Technical Note 17 — October 2013

Issued: August 2004 **Amended and reissued:** December 2006, April 2009, March 2012, June 2012, October 2013

Guidelines for the validation and verification of quantitative and qualitative test methods

Guidelines for the validation and verification of quantitative and qualitative test methods

Table of contents

Guidelines for the validation and verification of quantitative and qualitative test methods

1. Introduction

A method must be shown to be fit for purpose so that a facility's customers can have confidence in the results produced by its application. Method validation and verification provides objective evidence that a method is fit for purpose, meaning that the particular requirements for a specific intended use are fulfilled. *Note: the term 'method' includes kits, individual reagents, instruments, platforms and software.*

For these reasons, method validation and verification are essential requirements of accreditation to ISO/IEC 17025 and ISO 15189. Accordingly, facilities accredited to these Standards must demonstrate the validity of all methods used by validating all in-house and modified standard methods and verifying standard methods. Validation is always a balance between costs, risks and technical possibilities. The extent of validation required will depend on the status of the method under consideration and the needs relating to its intended application.

If a facility wishes to apply a standard method that has been extensively validated via collaborative studies, e.g. ASTM and (Gold) standard methods (such as Australian Standard methods and ISO Standard methods) consideration should be given to the extent of method verification that is required. Method verification studies are typically less extensive than those required for method validation. Nevertheless the facility would be expected to demonstrate the ability to achieve the published performance characteristics of the standard method under their own test conditions.

This Technical Note describes the aspects of a method that should be considered when undertaking method validation or method verification, and provides guidance on how they may be investigated and evaluated. It is intended to be applicable to most fields of testing. This guideline does not cover sampling in connection with the performance of a method. For some testing facilities, not all of the validation and verification approaches described in this document are relevant. In particular for facilities involved in subjective testing (e.g. forensic, non-destructive testing and mechanical testing facilities) the more applicable section in this document may be Section 5.

An IUPAC Technical Report (Thompson et al., 2002), and other publications by the ENFSI Standing Committee (QCC-VAL-001, 2006), the Laboratory of the Government Chemist, UK, (LGC, 2003) and B. Hibbert (Hibbert, 2004) are acknowledged as key sources for the information and guidance provided in this Technical Note. Users of this Technical Note should note that although there are many publications and methods for validating and verifying different methods, no one method is universally agreed and approaches other than those set forth in this guideline may be applicable and acceptable. The guideline cannot as such be regarded as a procedure for method validation or verification in connection with the facilities' compliance with the requirements of ISO/IEC 17025 and ISO 15189. It is the responsibility of the facility to choose the validation or verification procedure and protocol most suitable for the desired outcome. However, it is important to remember that the main objective of validation or verification of any testing method is to demonstrate that the method is suitable for its intended purpose. References and additional reading material listed at the end of this document may provide useful and further guidance on the verification and validation of methods.

A number of examples from different fields of testing have been provided throughout this document and are intended for guidance purposes only. They are by no means exhaustive and other approaches may be more appropriate based on individual circumstances.

For testing where a qualitative outcome is reported based on a numerical value it is expected that method validation or verification is in line with quantitative procedures.

2. Verification of previously validated methods

Methods published by organisations such as Standards Australia, ASTM, USEPA, ISO and IP have already been subject to validation by collaborative studies and found to be fit for purpose as defined in the scope of the method. Therefore, the rigour of testing required to introduce such a method into a facility is less than that required to validate an in-house method. The same applies to peer accepted methods published in scientific literature along with performance data. Where a facility uses a commercial test kit in which the methodology and reagents are unchanged from the manufacturer's instructions, the kit does not need to be independently revalidated in the testing facility. Essentially the facility only needs to verify that their

operators using their equipment in their laboratory environment can apply the method obtaining the same outcomes as defined in the validation data provided in the standard method. Verification of methods by the facility must include statistical correlation with existing validated methods prior to use.

It must be noted however, that the documentation for standardised methods of analysis published by standardisation bodies and recognised technical organisations (e.g. AOAC), etc. varies. In some cases there is no validation report as a basis for the method of analysis, or the performance characteristics are not – or only partially – validated. If this is the case, verification of the facility's ability to use the method of analysis is not directly possible and validation is necessary.

Verification under conditions of use is demonstrated by meeting system suitability specifications established for the method, as well as a demonstration of accuracy and precision or other method parameters for the type of method. Method performance may be demonstrated by:

- blanks, or un-inoculated media (e.g. in microbiology), to assess contamination;
- laboratory control samples (e.g. spiked samples for chemistry or positive culture controls for microbiology) to assess accuracy;
- duplicates to assess precision;
- calibration check standards analysed periodically in the analytical batch for quantitative analyses;
- monitoring quality control samples, usually through the use of control charts; and
- participation in a performance testing program provided that the tested material is representative of the method in terms of matrix, analytical parameters, concentration level(s), etc.

Minor modifications to previously validated in-house methods (e.g. using the same type of chromatographic column from a different manufacturer, use of a different non-selective growth medium, differences in details of sample dilutions as a consequence of expected counts or a slight change in a non-critical incubation temperature) should also be verified to demonstrate that there are no changes to the expected outcome.

The key parameters to consider in the verification process will depend on the nature of the method and the range of sample types likely to be encountered. A statistically significant number of samples must be used in the evaluation process and these must cover the full range of results for the intended use. The measurement of bias and measurement of precision are minimum requirements for methods that yield quantitative results. For trace analyses the facility should also confirm that the achievable limit of detection (LOD) and limit of quantitation (LOQ) are fit for purpose. For qualitative methods, correlation studies with existing validated methods or comparisons with known outcomes are required. For diagnostic methods, clinical sensitivity and selectivity (specificity) should also be evaluated in specific, local patient populations (e.g. hospital, community patients) wherever possible. Ideally the facility will be able to demonstrate performance in line with method specifications. If not, judgment should be exercised to determine whether the method can be applied to generate test results that are truly fit for purpose.

Full validation is required if a facility has reason to significantly modify a standard method. It is impossible to define what constitutes a major modification, other than to say one that will affect the tests results. Some examples might be: use of a different extraction solvent; use of HPLC instead of GLC; differences in the formulation of the selective/differential medium (e.g. addition of an alternative antibiotic); different antibiotic concentration to the base medium that is specified; a change to a critical incubation temperature or time (e.g. 3 days rather than 5 days incubation); or different confirmation procedure (e.g. use of an alternative suite of biochemical tests other than those specified).

Additional validation must also be considered if the customer requires specifications more stringent than those for which the standard method has been validated.

The decision tree illustrated in Appendix 1 is intended to provide further clarification on when to perform method validation or verification.

3. Method validation and validation parameters

Non-standard and in-house-developed methods require method validation. For facilities involved in medical testing, elements of methods endorsed 'research use only' or 'not for diagnostic use' must also be validated by the facility before use for diagnostic purposes as outlined in the NPAAC publication *Requirements for the Validation of In-House In-Vitro Diagnostic Devices (IVDs).* Facilities that have modified kit components or the manufacturer's procedures must demonstrate equivalence or superiority of the modified procedure by putting the process into routine use. The procedure must be treated as an in-house test for validation purposes as

per the NPAAC publications *Laboratory Accreditation Standards and Guidelines for Nucleic Acid Detection and Analysis* and *Requirements for the Validation of In-House In-Vitro Diagnostic Devices (IVDs).*

The method's performance characteristics are based on the intended use of the method. For example, if the method will be used for qualitative analysis, there is no need to test and validate the method's linearity over the full dynamic range of the equipment.

The scope of the method and its validation criteria should be defined and documented early in the process. These include but are not limited to the following questions:

a) Purpose of measurement (what is to be identified and why)?
b) What are the likely sample matrices?
c) Are there any interferences expected, and, if so, should they be determined?
d) What is the scope (what are the expected concentration levels or ranges)?
e) Are there any specific legislative or regulatory requirements?
f) Are there any specific equipment accommodation and environmental conditions that need to be considered?
g) What type of equipment is to be used? Is the method for one specific instrument, or should it be used by all instruments of the same type?
h) Method used for the preparation, sub-sampling, procedure and including equipment to be used?

The following tools can be used to demonstrate the ability to meet method specifications of performance:

1. *Blanks:* Use of various types of blanks enables assessment of how much of the analytical signal is attributable to the analyte and how much is attributable to other causes, e.g. interferences. Blanks can also be used in the measurement of Limit of Detection.
2. *Reference materials and certified reference materials:* Use of materials with known properties or quantity values can be used to assess the accuracy of the method, as well as obtaining information on interferences. When used as part of the measurement procedure, they are known as measurement standards. When placed periodically in an analytical batch, checks can be made that the response of the analytical process to the analyte is stable. *Note: the same measurement standard cannot be used both for calibration and measurement of bias.*
3. *Fortified (spiked) materials and solutions:* Recovery can be calculated from results of analyses of samples fortified with a reference material.
4. *Incurred materials:* These are materials in which the analyte of interest may be essentially alien, but has been introduced to the bulk at some point prior to the material being sampled. Incurred materials may be used as a substitute for fortified materials.
5. *Replication:* Repeated analyses allow assessment of the precision of a measurement.
6. *Statistical data analysis:* Statistical techniques are employed to evaluate accuracy, precision, linear range, limits of detection and quantification, and measurement uncertainty.

Validation studies can be divided into comparative and primary validations.

Comparative validation
Comparative (i.e. correlation or cross) validation is usually applied to bioanalytical methods and aims to demonstrate equivalent performance between two (or more) methods used to generate data within the same study or across different studies by comparing the validation parameters. An example of comparative validation would be a situation where an original validated bioanalytical method serves as the reference and the revised bioanalytical method is the comparator.

There is no single test of establishing method equivalence or numerical acceptance criteria for it. Generally, a method with the greatest sensitivity or highest recovery for the target analyte is the best. To determine if the alternative method mean is not statistically different from the reference method mean, a one way analysis of variance or a paired t-test by sample type and analyte concentration is performed. Comparative validation studies of qualitative methods involve the identification of operating characteristics of the method (e.g. sensitivity, selectivity, presumptive false positive and presumptive false negative).

Validation studies can be supported by additional technical studies sourced externally from the facility. The use of verifiable proficiency testing data could be considered.

When sample analyses within a single study are conducted at more than one site or more than one facility, cross-validation with spiked matrix standards and subject samples should be conducted at each site or facility to establish inter-laboratory reliability. Comparative validation should also be

considered when data generated using different analytical techniques (e.g. LC-MS-MS vs. ELISA) in different studies are included.

AS/NZS 4659 *Guide to Determining the Equivalence of Food Microbiology Test Methods* is an example of a protocol that can be used for comparative validation. This Standard provides guidance on the validation of qualitative, quantitative, confirmation and antibiotic tests.

Primary validation
For situations where comparative validation is not applicable (e.g. in-house-developed methods, standard methods that have been modified in such a way that the final result could be influenced, standard methods used outside the intended scope, use of an alternative isolation or detection principle, as well as rapid methods), primary validation must be undertaken prior to introducing the method. In such cases validation becomes an exploratory process with the aim of establishing operational limits and performance characteristics of the alternative, new or otherwise inadequately characterised method. It should result in numerical and / or descriptive specifications for the performance of the method.

