IAEA-TECDOC-1350

Development and use of reference materials and quality control materials

INTERNATIONAL ATOMIC ENERGY AGENCY IAEA

April 2003

The originating Section of this publication in the IAEA was:

Industrial Applications and Chemistry Section
International Atomic Energy Agency
Wagramer Strasse 5
P.O. Box 100
A-1400 Vienna, Austria

DEVELOPMENT AND USE OF REFERENCE MATERIALS AND
QUALITY CONTROL MATERIALS
IAEA, VIENNA, 2003
IAEA-TECDOC-1350
ISBN 92–0–103303–6
ISSN 1011–4289

Printed by the IAEA in Austria
April 2003

FOREWORD

The application of certified reference materials (CRMs) in analytical chemistry for quality control purposes is well recognized and recommended by a wide range of international, national and professional organizations. However, irrespective of the geographical region or the economic situation in laboratories, current practice in CRM application in many analytical sectors is not adequate. Therefore, the International Atomic Energy Agency (IAEA) organized a consultants meeting of a group of experts at its headquarters in Vienna in August 2001 to encourage quality assurance and quality control in nuclear analytical laboratories in Member States. This report is a result of the meeting.

The report summarizes current knowledge on correct use of commercially available CRMs and reference materials (RMs), and also acknowledges the limitations and restrictions analysts have to face if they want to apply quality control. For certain matrix types, CRMs might not be available at all, or the range of concentrations and/or analytes needed might not be certified. In many of the analytical laboratories in developing countries lack of financial resources restrict the comprehensive use of available CRMs that are largely prepared and commercialized in western countries. The concept of in-house RMs or quality control materials (QCMs) is advocated to supplement (*not* substitute) the use of CRMs for quality control purposes. On hand advice on how to select, prepare, characterize and use these QCMs is given from the experts' perspective. Several scenarios are described to make this concept widely applicable to: advanced laboratories with CRMs with validated analytical techniques available, laboratories with less experience and facilities, as well as cases were labile compounds and unstable matrices are involved. Each scenario considers different approaches to overcome the lack of appropriate CRMs and advise on the preparation of QCMs, which might fit the particular purpose.

This publication is intended to assist analytical chemists in their efforts to maintain good quality results and provide them with a tool to overcome situations where QA/QC could not be easily implemented. The report is a contribution to boost quality system implementation and finally encourage nuclear analytical laboratories to prepare themselves for formal accreditation. It is hoped that this initiative will add to the sustainability of nuclear applications in IAEA Member States.

The IAEA wishes to thank all the experts for their valuable contributions and the International Union of Pure and Applied Chemistry (IUPAC) for permission to use the article on Harmonized Guidelines for Internal Quality Control in Analytical Chemistry Laboratories that is annexed to this report.

The IAEA officer responsible for this publication was M. Rossbach of the Division of Physical and Chemical Sciences.

EDITORIAL NOTE

The use of particular designations of countries or territories does not imply any judgement by the publisher, the IAEA, as to the legal status of such countries or territories, of their authorities and institutions or of the delimitation of their boundaries.

The mention of names of specific companies or products (whether or not indicated as registered) does not imply any intention to infringe proprietary rights, nor should it be construed as an endorsement or recommendation on the part of the IAEA.

CONTENTS

1. INTRODUCTION

1.1. REQUIREMENTS FOR DEMONSTRATED QUALITY IN ANALYTICAL LABORATORIES

Testing and calibration laboratories, including analytical chemical laboratories, are continually requested to provide evidence on the quality of their operations. This is mandatory in cases where legislative limits are involved, e.g. in international trade, food and environmental analysis, clinical chemistry, etc. Demonstration of adequate quality is required also in research and development activities.

The general ISO definition of *"**quality**"* is given as *"totality of characteristics of an entity that bears on its ability to satisfy stated and implied needs"* [1].

For a chemical analytical laboratory, the 'entity' will in most cases be a measurement result. In a simplified form the quality requirements would then be represented in the form of reliable, comparable (traceable) results, accompanied with stated measurement uncertainty, produced in an agreed time.

The best and easiest way for laboratories to formally demonstrate their quality is to adhere to an appropriate international quality standard and obtain formal accreditation/certification. Various international or national standards have been prepared, some of them for specific scientific/technical fields, e.g. ISO 9000 series of standard, GLP, EN-45000 series, etc. However, basic quality requirements do no differ significantly. Due to a wide range of activities to which it can be applied and due to the well-established quality assessment structure, the ISO 17025 (1999) 'General requirements for competence of testing and calibration laboratories' [2] is commonly selected as a standard of choice whenever quality assurance in an analytical laboratory is to be demonstrated.

Quality assurance *comprises of all those planned and systematic actions undertaken by the organization necessary to provide adequate confidence that a product or service will satisfy given requirements for quality* [1]. In other words, quality assurance describes the overall measures that a laboratory uses to ensure the quality of its operations. Typical technical components of the laboratory quality assurance are listed in Table I.

TABLE I. TECHNICAL COMPONENTS OF QUALITY ASSURANCE ACTIVITIES IN AN ANALYTICAL LABORATORY

Suitable laboratory environment	Training procedures and records
Educated, trained and skilled staff	Requirements for reagents, calibrants and measurement standards
Suitable equipment, maintained and calibrated	Proper use of (certified) reference materials
Traceable calibrations	Procedures for checking and reporting results
Use of documented and validated methods	Proper storage and handling of samples
Quality control	Participation in proficiency tests

Together with two management requirements: internal audit and management review, quality control is forming the basic pillar of the quality system in an analytical laboratory.

Quality control: *Under this term we refer to operational techniques and activities that are used to fulfil requirements for quality* [1]. In contrast to quality assurance, which is aimed to assure the quality of laboratory operations, quality control is considered as a set of technical operations aimed to assure the reliability of the results for a specific set of samples (or batches of samples). Quality control practices and measures might and should be in place also when the other quality assurance

activities are not (yet) fully implemented. In this document quality control is understood as internal quality control. It describes measures that a specific laboratory takes to assure the quality of its results. Quality control should be distinguished from external quality control, such as proficiency tests, round robin analysis, etc. Although all of them support a laboratory quality assurance, it has to be appreciated that they are complementary activities, which normally cannot directly replace each other. At the same time it has to be realized that quality assurance and quality control activities will overlap in an operational quality system and a distinction will not always be so strict as in this document. Here it is done for the sake of easier understanding. The most common quality control measures for a chemical analytical laboratory are listed in Table II.

