyte

Actual mean level	Cut-off level (arbitrary conc. units)					
	3		3.5		4	
	FP	FN	FP	FN	FP	FN
3	0.00135	0.50000	0.00023	0.69146	0.00003	0.84134
4	0.00135	0.15866	0.00023	0.30854	0.00003	0.50000
5	0.00135	0.02275	0.00023	0.06681	0.00003	0.15866
6	0.00135	0.00135	0.00023	0.00621	0.00003	0.02275

Table 2 illustrates the effect of cut-off level on the false response rates at low levels of analyte. The levels are expressed in arbitrary concentration units and the blank is assumed to have a mean of 0 and a standard deviation of 1. The actual levels are assumed to have means as indicated and standard deviations of 1. For each cut-off level and actual level, the table entries show the proportion of results falling below the cut-off level (false negatives) and the proportion of blank results falling above the cut-off level (false positives). It can be seen that, for a given cut-off level, the false positive rate is constant but the false negative rate decreases, as would be expected, with increasing analyte concentration.

The second mechanism corresponds to the problem of committing type 1 (false positive) and type 2 (false negative) errors and the analyst must decide on a suitable balance between the two. Raising the cut-off level reduces the probability of obtaining false positives but increases that of obtaining false negatives – and conversely. These ideas are illustrated in Figure 1.

Estimation of the false response rates of a method should ideally be designed into the method validation studies. At this stage the analyte and detection technique would of course be known but a study should ensure that an adequate range of matrices, likely to be encountered in practice, is covered. A confirmatory detection technique will also need to be selected and a method incorporating it validated. Given that the number of false responses should ideally be low, the problem arises of how many samples to test to be reasonably sure of finding a non-zero number of false responses. One way of doing this is to model the problem as a set of Bernoulli trials – see below.

From published information (see, for example, Ferrara⁵) it is evident that false positive or negative rates can be as low as 0.5% and in some cases even lower. For a range of false response probabilities, Table 3 shows the number of samples that would need to be analysed in order to be certain, to at least the confidence levels indicated, of finding one or more false responses.

