OIV-MA-AS1-13 Harmonised guidelines for single laboratory validation of methods of analysis (technical report)

Synopsis

Method validation is one of the measures universally recognised as a necessary part of a comprehensive system of quality assurance in analytical chemistry. In the past ISO, IUPAC and AOAC INTERNATIONAL have co-operated to produce agreed protocols or guidelines on the "Design, Conduct and Interpretation of Method Performance Studies"[1] on the "Proficiency Testing of (Chemical) Analytical Laboratories"[2] on "Internal Quality Control in Analytical Chemistry Laboratories"[3] and on "The Use of Recovery Information in Analytical Measurement".[4] (from the usage of overlapping data in analytical measurements) The Working Group that produced these protocols/guidelines has now been mandated by IUPAC to prepare guidelines on the Single-laboratory Validation of methods of analysis. These guidelines provide minimum recommendations on procedures that should be employed to ensure adequate validation of analytical methods.

A draft of the guidelines has been discussed at an International Symposium on the Harmonisation of Quality Assurance Systems in Chemical Laboratory, the Proceedings from which have been published by the UK Royal Society of Chemistry.

Resulting from the Symposium on Harmonisation of Quality Assurance

Systems for Analytical Laboratories, Budapest, Hungary, 4-5 November 1999

held under the sponsorship of IUPAC, ISO and AOAC INTERNATIONAL

Contents

1. Introduction

1.1. Background

1.2. Existing protocols, standards and guides

2. Definitions and terminology

2.1.General

2.2. Definitions used in this guide

3. Method validation, uncertainty, and quality assurance

4. Basic principles of method validation

4.1. Specification and scope of validation

4.2. Testing assumptions

4.3. Sources of Error in Analysis

4.4. Method and Laboratory effects
5. Conduct of Validation Studies
6. Extent of validation studies
6.1. The laboratory is to use a “fully” validated method
6.2. The laboratory is to use a fully validated method, but new matrix is to be used
6.3. The laboratory is to use a well-established, but not collaboratively studied, method
6.4. The method has been published in the scientific literature together with some analytical characteristics
6.5. The method has been published in the scientific literature with no characteristics given or has been developed in-house
6.6. The method is empirical
6.7. The analysis is “ad hoc”
6.8. Changes in staff and equipment
7. Recommendations
8. Bibliography
Annex A: Notes on the requirements for study of method performance characteristics
.
A1 Applicability
A2.Selectivity
A3. Calibration and linearity
A3.1. Linearity and intercept
A3.2. Test for general matrix effect
A3.3. Final calibration procedure
A4. Trueness
A4.1. Estimation of trueness
A4.2. Conditions for trueness experiments
A4.3. Reference values for trueness experiments
A4.3.1. Certified reference materials (CRMs)
A4.3.2. Reference materials
A4.3.3. Use of a reference method
A4.3.4. Use of spiking/recovery

A5. Accuracy
A6. Recovery
A7. Concentration range
A8. Detection Limit
A9. Limit of determination or limit of quantification
A10. Sensitivity
A11. Ruggedness
A12.Fitness for trial purposes
A13. Matrix variation
A14.Measurement Uncertainty
Annex B Additional considerations for uncertainty estimation in validation studies
B1.Sensitivity analysis
B2. Judgement

1. Introduction

1.1. Background

Reliable analytical methods are required for compliance with national and international regulations in all areas of analysis. It is accordingly internationally recognised that a laboratory must take appropriate measures to ensure that it is capable of providing and does provide data of the required quality. Such measures include:

- using validated methods of analysis;
- using internal quality control procedures;
- participating in proficiency testing schemes; and
- becoming accredited to an International Standard, normally ISO/IEC 17025.

It should be noted that accreditation to ISO/IEC 17025 specifically addresses the establishment of traceability for measurements, as well as requiring a range of other technical and management requirements including all those in the list above.

Method validation is therefore an essential component of the measures that a laboratory should implement to allow it to produce reliable analytical data. Other aspects of the above have been addressed previously by the IUPAC Interdivisional Working Party on Harmonisation of Quality Assurance Schemes for Analytical Laboratories, specifically by preparing Protocols/Guidelines on method performance

(collaborative) studies,[1] proficiency testing,[2] and internal quality control.[3]

In some sectors, most notably in the analysis of food, the requirement for methods that have been "fully validated" is prescribed by legislation.[5,6] "Full" validation for an analytical method is usually taken to comprise an examination of the characteristics of the method in an inter-laboratory method performance study (also known as a collaborative study or collaborative trial). Internationally accepted protocols have been established for the "full" validation of a method of analysis by a collaborative trial, most notably the International Harmonised Protocol[1] and the ISO procedure.[7] These protocols/standards require a minimum number of laboratories and test materials to be included in the collaborative trial to validate fully the analytical method. However, it is not always practical or necessary to provide full validation of analytical methods. In such circumstances a "single-laboratory method validation" may be appropriate.

Single-laboratory method validation is appropriate in several circumstances including the following:

to ensure the viability of the method before the costly exercise of a formal collaborative trial;

to provide evidence of the reliability of analytical methods if collaborative trial data are not available or where the conduct of a formal collaborative trial is not practicable;

to ensure that "off-the-shelf" validated methods are being used correctly.

When a method is to be characterised in-house, it is important that the laboratory determines and agrees with its customer exactly which characteristics are to be evaluated. However, in a number of situations these characteristics may be laid down by legislation (e.g. veterinary drug residues in food and pesticides in food sectors). The extent of the evaluation that a laboratory undertakes must meet the requirements of legislation.

Nevertheless in some analytical areas the same analytical method is used by a large number of laboratories to determine stable chemical compounds in defined matrices. It should be appreciated that if a suitable collaboratively studied method can be made available to these laboratories, then the costs of the collaborative trial to validate that method may well be justified. The use of a collaboratively studied method considerably reduces the efforts which a laboratory, before taking a method into routine use, must invest in extensive validation work. A laboratory using a collaboratively studied method, which has been found to be fit for the intended purpose, needs only to demonstrate that it can achieve the performance characteristics stated in the method. Such a verification of the correct use of a method is much less costly than a full single laboratory validation. The total cost to the

Analytical Community of validating a specific method through a collaborative trial and then verifying its performance attributes in the laboratories wishing to use it is frequently less than when many laboratories all independently undertake single laboratory validation of the same method.

1.2. Existing Protocols, Standards and Guides

A number of protocols and guidelines[8-19] on method validation and uncertainty have been prepared, most notably in AOAC INTERNATIONAL, International Conference on Harmonisation (ICH) and Eurachem documents:

The Statistics manual of the AOAC, which includes guidance on single laboratory study prior to collaborative testing[13]

The ICH text[15] and methodology,[16] which prescribe minimum validation study requirements for tests used to support drug approval submission.

The Fitness for Purpose of Analytical Methods: A Laboratory Guide to Method Validation and Related Topics (1998)[12]

Quantifying Uncertainty in Analytical Measurement (2000)[9]

Method validation was also extensively discussed at a Joint FAO/IAEA Expert Consultation, December 1997, on the Validation of Analytical Methods for Food Controls, the Report of which is available[19].