TABLE II. QUALITY CONTROL ACTIVITIES MOST COMMONLY APPLIED IN ANALYTICAL LABORATORIES

Analysis of blanks	Analysis of measurement standards, calibrants and reference materials
Analysis of blind samples	Analysis of spiked samples
Analysis in duplicates	Recovery studies
Use of quality control samples	Use of control charts

The IUPAC, International Organization of Standards (ISO) and Association of Analytical Communities (AOAC) International have co-operated in the preparation of the 'Harmonized Guidelines for Internal Quality Control in Analytical Chemistry Laboratory' [3]. This document provides an excellent background information and practical guidance on the execution of quality control procedures. A copy of this document is attached as Annex 1. Quality control principles described in this document have been widely applied in laboratories around the world and have been often cited in the literature. They are also incorporated in quality assurance standards, including ISO 17025. However, the reader will find out that there is only a very general guidance provided on the preparation of QCMs and that the use of (certified) reference materials for the quality control purposes is encouraged whenever possible.

1.2. USE OF CRMs — WHY AND HOW

Certified reference material: *Reference material accompanied by a certificate, one or more of whose property values are certified by a procedure which establishes traceability to an accurate realization to the unit in which the property values are expressed, and for which each certified value is accompanied by an uncertainty at a stated level of confidence* [4].

CRMs are generally prepared in batches for which the property values are determined within the stated uncertainty limits by measurements on sample representative for the whole batch.

All CRMs lie within the definition of 'measurement standards' or 'etalons' given in the 'International Vocabulary of Basic and General Terms in Metrology' (VIM).

Reference material: *Material or substance one or more of whose property values are sufficiently homogeneous and well established to be used for the calibration of an apparatus, assessment of a measurement method, or assigning values to materials.*

A reference material may be in the form of a pure or mixed gas, liquid or solid. Examples are water for the calibration of viscometers, sapphire as a heat-capacity calibrant in calorimetry, and solutions used for calibration in chemical analysis [4].[1]

[1] In this document a distinction between CRMs and reference materials is almost negligible. When a CRM is entioned alone, it might be assumed that the same is valid also for an RM, and vice versa.

CRMs and RMs are being widely used in analytical chemistry. Their proper application provides the best information on (confidence in) the quality of the obtained measurement results. According to ISO Guide 33 'Uses of Certified Reference Materials' [5], they are applied for calibration of an apparatus, method validation, assessment of method and instrument performance, establishing traceability of the measurement results, and determining the uncertainty of these results.

Nomenclature in metrology, and specifically the nomenclature related to reference materials, is quite extensive. Despite or actually due to this fact, there is a certain confusion arising from the improper use of terms. Calibration standards, CRMs, RMs, quality control samples, reference standards, standard reference materials, etc. are often used with the same or very similar meaning. In some cases the RM producers claim a material to be a CRM although not all the requirements from the above definition are met. And even more frequently, the users misuse materials in the analytical process. The reason might be that the users are not aware of the differences or that they have no possibility to purchase an appropriate RM or CRM. It might also happen that the appropriate RM (CRM) is not available at all. For the sake of distinctness in this document, a material used for calibration of an instrument will be called calibrant. The usage of the term calibrant in this text is independent from the type of the material. It might be a pure substance CRM, a mixture or a composite CRM. The basic requirement for the calibrant is that it allows establishing traceability of the measurement result to a defined reference (SI Unit or other international or national standard). Additionally, uncertainty of the assigned property values in cases of calibrants would normally be small. An effort was also put in preparation of examples in this document to exactly define what kind of a material was meant when a specific term is used. This was necessary for the reasons already mentioned, and for another very important reason: In most cases analytical chemists would assume under the term 'certified reference material' or 'reference material' a (natural) matrix reference material.

Matrix (or compositional) reference materials: *A "natural" substance more representative of laboratory samples that has been chemically characterized for one or more elements, constituents, etc. with a known uncertainty* [6].[2]

Matrix reference materials are a specific type of reference materials. Within matrix reference materials different sub-groups exist, i.e. gaseous RMs, environmental and biological matrix RMs, alloys, coal RMs, etc. To a large extent differences in types of RMs are reflected in possible ways of characterization and certification of RMs and consequently in utilization of RMs in the analytical process. A large majority of matrix RMs is characterized through interlaboratory comparisons. Hence, assigned property values are established from the laboratory means. Traceability of these values can normally be claimed only to the respective laboratory intercomparison, and not to any other point of reference. Uncertainty of these values, normally expressed by confidence interval, only gives a measure of the scatter between the laboratory means. The uncertainties of results from individual laboratories are frequently not taken into account. Consequently, matrix RMs do not always fulfil the criteria for the established traceability of the assigned property values. The uncertainty associated with the assigned property values is not quantified as required by the ISO guidance [7]. Certain matrix RMs are prepared for calibration purposes, e.g. gaseous, stable isotope ratio RMs, alloys, etc. However, the majority of matrix RMs (presently available and in future) are and will be suitable for method validation, quality assurance and quality control purposes, but not for calibration.

Basic guidance on the use of CRMs, including matrix reference materials, is given in the ISO Guide 33 'Uses of certified reference materials' [5]. Proper selection and application of reference materials have also been extensively discussed at many conferences and symposia, e.g. BERM Symposia, and are continuously elaborated in the scientific literature. Two of the most recent books are The Use of Matrix Reference Materials in Environmental Analytical Process [8] and Reference Materials for Chemical Analysis [9]. Stricter requirements are continuously set for the production,

[2] This is not a standardized definition.

characterization and certification of reference materials. Producers also need to provide all the necessary information in the accompanying certificates or reports. However, before any reference material is selected and applied in the laboratory, it is the user's responsibility to become aware of the material's characteristics, advantages and limitations.

Particularly important are the instructions for the use of the CRM as stated in the certificate. The certified values do only apply if the material is strictly used according to these instructions. The user needs to follow closely to the recommendations given for storage of the material, eventual drying procedures, and observe the indicated shelf life of an RM. It is not justified to assume the validity of the reference values beyond the expiry date of a given material.

Besides well established (robust) property values — in analytical chemistry this would normally be a concentration, mass fraction, activity concentration, etc. — the most significant advantage of matrix reference material, when available and correctly selected, is its matrix and measurand (analyte) level match in comparison with the test material (sample). In addition, reference materials are normally well characterized for a large number of measurands (analytes) and also in respect to homogeneity of the material. This information is very useful in method development and method validation, providing a basis for estimation of accuracy and precision, as well as for the study of other statistical parameters, such as repeatability, reproducibility, linear range, limit of quantification, robustness and evaluation of eventual interferences. The use of validated — fit for purpose — analytical methods is a prerequisite for any laboratory which would like to claim and formally demonstrate its quality and provide confidence in its measurement results. Method validation might be considered as a large technical/scientific issue in analytical chemistry and a number of guidance documents has been prepared in the last few years. One of the most practically oriented is the EURACHEM Guide, The Fitness for Purpose of Analytical Methods – A Laboratory Guide to Method Validation and Related Topics [10]. Studying this document, the reader will get a very useful insight into the variety of applications of reference materials. In general, matrix reference materials are also very useful for practical assessment and quantification of sources of measurement uncertainty, which needs to be reported along with measurement results [7, 11]. Very often, the assessment of measurement uncertainty is an integral part of method validation.