The first step in method validation is to specify what you intend to identify or measure; both qualitatively describing the entity to be measured and the quantity (if applicable). A method is then validated against this specification and any customer requirements.

The second step in validation is to determine certain selected performance parameters. These are described below. *Note: the parameters for method validation have been defined in different working groups of national and international committees. Unfortunately, some of the definitions vary between the different organisations. This document uses definitions based on the International Vocabulary of Metrology (VIM) (JCGM200, 2008) in most cases as these now supersede all others. However, some interpretation may be required across different fields, e.g. for veterinary testing it may be more applicable to refer to the OIE Terrestrial Manual (2009) for most of the relevant terminology used here*.

The sample size for determining the performance parameters may vary across different fields of testing however it has to be such that it is large enough to produce statistically valid results with methods such as the Student's t-test for assessing accuracy. A minimum of 7 replicate analyses conducted at each concentration for each determination and each matrix type is recommended. In reality this number is often surpassed. Generally speaking, the more samples that are tested, the greater the number of degrees of freedom, the better the statistical basis for the measurement result in question.

Matrix variation is, in many sectors one of the most important but least acknowledged sources of error in analytical measurements (IUPAC, 2002). Hence, it may be important to consider the variability of the matrix due to the physiological nature of the sample. In the case of certain procedures, e.g. LC-MS-MS- based procedures, appropriate steps should be taken to ensure the lack of matrix effects throughout the application of the method, especially if the nature of the matrix changes from the matrix used during method validation (FDA, 2001).

Each step in the method should be investigated to determine the extent to which environmental, matrix, material, or procedural variables can affect the estimation of analyte in the matrix from the time of collection of the material up to and including the time of analysis. In addition to the performance parameters listed below, it may also be necessary to assess the stability of an analyte when conducting validation studies. For example, many solutes readily decompose prior to chromatographic investigations (e.g. during the preparation of the sample solutions, extraction, cleanup, phase transfer or storage of prepared vials in refrigerators or in an automatic sampler). Points which may need to be considered include the stability of analytes during sample collection and handling, after long-term and short-term storage, and after going through freeze and thaw cycles and the analytical process. Conditions used in stability experiments need to reflect situations likely to be encountered during actual sample handling and analysis. The procedure should also include an evaluation of analyte stability in stock solutions.

An example of a stability test performed as part a method validation plan for tin (Sn) in canned fruits is provided below:

Example:
Analyte (Sn) in standard solution:
A freshly prepared working standard is compared to one that has been made and stored. Measurements are made at intervals over a specified time period to determine Sn stability in solution.

Analyte (Sn) in matrix:

A canned fruit sample is run at specified time intervals over the time that the sample would be typically stored to see if Sn levels degrade or concentrate. Standards used for the ICP-OES calibration curve are monitored to ensure they have not degraded or expired.

Analyte (Sn) in sample digest:
A canned fruit sample with a known concentration of tin is digested and measured daily for a period of a week.

Performance Parameters:

3.1 Range in which the calibration equation applies (linearity of calibration)

The linearity of the calibration of an analytical procedure is its ability to induce a signal (response) that is directly proportional to the concentration of the given analytical parameter.

Determination of linearity is applied to a calibration equation and thus only covers instrumental measurement. Linearity can also be investigated for the method as a whole and thus becomes an investigation of trueness as a function of the concentration of the analyte.

For instrumental analyses, the following protocols (Thompson et al., 2002; LGC, 2003) are recommended for establishing the validity of the calibration model as part of method validation:

- there should be six or more calibration standards (including a blank or calibration standard with a concentration close to zero);
- the calibration standards should be evenly spaced over the concentration range of interest. Ideally, the different concentrations should be prepared independently, and not from aliquots of the same master solution;
- the range should encompass 0–150% or 50–150% of the concentration likely to be encountered, depending on which of these is the more suitable; and
- the calibration standards should be run at least in duplicate, and preferably triplicate or more, in a random order.

A simple plot of the data will provide a quick indication of the nature of the relationship between response and concentration. Classical least squares regression, usually implemented in a spreadsheet program, is used to establish the equation of the relation between the instrumental response (y) and the concentration (x) which for a linear model is $y = a + bx$, where a = y-intercept of best line fit, and b = the slope of best line fit. The standard error of the regression ($s_{y/x}$) is a measure of the goodness of fit. The use of the correlation coefficient derived from regression analysis as a test for linearity may be misleading (Mulholland and Hibbert, 1997), and has been the subject of much debate (Hibbert, 2005; Huber, 2004; Ellison, 2006; Van Loco et al, 2002). The residuals should also be examined for evidence of non-linear behaviour (Miller and Miller, 2000). Graphs of the fitted data and residuals should always be plotted and inspected to confirm linearity and check for outliers. *Note: if variance of replicates is proportional to concentration, a weighted regression calculation should be used rather than a 'classic' (i.e. non-weighted) regression.*

Statistics are also well known for methods, where calibrations may give curved fits (e.g. quadratic fits for ICP-AES and ICP-MS analyses). Examples are provided in Hibbert (2006) of how parameters and measurement uncertainty of equations that are linear in the parameters, such as a quadratic calibration, can be derived.

If the relationship does not follow the expected linear model over the range of investigation it is necessary to either eliminate the cause of non-linearity, or restrict the concentration range covered by the method to ensure linearity. In some cases it may be appropriate to use a non-linear function, but care must be exercised to properly validate the chosen model. In general the range of calibration should cover only the range of expected concentrations of test samples. There is no benefit of calibrating over a wider concentration range than necessary, as the measurement uncertainty from the calibration increases with the range.

Calibration data can be used to assess precision (indeed $s_{y/x}$ can be used as the repeatability of y, or $s_{y/x}$/b for that of x). To calculate precision the curve should be prepared at least three times. *Note: from this data the limit of quantitation can be calculated.*

For non-instrumental analyses, linearity can be determined by selecting different concentrations (low, medium and high levels) of standards. The lowest level should fall at approximately the limit of detection, the medium and high levels one and two levels higher respectively (additional intermediate levels may be added to improve precision). The results can then plotted in the form of a 'response-curve'. An example is provided below. For microbiological analyses, results from counts obtained must be converted to log values and plotted.

Example:
The range of values that constitute the linear operating range of a diagnostic assay may be determined by a dilution series in which a high positive serum is serially diluted in a negative serum. Each dilution is then run at the optimal working dilution in buffer and the results plotted. Serum standards and other reagents can be used to harmonise the assay with expected results gained from reference reagents of known activity. The in-house serum controls (used for normalisation of data) and additional secondary serum standards, such as low positive, high positive, and negative sera (used for repeatability estimates in subsequent routine runs of the assay), can be fitted to the response curve to achieve expected values for such sera.

3.1.1 Measuring interval

The measuring interval is normally derived from the linearity studies. The measuring interval of a method is defined as the interval between the upper and lower concentration (amounts) of analyte in the sample for which it has been demonstrated that the method has suitable levels of precision, accuracy and linearity (i.e. results will have an acceptable level of uncertainty). In practice, acceptable uncertainties may be achieved at concentrations greater than the upper limit (beyond the extent of the determined linear range). However, it is more prudent to consider the validated range, i.e. the range between the LOQ and the highest concentration studied during validation. *Note: in some fields, the term measuring interval is also referred to as "working interval", "working range", "measuring range", or "measurement range".*

3.1.2 Matrix effects

Once the range in which the calibration equation applies has been determined the effect of the matrix on the analyte recovery must be determined. As a rule, a sample contains not only the substance to be analysed (the analyte), but also further constituents (foreign substances, accompanying substances).The constituents of the matrix may have the potential to alter the results or create an enhanced or suppressed response from an instrument detector.

Matrix effects are notoriously variable in occurrence and intensity but some techniques are particularly prone to them (their interference can, under given circumstances, be so great that the recovery of the analyte deviates from 100% to a significant degree). For example, in chemical testing matrix enhancement is a well-recognised occasional phenomenon in pesticide residue analysis using gas-liquid chromatography. In the case of samples with a complex and/or inexactly known matrix - for example, foods - it is particularly difficult to estimate the potential influence of the foreign substances on the analysis (matrix effect). For pathology, consideration needs to be given to haemolysed, lipaemic and icteric samples.

If no matrix effects are apparent, it is preferable to prepare calibration standards as simple solutions of the analyte. If matrix effects are suspected, they may be investigated by making standard additions of the analyte to a typical sample extract solution. The suggested number of determinations for matrix effects is at least once or in duplicate at each of 3 concentrations in each sample matrix type.

To determine the matrix effect on an instrument response, the range of concentrations by standard addition should be the same as that of the matrix-free calibration so that the slopes of both calibration plots can be compared for significant difference. If the slopes are not significantly different (<10%), there is no need to compensate for matrix effects. However, it must be noted that standard addition does not compensate for additive matrix effects.

For non-instrumental methods recovery tests described in Section 3.4.2 of this document can be performed to determine if there are any matrix effects.

If the results obtained for matrix fortified standards are lower (or higher) than the results obtained for pure standards taken through the complete analysis, the results may be due to low recovery of analyte from the matrix material (or the presence of interferences when high recoveries are obtained) or may be due to matrix suppression or enhancement effects changing detector response. To check on these possibilities, the facility can compare the results obtained from pure standards, pure standards taken through the complete analysis, standards spiked into blank matrix extract and standards added to matrix prior to extraction. The following determinations can then be made:

- Pure standards versus standards taken through the analysis is indicative of any losses of analyte which are related to the method, while enhanced results may indicate reagent contamination.
- Pure standards taken through the analysis compared with pure standards added to extracted or digested extracts provides an indication of matrix enhancement or suppression effects on the detection system.
- Pure standards added to blank matrix after extraction or digestion, compared to pure standards fortified in matrix prior to extraction or digestion, provides an indication of losses of analyte during processing (AOAC flworkshop).

3.2 Selectivity

It is important to establish during method validation that the test method is measuring only what it is intended to measure. In other words, the methods must be free from interferences which could lead to an incorrect result. The selectivity of a method is the accuracy of its measurement in the presence of interferences such as competing non-target microorganisms, impurities, degradents and matrix components. The terms selectivity and specificity have often been used interchangeably. The term 'specific' generally refers to a method that produces a response for a single analyte only, while the term 'selective' refers to a method that provides responses for a number of entities that may or may not be distinguished from each other. If the response is distinguished from all other responses, the method is said to be selective. Since very few analytical methods respond to only one analyte, the use of the term selectivity is more appropriate than specificity.

Methods that employ highly specific determinative procedures, such as chromatography/mass spectrometry, have the capability to be very selective. However, methods based on colorimetric measurements may be affected by the presence of coloured sample co-extracts or compounds with chemical properties similar to the analyte. While it is impractical to consider every potential interferent, analysts should call on their knowledge and experience to consider those scenarios of most relevance.