The present 'Guidelines' bring together the essential scientific principles of the above documents to provide information which has been subjected to international acceptance and, more importantly, to point the way forward for best practice in single-laboratory method validation.

2. Definitions and terminology

2.1. General

Terms used in this document respect ISO and IUPAC definitions where available. The following documents contain relevant definitions:

i) IUPAC: Compendium of chemical terminology, 1987

ii) International vocabulary of basic and general terms in metrology. ISO 1993

2.2. Definitions used in this guide only:

Relative uncertainty: Uncertainty expressed as a relative standard deviation.

Validated range: That part of the concentration range of an analytical method which has been subjected to validation.

3. Method validation, uncertainty, and quality assurance

Method validation makes use of a set of tests which both test any assumptions on which the analytical method is based and establish and document the performance characteristics of a method, thereby demonstrating whether the method is fit for a particular analytical purpose. Typical performance characteristics of analytical methods are: applicability; selectivity; calibration; trueness; precision; recovery; operating range; limit of quantification; limit of detection; sensitivity; and ruggedness. To these can be added measurement uncertainty and fitness-for-purpose.

Strictly speaking, validation should refer to an 'analytical system' rather than an 'analytical method', the analytical system comprising a defined method protocol, a defined concentration range for the analyte, and a specified type of test material. For the purposes of this document, a reference to 'method validation' will be taken as referring to an analytical system as a whole. Where the analytical procedure as such is addressed, it will be referred to as 'the protocol'.

In this document method validation is regarded as distinct from ongoing activities such as internal quality control (IQC) or proficiency testing. Method validation is carried out once, or at relatively infrequent intervals during the working lifetime of a method; it tells us what performance we can expect the method to provide in the future. Internal quality control tells us about how the method has performed in the past. IQC is therefore treated as a separate activity in the IUPAC Harmonisation Programme.[3]

In method validation the quantitative characteristics of interest relate to the accuracy of the result likely to be obtained. Therefore it is generally true to say that method validation is tantamount to the task of estimating uncertainty of measurement. Over the years it has become traditional for validation purposes to represent different aspects of method performance by reference to the separate items listed above, and to a considerable extent these guidelines reflect that pattern. However, with an increasing reliance on measurement uncertainty as a key indicator of both fitness for purpose and reliability of results, analytical chemists will increasingly undertake measurement validation to support uncertainty estimation, and some practitioners will want to do so immediately. Accordingly, measurement uncertainty is treated briefly in Annex A as a performance characteristic of an analytical method, while Annex B provides additional guidance on some procedures not otherwise covered.

4. Bacis principles of method validation

4.1. Specification and scope of validation

Validation applies to a defined protocol, for the determination of a specified analyte and range of concentrations in a particular type of test material, used for a specified purpose. In general, validation should check that the method performs adequately for the purpose throughout the range of analyte concentrations and test materials to which it is applied. It follows that these features, together with a statement of any fitness-for-purpose criteria, should be completely specified before any validation takes place.

4.2. Testing assumptions

In addition to the provision of performance figures which indicate fitness for purpose and have come to dominate the practical use of validation data, validation studies act as an objective test of any assumptions on which an analytical method is based. For example, if a result is to be calculated from a simple straight line calibration function, it is implicitly assumed that the analysis is free from significant bias, that the response is proportional to analyte concentration, and that the dispersion of random errors is constant throughout the range of interest. In most circumstances, such assumptions are made on the basis of experience accumulated during method development or over the longer term, and are consequently reasonably reliable. Nonetheless, good measurement science relies on tested hypotheses. This is the reason that so many validation studies are based on statistical hypothesis testing; the aim is to provide a basic check that the reasonable assumptions made about the principles of the method are not seriously flawed.

There is an important practical implication of this apparently abstruse note. It is easier to check for gross departure from a reliable assumption than to 'prove' that a particular assumption is correct. Thus, where there is long practice of the successful use of a particular analytical technique (such as gas chromatographic analysis, or acid digestion methods) across a range of analytes and matrices, validation checks justifiably take the form of relatively light precautionary tests. Conversely, where experience is slight, the validation study needs to provide strong evidence that the assumptions made are appropriate in the particular cases under study, and it will generally be necessary to study the full range of circumstances in detail. It follows that the extent of validation studies required in a given instance will depend, in part, on the accumulated experience of the analytical technique used.

In the following discussion, it will be taken for granted that the laboratory is well practised in the technique of interest, and that the purpose of any significance tests is to check that there is no strong evidence to discount the assumptions on which the particular protocol relies. The reader should bear in mind that more stringent checks may be necessary for unfamiliar or less established measurement techniques.

4.3. Sources of Error in Analysis

Errors in analytical measurements arise from different sources[*] and at different levels of organisation. One useful way of representing these sources (for a specific concentration of analyte) is as follows[+]24:

- random error of measurement (repeatability);
- run bias ;
- laboratory bias;
- method bias;
- matrix variation effect.

Though these different sources may not necessarily be independent, this list provides a useful way of checking the extent to which a given validation study addresses the sources of error.

The repeatability (within-run) term includes contributions from any part of the procedure that varies within a run, including contributions from the familiar gravimetric and volumetric errors, heterogeneity of the test material, and variation in the chemical treatment stages of the analysis, and is easily seen in the dispersion of replicated analyses. The run effect accounts for additional day-to-day variations in the analytical system, such as changes of analyst, batches of reagents, recalibration of instruments, and the laboratory environment (e.g., temperature changes). In single-laboratory validation, the run effect is typically estimated by conducting a designed experiment with replicated analysis of an appropriate material in a number of separate runs. Between-laboratory variation arises from factors such as variations in calibration standards, differences between local interpretations of a protocol, changes in equipment or reagent source or environmental factors, such as differences in average climatic conditions. Between-laboratory variation is clearly seen as a reality in the results of collaborative trials (method performance studies) and proficiency tests, and between-method variation can sometimes be discerned in the results of the latter.

Generally, the repeatability, run effect and laboratory effect are of comparable magnitude, so none can safely be ignored in validation. In the past there has been a tendency for aspects to be neglected, particularly when estimating and reporting uncertainty information. This results in uncertainty intervals that are too tight. For example, the collaborative trial as normally conducted does not give the complete picture because contributions to uncertainty from method bias and matrix variation are not estimated in collaborative trials and have to be addressed separately (usually

by prior single-laboratory study). In single-laboratory validation there is the particular danger that laboratory bias also may be overlooked, and that item is usually the largest single contributor to uncertainty from the above list. Therefore specific attention must be paid to laboratory bias in single-laboratory validation.

In addition to the above-mentioned problems, the validation of a method is limited to the scope of its application, that is, the method as applied to a particular class of test material. If there is a substantial variation of matrix types within the defined class, there will be an additional source of variation due to within-class matrix effects. Of course, if the method is subsequently used for materials outside the defined class (that is, outside the scope of the validation), the analytical system cannot be considered validated: an extra error of unknown magnitude is introduced into the measurement process.