For almost all of the quality control activities mentioned above, use of CRMs or RMs is the most appropriate choice. Information obtained from their use would be the most extensive and reliable. However, there are also limitations and points for consideration in case of reference materials. Some of them are listed in Table III.

1.3. SUBSTITUTE CRMs FOR QUALITY CONTROL PURPOSES WITH IN-HOUSE QCM

There are situations emerging from Table III that require special attention and when preparation of the "internal quality control" samples might be considered:

- The appropriate reference material is not available at all (matching neither matrix nor measurand).
- RM is available, but too precious to be used for quality control of a larger number of analytical test runs.
- RM is available but is not stable for longer time in sense of matrix or measurand of interest.

TABLE III. POINTS OF CONSIDERATION REGARDING AVAILABILITY OF APPROPRIATE CRMs

Point	Consideration
Availability of CRMs	Is there appropriate CRM available in respect to: – matrix (chemical composition, physical properties and stability) – chemical form and level of measurand – presence of other substances and elements (eventual interferences)? **How well is this CRM comparable to the test material?**
Assigned property values	How where the property values of a specific CRM determined (characterization, certification)? Is there any quantitative information on element distribution in the material given? Homogeneity of the material? Are the assigned property values traceable to any stated reference, e.g. SI Unit? Is the traceability chain demonstrated? **Can this CRM be used for calibration?**
Uncertainty of the assigned property values	Is the uncertainty of the assigned property value given? What does it represent: – measurement uncertainty as a range of the assigned property value (according to the definition of uncertainty) – statistical parameter, e.g. confidence interval of the mean of laboratory means – does the information provided allow calculation of a standard uncertainty? **Can the uncertainty of the assigned property value of a CRM be further used, i.e. in assigning values to other materials?**
Costs of the CRM	How much CRM do we actually need for the intended purpose? What would be a related price?
CRM as a QCM	What is going to be assessed by the use of QCM and how often? How much QCM is needed? **Is it appropriate to prepare our own QCM?**

When one or more of the above situations appear, it might be appropriate to consider the preparation of an internal QCM and/or to strengthen other quality control measures. Different scenarios are given in this document. In all examples, the (high) price of commercially available reference materials is not considered as an acceptable excuse for not having appropriate RMs in place in an analytical laboratory. On the contrary, an RM, when available, should be used as a tool for assigning values to the QCM. However, the question remains when it would be appropriate (if at all) to develop and use an internal QCM instead of the commercially available RM? This question can only be answered when the intended use of a specific QCM is well defined.

1.4. INTENDED USE — PROPER USE OF QCMs

Results of quality control analysis are aimed to provide clear and fast information for the acceptance of analytical results obtained on a batch of test samples. Continuous use of control material allows estimation of the stability of performance of the instrument, including calibration, analytical procedure, analyst, and influence of environmental conditions. Beside well-established values of measurands of interest, a stability and appropriate homogeneity of the control material is required to allow reproducible use of the QCM. In contrast to CRMs or RMs, QCMs have closely defined scope. This would mean that the number of measurands for which the value is determined is normally small. Just those measurands routinely determined in tests samples need to be characterized. In an ideal situation, where one or more appropriate RMs are available, a method validation, including interference studies would be performed using RM and the QCM then continuously applied for monitoring purposes. Continuous use of QCMs allows the analyst to perform statistical analysis on results and to create control charts.

Two types of control charts that are most commonly used are the Shewhart Control Chart and Cumulative Sum (CUSUM) Control Chart. They are graphically presented in Figs 1 and 2.

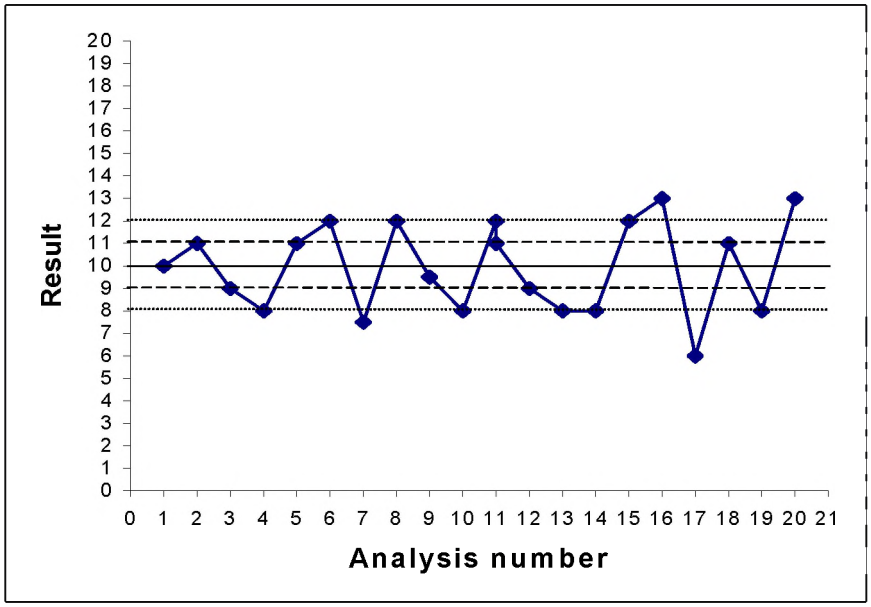

FIG. 1. Shewhart Control Chart (target value is 10; upper and lower warning limits are set at 11 and 9, respectively; action limits are set at 12 and 8).