If required, the effect of potential interferents may be checked by analysing samples to which known concentrations of the suspected interferents have been added (one of each analysed once should suffice). This must be carried out on samples containing the analyte over the concentration range expected in practice (single point tests are acceptable but various points with varying amounts of inhibitor will add more data about the interference of different amounts of substance). If the method is intended to quantify more than one analyte, each analyte must be tested to ensure that there is no interference. An examination of the effects of interferences needs to be conducted – does the presence of the interferent enhance or inhibit detection of quantification of the measurands? In principle, methods must be developed to provide a level of selectivity without significant interferences. If detection or quantification is significantly inhibited by interferences, further method development will be required, but minor effects can be tolerated and included in the estimation of bias.

For methods requiring a confirmation step, e.g. for samples with low concentrations of organic compounds (including pesticide residues and other organic contaminants in food and environmental samples, and drugs and their metabolites in body tissues and fluids), positive identification of the trace amounts of organic compounds is required. 'Confirmation' applies to both the identity and concentration of residues. Confirmation of analyte identity and concentration for positive samples may be achieved using a different detection system or column or using a specific detection system such as mass spectrometry or using an alternate analytical technique (e.g. DNA sequencing; gel electrophoresis). In such cases validation of the confirmatory technique must also be performed.

Some examples for determining the selectivity of a method and some of the factors which need to be considered are provided below. Some additional examples for determining both selectivity and sensitivity appear in section 3.3.

Example:
Analytical selectivity of a diagnostic assay (which is defined as the proportion of samples from known uninfected reference specimens that test negative in an assay) may be assessed by use of a panel of samples derived from specimens that have been exposed to genetically related organisms that may stimulate cross-reactive antibodies, or sera from specimens with similar clinical presentations. This 'near neighbour analysis' is useful in determining the probability of false-positive reactions in the assay. It is also appropriate to document a group specificity criterion that includes detection of the analyte of interest in sera from subjects that have experienced infections/exposure to an entire group or serotype of organism of

interest. It is also important to evaluate the analytical selectivity of the assay using samples from animals or humans (whichever is applicable) that have been vaccinated. It may be necessary for an assay to distinguish between live virus, vaccinated strains and viral fragments depending on the intended use of the assay. If the assay targets an antibody elicited by a virus, vaccination against that virus may produce an antibody that interferes with the assay's inferences about infection. Also, if the viral antigen used in the assay is derived from a whole-cell viral culture preparation, containing antigenic reagents (carrier proteins, etc.) in addition to the virus, a vaccinated animal or human may test falsely positive due to detection of non-viral antibodies.

Example:
For microbiological analysis, selectivity is the fraction of the total number of negative cultures or colonies correctly assigned in the presumptive inspection (ISO 13843:2000). It is the probability that a test result will be classified as negative by the test method given that the sample is negative for the specific organism (SANAS TG28-02, 2008). For an alternate qualitative microbiological method, the concern of its ability to detect a range of microorganisms that may be present in the test article is adequately addressed by growth promotion in the media for qualitative methods that rely upon growth to demonstrate presence or absence of microorganisms. However, for those methods that do not require growth as an indicator of microbial presence, the selectivity of the assay for microbes assures that extraneous matter in the test system does not interfere with the test.

For microbiological methods involving a confirmation step, a presumptive positive result is taken through the cultural procedure and confirmed to be a positive or determined to be a negative. In other words, the confirmatory steps allow the sample to be reclassified as a known positive or a known negative. As such, the selectivity rate of results after confirmation is always 100%.

3.3 Sensitivity

The sensitivity (or inclusivity) of a method is the rate of change of the measured response with change in the concentration (or amount) of analyte (or microorganism). A greater sensitivity usually means a lesser dynamic range but also a lesser measurement uncertainty. For instrumental systems, sensitivity is represented by the slope (b) of the calibration curve ($y = a + bx$) and can be determined by a classical least squares procedure (for linear fits), or experimentally, using samples containing various concentrations of the analyte. Sensitivity may be determined by the analysis of spiked or artificially contaminated samples or standards prepared in sample extract solutions (an initial check should suffice). A mean slope with a high positive (e.g. >1) or low negative linear value (e.g. <-1) indicates that on average the method is highly sensitive. The greater the sensitivity (slope/gradient of the calibration graph), the better a method is able to distinguish small changes in analyte (or microorganism) concentration. However, a high sensitivity usually means a small dynamic range. If the sensitivity changes with day-to-day operating conditions the consideration of sensitivity during method validation may be restricted to ensuring a satisfactory, linear response is regularly achievable within the required concentration range. Sensitivity should be checked as part of a facility's ongoing quality assurance and quality control procedures.

Some examples for determining the sensitivity of a method have been provided below. The microbiological example provided also includes an example for determining the selectivity.

Example:
Analytical sensitivity of a diagnostic assay can be assessed by quantifying the least amount of analyte that is detectable in the sample at an acceptable co-efficient of variation. This can be done by limiting dilutions of a standard of known concentration of the analyte. However, such an objective absolute measure is often impossible to achieve due to lack of samples or standards of known concentration or activity. Another approach is to use end-point dilution analysis of samples from known positive specimens, to define the penultimate dilution of sample in which the analyte is no longer detectable, or at least, is indistinguishable from the activity of negative sera. When the results for the assay under development are compared with other assay(s) run on the same samples, a relative measure of analytical sensitivity can be estimated.

Example:
In microbiological terms, sensitivity is the probability that a specific method will obtain a confirmed positive result given that the test sample is a known positive. A known positive refers to the confirmation of inoculated analyte. To demonstrate selectivity and sensitivity, a test sample should be inoculated with strains of the specific microorganisms under test as well as strains that are considered as potentially competitive. The evaluation of selectivity and sensitivity is not applicable to total viable count, yeast and mould count or similar total enumeration methods that are not directed at specific microorganisms. It is recommended the selectivity and sensitivity is established by the analysis of at least 30 pure strains of the specific microorganisms being studied and at least 20 pure strains of potentially competitive strains [AOAC OMA Program Manual, 2002).

This number may need to be increased (e.g. for *Salmonella* methods, this number of target analyte strains is increased to at least 100 strains that are selected to represent the majority of known serovars of *Salmonella)*.

For a binary classification test, sensitivity is defined as the ability of a test to correctly identify the true positive rate, whereas test specificity is defined as the ability of the test to correctly identify the true negative rate. For example, when applied to a medical diagnostic binary classification test, if 100 patients known to have a disease were tested, and 46 test positive, then the test has 46% sensitivity. If 100 patients with no disease are tested and 94 return a negative result, then the test has 94% specificity.

In addition to sensitivity and specificity, the performance of a binary classification test can be measured with positive predictive values (PPV) and negative predictive values (NPV). The positive prediction value answers the question "If the test result is positive, how well does that predict an actual presence of, for example, a disease?". It is calculated as (true positives) / (true positives + false positives); that is, it is the proportion of true positives out of all positive results. (The negative prediction value is the same, but for negatives.)

It is essential to note one important difference between the two concepts. That is, sensitivity and specificity are independent from the population in the sense that they don't change depending on what the proportion of positives and negatives tested are. Indeed, you can determine the sensitivity of the test by testing only positive cases. However, the prediction values are dependent on the population.

		Condition		
		Positive	Negative	
Test outcome	Positive	True Positive	False Positive	→ Positive predictive value
	Negative	False Negative	True Negative	→ Negative predictive value
		↓ **Sensitivity**	↓ **Specificity**	

A worked example for determining the specificity and sensitivity of a veterinary binary classification test (a faecal occult blood (FOB) screen test used in 206 dogs to look for bowel cancer) is provided below:

Example :

		Dogs with **bowel cancer**		
		Positive	Negative	
FOB outcome	Positive	TP = 4	FP = 18	→ Positive predictive value = TP / (TP + FP) = 4 / (4 + 18) = 4 / 22 **= 18%**
	Negative	FN = 2	TN = 182	→ Negative predictive value = TN / (FN + TN) = 182 / (2 + 182) = 182 / 184 **≈ 98.9%**
		↓ **Sensitivity** = TP / (TP + FN) = 4 / (4 + 2) = 4 / 6 **≈ 66.67%**	↓ **Specificity** = TN / (FP + TN) = 182 / (18 + 182) = 182 / 200 **= 91%**	

Calculations:
False positive rate (%) = FP / (FP + TN) = 18 / (18 + 182) = 9% = 1 − specificity
False negative rate (%) = FN / (TP + FN) = 2 / (4 + 2) = 33% = 1 − sensitivity

Power = sensitivity = 1 − false negative rate
Likelihood ratio positive = sensitivity / (1 − specificity) = 66.67% / (1 − 91%) = 7.4
Likelihood ratio negative = (1 − sensitivity) / specificity = (1 − 66.67%) / 91% = 0.37

Hence with large numbers of false positives and few false negatives, a positive FOB screen test is in itself poor at confirming cancer (PPV=18%) and further investigations must be undertaken, it will, however, pick up 66.7% of all cancers (the sensitivity). However as a screening test, a negative result is very good at reassuring that the dog does not have cancer (NPV=98.9%) and at this initial screen correctly identifies 91% of those who do not have cancer (the specificity).

3.4 Accuracy

Accuracy is a property of a single measurement result and is influenced by both random and systematic errors, it is therefore not included as such in the method validation itself. The accuracy of a measurement result describes how close the result is to its true value and therefore includes the effect of both precision and trueness (expressed in the form of bias).

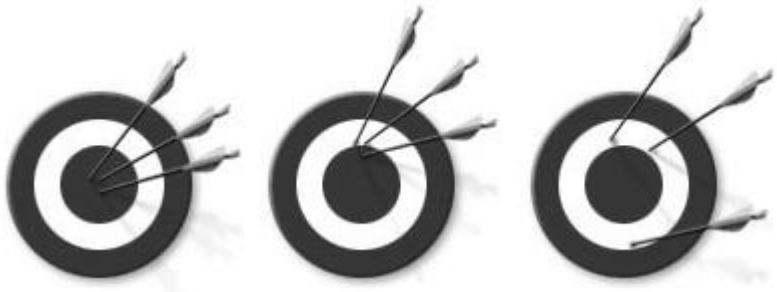
Precision relates to the repeatability and / or reproducibility condition of the measurement “getting the same measurement each time”. In other words the closeness of results of multiple analyses to each other. Precision is therefore a measure of the spread of repeated measurement results and depends only on the distribution of random errors – it gives no indication of how close those results are to the true value.

A measurement system can be accurate but not precise, precise but not accurate, neither, or both. For example, if an experiment contains a systematic error, then increasing the sample size generally increases precision but does not improve accuracy. Eliminating the systematic error improves accuracy but does not change precision. However, the validity of a measurement system is dependent on ‘fitness for purpose’ and hence the accuracy and precision need to be assessed against the validation criteria. There is no absolute scale.

We can never make a perfect measurement. The best we can do is to come as close as possible within the limitations of the measuring instruments.

The illustration below is used in an attempt to demonstrate the difference between the two terms, accuracy and precision.

Suppose you are aiming at a target, trying to hit the bull's eye (the center of the target) with darts. Here are some representative patterns of darts in the target:



A B C

A Accurate and precise. The darts are tightly clustered and their average position is the center of the bull's eye.
B Precise, but not accurate. The darts are clustered together but did not hit the intended mark.
C Accuracy and precision are poor. The darts are not clustered together and are not near the bull's eye.