Probability	Confidence Level	
	95%	99%
0.005	598	919
0.01	299	459
0.05	59	90

Table 3: Minimum number of analyses to find one or more false responses

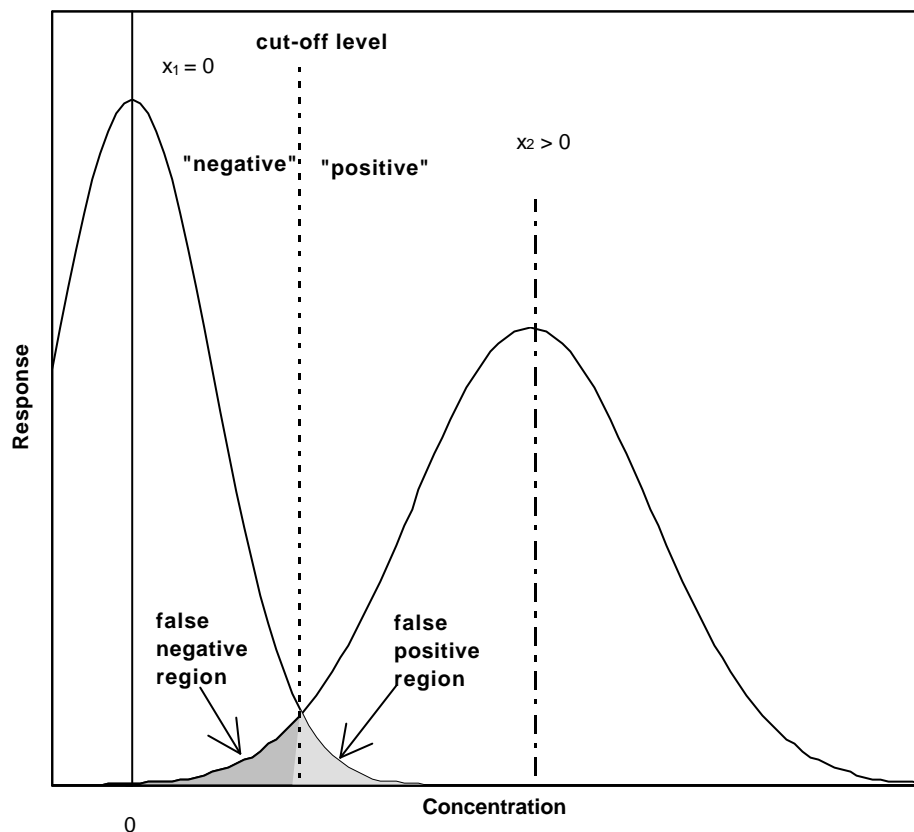


Figure 1: False response rates from distributions

In attempting to determine false response rates experimentally for a new method/analyte, the analyst is faced with a dilemma. On the one hand, since, for a given method, he does not know the false response rate of interest – this is what he is trying to determine – he cannot decide on an appropriate number of samples to analyse in order to be reasonably sure of detecting a false response. On the other hand, if he simply kept on testing until the first false response occurred this would not necessarily give a true picture of the false response rate (a false response could occur in the first experiment and then not again for a further 500 experiments!).

To get round this problem, it is suggested that the analyst decides in advance on tolerable levels for the two false response rates. For a chosen confidence level, he can then calculate, via a binomial distribution, the number of experiments needed to find one or more false responses. This approach is not guaranteed to produce an exact figure for the false response rate but it will at least place a bound on it. For example, if the analyst decides that a 5% false positive rate is acceptable and, if after performing 59 experiments (Table 1) covering the likely range of matrices, no false positives are found, then he can be reasonably certain that the true false positive rate is not greater than 5%. It is further recommended, as a quality control measure, that the samples be interspersed with blanks and standards containing the target analyte just above and below the method detection limit. When, as is usual, observed responses which correspond with expectation are not confirmed, the analyst should be aware that some of them may be false responses. When all observations are not confirmed, calculated false response rates should be treated with caution. It should always be remembered that false response rates cannot be viewed as exact values since they depend very much upon the vagaries of the population being sampled and also upon the method of sampling.

From Table 3, it can be seen that, for low false response rates, it may be impractical to analyse a sufficient number of samples to detect a false response. Accordingly, if a test is cheap to operate and/or is intended to be used with high sample numbers, *e.g.* as a drugs screening test, it may be preferable to establish first that the false response rate does not exceed an upper limit, say 5%, by

experiment, and then to refine this figure in the light of experience with further samples. Where sample numbers are likely to be relatively low and/or the test is expensive to apply, there may be little choice but to run the test in parallel with a confirmatory test (on all results!) and, from time to time, recalculate the false responses rates based on the experimental results.

16 Additional Costs

Positive test results are routinely confirmed by an independent method wherever the analyte is normally expected to be absent from the sample matrix. Negative test results are not usually confirmed, since this would add to costs. Similarly, if an analyte is normally expected to be present, a negative result would be confirmed but not a positive one. In adopting this policy analysts make the assumption that, in the first case, the negative results found are true negatives, and, in the second case, that the positive results found are true positives. This assumption may, on occasion, be incorrect but the analyst will never know. The point here is that, in order to calculate, say, a false positive rate, it is necessary to know the number of true negatives. Thus, if false response rates are to be determined reasonably accurately then all test results must be independently confirmed and additional costs must therefore be incurred.

17 The Limit of Detection (LOD)

The Limit of Detection for a qualitative analytical method, with respect to a given analyte, is found by applying the method to samples containing progressively smaller amounts of the analyte until the criteria for reliable detection are no longer met⁶. The concentration of analyte corresponding to this point is then the Limit of Detection of the method for the particular analyte.

18 Selectivity

Selectivity, in the sense in which this term is usually employed in analytical chemistry, refers to the ability of a method to discriminate between different components of a sample. It is particularly important when several components of a sample are similar with respect to the property being measured.

19 Relevance

In many cases where qualitative analysis is performed there is a requirement for confirmatory tests to be applied. This is particularly so when analysis is being performed for forensic purposes, for medical diagnostic purposes or where important financial or safety-critical consequences hinge on the result. In short, where the perceived consequences of an incorrect identification are seen as serious, confirmatory tests will be carried out as a matter of course. In these situations the analyst will be as certain as he can be of the reported identification and any measure of identification certainty should be so high as to be effectively redundant. A quality metric associated with a measurement/classification is only of use when it has the potential to influence a decision based on the measurement/classification.