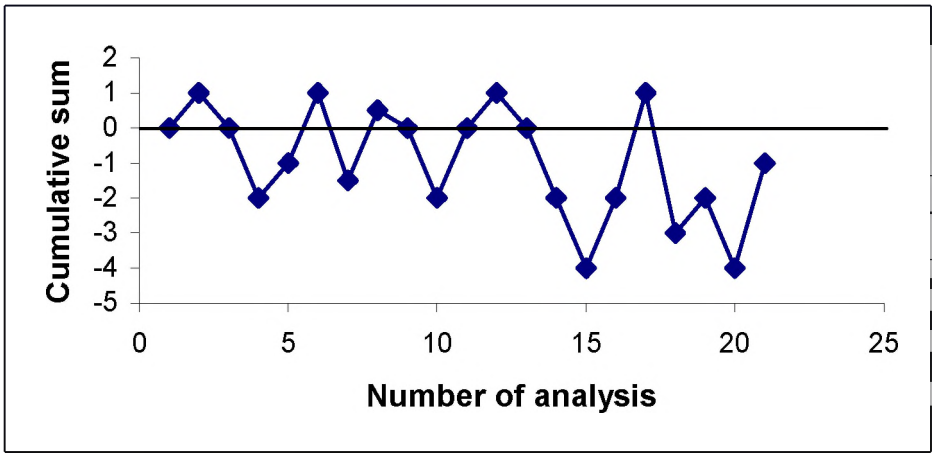

FIG. 2. Cumulative Sum (CUSUM) Control Chart on the same set of data as for the Shewhart Control Chart above.

The theory on control chart for statistical control on analytical performance is explained in many textbooks. The American Society for Testing Materials (ASTM) Manual on Presentation of Data and Control Chart Analysis is suggested for the further reading [11]. Further considerations regarding the use of control charts are given in Sections 2.10 and 5.9.

2. DEVELOPMENT OF QCMs UNDER SCENARIO 1: CRMs AVAILABLE, VALIDATED METHOD UNAVALABLE

2.1. GENERAL CONSIDERATIONS

Production of reliable data is the raison d'être of any analytical laboratory. Data reliability is assured only when appropriate QC/QA regimen is strictly adhered to in the overall analytical process. A key component of the QC/QA procedure is the use of RMs to validate the analytical method, and to ensure that data produced is traceable to a fundamental standard. In order to maintain the long term reliability of the measurement process, it is essential that carefully characterized QCMs are analysed together with the samples. Although the latter role has been suggested for RMs, such volume use of what is an expensive resource is not cost effective. QCMs like RMs must have similar characteristics as the analysed samples but, in contrast to RMs, they are used whenever a sample batch is analysed. Use of QCMs not only provides a check that the measurement process is under statistical control but forms part of an unbroken chain linking the analytical results to a primary standard.

Indeed, QCMs come into their own when no appropriate RM is available for a given application. Consequently, development of QCMs complements the use of RMs in the analytical laboratory. For QCMs to fulfil the desired role of ensuring the reliability of analytical data, the materials have to be characterized for the analytes of interest, and an indication of the measurement uncertainty provided. Of crucial importance in the characterization of QCMs is the availability of validated methods. However, in the absence of validated methods, QCMs can still be produced. Fig. 3 summarizes the various possibilities.

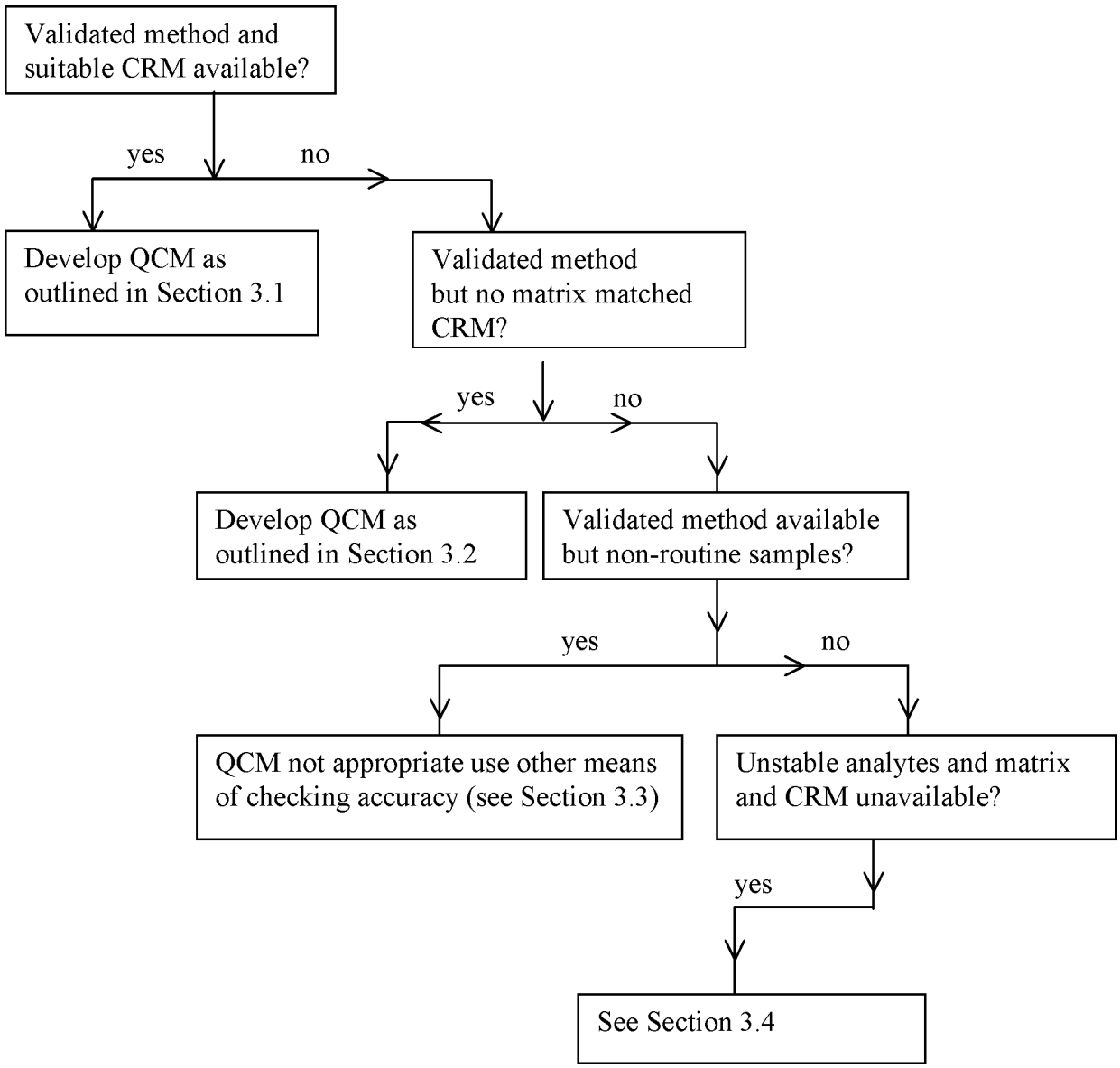

FIG. 3. Decision tree for preparation and use of QCMs.