Accuracy is also used as a statistical measure of how well a binary classification test correctly identifies or excludes a condition. That is, the accuracy is the proportion of true results (both true positives and true negatives) in the population.

		Condition		
		True	False	
Test	Positive	True positive	False positive	→ Positive

outcome				predictive value
	Negative	False negative	True negative	→ Negative predictive value
		↓ **Sensitivity**	↓ **Specificity**	**Accuracy**

3.4.1 Precision

ISO 3534-1(2006) describes precision as a measure of the closeness (degree of scatter) between independent test results obtained under stipulated conditions (stipulated conditions can be, for example, repeatability, intermediate precision, or reproducibility). The required precision is determined by the role the test results are going to play in making a decision. *Note: "Independent test results" means results obtained in a manner not influenced by any previous results on the same or similar test object but also the test materials have been independently prepared and so are considered random samples from the population that is being measured.*

Precision is usually expressed numerically by measures of imprecision, such as standard deviation (less precision is reflected by a larger standard deviation) or relative standard deviation (co-efficient of variance) of replicate results. However, other appropriate measures may be applied. In order for the stated precision to truly reflect the performance of the method under normal operating conditions, it must be determined under such conditions. Test materials should be typical of samples normally analysed. Sample preparation should be consistent with normal practice and variations in reagents, test equipment, analysts and instrumentation should be representative of those normally encountered. Samples should be homogeneous, however, if it is not possible to obtain a homogeneous sample precision may be investigated using artificially prepared samples or a sample solution.

Precision may vary with analyte concentration. This should be investigated if the analyte concentration is expected to vary by more than 50% of an average value. For some tests, it may be appropriate to determine precision at only one or two concentrations of particular significance to the users of test data, e.g. a production quality control (QC) specification or regulatory limit.

For single-laboratory validation, the best measure of precision is obtained by replicate analyses of independently prepared test portions of a laboratory sample, certified reference material (CRM) or reference material (RM), under normal longer term operating conditions. Usually this will involve the determination of intra-laboratory reproducibility as described below.

If data is available from precision experiments carried out on different samples, possibly at different times and there is no significant difference between the variances from each data set, the data may be combined to calculate a pooled standard deviation.

For a binary classification test precision is defined as the proportion of the true positives against all the positive results (both true positives and false positives). An accuracy of 100% means that the measured values are exactly the same as the given values.

$$\text{Precision} = \frac{\text{numberof truepositives}}{\text{numberof truepositives+ falsepositives}}$$

For comparing precision of two methods, the F-test is recommended. If the calculated test statistic (F = var1/var2) exceeds the critical value obtained from the statistical table, a significant difference exists between the variances of the methods and the null hypothesis is rejected.

For precision, two conditions of measurement, repeatability and reproducibility, are commonly quoted. AS 2850 (1986) provides guidance on this aspect of method validation. As repeatability and reproducibility will vary typically within the measuring range of a method of analysis, these should be determined at several concentration levels, with one of these levels being close to the lower value of the measuring range. With variations within the matrix in question, determinations at several levels will be necessary.

3.4.1.1 Repeatability

Repeatability is the precision estimate obtained when measurement results are produced in one facility and tests are performed on identical test items during a short interval of time by one operator using the same equipment under conditions that are as constant as possible (e.g. incubation time and temperature). It can

be expressed as standard deviation (*s*), variance (s^2), probability distribution function, etc for a suitable number of measurements made under repeatability conditions. If the method of analysis involves instrumental techniques the instrument repeatability needs to be determined in addition to method repeatability (examples are provided below). Repeatability gives an indication of the short-term variation in measurement results and is typically used to estimate the likely difference between replicate measurement results obtained in a single batch of analysis. However, it underestimates the spread of results that can be expected under normal operating conditions over the longer term

The repeatability standard deviation, variance, probability distribution function, etc must be determined with at least 6 degrees of freedom. This can be achieved for example, by analysing 7 times in a series with one test item (df=6), 4 times in a series with 2 test items (df=6), 3 times in a series with 3 test items (df=6), etc (Zar, 1974).

Instrumental repeatability may be determined by the injection of the standard solutions that are used to prepare the working calibration curve as well as an incurred or fortified sample at each of the spike levels 7 times. These injections should be done in random order to minimise bias. Calculate mean, standard deviation and percent relative standard deviation.

Method repeatability may be determined by preparing pools of sample material with levels of the analyte(s) at or near the concentrations used for method recovery studies. This may be done by using incurred material or by fortifying material (blank or incurred) with the required amount of the analyte(s). Replicate extracts are prepared of each of these samples and analysed by one analyst on the same day. Calculate mean, standard deviation and percent relative standard deviation.

For microbiological analyses, repeatability can also be assessed in those cases where samples are plated out in duplicate.

3.4.1.2 Reproducibility

Reproducibility is the precision estimate obtained when a series of measurements are made under more variable conditions, i.e. the same method on identical test items used by different operators with different equipment in different facilities at different times. It can be expressed as standard deviation (s), variance, probability distribution factor, etc. of suitable number of determinations on identical specimens analysed over several days with at least two different calibration standards. 'Intra-laboratory reproducibility', 'within-laboratory reproducibility' or 'intermediate precision' are terms used to describe the precision relating to reproducibility conditions restricted to a single facility. This involves making replicate measurements on different days, under conditions which mirror, as far as possible, the conditions of routine use of the method (e.g. measurements made by different analysts using different sets of equipment).

The reproducibility standard deviation can be determined by several samples run in the series or as the pooled standard deviation of a number of multiple double determinations run over several series. The number of degrees of freedom achieved with various combinations of numbers of series and numbers of specimens in each series is set out in the table below:

Number of samples per series	Number of series	Degrees of freedom for repeatability standard deviation	Degrees of freedom for facility reproducibility standard deviation
7	1	6	Not determined
4	2	6	7
3	3	6	8
2	6	6	11
n	*m*	(*n*-1)**m*	*n***m*-1

If the test method is to be used for the analysis of a range of sample type (e.g. different analyte concentrations or sample matrices) then the precision will need to be evaluated for a representative range of samples. For example, it is common for the precision of a test method to deteriorate as the concentration of the analyte decreases.

3.4.2 Trueness

Measurement trueness describes the closeness of agreement between the average of an infinite number of replicate measured quantity values and a reference quantity value. Lack of trueness indicates systematic error. Bias is a quantitative expression of trueness. The trueness of a result improves as bias decreases.

The bias of a measurement result may be seen as the combination of the bias of the method itself, laboratory bias and the bias attributable to a particular analytical run (analytical recovery). G.E. O'Donnell and D.B. Hibbert (2005) explain how the different biases are randomised as higher levels of communal trials are used to give a standard deviation.

Recovery tests are used to assess bias. Bias is measured by the detection of a known amount of an analytical parameter added to a specimen and included throughout the method of analysis. After deducting any detected content of the analytical parameter in question in the original specimen without addition, the recovery percentage can be calculated as a percentage of the amount added. A substantial departure from the accepted reference value is manifested in a high level of bias.

To take account any variation between runs, bias must be determined over several days and preferably throughout the measuring range and, if applicable, through the use of a suitable combination of different specimens. As the method of analysis cannot be expected to have the same bias throughout the measuring range (e.g. with non-linear calibration curves), several concentration levels must be incorporated in the determination of bias (at least one determination at a high level and a low level). Otherwise the facility must be able to prove that the method of analysis employed has the same trueness throughout the measuring range.

When certified reference materials (CRMs) are available to match the matrices and values of laboratory samples, they present the best option for estimating bias. Ideally, several CRMs with appropriate matrices and analyte concentrations should be measured.

However, for most test methods, suitable CRMs are not available, and alternatives are required to estimate bias. If suitable CRMs are not available, recovery may be estimated by analysing a reference material (RM) (provided they are matrix matched with the samples to be tested and sufficiently characterised with respect to the analytes of interest). Materials characterised by restricted collaborative testing may be suitable for the purpose.

If neither suitable CRMs nor RMs are available, bias may be investigated by the analysis of matrix blanks or samples unfortified and fortified (i.e. artificially contaminated) with the analyte of interest at a range of concentrations. In this instance the recovery (*R*) is calculated from the difference between the results obtained before and after spiking as a fraction of the added amount.

$$R = \frac{c_1 - c_2}{c_3}$$

Where: c_1 = measured concentration in fortified sample
c_2 = measured concentration in unfortified sample
c_3 = concentration of fortification

Note: Recovery is expressed as a percentage by multiplying the result by 100.

For some tests, e.g. pesticide residue analysis, facilities may be able to spike samples that have been determined not to contain detectable residues of the analyte(s) of interest. However, for many tests, it will be necessary to spike samples that contain natural concentration(s) of analyte(s).

In such cases, bias is estimated from the difference between results obtained for analysis of the sample in its spiked and original states. Caution is advised when evaluating bias from the analysis of spiked samples since the recovery may be better for spiked analyte compared to 'native' analyte, or incurred residues/contaminants. For example, whilst spiking drinking water with fluoride would allow a reliable estimate of recovery the same may not be true for spiking a soil with organochlorine pesticides. This is largely due to different extraction efficiencies for 'added' and 'native' analytes. If possible, spiked recovery data should be substantiated by some means; for example, participation in proficiency testing trials involving natural samples or samples with incurred residues/contamination.

In some cases, facilities will have to rely solely on spiked or artificially contaminated recovery data to estimate bias. In such instances, it should be noted that while a 100% recovery does not necessarily indicate trueness, a poor recovery definitely indicates bias, albeit a possible underestimate of the total bias.

For microbiological analyses, the inoculating organisms must represent different genera, species and/or toxin-producing microorganisms that are intended to be included in the method applicability statement. The choice of cultures should be broad enough to represent inherent variation in the microorganisms of interest.

A reference method (usually a recognised international or national standard method) with a known bias may also be used to investigate the bias of another method. Typical samples covering the range of matrices and analyte concentrations relevant to proposed testing programs are analysed by both methods. The significance of the bias of the test method may be estimated by statistical analysis (a t-test) of the results obtained.

3.5 Limit of detection and limit of quantitation

The determination of the limit of detection (LOD) or limit of quantitation (LOQ) is normally only required for methods intended to measure analytes at concentrations close to zero. There is no need to estimate the LOD or LOQ for methods that will always be applied to measure analyte concentrations much greater than the LOQ. However, the estimates often have great importance for trace and ultra-trace methods where concentrations of concerns are often close to the LOD and LOQ and results reported as 'not detected' may nevertheless have significant impact on risk assessments or regulatory decisions.

Separate determinations of limit of detection and limit of quantitation may be required for different matrices.

3.5.1 The limit of detection (LOD)

For most modern analytical methods the LOD may be divided into two components, the method detection limit (MDL) and instrumental detection limit (IDL)

The MDL is a term that should be applied to extraction and analysis methods developed for the analysis of specific analytes within a matrix. The MDL can be defined as the smallest amount or concentration of an analyte that can be reliably detected or differentiated from the background for a particular matrix (by a specific method). In other words, the LOD is the lowest value measured by a method that is greater than the uncertainty associated with it (Taylor, 1989). All matrix interferences must be taken into account when determining the MDL. Applied to microbiological tests, the MDL is the lowest number of microorganisms that can be detected.