Identification certainty is of use where the correctness of a presumptive identification is not critical but where the cost of confirmation is high. The end user of the result can see that the classification does not purport to be completely accurate and is further provided with a quantification of the degree of doubt attached to it. Alternatively, the analyst can use an identification certainty value [for a

classification], in conjunction with a rule or a set of criteria, to decide whether to carry out a confirmatory analysis.

20 Terminology and definitions

There is an issue surrounding the terminology used to describe the various method performance measures discussed. For example, for most analytical chemists, the term *sensitivity* is understood to refer to the rate of change of a response variable with respect to a control variable. The sense in which this term is used here, however, is quite different; it is the fraction of true positive results which respond as positives. It may be that for such multiple use terms the context can be taken to supply the meaning.

Definitions, on the other hand, cannot always be inferred from context. The definition of *false positive rate*, for example, used here is not intuitively obvious. Table 4 provides 4 different definitions of a false positive rate – all of them equally plausible – but only the last one corresponds to convention within the current context.

Table 4: Alternative definitions for the false positive rate

Formula	Definition
$\frac{FP}{P_{obs}}$	The fraction of observed positive results which are false.
$\frac{FP}{P_{obs} + N_{obs}}$	The fraction of all results which are false positives.
$\frac{FP}{TP + FN}$	The number of false positive results obtained for each true positive result.
$\frac{FP}{TN + FP}$	The fraction of true negative results which respond as positive.

Where: P_{obs} = number of observed positive results (true + false); N_{obs} = number of observed negative results (true + false).

There is thus ample scope for confusion when the terms discussed here are employed by different analytical sectors. AOAC International defines the false positive/negative rates in the clinical chemistry sense used here but the definitions are not stated in *Official Methods of Analysis*⁷] their publication to which analysts will most likely have access. The other main international body to which analysts might turn for guidance is IUPAC but the IUPAC Commission on Analytical Nomenclature in their recent 1995 report⁸ did not address this area of terminology at all. This may be a topic that EURACHEM could add to their other work on the clarification of terminology.

21 Summary

The fundamental measures of reliability for methods of qualitative analysis providing presumptive results are the false positive and false negative rates. A number of other measures can be calculated directly from these or from their basic component parts (the numbers of true and false positive and negative results observed in a sufficiently long series of trials).

If reliability is to be expressed in terms of false response rates then both the false positive and false negative rates must be quoted if a true picture of method performance is to be obtained.

Alternatively, both the sensitivity and specificity could be quoted. The Efficiency and Youden indices individually combine all of the information carried by the false positive and false negative rates (and also by the sensitivity and specificity); thus only one of these unitary measures need be quoted in place of one of the other associated pairs. The Likelihood Ratio also provides a single measure of method performance

Current practice is to subject samples to confirmatory tests only when the presumptive result is contrary to what would be expected from a reference population. This practice is driven by economic considerations but can lead to erroneous results being reported when an expected, and hence unconfirmed, result is in fact wrong.

Part 3: Examples

Example 1: A Reported Study on Identification Certainty in Mass Spectrometry Using Database Searching

Mass spectrometry, particularly in combination with a chromatographic separation stage, is a powerful tool that can aid in the identification of unknown compounds. For most purposes, low resolution mass spectrometry using electron impact (EI) ionisation is the method of choice when identification, as opposed to quantification, is required. A mass spectrum can contain many ions, not all of which are useful for diagnostic purposes, and this raises the question of whether there is a minimum number of ions which would be sufficient to ensure an unequivocal identification. Not much work on this question has been reported but a recent study by Webb and Carter [6] indicates that, generally, three ions may be sufficient.

The Webb and Carter study discusses earlier works by Sphon who investigated the minimum number of ions that needed to be monitored in order to produce an unambiguous identification of diethylstilboestrol (DES). Data relating to Sphon's most recent study, based on a mass spectral library containing about 270,000 entries, is presented in Table 1.

Table 1: Diagnostic Ions for DES

Ion, <i>m/z</i>	% RA Range	Matc hes
268	1-100	9995
268	1-100	
239	1-100	5536
268	90-100	
239	10-90	46
268	90-100	
239	50-70	9
268	90-100	
239	50-90	15
145	5-90	
268	90-100	
239	50-70	1
145	45-65	

RA = Relative Abundance

Table 1 shows that, when the relative abundance of each ion is considered, the number of matches occurring in a database can be reduced dramatically.