Although QCMs are lower than RMs in the hierarchy of QC materials, their production should nevertheless meet rigorous standards. Discussed in this section are the steps in the selection preparation, storage, characterization and evaluation of data in the production of QCMs when validated analytical methods are available.

2.2. AVAILABILITY OF VALIDATED METHODS

It is good practice for laboratories engaged in both routine and non-routine measurements to use validated methods. Application of validated methods not only ensures that the analytical results are fit for the purpose for which they are produced, but also makes it possible to issue statements based on their statistical behaviour. A thorough validation process should provide information about

the analytical bias and figures of merit of the method. The former can be established using either a CRM, comparison with an independent validated method, or recovery tests. In practice, use of a CRM to check accuracy ensures that the results produced for the QCM are reliable and traceable to the fundamental standard to which the CRM is linked.

Since validated methods are very well characterized in terms of their analytical performance they should be the first choice for the chemical characterization of QCMs. Other means of obtaining measurement data on the QCMs are explored in some parts of this report.

2.3. MATERIAL SELECTION

The choice of material for QCM preparation must meet similar criteria as those set out for RMs (1–3). Since QCMs are prepared in-house or for a specific purpose the material used can be closely matched in terms of matrix and concentration levels of the analytes to the samples analysed. It is inappropriate to use materials that differ from the sample by more than one order of magnitude in concentration of any of the major constituents. Not only must the samples and QCMs have similar concentration ratio of the matrix constituents but also the levels of the analytes must be close. To meet this criterion, it is essential that the QCM is prepared from the same biological tissue, plant genus, food material, water, soil, etc if these are available.

QCMs can be prepared from pooled samples in cases where a given sample type is analysed routinely. For example in the determination of lead in blood, QCMs can be prepared from pooled blood samples to cover the concentration range, i.e. high, medium and low. In contrast, where different soil types are analysed non-routinely, a case can be made for preparing a representative matrix in which the soil types are pooled. On the other hand, a single soil type whose properties closely match those of a group of soils can be used. If surplus materials of samples are mixed to form a composite QCM, great care must be taken to ensure thorough mixing to guaranty homogeneous analyte distribution in the resulting materials.

A special case can be encountered when the same matrix with different concentrations has to be analysed frequently such as a biomonitor from a pristine and a contaminated site or a food commodity of different origin. In such a case it is well suited to prepare two QCMs, one with low level and one with elevated level of analytes. Not only might these materials be ideally suited with regard to "matrix matching" but also they could cover the entire range of concentrations met during the sample analysis. In extreme cases these materials could be used (through mixtures of the high and low level materials) to create calibration curves for methods such as AAS or voltammetry. The IAEA has prepared two batches of a single cell algae (IAEA 336 and IAEA 413) with natural and elevated levels of elements, which are currently under certification.

2.4. IDENTIFICATION AND SELECTION OF ANALYTES

Once the material has been chosen, it is important to consider the number of analytes that can be determined accurately given the facilities and capabilities of the laboratory. It may in some cases be useful to enlist the co-operation of a reference laboratory. The use of another laboratory may not be necessary if validated methods are available. In the absence of such methods, other means of establishing accuracy such as recovery tests, use of independent methods, and appropriate reference materials should be considered. Whichever approach is chosen, it would be useful in selecting the sample preparation methods such that the form of the analytes in the material are preserved. In addition the analytical method(s) must be robust enough so that changes in the chemical speciation of the analytes do not lead to inaccurate determinations.

Careful consideration must be given to the stability of the material and integrity of the analytes. For production to be cost effective, QCMs, in contrast to RMs, must be subject to the minimum of pretreatment. Ideally, QCMs should be in the form or close to the form in which the samples are analysed.

Visually, the samples and QCMs must be indistinguishable so that laboratory staff do not accord special treatment to the handling of the QCMs. Furthermore, concentration levels of the analytes of interest should be typical of those in the samples in order that measurement uncertainties are comparable. Indeed, special treatment such as freeze drying to improve the storage potential of the sample only serves to increase preparation costs.

2.5. AMOUNT OF MATERIAL

Since QCMs are produced for immediate use, the quantities to be prepared are solely determined by the availability of storage space, length of time the materials can be stored and facilities available for material preparation.

Over and above the quantities used in the laboratory, enough material must be collected to allow for chemical and physical characterization. Physical characterization particularly homogeneity testing is essential for determining whether the total analytical uncertainty fall within acceptable limits. Total uncertainty is made up of measurement and sampling errors with negligible contribution from the latter when the material is homogeneous. However, as material homogeneity increases sampling error makes a greater contribution to the overall analytical uncertainty. Consequently, weight or volume aliquot has to be prescribed so that the sampling error is within acceptable limits. Indeed if the samples and QCMs are closely matched, then the sampling error for a given amount of material would be already known. Therefore, the aim is to prepare the QCM in the same manner as the sample in order to achieve similar results.

The cost of preparing QCMs would depend on the extent of material pretreatment. Whereas some materials may require minimum treatment such as pooling samples and dividing them into suitable aliquots for storage, others may be subjected to multisteps such as comminution, homogenization, and drying before samples are split up in aliquots. Essentially, the nature of pretreatment must be similar to that used for the analytical samples.

Enough QCM to take up the extra capacity in the devices already available in the laboratory should be prepared. The additional costs in preparing the QCM are incurred in the time taken to process larger quantities of samples. As a guide enough QCM should be prepared to enable weekly or monthly control charts to be completed (see Sections 2.10 and 6.9).

Since QCMs must closely resemble the samples being analysed, the materials for their preparation should be obtained from the same sources. Surplus material from regular sampling exercises can be pooled together to prepare QCM. For example, more samples then required for analyses can be obtained from a production line. If facilities exist for processing large sample quantities then materials can be bought but making sure that the essential properties such as the matrix and analyte concentration levels closely match those of the samples. An example of this is the preparation QCMs from fish, food and meat products.

2.6. MATERIAL PREPARATION

Since QCMs must be indistinguishable from the samples, pretreatment steps must be similar. Non-perishable materials can be brought out of storage and prepared like the samples. In contrast, perishable biological and food materials should be prepared and stored in a form suitable either for the penultimate step before analyte determinations or direct analyses.

It is essential that the candidate QCM is not contaminated with the analytes of interest during the preparation steps. In order to reduce the risk of contamination, the same facilities used for preparing the samples should be used for the QCMs. The tried and tested techniques and methods for sample preparation should be applied to the QCMs. Extra cost should not be incurred in obtaining separate equipment for the preparation of the QCMs.

Whereas it is acceptable to stabilize RMs, to stop them from deteriorating in long term storage, the addition of preservatives, irradiation treatment, freeze drying, etc are not recommended for QCMs because these should be used up within the period of their natural shelf life.