It is also important for analysts to take account of the way in which method performance varies as a function of the concentration of the analyte. In most instances the dispersion of results increases absolutely with concentration and recovery may differ substantially at high and low concentrations. The measurement uncertainty associated with the results is therefore often dependent on both these effects and on other concentration-dependent factors. Fortunately, it is often reasonable to assume a simple relationship between performance and analyte concentration; most commonly that errors are proportional to analyte concentration.[*] However, where the performance of the method is of interest at substantially different concentrations, it is important to check the assumed relationship between performance and analyte concentration. This is typically done by checking performance at extremes of the likely range, or at a few selected levels. Linearity checks also provide information of the same kind.

4.4. Method and Laboratory effects

It is critically important in single-laboratory method validation to take account of method bias and laboratory bias. There are a few laboratories with special facilities where these biases can be regarded as negligible, but that circumstance is wholly exceptional. (However, that if there is only one laboratory carrying out a particular analysis, then method bias and laboratory bias take on a different perspective). Normally, method and laboratory effects have to be included in the uncertainty budget, but often they are more difficult to address than repeatability error and the run effect. In general, to assess the respective uncertainties it is necessary to use information gathered independently of the laboratory. The most generally useful sources of such information are (i) statistics from collaborative trials (not available in many situations of single-laboratory method validation), (ii) statistics from proficiency

tests and (iii) results from the analysis of certified reference materials.

Collaborative trials directly estimate the variance of between-laboratory biases. While there may be theoretical shortcomings in the design of such trials, these variance estimates are appropriate for many practical purposes. Consequently it is always instructive to test single-laboratory validation by comparing the estimates of uncertainty with reproducibility estimates from collaborative trials. If the single-laboratory result is substantially the smaller, it is likely that important sources of uncertainty have been neglected. (Alternatively, it may be that a particular laboratory in fact works to a smaller uncertainty than found in collaborative trials: such a laboratory would have to take special measures to justify such a claim.) If no collaborative trial has been carried out on the particular method/test material combination, an estimate of the reproducibility standard deviation σ_H at an analyte concentration c above about 120 ppb can usually be obtained from the Horwitz function, $\sigma_H = 0.02c^{0.8495}$, with both variables expressed as mass fractions. (The Horwitz estimate is normally within a factor of about two of observed collaborative study results). It has been observed that the Horwitz function is incorrect at concentrations lower than about 120 ppb, and a modified function is more appropriate.[21, 25] All of this information may be carried into the single-laboratory area with minimum change.

Statistics from proficiency tests are particularly interesting because they provide information in general about the magnitude of laboratory and method biases combined and, for the participant, information about total error on specific occasions. Statistics such as the robust standard deviation of the participants results for an analyte in a round of the test can in principle be used in a way similar to reproducibility standard deviations from collaborative trials, i.e., to obtain a benchmark for overall uncertainty for comparison with individual estimates from single-laboratory validation. In practice, statistics from proficiency tests may be more difficult to access, because they are not systematically tabulated and published like collaborative trials, but only made available to participants. Of course, if such statistics are to be used they must refer to the appropriate matrix and concentration of the analyte. Individual participants in proficiency testing schemes can also gauge the validity of their estimated uncertainty by comparing their reported results with the assigned values of successive rounds[26]. This, however, is an ongoing activity and therefore not strictly within the purview of single-laboratory validation (which is a one-off event).

If an appropriate certified reference material is available, a single-laboratory test allows a laboratory to assess laboratory bias and method bias in combination, by

analysing the CRM a number of times. The estimate of the combined bias is the difference between the mean result and the certified value.

Appropriate certified reference materials are not always available, so other materials may perforce have to be used. Materials left over from proficiency tests sometimes serve this purpose and, although the assigned values of the materials may have questionable uncertainties, their use certainly provides a check on overall bias. Specifically, proficiency test assigned values are generally chosen to provide a minimally biased estimate, so a test for significant bias against such a material is a sensible practice. A further alternative is to use spiking and recovery information[4] to provide estimates of these biases, although there may be unmeasurable sources of uncertainty associated with these techniques.

Currently the least recognised effect in validation is that due to matrix variation within the defined class of test material. The theoretical requirement for the estimation of this uncertainty component is for a representative collection of test materials to be analysed in a single run, their individual biases estimated, and the variance of these biases calculated. (Analysis in a single run means that higher level biases have no effect on the variance. If there is a wide concentration range involved, then allowance for the change in bias with concentration must be made.) If the representative materials are certified reference materials, the biases can be estimated directly as the differences between the results and the reference values, and the whole procedure is straightforward. In the more likely event that insufficient number of certified reference materials are available, recovery tests with a range of typical test materials may be resorted to, with due caution. Currently there is very little quantitative information about the magnitude of uncertainties from this source, although in some instances they are suspected of being large.

5. Conduct of Validation Studies

The detailed design and execution of method validation studies is covered extensively elsewhere and will not be repeated here. However, the main principles are pertinent and are considered below:

It is essential that validation studies are representative. That is, studies should, as far as possible, be conducted to provide a realistic survey of the number and range of effects operating during normal use of the method, as well as to cover the concentration ranges and sample types within the scope of the method. Where a factor (such as ambient temperature) has varied representatively at random during the course of a precision experiment, for example, the effects of that factor appear directly in the observed variance and need no additional study unless further method

optimisation is desirable.

In the context of method validation, “representative variation” means that the factor must take a distribution of values appropriate to the anticipated range of the parameter in question. For continuous measurable parameters, this may be a permitted range, stated uncertainty or expected range; for discontinuous factors, or factors with unpredictable effects such as sample matrix, a representative range corresponds to the variety of types or “factor levels” permitted or encountered in normal use of the method. Ideally, representativeness extends not only to the range of values, but to their distribution. Unfortunately, it is often uneconomic to arrange for full variation of many factors at many levels. For most practical purposes, however, tests based on extremes of the expected range, or on larger changes than anticipated, are an acceptable minimum.

In selecting factors for variation, it is important to ensure that the larger effects are ‘exercised’ as much as possible. For example, where day to day variation (perhaps arising from recalibration effects) is substantial compared to repeatability, two determinations on each of five days will provide a better estimate of intermediate precision than five determinations on each of two days. Ten single determinations on separate days will be better still, subject to sufficient control, though this will provide no additional information on within-day repeatability.

Clearly, in planning significance checks, any study should have sufficient power to detect such effects before they become practically important (that is, comparable to the largest component of uncertainty).

In addition, the following considerations may be important:

Where factors are known or suspected to interact, it is important to ensure that the effect of interaction is accounted for. This may be achieved either by ensuring random selection from different levels of interacting parameters, or by careful systematic design to obtain ‘interaction’ effects or covariance information.

In carrying out studies of overall bias, it is important that the reference materials and values are relevant to the materials under routine test.