The LOD of a method should not be confused with the lowest instrumental response. The use of a signal to noise ratio for an analytical standard introduced to an instrument is a useful indicator of instrument performance but an inappropriate means of estimating the LOD of a method. A large majority of analyses these days are performed on analytical instruments. Each of these instruments has a limitation on the amount of an analyte that they can detect. This limitation can be expressed as the instrument detection limit (IDL), which may be defined as the smallest amount of an analyte that can be reliably detected or differentiated from the background on an instrument (i.e. instrumental noise). As the sensitivity increases, the IDL decreases, and as the instrumental noise decreases, so does the IDL.

Several approaches for determining the detection limit are possible. Approaches other than those listed below may be acceptable.

3.5.1.1 LOD based on visual evaluation

Visual evaluation may be used for non-instrumental methods but may also be used with instrumental methods. It is also useful for establishing the LOD for qualitative measurements.

The detection limit is determined by the analysis of sample blanks samples with known concentrations of analyte and by establishing the minimum level at which the analyte can be reliably detected.

Sample blanks are spiked with an analyte at a range of concentration levels. At each concentration level, it will be necessary to measure approximately 7 independent replicates (as mentioned previously, in reality this number is often surpassed). Measurement of the replicates at various levels should be randomised. A response curve of percentage positive (or negative) results versus concentration should be constructed from the data, from which it should be possible to determine, by inspection, the threshold concentration at which the test becomes unreliable.

The limit of detection applied to microbiological tests, is the presence or absence of 1cfu in a fixed amount of sample. Although it can be investigated and proven in theory, in practice, this is very difficult to prove because of the presence of competing microorganisms as well as sample effects on the organisms. It is

therefore satisfactory to use test samples contaminated with 5-10 cfu to determine the limit of detection (SANAS TG 28-02, 2008).
Example:
One method to demonstrate the limit of detection for a microbiological quantitative assay is to evaluate two methods (alternative and compendial) by inoculation with a low number of challenge microorganisms (not more than 5 cfu per unit) followed by a measurement of recovery. The level of inoculation should be adjusted until at least 50% of the samples show growth in the compendial test. It is necessary to repeat this determination several times, as the limit of detection of an assay is determined from a number of replicates. The ability of the two methods to detect the presence of low numbers of microorganisms can be demonstrated using the Chi square test.

3.5.1.2 LOD based on the standard deviation of the blank

The detection limit may be determined by the analysis of a large number of blanks ($n \geq 20$ is recommended). Where independent sample blanks are measured once each ($n \geq 10$ is recommended) and independent sample blanks fortified at lowest acceptable concentration are measured once each ($n \geq 10$ is recommended). The LOD is expressed as mean sample blank value plus three standard deviations (+ 3*s)*.

3.5.1.3 LOD based on the range in which the calibration equation applies

If data on samples near or at the LOD are not available, parameters of the calibration equation can be used to estimate the instrumental LOD. Using the estimate of LOD as the blank plus three standard deviations of the blank, the instrument response to a blank is taken as the intercept of the calibration (*a*), and the standard deviation of the instrument response is taken as the standard error of the calibration ($s_{y/x}$). Therefore from the calibration equation if $y_{LOD} = a + 3\ s_{y/x} = a + bx_{LOD}$, then $x_{LOD} = 3\ s_{y/x}/b$. This equation is widely used in analytical chemistry. However because this is an extrapolation, the results cannot be as reliable as those from experiments made near the expected LOD, and it is recommended that samples with concentration near the estimated LOD be analysed to confirm that they can be detected with the appropriate probability.

3.5.1.4 LOD based on signal-to-noise

In the case of instrumental analytical procedures that exhibit background noise, a common approach is to compare measured signals from samples with known low concentrations of analyte with those of blank samples and establishing the minimum concentration at which the analyte can reliably be detected. Typically acceptable signal-to-noise ratios are 2:1 or 3:1.

3.5.2 The limit of quantitation (LOQ)

The limit of quantitation is also referred to as limit of determination, however, the term limit of quantitation is preferred to differentiate it from LOD. Similarly to the LOD, the LOQ can be divided into two components, method quantitation limit (MQL) and instrumental quantitation limit (IQL).

The MQL can be defined as the smallest amount of analyte that can be reliably identified and quantified with a certain degree of reliability within a particular matrix (by a specific method).

The IQL may be defined as the smallest amount of an analyte that can be reliably identified and quantified by the instrument.

The quantitation limit is expressed as the concentration of analyte (e.g. percentage, parts per billion) in the sample. Various conventions have been applied to estimating the LOQ. Depending on the level of certainty required (e.g. whether or not the analysis is for legal purposes, the target measurement uncertainty and acceptance criteria), the most common recommendation is to quote the LOQ as the blank value plus 10 times the repeatability standard deviation, or 3 times the LOD (which gives largely the same figure) or as 50% above the lowest fortification level used to validate the method. For greater certainty the LOQ can be quoted as ten times the LOD. Other factors can be used within certain testing fields and the facility shall in that case refer to the factor that is current within the testing field in question.

The limit of quantitation for a defined matrix and method may vary between facilities or within the one facility from time to time because of different equipment, techniques and reagents.

3.5.2.1 The limit of reporting (LOR)

The LOR is the practical limit of quantitation at or above the LOQ.

3.6 Ruggedness

The ruggedness (a measure of robustness) of a method is the degree to which results are unaffected by minor changes from the experimental conditions described in the method, for example, small changes in temperature, pH, reagent concentration, flow rates, extraction times, composition of mobile phase. Ruggedness testing provides an indication of the methods reliability during normal usage. The aim of ruggedness testing is to identify and, if necessary, better control method conditions that might otherwise lead to variation in measurement results, when measurements are carried out at different times or in different facilities. It can also be used to identify factors which need to be addressed to improve precision and bias. Ruggedness is investigated by measuring the effects of small, planned changes to the method conditions on the mean of the results. Trials need to be conducted with the aid of blank specimens, (certified) reference materials, specimens of known composition, etc. To evaluate the degree of ruggedness, significance testing may be carried out. In some cases, information may be available from studies conducted during in-house method development. Intra-laboratory reproducibility investigations, by their nature, take into account some aspects of a method's ruggedness.

Ruggedness testing can be carried out by considering each effect separately, by repeating measurements after varying a particular parameter by a small amount and controlling the other conditions appropriately (single variable tests). However, this can be labour intensive as a large number of effects may need to be considered. Experimental designs are available, which allow several independent factors to be examined simultaneously. Hibbert (2007) describes Plackett-Burman experimental designs that provide an economical and efficient approach whereby ($4n$ -1) variables are evaluated by conducting only $4n$ analyses. Both approaches assume independence of effects.

In practice, an experienced analyst will be able to identify those method parameters with the potential to affect results and introduce controls, e.g. specified limits for temperature, time or pH ranges, to guard against such effects.

3.7 Measurement Uncertainty

Measurement Uncertainty (MU) is the property of a measurement result, not a method. However, if a method is under sufficient statistical control indicative estimates of MU of typical measurement results can be quoted.

MU is defined as a parameter, associated with the result of a measurement, which characterises the dispersion of the values that could reasonably be attributed to the measurand (JCGM200, 2008). Knowledge of MU is necessary for the effective comparison of measurements and for comparison of measurements with specification limits. ISO/IEC 17025 and ISO 15189 require that facilities estimate and, where applicable, report the MU associated with results. Therefore the estimation of MU may be considered an essential requirement of method validation.

Numerous references are available that present different approaches for the estimation of MU, and hence MU is not dealt with at length in this document. ISO has published guidelines on the estimation of MU (ISO/IEC Guide 98-3, 2008) and an interpretative document by Eurachem/CITAC describes how they may be applied to analytical measurements (Eurochem/CITAC, 2000). These documents have now been supplemented by guidelines and examples from a number of other sources (UKAS, 2000; ILAC, 2002; APLAC, 2003; Magnusson et al., 2003; ISO/TS, 2004; Nordtest, 2005; Eurolab, 2007) aiming to provide facilities with more practical examples and simpler approaches which may be used to calculate reasonable estimates of MU. Excellent examples are also available from the website www.measurementuncertainty.org.

The information gained from other aspects of method validation, as described above, can provide a large contribution to a measurement uncertainty budget (but other components must be assessed too). These data can be supplemented with data from regular QC checks once the method is operational and data resulting from participation in relevant proficiency testing trials. Estimates may also be based on, or partly based on, published data and professional judgment. As with all aspects of method validation, estimates of MU should be fit-for-purpose. The required rigour for estimates will vary according to the rationale for testing; the principle being that estimates should be reasonable for the intended purpose. A reasonable estimate of MU may be obtained from consideration of long-term precision (intra-laboratory reproducibility) and bias. In some instances, other significant contributors to MU, e.g. purity of standards and metrological traceability, which may not be covered by these parameters, may need to be included in the estimation.

3.8 Summary

The tables below summarise the performance characteristic that should be considered when planning method validation and method verification investigations for both quantitative and qualitative methods and also include brief notes on how each performance characteristic may be determined.

	Validation		**Verification**	
Characteristics to be evaluated	Quantitative Method	Qualitative Method	Quantitative Method	Qualitative Method
Limit of detection* and quantitation	✓	-	✓	-
Sensitivity	✓	✓	✓	✓
Selectivity	✓	✓	✓	✓
Range in which the calibration equation applies (linearity of calibration)	✓	-	✓	-
Measuring interval	✓	-	✓	-
Matrix effects	✓	✓	✓	✓
Trueness; bias	✓	✓	✓	✓
Precision (repeatability and reproducibility) / Accuracy	✓	✓	✓	✓
Ruggedness	✓	✓	-	-
Measurement Uncertainty (MU)	✓	-	(1)	-

* Required for methods intended to measure analytes at concentrations close to zero
✓ Signifies that this characteristic is normally evaluated
\- Signifies that this characteristic is not normally evaluated
(1) If a well-recognised test method specifies limits to the values of the major sources of MU and the form of presentation of calculated results, the facility is considered to have satisfied the requirements of ISO/IEC 17025 or ISO 15189.

Characteristics to be evaluated	**Procedures which may be followed**
Limit of detection and quantitation	Replicate analysis at multiple concentrations including a concentration close to zero (graphical method), or replicate analysis at a concentration estimated to be equal to twice the LOQ (statistical method). Use blanks and a range of standards or samples containing low concentrations of (naturally occurring or artificially contaminated) analytes. Separate determinations may be required for different matrices.
Sensitivity	Analysis of spiked or artificially contaminated samples or standards prepared in sample extract solutions. Initial check for satisfactory gradient for plot of response vs concentration. (More appropriately a QC issue following initial check).
Selectivity	Analysis of reagent and matrix blanks, standards and matrix samples spiked with standards (in working range) to which known concentrations of suspected interferents have been added.
Range in which the calibration equation applies (linearity of calibration)	Duplicate measurements of standards evenly spaced over expected concentration range of samples.
Measuring interval	Evaluation of bias and possibly LOQ determinations
Matrix effects	Analysis of matrix blanks or matrix spiked with standards (at least once or in duplicate at each of 3 concentrations in each sample matrix type).
Trueness; bias	Analysis of replicates. Reference samples should be matrix and concentration matched with samples.