2.7. STORAGE

Proper storage is essential for the successful deployment of QCMs. Indeed, in planning the production of a QCM careful consideration must be given to the natural shelf life of the material under conditions available in the laboratory. Some time and effort must be invested in investigating the conditions and length of time under which a material is best stored without compromising its integrity.

QCMs that are best stored frozen should be first divided up into suitable aliquots so that once the material is defrosted it is used up. It is important to ensure that materials stored in bulk do not undergo changes which invalidate any recommendations on taking test portions.

2.8. PHYSICAL CHARACTERIZATION

Extensive physical characterization that is the hallmark for preparing RMs is not needed for QCMs since large quantities, and long term storage are not contemplated. However, it may be necessary initially to examine the physical properties such as moisture content and particle size to confirm that these are similar to those of the samples.

2.9. CHEMICAL CHARACTERIZATION

The key to the successful preparation of a QCM is accurate determination of the analyte(s). Not only must the data on the mean concentration levels be accurate but also some information on the likely spread of the results must be available. Both the mean concentration and the standard deviation of the measurements are required for constructing the control charts, on which data from the use of the QCM will be plotted (see Section 3.1.9).

Availability of validated methods for determination of the analytes is essential for the chemical characterization of the material. Although the reliability of validated methods is already well established, it is essential to test how well the methods perform in the laboratory using CRMs. Results obtained using the methods can be compared with the certified values to establish method bias or systematic error. Repeated determinations on the same CRM will give an indication of random error. The limits of both types of error beyond which the data produced is unacceptable can be set on the basis of the performance characteristics of the methods.

Once it has been established that the methods produce acceptable results, QCMs can be characterized. It is necessary that the CRMs are used again at this stage to ensure reliable results are obtained for the candidate QCMs. The QCM samples should be interspersed with the CRM during analyses so that both are subjected to similar measurement conditions. This is essential if the link between the CRMs and QCMs are to be maintained, and the traceability chain is unbroken.

2.10. STATISTICAL EVALUATION

Data obtained from the determination of an analyte in the QCM are used to construct a control chart. Marked on the y-axis of the control chart is the mean value of the analyte and boundaries corresponding to ± 2 and ± 3 times the standard deviation of the measurements, respectively. The former is called the warning and the latter the action limits, respectively. On the x-axis, the dates on which the analyses were performed are entered. If the results entered on the control chart cluster within ± 2s then the method is said to be under statistical control. Results outside ± 2s but within ± 3s indicate that the analytical method may not be functioning as required. Consequently, subsequent measurements have to be scrutinized for evidence of systematic error. Values that fall outside ± 3s

indicate that the method is not under statistical control. The results obtained can be said with 99% confidence that they do not belong to the population data set for the QCM. In order for reliable statistical statements to be made about the analytical data from the analyses of the QCM, it is essential that the mean and standard deviation values used to set up the control chart are obtained from a good size sample population. Therefore the number of replicate determinations on the QCM is critical. Initially, at least ten independent replicate analyses of the candidate QCM must be carried out. The data obtained from this exercise can be used to calculate the mean and standard deviation values. These values can be adjusted as more data become available during the deployment of the QCM. This can only be done when enough material is available to cover a period of months or a year.

3. DEVELOPMENT OF QCMs UNDER SCENARIO 2: VALIDATED METHOD AVAILABLE, NATURAL MATRIX (C)RM UNAVAILABLE

3.1. INTRODUCTION

Within a laboratory's quality control programme, incorporation of appropriate, compositionally-similar RMs (defined as any material, device or physical system for which definitive numerical values can be associated with specific properties and that is used to calibrate a measurement process) is a valuable, cost effective aspect of a good quality control programme, and a way of transferring accuracy from well defined methods of analysis to the laboratory [13–17]. Results obtained with the RM taken concurrently through the analysis with actual samples are compared with the certified values. Closeness of agreement indicates performance of the analytical method and may suggest the need for modifications to reduce errors. This important component of quality control is not possible in this scenario 2 as the assumption is that no natural matrix RM is available.

A previous section in this report discussed in depth the in-house development preparation as well as physical and chemical characterization of QCMs. It dealt with QCM chemical characterization under the situation of the case (scenario 1) where (1) a validated method of analysis is available and (2) appropriate CRMs are available. The chemical characterization process involved material analysis using, essentially one selected analytical method, with confirmation of the measurement process by incorporation of CRMs or RMs into the measurement scheme.

In this scenario (scenario 2), the significant difference is that suitable natural matrix CRMs are unavailable for confirmation of the measurement process. Thus the assumptions are: (1) a validated method of analysis is available and (2) appropriate natural matrix CRMs are **not** available. A discussion is presented of factors to be considered in the characterization of in-house QCMs under the conditions noted. Some guidelines are offered regarding approaches to the many considerations required for such an endeavour.

3.2. GENERAL CONSIDERATIONS

QCM preparative considerations are identical to those described previously in this document. The chemical characterization approach, however, must be modified somewhat to account for the unavailability of natural matrix CRMs to control the analytical procedure. The central approach here is to use a principal method, validated carefully and with certainty, and demonstrated to be appropriate and applicable to the task at hand, for the chemical characterization of the QCM, with analyses backed up by use of a second (or more) method, independent in theory and practice. Use of the two such methods replaces the absence of the CRM in order to assess the analytical process for the verification of the QCM. T-tests to compare the averages as well as F-tests to compare the variances are performed on the replicate analytical results. If the results are not significantly different (or if different but deemed to be suitable for pooling, as is discussed later) they can be averaged and the variances pooled to determine the target value of the analyte and the associated uncertainty. If results of methods are deemed to be significantly different, then an investigation into the reasons for

the differences is conducted, and, if possible, a third, different method is applied. Combined results from the two (or three) methods are considered as target values for the QCM.

3.3. DETAILED CONSIDERATIONS OF STEPS RECOMMENDED FOR THE DEVELOPMENT OF IN-HOUSE RMs/QCMs

As presented at the Consultants Meeting and published as an individual contribution in the report of that meeting [18], in general, a number of factors should be considered in the development of food-based and in fact any biological in-house QCMs. Some guidelines are offered therein regarding a preliminary draft recommendation of the sequence of 26 steps (Table IV) to be considered in developing a QCM for a range of natural matrices and analytes. The steps are based on a perusal of many original publications and reports compiling preparatory and measurement details on biological materials [19–22], details by this author on his long term preparatory and measurement activities [23–24], in the many internal reports from IRMM/BCR on RM preparation and European Commission Reports on interlaboratory analyses of BCR materials and guidelines from BCR [25] and ISO [26–27]. Although all steps are common to both scenarios 1 and 2, modifications to many of the steps (nos. 3, 8–24) are required for scenario 2. Consequently, details of these steps are presented below, with particular emphasis being placed on modifications in light of the particular, specific requirements of scenario 2.