Although Sphon and others have recommended the use of a minimum of three ions for identification, the European Union (EU) requires a minimum of four ions when testing for veterinary drugs residues in cattle. The more stringent EU requirement has sometimes proved difficult to achieve and, it is suspected, has led to a number of false negative results being reported.

Webb and Carter, in a similar study based on a NIST database containing 62235 spectra, confirmed the results of Sphon and extended these through the inclusion of additional compounds of interest in the forensic and agro-chemical fields. A subset of their results is presented in Table 2.

Table 2: Diagnostic Ions Used in Webb and Carter Study

Compound	Ion, m/z	% RA Range	Matches
Heroin	369	1-100	1672
	369	1-100	
	327	1-100	526
	369	45-85	43
	369	45-85	
	327	60-100	1
DES	268	1-100	3597
	268	1-100	
	239	1-100	1597
	268	55-95	83
	268	55-95	
	239	30-70	4
	268	55-95	
239	30-70	1	
145	60-100		
DDT	352	1-100	1242
	352	1-100	
	235	1-100	234
	352	1-40	1140
	352	1-40	
	235	60-100	1
	352	1-40	
	235	1-100	7
	237	48-88	
	352	1-40	
235	60-100	1	
237	48-88		

RA = Relative Abundance

Using the criteria of De Ruig *et al*, the number of possible ions in a mass spectrum is 300 (from the m/z range 180-480). The number of combinations of 300 objects taken 3 at a time is $\frac{300!}{(300-3)!3!}$
= 4,455,100. Hence, for any three peaks, the chance match probability is taken to be ~1: 4.5×10^6 . The approach of De Ruig takes no account of the intensity information in a mass spectrum however, in the studies described above, such information has been utilised in order to reduce the number of matches to one. The effective chance match probability is therefore very much lower than the figure calculated.

Example 2: Chance Matches When Using an IR Database

The use of database statistics in evaluating criteria for qualitative analysis has been investigated by several authors. De Ruig *et al* [4] proposed criteria to be met by identification methods employed in veterinary drug residue identification. The authors give indicative values of chance match probabilities based on a simple binomial model. Ellison *et al* [5], commenting on this paper, noted that a hypergeometric distribution was a more appropriate model. The latter authors focused on the occurrence of chance matches when an infrared spectrum is compared against a spectral library.

The library used by Ellison *et al* was the Sadtler library containing spectra on 59,626 different materials. A random subset of thirty compounds was selected from this library and the number of peaks, q , in the range 500 - 1800 cm^{-1} noted for each compound. It was determined that the average number of peaks per spectrum in the region 500 - 1800 cm^{-1} , m , was 16. The spectral resolution available was 4 cm^{-1} and this implied the existence of $1300/4 = 325$, p , discrete peak positions in the 500 - 1800 cm^{-1} region. For each different spectrum in the chosen subset, the entire database was searched twice – first for a minimum of three matching peaks and the second time for a minimum of six matching peaks. The number of matches for each compound was compared with the number predicted on the basis of a hypergeometric distribution. For $n \geq 3$ the number of observed matches was about twice the predicted number. For $n \geq 6$, although the number of matches was considerably lower, as would be expected, the observed matches exceeded the predicted by a factor of ten. Part of the data for six peak matches is presented in Table 1.

Table 1: Chance matches against six peaks in an IR database

Compound	Peaks in range	Chance match probability	Pred. matches	Obs. matches
1-Chloro-3-(1-naphthoxy)-2-propanol	23	3.19×10^{-4}	19	192
α -Cyano-methyl ester- cinnamic acid	17	5.03×10^{-5}	3	29
Phenyl ν -triazolo-[1,5- α]-pyridin-3-yl ketone	24	4.19×10^{-4}	25	190
Benzo- β -thiophene-6-acrylic acid	20	1.34×10^{-4}	8	52
3-((Dipropylamino)methyl)1-5-nitroindole	17	5.03×10^{-5}	3	29
2-Mesityl-5-phenyl-oxazole	22	2.52×10^{-4}	15	99
<i>p</i> -Hydroxy-benzoic acid	18	6.71×10^{-5}	4	44
Caproic acid, isobutyl ester	8	1.36×10^{-7}	0	1
1-Bromoadamantane	10	9.64×10^{-7}	0	1
Phenyl propyl ether	17	5.03×10^{-5}	3	47

The calculated chance match probabilities for six peak matches were in the range 10^{-8} - 10^{-10} . The chance match probability for a compound when multiplied by the number of entries in the database gives an estimate of the number of compounds fitting the search criteria. In the case of two of the

compounds in Table 1, *viz.* Caproic acid, isobutyl ester and 1-Bromadamantane, the search criteria produce a single match and hence would appear to be adequate if these compounds are suspected. For the remaining compounds, many more matches are produced which indicates a requirement for more stringent criteria.