6. Extent of validation studies

The extent to which a laboratory has to undertake validation of a new, modified or unfamiliar method depends to a degree on the existing status of the method and the competence of the laboratory. Suggestions as to the extent of validation and verification measures for different circumstances are given below. Except where stated, it is assumed that the method is intended for routine use.

6.1. The laboratory is to use a “fully” validated method

The method has been studied in a collaborative trial and so the laboratory has to verify that it is capable of achieving the published performance characteristics of the method (or is otherwise able to fulfil the requirements of the analytical task). The laboratory should undertake precision studies, bias studies (including matrix variation studies), and possibly linearity studies, although some tests such as that for ruggedness may be omitted.

6.2. The laboratory is to use a fully validated method, but new matrix is to be used

The method has been studied in a collaborative trial and so the laboratory has to verify that the new matrix introduces no new sources of error into the system. The same range of validation as the previous is required.

6.3. The laboratory is to use a well-established, but not collaboratively studied, method

The same range of validation as the previous is required.

6.4. The method has been published in the scientific literature together with some analytical characteristics

The laboratory should undertake precision studies, bias studies (including matrix variation studies), ruggedness and linearity studies.

6.5. The method has been published in the scientific literature with no characteristics given or has been developed in-house

The laboratory should undertake precision studies, bias studies (including matrix variation studies), ruggedness and linearity studies.

6.6. The method is empirical

An empirical method is one in which the quantity estimated is simply the result found on following the stated procedure. This differs from measurements intended to assess method-independent quantities such as the concentration of a particular analyte in a sample, in that the method bias is conventionally zero, and matrix variation (that is , within the defined class) is irrelevant. Laboratory bias cannot be ignored, but is likely to be difficult to estimate by single-laboratory experiment. Moreover, reference materials are unlikely to be available. In the absence of collaborative trial data some estimate of interlaboratory precision could be obtained from a specially designed ruggedness study or estimated by using the Horwitz function.

6.7. The analysis is “ad hoc”

“Ad hoc” analysis is occasionally necessary to establish the general range of a value, without great expenditure and with low criticality. The effort that can go into validation is accordingly strictly limited. Bias should be studied by methods such as recovery estimation or analyte additions, and precision by replication.

6.8. Changes in staff and equipment

Important examples include: change in major instruments; new batches of very variable reagents (for example, polyclonal antibodies); changes made in the laboratory premises; methods used for the first time by new staff; or a validated method employed after a period of disuse. Here the essential action is to demonstrate that no deleterious changes have occurred. The minimum check is a single bias test; a "before and after" experiment on typical test materials or control materials. In general, the tests carried out should reflect the possible impact of the change on the analytical procedure.

7. Recommendations

The following recommendations are made regarding the use of single-laboratory method validation:

Wherever possible and practical a laboratory should use a method of analysis that has had its performance characteristics evaluated through a collaborative trial conforming to an international protocol.

Where such methods are not available, a method must be validated in-house before being used to generate analytical data for a customer.

Single-laboratory validation requires the laboratory to select appropriate characteristics for evaluation from the following: applicability, selectivity, calibration, accuracy, precision, range, limit of quantification, limit of detection, sensitivity, ruggedness and practicability. The laboratory must take account of customer requirements in choosing which characteristics are to be determined.

Evidence that these characteristics have been assessed must be made available to customers of the laboratory if required by the customer.

8. References

- "Protocol for the Design, Conduct and Interpretation of Method Performance Studies", W Horwitz, Pure Appl. Chem., 1988, 60, 855 864, revised W. Horwitz, Pure Appl. Chem., 1995, 67, 331-343.
- "The International Harmonised Protocol for the Proficiency Testing of (Chemical) Analytical Laboratories", M Thompson and R Wood, Pure Appl. Chem., 1993, 65, 2123-2144. (Also published in J. AOAC International, 1993, 76, 926-940.
- "Harmonised Guidelines For Internal Quality Control in Analytical Chemistry Laboratories", Michael Thompson and Roger Wood, J. Pure & Applied

Chemistry, 1995, 67(4), 49-56.

- "Harmonised Guidelines for the Use of Recovery Information in Analytical Measurement", Michael Thompson, Stephen Ellison, Ales Fajgelj, Paul Willetts and Roger Wood, J. Pure & Applied Chemistry, 1999, 71(2), 337-348.
- "Council Directive 93/99/EEC on the Subject of Additional Measures Concerning the Official Control of Foodstuffs", O. J., 1993, L290.
- "Procedural Manual of the Codex Alimentarius Commission, 10th Edition", FAO, Rome, 1997.
- "Precision of Test Methods", Geneva, 1994, ISO 5725, Previous editions were issued in 1981 and 1986.
- "Guide to the Expression of Uncertainty in Measurement", ISO, Geneva, 1993.
- "Quantifying Uncertainty in Analytical Measurement", EURACHEM Secretariat, Laboratory of the Government Chemist, Teddington, UK, 1995, EURACHEM Guide (under revision).
- "International vocabulary of basic and general terms in metrology" ISO, Geneva 1993
- "Validation of Chemical Analytical Methods", NMKL Secretariat, Finland, 1996, NMKL Procedure No. 4.
- "EURACHEM Guide: The fitness for purpose of analytical methods. A Laboratory Guide to method validation and related topics", LGC, Teddington 1996. Also available from the EURACHEM Secretariat and website.
- "Statistics manual of the AOAC", AOAC INTERNATIONAL, Gaithersburg, Maryland, USA, 1975
- "An Interlaboratory Analytical Method Validation Short Course developed by the AOAC INTERNATIONAL", AOAC INTERNATIONAL, Gaithersburg, Maryland, USA, 1996.
- "Text on validation of analytical procedures" International Conference on Harmonisation. Federal Register, Vol. 60, March 1, 1995, pages 11260
- "Validation of analytical procedures: Methodology" International Conference on Harmonisation. Federal Register, Vol. 62, No. 96, May 19, 1997, pages 27463-27467.

- "Validation of Methods", Inspectorate for Health Protection, Rijswijk, The Netherlands, Report 95-001.
- "A Protocol for Analytical Quality Assurance in Public Analysts' Laboratories", Association of Public Analysts, 342 Coleford Road, Sheffield S9 5PH, UK, 1986.
- "Validation of Analytical Methods for Food Control", Report of a Joint FAO/IAEA Expert Consultation, December 1997, FAO Food and Nutrition Paper No. 68, FAO, Rome, 199
- "Estimation and Expression of Measurement Uncertainty in Chemical Analysis", NMKL Secretariat, Finland, 1997, NMKL Procedure No. 5.
- M Thompson, PJ Lowthian, J AOAC Int, 1997, 80, 676-679
- IUPAC recommendation: "Nomenclature in evaluation of analytical methods, including quantification and detection capabilities" Pure and Applied Chem." 1995, 67 1699-1723
- ISO 11843. "Capability of detection." (Several parts). International Standards Organisation, Geneva.
- M. Thompson, Analyst, 2000, 125, 2020-2025
- "Recent trends in inter-laboratory precision at ppb and sub-ppb concentrations in relation to fitness for purpose criteria in proficiency testing" M Thompson, Analyst, 2000, 125, 385-386.
- "How to combine proficiency test results with your own uncertainty estimate - the zeta score", Analytical Methods Committee of the Royal Society of Chemistry, AMC Technical Briefs, editor M. Thompson, AMC Technical Brief No. 2, www.rsc.org/lap/rsccom/amc

Annex A: Notes on the requirements for study of method performance characteristics

The general requirements for the individual performance characteristics for a method are as follows.