TABLE IV. RECOMMENDED STEPS FOR THE DEVELOPMENT OF IN-HOUSE QCMs[a]

1.	Nomenclature and definitions of reference and control materials
2.	Nomenclature and definitions of `Certified Values' and related concentration terms
3.	Overall measurement system
4.	Material preparation
5.	Physical characterization
6.	Material stability
7.	Material homogeneity
8.	Analytical characterization (certification) philosophy — approaches to the establishment of concentration values
9.	Definition of analytical methods
10.	Selection of measurands for characterization
11.	Performance of analytical methods
12.	Selection of analytical methodologies
13.	Selection of analysts/laboratories
14.	Selection of statistical protocols, uncertainty statements
15.	In-house (initiating/coordinating laboratory) characterization
16.	Co-operative interlaboratory analytical characterization campaign

17.	Data quality control of in-house and interlaboratory analyses
18.	Critical evaluation of the methods used by co-operators
19.	Evaluation of data on technical merits
20.	Evaluation and selection of multilaboratory/multimethod analytical data
21.	Statistical treatment of data
22.	Calculation of concentration values and associated uncertainties
23.	Reporting of results and information
24.	Publication of protocol followed
25.	Testing and applying this protocol
26.	Future status of reference and informational values

[a] Consideration (with necessary modifications) of steps 3 and 8–24 are required for scenario 2.

3.3.1. Discussion of relevant steps

Step 3. Overall measurement system

For analytical values, for the characterization exercise to be meaningful, the measurement process must produce precise numerical values of the property under analysis that are free of, or corrected for, all known systematic errors within agreed or practical limits required for the end use of the material; such values are also related to the "true value". Existing RMs can be used within an accuracy-based measurement system to serve as vehicles for transfer of accuracy of a definitive method to the measurement process and the numerical data generated there from. In the absence of CRMs the approach is to use a principal method, validated carefully and with certainty, and demonstrated to be appropriate and applicable to the task on hand of chemical characterization of the QCM, with analyses backed up by use of a second method, independent in theory and practice. Use of the two such methods replaces the quality control step utilizing CRMs in order to assess the analytical process for the verification of the QCM. T-tests to compare averages as well as F-tests to compare variances are performed on the replicate analytical results. If the results are not significantly different, or if different, deemed to be acceptable, they can be averaged and the variances pooled to determine the target value of the analyte and the associated uncertainty. If results of methods are significantly different, or deemed to be unacceptably different, then an investigation into the reasons for the differences is conducted, and, if possible, a third, different method is applied. Combined results from the two (or three) methods are considered as target values for the QCM.

Step 8. Analytical characterization (certification) philosophy

Probably the most difficult and challenging task of the QCM development process is analytical characterization (certification), i.e. the process of obtaining concentration data that approach as closely as possible the "true value", together with uncertainty limits. Chemical characterization for quantification or certification purposes encompasses analyte selection based on nutritional, toxicological and environmental significance as well as availability of suitable analytical methodologies and analysts. It includes selection of certification protocols based on definitive, reference and validated methodologies, selection of expert analysts applying conceptually different approaches, selection, development, assessment and validation of methodologies and adaptation of

statistical protocols for data analysis [16–18]. The literature on RM certification indicates that there are two broad types of approaches for the characterization of RMs: (1) statistical and (2) measurement. The statistical approach relies on the in-depth application of statistical calculations to a body of, often widely scattered and discordant, analytical results obtained from diverse exercises. The approach based on measurement emphasizes laboratory measurement aspects and deals more in detail with various diverse analytical measurement possibilities to generate a coherent dataset, followed by necessary minimal calculations. This is elaborated in Section 3.2.4, Approaches to the Characterization/Certification of Reference Materials. Major approaches to characterization/ certification may be classified as:

(1) Definitive method — one organization
(2) Independent reference methods — one organization
(3) Independent reference and validated methods by selected expert analysts — multiple organizations and laboratories
(4) Volunteer analysts, various methods — multiple organizations and laboratories
(5) Method-specific — characterization by a specific, validated method by selected expert or experienced analysts — multiple organizations and laboratories.

Of these, approaches 1 to 4 are considered possible for the generation of the usually demanded, method-independent values for control materials, with approaches 1 to 3 deemed to be the most viable. Application of any of the five approaches is, of course, feasible without the incorporation of CRMs with the risk of reduced reliability. The QCM developer will have to determine which approaches are within his reach and appropriate for the venture; it is expected that 2 and 3 are the most likely in the normal course.

Step 9. Definition of analytical methods

Three methodology terms should be kept in mind as they are intimately integrated into the measurement system and the first two are utilized in the characterization of QCMs. A *definitive method* of chemical analysis is one that has a valid and well described theoretical foundation, has been experimentally evaluated to lead to negligible systematic errors and a high level of precision. Definitive methods provide the fundamental basis for accuracy in chemical analysis. Such methods usually require highly skilled personnel, are time-consuming as well as expensive to perform. A *reference method* is a method of proven and demonstrated accuracy established by direct comparison with a definitive method or with a primary RM. Since reference methods may also be moderately sophisticated, their use may not always be possible. Reference methods can be used to produce secondary reference materials, and control the accuracy of quality assurance procedures. The term *field method* denotes any method of chemical analysis used in applications requiring large numbers of measurements on a routine basis usually with automated instrument systems capable of producing highly precise (but not necessarily accurate) data. Definitive and reference methods of analysis are generally considered for high quality work, including that for the characterization of RMs and QCMs.

Step 10. Selection of analytes for characterization

This is of course specific to the requirements of the laboratory developing the QCM. Those of interest, including others of potential interest in food analysis may be selected for simultaneous cost effective characterization.

Step 11. Performance of analytical methods

Analytical procedures are subject to many sources of error starting with sampling and sample preparation and ending with the calculation and recording of the results. Their accuracy, systematic error and precision, cannot readily be evaluated by means of any single test. However, a substantial part of the whole procedure can generally be tested by the use of appropriate analytical QCMs,

especially certified RMs. There is a need to ensure good performance for the purpose of QCM characterization. In absence of natural matrix CRMs, resort must be made to other tests, as mentioned in this section, in order to establish method performance and validity.