As stressed in the main body of this guide, reference databases – of which spectral libraries are one type – cannot be used to obtain information on false response rates. It is the responsibility of the analyst to decide which, if any, of a set of matches corresponds to an unknown.

Example 3: Sample databases for assessing drug identification performance

The use of sample databases for obtaining the relevant probabilities for a Bayesian analysis has been reported in the literature. S.D. Ferrara *et.al.*¹⁰, in testing for drugs of abuse in urine, assembled a database containing information on drug types, analytical techniques, false response rates for the techniques, and prevalence of the drug. For the authors' laboratory, Table 1 summarises part of this data for EMIT, an immunochemical technique.

Table 1: Bayes Probabilities for EMIT Technique

Description	Probability	Opiates	Methadone	Cocaine
Prevalence	$P(A)$	0.44	0.26	0.20
False Positive Rate	$P(e \neg A)$	0.028	0.004	0.009
False Negative Rate	$P(\neg e A)$	0.069	0.018	0.056

In the case of methadone, the Bayesian posterior probability is 0.988. In other words the analyst can be over 98% certain that a positive response for methadone genuinely indicates the presence of this drug.

Table 2 shows similar data for a different, non-immunochemical, technique. Note that the false positive rate for cocaine by this technique is reported as zero. It is debatable however whether the false response rates for such screening tests can truly be zero. In this case no false positives were found but, had more samples been analysed, it is possible that one or more false positives would have appeared.

Table 2: Bayes Probabilities for Toxi-Lab Technique

Description	Probability	Opiates	Methadone	Cocaine
Prevalence	$P(A)$	0.44	0.26	0.20
False Positive Rate	$P(e \neg A)$	0.038	0.012	0.000
False Negative Rate	$P(\neg e A)$	0.276	0.179	0.247

Considering methadone again, the Bayesian posterior probability is 0.960. This is a high probability though slightly less convincing than that produced by the EMIT test. If both tests are performed, and a positive response obtained in each case, then the combined Bayesian probability becomes 0.999

In this example, reliable prior probabilities are available in the form of prevalence values. Had these not been to hand, or if the analyst had preferred not to use them, likelihood ratios could have been used instead; the corresponding values being 246 (EMIT) and ~68 (Toxi-Lab). The combined likelihood ratio then being 16,830.

In all cases GC-MS was used as a reference technique to establish the false response rates. The particular database referred to here is quite comprehensive for the analytes of interest to the authors, and has clearly been designed to permit a Bayesian analysis of the data. There are inevitably some

missing values but, as more data is added, these should be reduced in number and the accuracy of predictions further improved.

A further advantage of a database set up to record Bayesian input data for several different techniques is the information it provides to enable one to optimise analytical performance. For example, by selecting a screening method with a low false positive rate this should minimise the costs of expensive confirmatory analyses. However, other factors also need to be taken into account such as the limit of detection of a technique, its false negative rate, and the speed of analysis.

References

1. de Ruig W.G; Dijkstra G.; Stephany R.W. *Anal.Chim.Acta* **1989**, 223, 277-82.
2. Milman B.L.; Konopelko L.A. *Fresenius.J.Anal.Chem.* **2000**, 367, 621-28.
3. Ellison S.L.R.; Gregory S.; Hardcastle W.A. *Analyst* **1998**, 123, 1155-61.
4. *Guide to the Expression of Uncertainty in Measurement*, ISO: 1993.
5. Ferrara S.D; Tedeschi L.; Frison G.; Brusini G.; Castagna F.; Bernadelli B.; Soregaroli D. *J.Anal.Toxicol.* **1994**, 18, 278.
6. *The Fitness for Purpose of Analytical Methods*, 1.0 ed.; Eurachem: 1998.
7. *AOAC Official Methods of Analysis*, 16th ed.; AOAC: 1995.
8. *Pure Appl.Chem.* **1995**, 67, 1699.