A.1 Applicability

After validation the documentation should provide, in addition to any performance specification, the following information:

- the identity of the analyte, including speciation where appropriate (Example: 'total arsenic'

- the concentration range covered by the validation (Example: '0-50 ppm');
- a specification of the range of matrices of the test material covered by the validation (Example: 'seafood');
- a protocol, describing the equipment, reagents, procedure (including permissible variation in specified instructions, e.g., 'heat at 100 □ 5□ for 30 □ 5 minutes'), calibration and quality procedures, and any special safety precautions required;
- the intended application and its critical uncertainty requirements (Example: 'The analysis of food for screening purposes. The standard uncertainty u(c) of the result c should be less than 0.1□c.').

A.2 Selectivity

Selectivity is the degree to which a method can quantify the analyte accurately in the presence of interferents. Ideally, selectivity should be evaluated for any important interferent likely to be present. It is particularly important to check interferents which are likely, on chemical principles, to respond to the test. For example, colorimetric tests for ammonia might reasonably be expected to respond to primary aliphatic amines. It may be impracticable to consider or test every potential interferent; where that is the case, it is recommended that the likely worst cases are checked. As a general principle, selectivity should be sufficiently good for any interferences to be ignored.

In many types of analysis, selectivity is essentially a qualitative assessment based on the significance or otherwise of suitable tests for interference. However, there are useful quantitative measures. In particular, one quantitative measure is the selectivity index b_{an}/b_{int}, where b_{an} is the sensitivity of the method (slope of the calibration function) and b_{int} the slope of the response independently produced by a potential interferent, provides a quantitative measure of interference. b_{int} can be determined approximately by execution of the procedure on a matrix blank and the same blank spiked with the potential interferent at one appropriate concentration. If a matrix blank is unavailable, and a typical material used instead, b_{int} can be estimated from such a simple experiment only under the assumption that mutual matrix effects are absent. Note that b_{int_t} is more easily determined in the absence of the analyte because the effect might be confused with another type of interference when the sensitivity of the analyte is itself affected by the interferent (a matrix effect).

A.3 Calibration and linearity

With the exception of gross errors in preparation of calibration materials, calibration

errors are usually (but not always) a minor component of the total uncertainty budget, and can usually be safely subsumed into various categories estimated by "top-down" methods. For example random errors resulting from calibration are part of the run bias, which is assessed as a whole, while systematic errors from that source may appear as laboratory bias, likewise assessed as a whole. Never-the-less, there are some characteristics of calibration that are useful to know at the outset of method validation, because they affect the strategy for the optimal development of the procedure. In this class are such questions as whether the calibration function plausibly (a) is linear, (b) passes through the origin and (c) is unaffected by the matrix of the test material. The procedures described here relate to calibration studies in validation, which are necessarily more exacting than calibration undertaken during routine analysis. For example, once it is established at validation that a calibration function is linear and passes through the origin, a much simpler calibration strategy can be used for routine use (for example, a two point repeated design). Errors from this simpler calibration strategy will normally be subsumed into higher level errors for validation purposes.

A3.1. Linearity and intercept

Linearity can be tested informally by examination of a plot of residuals produced by linear regression of the responses on the concentrations in an appropriate calibration set. Any curved pattern suggests lack of fit due to a non-linear calibration function. A test of significance can be undertaken by comparing the lack-of-fit variance with that due to pure error. However, there are causes of lack of fit other than nonlinearity that can arise in certain types of analytical calibration, so the significance test must be used in conjunction with a residual plot. Despite its current widespread use as an indication of quality of fit, the correlation coefficient is misleading and inappropriate as a test for linearity and should not be used.

Design is all-important in tests for lack of fit, because it is easy to confound nonlinearity with drift. Replicate measurements are needed to provide an estimate of pure error if there is no independent estimate. In the absence of specific guidance, the following should apply:

- there should be six or more calibrators;
- the calibrators should be evenly spaced over the concentration range of interest;
- 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;
- the calibrators should be run at least in duplicate, and preferably triplicate or

more, in a random order.

After an exploratory fit with simple linear regression, the residuals should be examined for obvious patterns. Heteroscedasticity is quite common in analytical calibration and a pattern suggesting it means that the calibration data are best treated by weighted regression. Failure to use weighted regression in these circumstances could give rise to exaggerated errors at the low end of the calibration function.

The test for lack of fit can be carried out with either simple or weighted regression. A test for an intercept significantly different from zero can also be made on this data if there is no significant lack of fit.

A3.2. Test for general matrix effect

It simplifies calibration enormously if the calibrators can be prepared as a simple solution of the analyte. The effects of a possible general matrix mismatch must be assessed in validation if this strategy is adopted. A test for general matrix effect can be made by applying the method of analyte additions (also called “standard additions”) to a test solution derived from a typical test material. The test should be done in a way that provides the same final dilution as the normal procedure produces, and the range of additions should encompass the same range as the procedure-defined calibration validation. If the calibration is linear the slopes of the usual calibration function and the analyte additions plot can be compared for significant difference. A lack of significance means that there is no detectable general matrix effect. If the calibration is not linear a more complex method is needed for a significance test, but a visual comparison at equal concentrations will usually suffice. A lack of significance in this test will often mean that the matrix variation effect [Section A13] will also be absent.

A3.3. Final calibration procedure

The calibration strategy as specified in the procedure may also need to be separately validated, although the errors involved will contribute to jointly estimated uncertainties. The important point here is that evaluation uncertainty estimated from the specific designs for linearity etc., will be smaller than those derived from the simpler calibration defined in the procedure protocol.

A.4 Trueness

A4.1. Estimation of trueness

Trueness is the closeness of agreement between a test result and the accepted reference value of the property being measured. Trueness is stated quantitatively in terms of “bias”; with smaller bias indicating greater trueness. Bias is typically determined by comparing the response of the method to a reference material with the

known value assigned to the material. Significance testing is recommended. Where the uncertainty in the reference value is not negligible, evaluation of the results should consider the reference material uncertainty as well as the statistical variability.

A4.2. Conditions for trueness experiments

Bias can arise at different levels of organisation in an analytical system, for example, run bias, laboratory bias and method bias. It is important to remember which of these is being handled by the various methods of addressing bias. In particular:

The mean of a series of analyses of a reference material, carried out wholly within a single run, gives information about the sum of method, laboratory and run effect for that particular run. Since the run effect is assumed to be random from run to run, the result will vary from run to run more than would be expected from the observable dispersion of the results, and this needs to be taken into account in the evaluation of the results (for example, by testing the measured bias against the among-runs standard deviation investigated separately).