Characteristics to be evaluated	**Procedures which may be followed**
Precision (repeatability and reproducibility) / Accuracy	Replicate analysis for each sample matrix type (if possible selected to contain analytes at concentrations most relevant to users of test results) under stipulated conditions. For comparing precision of two methods, the F-test is recommended. For accuracy, compare each mixture's true value vs. the measured result.
Ruggedness	Introduce appropriate limits to method parameters likely to impact results if not carefully controlled. Investigate if necessary: i) single variable tests (test and re-test with small change to one method parameter); ii) multi variable tests (Plackett-Burman designed experiment).
Measurement Uncertainty (MU)	Calculate a reasonable, fit-for-purpose estimate of MU. Ensure estimates are aligned with the concentration(s) most relevant to the users of results.

Not all parameters need to be assessed for all methods. The rigour of validation should be sufficient to ensure that test results produced by a method are technically sound and will satisfy the client's needs. Well-planned method validation studies will be based on a clear understanding of the specific requirements for the method in use. Within this framework, carefully designed experiments will provide information to satisfy more than one of the parameters.

Facilities need to keep comprehensive records of method validation, including the procedures used for validation, the results obtained and a statement as to whether the method is fit for purpose.

Facilities must also continually check that a method of analysis meets the values for the performance characteristics documented in connection with validation. This can be achieved by for example tracking the behaviour of internal quality control samples over time. Should the method no longer produce results consistent with the original validation data, the method may be rendered unfit for its intended purpose.

4. Validation of non-routine methods

Frequently, a specific method is used for only a few sample analyses. The question should be raised as to whether this method also needs to be validated using the same criteria as recommended for routine analysis. In this case, the validation may take much more time than the sample analysis and may be considered inefficient, because the cost per sample will increase significantly. The answer is quite simple: Any analysis is worthwhile only if the data are sufficiently accurate; otherwise, sample analysis is pointless. The suitability of an analysis method for its intended use is a prerequisite to obtaining accurate data; therefore, only validated methods should be used to acquire meaningful data. However, depending on the situation, the validation efforts can be reduced for non-routine methods. The Eurachem/CITAC *Guide to quality in analytical chemistry: An aid to accreditation* (2002) includes a chapter on how to treat non-routine methods. The recommendation is to reduce the validation cost by using generic methods, for example, methods that are broadly applicable. A generic method could, for example, be based on capillary gas chromatography or on reversed phase gradient HPLC. With little or no modification, the method can be applied to a large number of samples. The performance parameters should have been validated on typical samples characterised by sample matrix, compound types and concentration range.

If, for example, a new compound with a similar structure in the same matrix is to be analysed, the validation will require only a few key experiments. The documentation of such generic methods should be designed to easily accommodate small changes relating to individual steps, such as sample preparation, sample analysis or data evaluation.

The method's operating procedure should define the checks that need to be carried out for a novel analyte in order to establish that the analysis is valid. Detailed documentation of all experimental parameters is important to ensure that the work can be repeated in precisely the same manner at any later date.

5. Validation and verification of subjective methods

According to ISO/IEC 17025 all technical procedures used by a facility must be fully validated or verified before use. Not only does this requirement apply to analytical methods but also to methods where the results are based on interpretive decisions or a degree of subjectivity (e.g. handwriting, firearms and audio analysis in a forensic laboratory, microscopic identifications, non-destructive testing methods, mechanical testing methods, etc.). This guideline is intended to provide a level of consistency, reliability and harmonisation for

the validation and verification of not only analytical methods as covered under sections 2 and 3 of this document but also for test methods based on interpretive decisions (i.e. a subjective test method).
A subjective test method is one where the experience and expertise of the analyst(s) or examiner(s) is a significant determinant on whether or not samples or test items are said to meet certain criteria. The assessment is based on qualitative data and the individual's expertise in interpreting that information. The validation of such methods is more challenging and less proscriptive than it is for analytical measurements.

Validation parameters such as linearity or limit of detection are not relevant in subjective tests. Precision and accuracy could be relevant and may need to be considered.

Parameters which should be considered are:

- Repeatability/reproducibility; and
- Probability of detection and/or potential error rate.*see note 1 below

The following mitigating or control measures may also be relevant:

- Critical (or risk) analysis;
- Collaborative exercises such as inter- and intra-laboratory comparison testing;*see note 2 below
- Quality control;
- Competency levels of staff involved; and
- Level of acceptance of test method.

The following questions should also be considered:

Has the technique been tested in actual field conditions (and not just in a facility), if applicable? Can the probability of detection and error rate be reliably estimated?

Note 1: Facilities may not have the means of **reliably estimating probability of detection or potential error rate for interpretive/comparative methods. This is due to the influence of the individual examiner on these parameters.*

**Note 2: Inter-laboratory comparisons may lead to heightened uncertainty.*

Validation Parameters

5.1 Repeatability/reproducibility (i.e. reliability)

Subjective test methods can be said to be valid if the analyst or examiner repeatedly obtains correct results for positive and negative known tests. Evidence of repeatability could include providing evidence of:
-The level of agreement between two different analysts or examiners who have assessed the same evidence of performance for a particular method (i.e. inter-analyst or examiner reliability); or
-The level of agreement of the same analyst or examiner who has assessed the same evidence of performance but at different times (i.e. intra-analyst or examiner reliability).

Other forms of reliability that should be considered include the internal consistency of a test (i.e. internal reliability) as well as the equivalence of alternative assessment tests (i.e. parallel forms).

5.2 Probability of detection and/or potential error rate

When validating or verifying subjective qualitative test methods the probability of detection (POD) and potential rate of error need to be considered. In addition to the POD and potential rate of error, sensitivity and selectivity are also identified as critical performance characteristics in the validation of qualitative identity methods. The POD and potential error rate are expressed as a percentage. The POD is the likelihood that a test methodology will correctly identify the measurand. Empirical determination of POD for new situations is not always practical due to lack of time, test specimens and other limited resources. Physical models can provide a useful alternative to empirical determination for quantifying POD in situations when data is limited but may not account for all factors and sources of variability. Statistical methods can be used to quantify and adjust for systematic biases and sources of variability that are not accounted for in the physical model. Hence, combining physical models with an empirical adjustment has the potential to provide a workable methodology.

The potential error rate, often determined through the participation in proficiency testing studies, is also known both as the "level of confidence" of the hypothesis test and as the level of statistical significance of the test's results. It has no uniform definition. It can mean false positive rate, which is the percentage of false positives divided by the combination of true negatives and false positives; percentage false positives, which is the percentage of false positives divided by the combination of true positives and false positives; or any number of different possibilities that may vary greatly even within one sample.

Control Measures

5.3 Critical (or risk) analysis

A critical analysis (or risk analysis) needs to be performed. This includes identification of those possible sources of errors and process variations. Once the sources are identified, their impact on the process need to be evaluated. The probability of detection (POD), the possible occurrence of the problem (O), and the gravity of the resulting consequence (G) are evaluated. The critical parameter (C) can then be calculated (C=POD*O*G), and represents a quantitative measure of the criticality of each possible source of non-conformity. Appropriate solutions can be proposed for each possible source identified to reduce the critical (C) to an acceptable level, by either improving the detectability of the problem (and reduce factor POD) and/or reducing the possible occurrence of the problem (and reduce factor O). The calculation of the critical factor (C) can be performed again, considering the solutions proposed (Ethier, 2005).

5.4 Collaborative exercises

Participation in collaborative exercises such as intra- and inter-laboratory comparison testing is important to determine the limits of the methods. It is also an important source of error rate determinations.

5.5 Quality control

Any subjective test must have built in quality control procedures. Test methods should include procedures for the inclusion of negative and positive controls whenever possible. For example, the result of a subjective test can be checked by other competent analysts where practicable. The check should be carried out without access to the opinion of the first analyst. The results of such checking can be collated to demonstrate the reliability of the method. Care must be taken, however as the reliability may be affected by the expertise of the analyst.

5.6 Competence of staff

The real purpose of validation of subjective methods is to verify the likelihood of obtaining a true result using the test in question. Subjective tests are by nature qualitative evaluations. The expertise of the analyst is a vital component. Therefore validation must depend to a large extent on testing the expertise of the examiner. Ongoing verification of competency is needed.

5.7 Level of acceptance of the test method

Well established and accepted test methods may not always require the same level of method validation or verification. Relevant considerations might include whether or not the technique has been subject to peer review and published in a recognised journal. Whether or not standards exist for the control of the technique's operation and if the technique has been generally accepted within the relevant scientific community.

6. Glossary of terms

Accuracy, Measurement Accuracy, Accuracy of Measurement - Closeness of agreement between a measured quantity value and a true quantity value of a measurand (JCGM200:2008).

Analyte - The component of a sample or test item which embodies a quantity or quality that is ultimately determined directly or indirectly. The term 'analyte' in this document is applied to any substance or material to be analysed (e.g. blood components, chemical constituents, microorganisms, etc).

Bias, Measurement Bias - Estimate of a systematic measurement error (JCGM200:2008).

Binary Classification Test - Is the task of classifying the members of a given set of objects into two groups on the basis of whether they have some property or not (e.g. applied to medical testing a binary classification tests is used to determine if a patient has certain disease or not - the classification property is the disease). May also be referred to as a presence/absence test, or a positive/negative test.

Blank - A blank value is obtained as a result of analysis of a specimen which does not, as far as possible, contain the analyte(s) in question. Use of various types of blanks (to which no analytes have been added) enables assessment of how much of the measured instrument response is attributable to the analyte and how much to other causes. Various types of blank are available to the user: **Reagent blanks**: Reagents used during the analytical process (including solvents used for extraction or dissolution) are analysed in isolation in order to see whether they contribute to the measurement signal. The measurement result arising from the analyte can then be corrected accordingly. **Sample blanks**. These are essentially matrices with no analyte. They may be difficult to obtain but such materials give the best estimate of the effects of interferences that would be encountered in the analysis of test samples (Eurachem, 1998).

Calibration - Operation that, under specified conditions in a first step, establishes a relation between the quantity values with measurement uncertainties provided by measurement standards and corresponding indications with associated measurement uncertainties and, in a second step, uses this information to establish a relation for obtaining a measurement result from an indication. A calibration may be expressed by a statement, calibration function, calibration diagram, calibration curve, or calibration table. In some cases, it may consist of an additive or multiplicative correction of the indication with associated measurement uncertainty. Calibration should not be confused with adjustment of a measuring system, often mistakenly called "self-calibration", nor with verification of calibration (JCGM200:2008).

Certified Reference Material - Reference material, accompanied by documentation issued by an authoritative body and providing one or more specified property values with associated uncertainties and traceabilities, using valid procedures (JCGM200:2008).

Degrees of Freedom - The number of independent determinations (estimates) of a given statistical magnitude (e.g. mean or standard deviation) that can be performed on the basis of a given data set.