The mean of repeated analyses of a reference material in several runs, estimates the combined effect of method and laboratory bias in the particular laboratory (except where the value is assigned using the particular method).

A4.3. Reference values for trueness experiments

A.4.3.1. Certified reference materials (CRMs)

CRMs are traceable to international standards with a known uncertainty and therefore can be used to address all aspects of bias (method, laboratory and within-laboratory) simultaneously, assuming that there is no matrix mismatch. CRMs should accordingly be used in validation of trueness where it is practicable to do so. It is important to ensure that the certified value uncertainties are sufficiently small to permit detection of a bias of important magnitude. Where they are not, the use of CRMs is still recommended, but additional checks should be carried out.

A typical trueness experiment generates a mean response on a reference material. In interpreting the result, the uncertainty associated with the certified value should be taken into account along with the uncertainty arising from statistical variation in the laboratory. The latter term may be based on the within-run, between-run, or an estimate of the between-laboratory standard deviation depending on the intent of the experiment. Statistical or materials. Where the certified value uncertainty is small, a Student's t test is normally carried out, using the appropriate precision term.

Where necessary and practicable, a number of suitable CRMs, with appropriate matrices and analyte concentrations, should be examined. Where this is done, and the uncertainties on the certified values are smaller than those on the analytical results, it

would be reasonably safe to use simple regression to evaluate the results. In this way bias could be expressed as a function of concentration, and might appear as a non-zero intercept ("transitional" or constant bias) or as a non-unity slope ("rotational" or proportional bias). Due caution should be applied in interpreting the results where the range of matrices is large.

A.4.3.2. Reference materials

Where CRMs are not available, or as an addition to CRMs, use may be made of any material sufficiently well characterised for the purpose (a reference material[10]), bearing in mind always that while insignificant bias may not be proof of zero bias, significant bias on any material remains a cause for investigation. Examples of reference materials include: Materials characterised by a reference material producer, but whose values are not accompanied by an uncertainty statement or are otherwise qualified; materials characterised by a manufacturer of the material; materials characterised in the laboratory for use as reference materials; materials subjected to a restricted round-robin exercise, or distributed in a proficiency test. While the traceability of these materials may be questionable, it would be far better to use them than to conduct no assessment for bias at all. The materials would be used in much the same way as CRMs, though with no stated uncertainty any used in much the same way as CRMs, though with no stated uncertainty any significance test relies wholly on the observable precision of results.

A.4.3.3. Use of a reference method

A reference method can in principle be used to test for bias in another method under validation. This is a useful option when checking an alternative to, or modification of, an established standard method already validated and in use in the laboratory. Both methods are used to analyse a number of typical test materials, preferably covering a useful range of concentration fairly evenly. Comparison of the results over the range by a suitable statistical method (for example, a paired t-test, with due checks for homogeneity of variance and normality) would demonstrate any bias between the methods.

A.4.3.4. Use of spiking/recovery

In the absence of reference materials, or to support reference material studies, bias can be investigated by spiking and recovery. A typical test material is analysed by the method under validation both in its original state and after the addition (spiking) of a known mass of the analyte to the test portion. The difference between the two results as a proportion of the mass added is called the surrogate recovery or sometimes the marginal recovery. Recoveries significantly different from unity indicate that a bias is

affecting the method. Strictly, recovery studies as described here only assess bias due to effects operating on the added analyte; the same effects do not necessarily apply to the same extent to the native analyte, and additional effects may apply to the native analyte. Spiking/recovery studies are accordingly very strongly subject to the observation that while good recovery is not a guarantee of trueness, poor recovery is certainly an indication of lack of trueness. Methods of handling spiking/recovery data have been covered in detail elsewhere.[4]

A.5 Precision

Precision is the closeness of agreement between independent test results obtained under stipulated conditions. It is usually specified in terms of standard deviation or relative standard deviation. The distinction between precision and bias is fundamental, but depends on the level at which the analytical system is viewed. Thus from the viewpoint of a single determination, any deviation affecting the calibration for the run would be seen as a bias. From the point of view of the analyst reviewing a year's work, the run bias will be different every day and act like a random variable with an associated precision. The stipulated conditions for the estimation of precision take account of this change in view point.

For single laboratory validation, two sets of conditions are relevant: (a) precision under repeatability conditions, describing variations observed during a single run as expectation 0 and standard deviation σ_r, and (b) precision under run-to-run conditions, describing variations in run bias δ_{run} as expectation 0, standard deviation σ_{run}. Usually both of these sources of error are operating on individual analytical results, which therefore have a combined precision $\sigma_{tot} = (\frac{\sigma_r^2}{n} + \sigma_{run}^2)^{1/2}$, where n is the number of repeat results averaged within a run for the reported result. The two precision estimates can be obtained most simply by analysing the selected test material in duplicate in a number of successive runs. The separate variance components can then be calculated by the application of one-way analysis of variance. Each duplicate analysis must be an independent execution of the procedure applied to a separate test portion. Alternatively the combined precision σ_{tot} can be estimated directly by the analysis of the test material once in successive runs, and estimating the standard deviation from the usual equation. (Note that observed standard deviations are generally given the symbol s, to distinguish them from standard deviations □).

It is important that the precision values are representative of likely test conditions. First, the variation in conditions among the runs must represent what would normally happen in the laboratory under routine use of the method. For instance, variations in reagent batches, analysts and instruments should be representative. Second, the test

material used should be typical, in terms of matrix and (ideally) the state of comminution, of the materials likely to encountered in routine application. So actual test materials or, to a lesser degree, matrix-matched reference materials would be suitable, but standard solutions of the analyte would not. Note also that CRMs and prepared reference materials are frequently homogenised to a greater extent than typical test materials, and precision obtained from their analysis may accordingly under-estimate the variation that will be observed for test materials.

Precision very often varies with analyte concentration. Typical assumptions are i) that there is no change in precision with analyte level, or ii) that the standard deviation is proportional to, or linearly dependent on, analyte level. In both cases, the assumption needs to be checked if the analyte level is expected to vary substantially (that is, by more than about 30% from its central value). The most economical experiment is likely to be a simple assessment of precision at or near the extremes of the operating range, together with a suitable statistical test for difference in variance. The F-test is appropriate for normally distributed error.

Precision data may be obtained for a wide variety of different sets of conditions in addition to the minimum of repeatability and between-run conditions indicated here, and it may be appropriate to acquire additional information. For example, it may be useful to the assessment of results, or for improving the measurement, to have an indication of separate operator and run effects, between or within-day effects or the precision attainable using one or several instruments. A range of different designs and statistical analysis techniques is available, and careful experimental design is strongly recommended in all such studies.

A.6 Recovery

Methods for estimating recovery are discussed in conjunction with methods of estimating trueness (above).

A.7 Range

The validated range is the interval of analyte concentration within which the method can be regarded as validated. It is important to realise that this range is not necessarily identical to the useful range of the calibration. While the calibration may cover a wide concentration range, the remainder of the validation (and usually much more important part in terms of uncertainty) will cover a more restricted range. In practice, most methods will be validated at only one or two levels of concentration. The validated range may be taken as a reasonable extrapolation from these points on the concentration scale.

When the use of the method focuses on a concentration of interest well above the detection limit, validation near that one critical level would be appropriate. It is

impossible to define a general safe extrapolation of this result to other concentrations of the analyte, because much depends on the individual analytical system. Therefore the validation study report should state the range around the critical value in which the person carrying out the validation, using professional judgement, regards the estimated uncertainty to hold true.

When the concentration range of interest approaches zero, or the detection limit, it is incorrect to assume either constant absolute uncertainty or constant relative uncertainty. A useful approximation in this common circumstance is to assume a linear functional relationship, with a positive intercept, between uncertainty u and concentration c, that is of the form

$$u(c) = u_0 + \theta_c$$

where □ is the relative uncertainty estimated a some concentration well above the detection limit. u_0 is the standard uncertainty estimated for zero concentration and in some circumstances could be estimated as $C_L/3$. In these circumstances it would be reasonable to regard the validated range as extending from zero to a small integer multiple of the upper validation point. Again this would depend on professional judgement.

A.8 Detection Limit

In broad terms the detection limit (limit of detection) is the smallest amount or concentration of analyte in the test sample that can be reliably distinguished from zero.[22,23] For analytical systems where the validation range does not include or approach it, the detection limit does not need to be part of a validation.

Despite the apparent simplicity of the idea, the whole subject of the detection limit is beset with problems outlined below:

There are several possible conceptual approaches to the subject, each providing a somewhat different definition of the limit. Attempts to clarify the issue seem ever more confusing.

Although each of these approaches depends on an estimate of precision at or near zero concentration, it is not clear whether this should be taken as implying repeatability conditions or some other condition for the estimation.

Unless an inordinate amount of data is collected, estimates of detection limit will be subject to quite large random variation.

Estimates of detection limit are often biased on the low side because of operational factors.

Statistical inferences relating to the detection limit depend on the assumption of

normality, which is at least questionable at low concentrations.

For most practical purposes in method validation, it seems better to opt for a simple definition, leading to a quickly implemented estimation which is used only as a rough guide to the utility of the method. However, it must be recognised that the detection limit as estimated in method development, may not be identical in concept or numerical value to one used to characterise a complete analytical method. For instance the "instrumental detection limit", as quoted in the literature or in instrument brochures and then adjusted for dilution, is often far smaller than a "practical" detection limit and inappropriate for method validation.

It is accordingly recommended that for method validation, the precision estimate used (σ_0) should be based on at least 6 independent complete determinations of analyte concentration in a typical matrix blank or low-level material, with no censoring of zero or negative results, and the approximate detection limit calculated as $3\sigma_0$. Note that with the recommended minimum number of degrees of freedom, this value is quite uncertain, and may easily be in error by a factor of two. Where more rigorous estimates are required (for example to support decisions based on detection or otherwise of a material), reference should be made to appropriate guidance (see, for example, references 22-23).

A.9 Limit of determination or limit of quantification

It is sometimes useful to state a concentration below which the analytical method cannot operate with an acceptable precision. Sometimes that precision is arbitrarily defined as 10% RSD, sometimes the limit is equally arbitrarily taken as a fixed multiple (typically 2) of the detection limit. While it is to a degree reassuring to operate above such a limit, we must recognise that it is a quite artificial dichotomy of the concentration scale: measurements below such a limit are not devoid of information content and may well be fit for purpose. Hence the use of this type of limit in validation is not recommended here. It is preferable to try to express the uncertainty of measurement as a function of concentration and compare that function with a criterion of fitness for purpose agreed between the laboratory and the client or end-user of the data.

A.10 Sensitivity

The sensitivity of a method is the gradient of the calibration function. As this is usually arbitrary, depending on instrumental settings, it is not useful in validation. (It may be useful in quality assurance procedures, however, to test whether an instrument is performing to a consistent and satisfactory standard.)

A.11 Ruggedness

The ruggedness of an analytical method is the resistance to change in the results

produced by an analytical method when minor deviations are made from the experimental conditions described in the procedure. The limits for experimental parameters should be prescribed in the method protocol (although this has not always been done in the past), and such permissible deviations, separately or in any combination, should produce no meaningful change in the results produced. (A "meaningful change" here would imply that the method could not operate within the agreed limits of uncertainty defining fitness for purpose.) The aspects of the method which are likely to affect results should be identified, and their influence on method performance evaluated by using ruggedness tests.

The ruggedness of a method is tested by deliberately introducing small changes to the procedure and examining the effect on the results. A number of aspects of the method may need to be considered, but because most of these will have a negligible effect it will normally be possible to vary several at once. An economical experiment based on fractional factorial designs has been described by Youden[13]. For instance, it is possible to formulate an approach utilising 8 combinations of 7 variable factors, that is to look at the effects of seven parameters with just eight analytical results. Univariate approaches are also feasible, where only one variable at a time is changed.

Examples of the factors that a ruggedness test could address are: changes in the instrument, operator, or brand of reagent; concentration of a reagent; pH of a solution; temperature of a reaction; time allowed for completion of a process etc.

A.12 Fitness for Purpose

Fitness for purpose is the extent to which the performance of a method matches the criteria, agreed between the analyst and the end-user of the data, that describe the end-user's needs. For instance the errors in data should not be of a magnitude that would give rise to incorrect decisions more often than a defined small probability, but they should not be so small that the end-user is involved in unnecessary expenditure. Fitness for purpose criteria could be based on some of the characteristics described in this Annex, but ultimately will be expressed in terms of acceptable total uncertainty.

A.13 Matrix variation

Matrix variation is, in many sectors, one of the most important but least acknowledged sources of error in analytical measurements. When we define the analytical system to be validated by specifying, amongst other things, the matrix of the test material, there may be scope for considerable variation within the defined class. To cite an extreme example, a sample of the class "soil" could be composed of clay, sand, chalk, laterite (mainly Fe_2O_3 and Al_2O_3), peat, etc., or of mixtures of these. It is easy to imagine that each of these types would contribute a unique matrix effect on an

analytical method such as atomic absorption spectrometry. If we have no information about the type of soils we are analysing, there will be an extra uncertainty in the results because of this variable matrix effect.

Matrix variation uncertainties need to be quantified separately, because they are not taken into account elsewhere in the process of validation. The information is acquired by collecting a representative set of the matrices likely to be encountered within the defined class, all with analyte concentrations in the appropriate range. The material are analysed according to the protocol, and the bias in the results estimated. Unless the test materials are CRMs, the bias estimate will usually have to be undertaken by means of spiking and recovery estimation. The uncertainty is estimated by the standard deviation of the biases. (Note: This estimate will also contain a variance contribution from the repeat analysis. This will have a magnitude if spiking has been used. If a strict uncertainty budget is required, this term should be deducted from the matrix variation variance to avoid double accounting.)