Facility - It is recognised that not all testing or calibration activities are performed in a 'laboratory'. Accordingly, the expressions 'facility' and 'laboratory' are used interchangeably throughout this document.

False Negatives - A negative outcome of a binary classification test when the true outcome is positive.

False Positives - A positive outcome of a binary classification test when the true outcome is negative.

Fitness for Purpose - Degree to which data produced by a measurement process enables a user to make technically and administratively correct decisions for a stated purpose (IUPAC, 2000).

Fortified Sample - An artificially contaminated sample to which a known concentration of analyte has been added (also known as a **spiked** sample). Used to test the accuracy (especially the efficiency of *recovery*) of an analytical method.

Gold Standard - Term often applied in medical and veterinary fields of testing. Refers to a diagnostic test method or benchmark that is regarded as definitive. (See *Standard Method* definition).

Limit of Detection - Measured quantity value, obtained by a given measurement procedure, for which the probability of falsely claiming the absence of a component in a material is β, given a probability α of falsely claiming its presence (JCGM200:2008). *Note: IUPAC recommends default values for α and β equal to 0.05.*

Limit of Quantitation, Limit of Determination - Refers to the smallest analyte concentration or mass, which can be quantitatively analysed with a reasonable reliability by a given procedure.

Intermediate Precision, Intermediate Measurement Precision - Measurement precision under a set of intermediate precision conditions (JCGM200:2008).

Intermediate Precision Condition of Measurement, Intermediate Precision Condition - Condition of measurement, out of a set of conditions that includes the same measurement procedure, same location, and replicate measurements on the same or similar objects over an extended period of time, but may include other conditions involving changes (JCGM200:2008).

Matrix - The predominant material, component or substrate which contains the analyte of interest.

Matrix Effects - The direct or indirect alteration or interference in response due to the presence of unintended analytes (for analysis) or other interfering substances in the sample.

Matrix Verification - When a facility needs to apply a standard method or peer reviewed method to matrices which are generically different from the scope of what the method was originally intended for, it is necessary to verify this method is appropriate for the new matrix. Consideration must be given to the properties of the new matrix before a validation protocol is undertaken. For example, the nature of the physical properties of the matrix or the presence of a biostatic agent, which may affect the recovery of the determinant, must be considered and appropriate action taken.

Measurand - quantity intended to be measured (JCGM200:2008).

Measurement Uncertainty - Non-negative parameter characterising the dispersion of the quantity values being attributed to a measurand, based on the information used (JCGM200:2008).

Measuring interval, working interval - Set of values of quantities of the same kind that can be measured by a given measuring instrument or measuring system with specified instrumental uncertainty, under defined conditions (JCGM200:2008). *Note: The lower limit of a measuring interval should not be confused with detection limit.*

Method Validation - The process of establishing the performance characteristics and limitations of a method and the identification of the influences which may change these characteristics and to what extent. Which analytes can it determine in which matrices in the presence of which interferences? Within these conditions what levels of precision and accuracy can be achieved? The process for verifying that a method is fit for purpose; i.e. for use of solving a particular analytical problem (Eurachem, 1998).

Method Verification - When accreditation is first sought for a standard method verification data must be generated. (Where a new version of a standard method is introduced a training module is required. Refer to Appendix 1). Verification is the process of demonstrating the performance criteria included in the method can be met by the facility prior to introducing them for routine use.

Microbiological Testing - The term ‘microbiological testing’ referred to in the body of this document covers the testing of samples for microorganisms and can be applied to all fields of testing (e.g. Biological, Veterinary, Medical Testing fields and to some extent the Forensic and Chemical Testing fields).

Platform - The term ‘platform’ typically refers to a computer’s operating system. The term is often used when referring to what kind of computer systems a certain software program will run on.

Positive and Negative Predictive Values - The terms ‘positive and negative predictive values’ is often applied to diagnostic binary classification tests and refers to the precision rate. For example, the **positive predictive value** (ratio of true positives to combined true and false positives) is the proportion of patients with positive test results who are correctly diagnosed. It is the most important measure of a diagnostic method as it reflects the probability that a positive test reflects the underlying condition being tested for. The **negative predictive value** is the proportion of patients with negative test results who are correctly diagnosed (ratio of true negatives to combined true and false negatives).

Precision, Measurement Precision - Closeness of agreement between indications or measured quantity values obtained by replicate measurements on the same or similar objects under specified conditions.

Measurement precision is used to define measurement repeatability, intermediate measurement precision, and measurement reproducibility (JCGM200:2008).

Pure Strain - A pure strain refers to the descendants of a single isolation in a pure culture and are usually made up of a succession of cultures ultimately derived from a single colony. A **pure culture** is a culture of cells or multicellular organisms growing in the absence of other species or types. A pure culture may originate from a single cell or single organism, in which case the cells are genetic clones of one another.

Qualitative test results - Results of tests not numerically derived (e.g. visual examinations or **binary classification tests** such as absence/presence, positive/negative, reactive/non-reactive, etc). Qualitative test results based on a numerical outcome, e.g. based on thresholds, are often described as semi-quantitative or semi-qualitative and it is expected that method validation or verification is in line with quantitative procedures.

Quantitative test results - Numerically derived test results.

Recovery - The extraction efficiency of an analytical process, reported as (a percentage of) the known amount of analyte carried through the sample extraction and processing steps of the method.

Reference Material - Material, sufficiently homogenous and stable with reference to specified properties, which has been established to be fit for its intended use in measurement of nominal properties (JCGM200:2008).

Relative Standard Deviation - (RSD or %RSD) is the absolute value of the coefficient of variation ($s / \bar{x}$). It is often expressed as a percentage. A similar term that is sometimes used is the relative variance which is the square of the coefficient of variation. The relative standard deviation is widely used to express precision and repeatability.

Repeatability, Measurement Repeatability - Measurement precision under a set of repeatability conditions of measurement (JCGM200:2008).

Repeatability Condition of Measurement, Repeatability Condition - Condition of measurement, out of a set of conditions that includes the same measurement procedure, same operators, same measuring system, same operating conditions and same location, and replicate measurements on the same or similar objects over a short period of time. Measurement precision under a set of repeatability conditions of measurement (JCGM200:2008).

Reproducibility, Measurement Reproducibility - Measurement precision under reproducibility conditions of measurement (JCGM200:2008).

Reproducibility Condition of Measurement, Reproducibility Condition - Condition of measurement, out of a set of conditions that includes different operators, measuring systems, locations, and replicate measurements on the same or similar objects (JCGM200:2008).

Ruggedness/Robustness - The degree of independence of the method of analysis from minor deviations in the experimental conditions of the method of analysis.

Selectivity, Selectivity of a Measuring System - Property of a measuring system, used with a specified measurement procedure, whereby it provides measured quantity values for one or more measurands such that the values of each measurand are independent of other measurands or other quantities in the phenomenon, body, or substance being investigated (JCGM200:2008).

Sensitivity, Sensitivity of a Measuring System - Quotient of the change in an indication of a measuring system and the corresponding change in a value of a quantity being measured (JCGM200:2008).

Stability - The (chemical) stability of an analyte in a given matrix under specific conditions for given time intervals.

Standard Deviation - A measure of how values are dispersed about a mean in a distribution of values. The standard deviation σ for the whole population of *n* values is given by:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}}$$

In practice we usually analyse a statistical sample and not the whole population. The standard deviation *s* for a randomly-selected sample is given by:

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Standard Method - According to the British Standards Institution, a standard is a published: "specification that establishes a common language, and contains a technical specification or other precise criteria and is designed to be used consistently, as a rule, a guideline, or a definition". (See *Gold Standard* definition).

Target Measurement Uncertainty - Measurement uncertainty specified as an upper limit and decided on the basis of the intended use of measurement results (JCGM200:2008). Represents the greatest allowed uncertainty for a particular usage of that result.

Test Method - A body of procedures and techniques for performing an activity. The terms 'test method' and 'procedure' are used interchangeably throughout this document and apply to all methods of testing including examination procedures, such as fingerprint examinations, and reviews.

Training Module - The term training module refers to the action required to be taken when a new version of a method for which a facility is currently accredited is introduced. Training modules require staff to be advised of the changes in the new version of the method and for this to be recorded in staff training records. A training module takes into consideration the competency of the facility staff to perform the technique and familiarity with e.g. the target organism(s).

Trueness, Measurement Trueness, Trueness of Measurement - Closeness of agreement between the average of an infinite number of replicate measured quantity values and a reference value (JCGM200:2008).

True Negative - A negative outcome of a binary classification test when the true outcome is negative.

True Positive - A positive outcome of a binary classification test when the true outcome is positive.

Validation - Confirmation by examination and provision of objective evidence that the particular requirements for a specific intended use are fulfilled (ISO 9000:2005). (See *Method Validation* definition*).*
Comparative validation aims to demonstrate equivalent performance between two (or more) methods. There is no single test of establishing method equivalence or numerical acceptance criteria for it. Generally, a method with the greatest sensitivity or highest recovery for the target microorganism/analyte is the best. For situations where comparative validation is not applicable (e.g. use of a matrix significantly different to those specified in the scope of the method or use of an alternative isolation or detection principle), ***primary validation*** must be undertaken prior to introducing the method. In such cases validation becomes an exploratory process with the aim of establishing operational limits and performance characteristics of the alternative, new or otherwise inadequately characterised method. It should result in numerical and descriptive specifications for the performance of the method.

Verification - Confirmation by examination and provision of objective evidence that specified requirements have been fulfilled (ISO 9000:2005).

7. References

AOAC flworkshop, *Guidelines for single laboratory validation (SLV) of chemical methods for metals in food,* www.flworkshop.com/metals/AOAC%20SLV%20Metals%20protocol.doc

AOAC International (2002), *AOAC International methods committee guidelines for validation of qualitative and quantitative food microbiological official methods and analysis*, OMA Program Manual, www.aoac.org

AOAC International (2007), *How to meet ISO/IEC 17025 requirements for method verification*, ALACC Guide, www.aoac.org/alacc_guide_2008.pdf.

AS 2850 (1986), *Chemical analysis-interlaboratory test programs for determining precision of analytical method(s)-Guide to the planning and conduct.*

AS/NZS 4659 (Parts 1-4), *Guide to determining the equivalence of food microbiology test methods.*

Ellison, S. L. R. (2006), *In defense of the correlation coefficient*. Accred. Qual. Assurance, Issue No. 11, pp 146 - 152.

ENFSI Standing Committee for Quality and Competence (QC) (2006), *Validation and Implementation of (New) Methods*, QCC-VAL-001, Issue No. 001, www.enfsi.eu

Ethier, J. (2005), *Procedure for the validation of biological active pharmaceutical ingredients (APIs) manufacturing processes*, The Official Journal of ISPE, Vol. 25 No. 2

Eurachem/CITAC (2002), *Guide to quality in analytical chemistry: An aid to accreditation*, www.eurachem.org/guides/pdf/CITAC%20EURACHEM%20GUIDE.pdf

Eurachem Guide (1998), *The fitness for purpose of analytical Methods. A laboratory guide to method validation and related topics*, Edition 1.0, http://eurachem.ul.pt/guides/mval.html

Food and Drug Administration (FDA) (2001), *Guidance for industry – Bioanalytical method validation*

Hibbert, D. B. (2005), *Further comments on the (miss-) use of r for testing the linearity of calibration*, Accreditation and Quality Assurance, Issue 10, pp 300-301

Hibbert, D.B. (2006), *The uncertainty of a result from a linear calibration*, Analyst, 131, pp 1273-1278

Hibbert, D.B. (2004), *Method validation,* in Encyclopedia of Analytical Science, 2nd Edition, Quality Assurance, Elsevier Ltd.

Huber, W. (2004), *On the use of the correlation coefficient r for testing the linearity of calibration functions*, Accreditation and Quality Assurance, Issue No 9, pp 726

ISO 3534-1 (2006), *Statistics-vocabulary and symbols, Part 1: General statistical terms and terms used in probability*

ISO 9000 (2005), *Quality management systems -- Fundamentals and vocabulary*

ISO 13843 (2005), *Water quality – Guidance on validation of microbiological methods.*

ISO/IEC Guide 98-3 (2008), *Uncertainty of measurement – Part 3: Guide to the expression of uncertainty in measurement (GUM 1995)*

IUPAC (2002), *Harmonised guidelines for single laboratory validations of methods of analysis,* Pure Appl. Chem. 74(5), pp 835-855, http://www.iupac.org/publications/pac/2002/7405/x0835.html

JCGM:200 (2008), *International Vocabulary of Metrology (VIM) – Basic and general concepts and associated terms,* 3rd edition, BIPM, Sèvres, www.bipm.org/vim2008

Miller, J.C. and Miller, J.N. (2000), *Statistics and chemometrics for analytical chemistry,* 4th Edition, Prentice Hall, ISBN 0-13-022888-5, and/or VAMSTAT II *Statistics Training for Valid Analytical Measurements* CD-ROM, www.vam.org.uk

LGC (2003), *In-house method validation - A guide for chemical laboratories*, Laboratory of the Government Chemist (LGC), UK.

Mulholland, M.I. and Hibbert, D.B. (1997), *Linearity and the limitation of least squares calibration*, Journal of Chromatography A, 76273-8

National Pathology Accreditation Advisory Council (NPAAC) (2006), *Laboratory Accreditation Standards and Guidelines for Nucleic Acid Detection and Analysis*

National Pathology Accreditation Advisory Council (NPAAC) (2007), *Requirements for the development and use of in-house In Vitro Diagnostic Devices (IVDs)*

National Pathology Accreditation Advisory Council (NPAAC) (2007), *Requirements for the estimation of measurement uncertainty*

O'Donnell, G.E. and Hibbert, D.B. (2005), *Treatment of bias in estimating measurement uncertainty*, The Analyst, Issue No. 130, pp 721-729

OIE Terrestrial Manual (May 2009), *Principles and methods of validation of diagnostic assays for infectious diseases,* Chapter 1.1.4/5,www.oie.int/fileadmin/Home/eng/Health_standards/tahm/1.1.04_VALID.pdf"

SANAS Accreditation Council (2008), *Guidelines on validation and quality assurance in microbiological testing*, TG28-02.

Taylor, J.K. (1989), *Quality assurance of chemical measurements*, sixth edition, Lewis Publishers, ISBN 0-87371-097-5, page 79.

Thompson M., Ellison S.L.R. and Wood R., UPAC Interdivisional Working Party (2002), *Harmonised guidelines for single-laboratory validation of methods of analysis* (IUPAC Technical Report), Pure Appl. Chem., 74(5), pp 835-855

Van Loco, J., Elskens, M., Croux, C., Beernaert, H.(2002), *Linearity of calibration curves: use and misuse of the correlation coefficient*, Accred. Qual. Assur., Issue No. 7, pp 281-285.

Youden, W.J. and Steiner, E.H. (1975), *Statistical manual of the Association of Official Analytical Chemists*, AOAC, ISBN 0-935584-15-3

Zar, J.H. (1996), *Biostatistical analysis*, third edition, Prentice-Hall Inc.

8. Additional reading

APLAC (2003), *Interpretation and guidance on the estimation of uncertainty of measurement in testing*, APLAC TC 005, APLAC, www.aplac.org.

Best Practice for Microbiological Methodology (BPMM). Final Report at http://www.fda.gov/ucm/groups/fdagov-public/@fdagov-foods-gen/documents/document/ucm088708.pdf with supplements at http://www.fda.gov/Food/ScienceResearch/LaboratoryMethods/ucm124900.htm

EURACHEM Guide (1996), *The fitness for purpose of analytical methods*. *A laboratory guide to method validation and related topics*, LGC, Teddington. Also available from the EURACHEM Secretariat and website.

Eurachem/CITAC (2000), *Quantifying uncertainty in analytical measurement,* Eurachem/CITAC Guide CG4, 2nd Edition, ISBN 0-948926-15-5, www.eurachem.ul.pt.

Eurachem/CITAC (2003), *Traceability in chemical measurement,* www.eurachem.ul.pt

Eurolab (2007), *Measurement uncertainty revisited: Alternative approaches to uncertainty evaluation,* Technical Report No 1/2007, www.eurolab.org

Garfield, F.M., Klestan, E. and Hirsch, J. (2000), *Quality assurance principles for analytical laboratories*, 3rd Edition, AOAC International, ISBN-0-935584-70-6.

Hibbert, D. B. (2007), *Quality assurance for the analytical chemistry laboratory,* Oxford University Press, New York

ILAC (2002), *Introducing the concept of uncertainty of measurement in testing in association with the application of the Standard ISO/IEC 17025*, ILAC G17, www.ilac.org.

ISO 16140 (2003), *Microbiology of food and animal feeding stuffs – Protocol for the validation of alternative methods*

ISO/TS 21748:2004, *Guidance for the use of repeatability, reproducibility and trueness estimates in measurement uncertainty estimation*

Magnusson, B., Naykki, T., Hovind, H. and Krysell, M. (2003), *Handbook for calculation of measurement uncertainty in environmental laboratories*, NORDTEST Report TR537

Report of a Joint FAO/IAEA Expert Consultation (1997), *Validation of analytical methods for food control*, FAO Food and Nutrition Paper No. 68, FAO

Thompson. M. and Wood, R. (1995), *Harmonised guidelines for internal quality control in analytical chemistry laboratories,* Pure Appl. Chem., 67 (4), pp 49-56

Thompson, M., Ellison, S., Fajgelj, A., Willetts, P. and Wood, R. (1999), *Harmonised guidelines for the use of recovery information in analytical measurement*, Pure Appl. Chem., 71 (2), pp 337-348

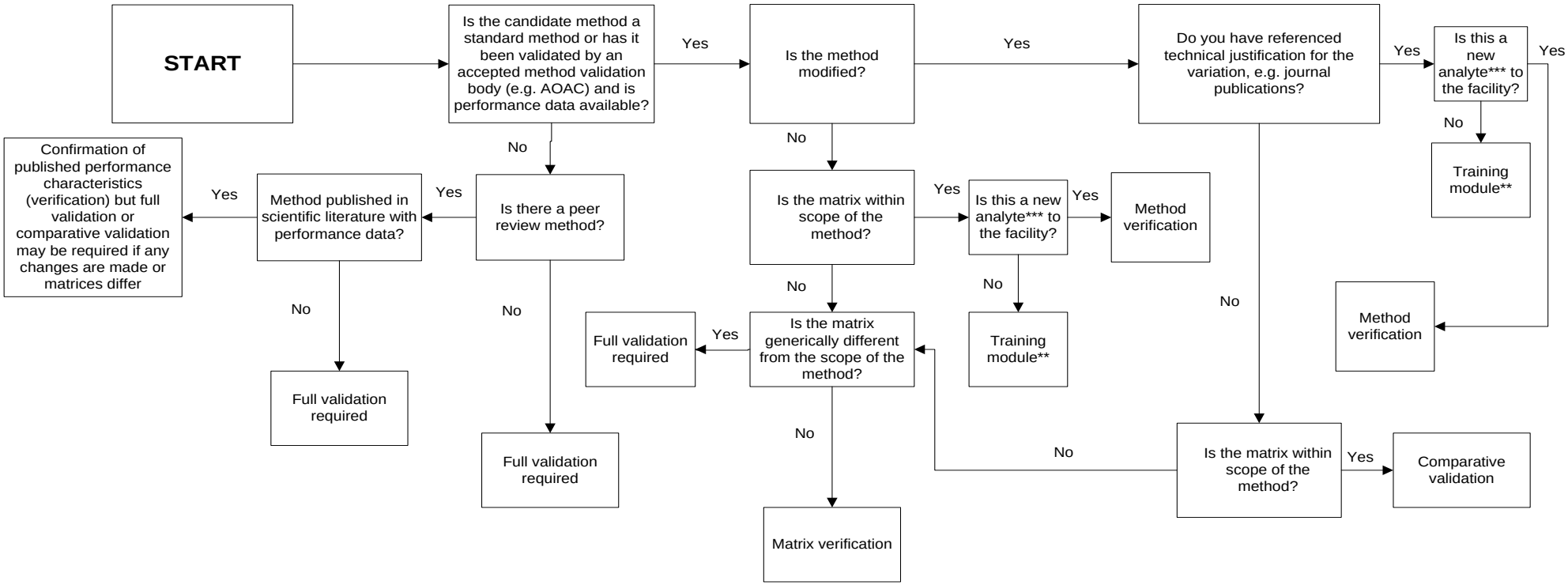
NMKL Secretariat, Finland (1996), *Validation of chemical analytical methods,* NMKL Procedure No. 4

Nordtest (2005), *Handbook for calculation of measurement uncertainty in environmental laboratories*, Report TR 537, www.nordtest.org

Sub-committee for Animal Health and Laboratory Standards (SCAHLS), http://www.scahls.org.au

UKAS (2000), *The expression of uncertainty in testing*, UKAS LAB12, 1st Edition, www.ukas.com.

Appendix 1. Method validation and verification decision tree



*It has been found in some cases (e.g. veterinary microbiological testing) that a specific test kit performs differently under local environmental conditions, to that of the original environmental conditions it was subjected to during primary validation. In such cases, the facility should conduct the validation to prove the kit performs under local environmental conditions.

**This covers the introduction of a new version of a standard method. This may for example have differences in agar, temperatures, reactions, etc which need to be explained.

***In some cases where the analyte is not a new analyte to the facility but a different technique is employed for the analysis of the analyte, performing a training module may not be adequate and method verification may be required. For example in biological testing, AS5013 has 2 methods (AS5013.3 and AS5013.4) for the enumeration of coliforms. Experience in 1 method does not mean that the facility is necessarily competent to the perform the other.

For ad-hoc or special analyses, the extent of validation will be limited by circumstance.

Amendments

The table below provides a summary of changes made to the document with this issue.

Section	Amendment
6	Definitions added to glossary of terms for ‘training module’, ‘matrix verification’, ‘platforms’, ‘primary validation’ and ‘comparative validation’.
Appendix 1	Decision tree updated to exclude certain areas where it may not be adequate to perform a training module alone and method verification is required.