A.14 Measurement Uncertainty

The formal approach to measurement uncertainty estimation calculates a measurement uncertainty estimate from an equation, or mathematical model. The procedures described as method validation are designed to ensure that the equation used to estimate the result, with due allowance for random errors of all kinds, is a valid expression embodying all recognised and significant effects upon the result. It follows that, with one caveat elaborated further below, the equation or 'model' subjected to validation may be used directly to estimate measurement uncertainty. This is done by following established principles, based on the 'law of propagation of uncertainty' which, for independent input effects is

$$u(y(x_1,x_2,\ldots)) = \sqrt{\sum_{i=1,n} c_i^2 u(x_i)^2}$$

where $y(x_1,x_2,\ldots x_n)$ is a function of several independent variables x_1,x_2.., and c_i is a sensitivity coefficient evaluated as $c_i = \partial y/\partial xi$, the partial differential of y with respect to x_i. $u(x_i)$ and u(y) are standard uncertainties, that is, measurement uncertainties expressed in the form of standard deviations. Since $u(y(x_1,x_2,\ldots))$ is a function of several separate uncertainty estimates, it is referred to as a combined standard uncertainty.

To estimate measurement uncertainty from the equation $y=f(x_1,x_2\ldots)$ used to calculate the result, therefore, it is necessary first, to establish the uncertainties $u(x_i)$ in each of the terms x_1,x_2 etc. and second, to combine these with the additional terms required

to represent random effects as found in validation, and finally to take into account any additional effects. In the discussion of precision above, the implied statistical model is

$$y = f(x_1,x_2...) + \delta_{run} + e$$

where e is the random error for a particular result. Since δ_{run} and e are known, from the precision experiments, to have standard deviations σ_{run} and σ_r respectively, these latter terms (or, strictly, their estimates s_{run} and s_r) are the uncertainties associated with these additional terms. Where the individual within-run results are averaged, the combined uncertainty associated with these two terms is (as given previously) $s_{tot} = (\frac{s_r^2}{n} + s_{run}^2)^{1/2}$. Note that where the precision terms are shown to vary with analyte level, the uncertainty estimate for a given result must employ the precision term appropriate to that level. The basis for the uncertainty estimate accordingly follows directly from the statistical model assumed and tested in validation. To this estimate must be added any further terms as necessary to account for (in particular) inhomogeneity and matrix effect (see section A13). Finally, the calculated standard uncertainty is multiplied by a 'coverage factor', k, to provide an expanded uncertainty, that is, "an interval expected to encompass a large fraction of the distribution of values that may be attributed to the measurand"[8]. Where the statistical model is well established, the distribution known to be normal, and the number of degrees of freedom associated with the estimate is high, k is generally chosen to be equal to 2. The expanded uncertainty then corresponds approximately to a 95% confidence interval.

There is one important caveat to be added here. In testing the assumed statistical model, imperfect tests are perforce used. It has already been noted that these tests can not prove that any effect is identically zero; they can only show that an effect is too small to detect within the uncertainty associated with the particular test for significance. A particularly important example is the test for significant laboratory bias. Clearly, if this is the only test performed to confirm trueness, there must be some residual uncertainty as to whether the method is indeed unbiased or not. It follows that where such uncertainties are significant with respect to the uncertainty calculated so far, additional allowance should be made.

In the case of an uncertain reference value, the simplest allowance is the stated uncertainty for the material, combined with the statistical uncertainty in the test applied. A full discussion is beyond the scope of this text; reference 9 provides further detail. It is, however, important to note that while the uncertainty estimated directly from the assumed statistical model is the minimum uncertainty that can be associated

with an analytical result, it will almost certainly be an underestimate; similarly, an expanded uncertainty based on the same considerations and using k=2 will not provide sufficient confidence.

The ISO Guide[8] recommends that for increased confidence, rather than arbitrarily adding terms, the value of k should be increased as required. Practical experience suggests that for uncertainty estimates based on a validated statistical model, but with no evidence beyond the validation studies to provide additional confidence in the model, k should not be less than 3. Where there is strong reason to doubt that the validation study is comprehensive, k should be increased further as required.

Annex B. Additional considerations for Uncertainty estimation in validation studies

B.1 Sensitivity analysis

The basic expression used in uncertainty estimation

$$u(y(x_1,x_2,\ldots)) = \sqrt{\sum_{i=1,n} c_i^2\, u(x_i)^2}$$

requires the 'sensitivity coefficients' c_i. It is common in uncertainty estimation to find that while a given influence factor x_i has a known uncertainty $u(x_i)$, the coefficient c_i is insufficiently characterised or not readily obtainable from the equation for the result. This is particularly common where an effect is not included in the measurement equation because it is not normally significant, or because the relationship is not sufficiently understood to justify a correction. For example, the effect of solution temperature T_{sol} on a room temperature extraction procedure is rarely established in detail.

Where it is desired to assess the uncertainty in a result associated with such an effect, it is possible to determine the coefficient experimentally. This is done most simply by changing x_i and observing the effect on the result, in a manner very similar to basic ruggedness tests. In most cases, it is sufficient in the first instance to choose at most two values of x_i other than the nominal value, and calculate an approximate gradient from the observed results. The gradient then gives an approximate value for c_i. The term $c_i.u(x_i)$ can then be determined. (Note that this is one practical method for demonstrating the significance or otherwise of a possible effect on the results).

In such an experiment, it is important that the change in result observed be sufficient for a reliable calculation of x_i. This is difficult to predict in advance. However, given a permitted range for the influence quantity x_i, or an expanded uncertainty for the quantity, that is expected to result in insignificant change, it is clearly important to

assess c_i from a larger range. It is accordingly recommended that for an influence quantity with an expected range of $\pm$ a, (where $\pm$ a might be, for example, the permitted range, expanded uncertainty interval or 95% confidence interval) the sensitivity experiment employ, where possible, a change of at least 4a to ensure reliable results.

B.2 Judgement

It is not uncommon to find that while an effect is recognised and may be significant, it is not always possible to obtain a reliable estimate of uncertainty. In such circumstances, the ISO Guide makes it quite clear that a professionally considered estimate of the uncertainty is to be preferred to neglect of the uncertainty. Thus, where no estimate of uncertainty is available for a potentially important effect, the analyst should make their own best judgement of the likely uncertainty and apply that in estimating the combined uncertainty. Reference 8 gives further guidance on the use of judgement in uncertainty estimation.

---

[*] Sampling uncertainty in the strict sense of uncertainty due to the preparation of the laboratory sample from the bulk target is excluded from consideration in this document. Uncertainty associated with taking a test portion from the laboratory sample is an inseparable part of measurement uncertainty and is automatically included at various levels of the following analysis.

[+] Many alternative groupings or 'partitions of error' are possible and may be useful in studying particular sources of error in more detail or across a different range of situations. For example, the statistical model of ISO 5725 generally combines laboratory and run effects, while the uncertainty estimation procedure in the ISO GUM is well suited to assessing the effects of each separate and measurable influence on the result.

[*] This may not be applicable at concentrations less than 10 times the detection limit.