Step 12. Selection of analytical methodologies

If a large number of analyte/material combinations are required for which concentration values are targeted for the QCM(s), this will necessitate a large number of analytical methods. Basically, the reason is fundamentally inherent methodological limitations. In particular, in absence of CRMs for method and procedure verification, there is an important requirement for additional suitable methods. In work on the elemental certification of 12 RMs for 303 assigned values [23], overall, 13 major classes of methods were used including the usual currently used single and multielement instrumental techniques ranging from atomic absorption and emission spectrometry, mass spectrometry, neutron activation analysis and electrochemically-based techniques to the classical Kjeldahl method for nitrogen, light absorption spectrometry, fluorometry and gravimetry. Purposely, an attempt was made to get wide ranging techniques and procedures with different sample preparation steps, including no decomposition as in instrumental neutron activation analysis and particle induced X ray emission spectrometry, as well as different detection/measurement techniques.

Analytical methods should include nuclear methods [31, 32]. In the above work [23, 31] six different variants of neutron activation analysis (NAA) methods were employed including: instrumental neutron activation analysis, instrumental neutron activation analysis with acid digestion, neutron activation analysis with radiochemical separation, neutron capture prompt gamma activation analysis, epithermal instrumental neutron activation analysis, and neutron activation analysis with preconcentration. Methods based on NAA were found to rank typically in the middle of the range with the three other major analytical methods (atomic absorption spectrometry, atomic emission spectrometry, mass spectrometry) with respect to precision. NAA methods, however, distinguished themselves by often exhibiting superior accuracy. These facts, together with the need for no sample treatment in the case of INAA, the version used in the vast majority of NAA applications, make contributions by NAA methodologies, extremely valuable to RM/QCM characterization. The QCM developer will have to determine whether access to such methods are within his reach and appropriate for the venture.

Step 13. Selection of analysts/laboratories

Following selection of appropriate and desired methods of analysis for characterization of the QCM, it is highly probable that not all methods and requisite expertise will be available in-house. This will especially be so if natural matrix CRMs are unavailable for verification of in-house methods and procedures. The QCM developer will have to consider engaging the participation of analysts in external laboratories. Participating analysts and laboratories should be selected on the basis of their established capabilities (competence, experience, motivation, healthy scepticism concerning results obtained). The reliability of analyses seems to depend much more on the analyst than on technique.

Step 14. Selection of statistical protocols, uncertainty statements

Statistical protocols for homogeneity testing, in-house and outside laboratory analyses, dealing with aberrant data and calculation of assigned values and associated uncertainties must be selected. Analysis of variance (ANOVA) and variance component calculations would be typical; various plots (e.g. concentration versus unit, concentration versus laboratory number, concentration versus observation number) could be made for inspecting, assessing and selecting results for use in calculating reference and informational concentration values and uncertainties. It is recommended that input from a statistician be available.

Step 15. In-house (initiating/co-ordinating laboratory) characterization

Depending on the extent and complexity of the QCM project, the initiating laboratory will make either a major or minor contribution to the analytical characterization effort. This will principally depend on locally available methodological, instrumental and technical competencies. The laboratory should be involved at least in preliminary analyses, homogeneity studies and contributing data to final reference values.

Step 16. Co-operative interlaboratory analytical characterization campaign

Should a large number of materials and a wide range of analytes be involved and there be lack of requisite techniques in the initiating laboratory, involvement of outside analysts will be necessary. Following selection of targeted and desired methods of analysis, analysts should be selected on the basis of their established capabilities. A conscious attempt should be made to get wide ranging techniques and procedures including different sample preparation steps. The initiating laboratory will coordinate the overall preparation and characterization effort. Clear, sufficiently detailed instructions and forms for reporting methodological details and analytical results are to be provided.

Step 17. Data quality control of in-house and interlaboratory analyses

Usual data quality control procedures should be in place in the initiating laboratory and a request for the same should be made to outside participants. This generally includes emphasis on the importance to the undertaking of the simultaneous incorporation of appropriate RMs into the scheme of analysis, which, of course, is omitted in absence of CRMs. Reliability of the data generated in the characterization exercise is related to the concept of 'traceability', that is the relating of acquired data to a national or international reference through an unbroken chain of comparisons all having stated uncertainties [33].

Step 18. Critical evaluation of the methods used by co-operators

Critical evaluation of analytical methods and procedures used is complementary to, but independent, from evaluation of submitted results. This step becomes even more significant in the absence of CRMs. Complete descriptions of methods followed may be submitted as scientific journal articles or laboratory notes. Evaluation relies on the initiating analyst's experience and his interpretation of the validity and appropriateness of the applied methods and procedural details as they relate to the specific matrices and measurands under investigation. Many combinations of sample treatment, and detection and measurement schemes will lead to multiple variants of each method; each must be considered.

Step 19. Evaluation of data on technical merits

This item, closely related to step 18, deals with the evaluation of the participant-submitted results in light of the methods used, specific procedures followed and any reports of difficulties or particularities during the conduct of the analysis. A good and interesting approach is ingrained in the RM development activities of the European Commission (BCR) whereby meetings between organizers of the RM programme and participating analysts from member countries are held for dialogue to critically assess results.

Step 20. Evaluation and selection of multilaboratory/multimethod analytical data

Information to be requested from participating analysts to aid in data evaluation can include: analytical method; brief details on sample preparation, instrumentation used and detection limit (and definition) for each analyte; analysts involved in analyses; the number of instrumental readings taken to give each mean concentration; calibration; instrumental precision; unusual occurrences observed during the work; nominal subsample masses taken; concentration results. Similar information,

separately provided for materials utilized for quality control, is absent when CRMs are unavailable. Analytical results are perused and requests for clarification, remeasurement or additional information are made as required. Prior to final calculations of reference and information values, the analytical concentration results are carefully inspected using technical, statistical (variation and bias) and judgement criteria to remove aberrant, outlying or non-representative data.

Step 21. Statistical treatment of data

Dealing with outliers is an early order of business, carried out in perhaps three stages: (1) deletion of obviously erroneous, aberrant or outlying data (2) inspection of concentration versus laboratory number plots and deletion of all data for an analyte/matrix combination with excessive within-laboratory variation or systematic errors (bias) relative to data from the other laboratories (confirmation of rejection by noting performance with certified RMs is not possible in absence of CRMs) (3) repetition of (2) and rejection of additional individual outliers or entire sets from a laboratory when their retention has a serious impact on final uncertainty (spread of accepted results). Outlier rejection criteria can include the following considerations: (1) poor within-laboratory precision compared to that of other laboratories, (2) poor within-subsample precision (within-laboratory instrumental precision) compared with similar parameters of other laboratories, (3) laboratory systematic error judged by deviation of laboratory mean from overall mean, (4) accuracy, based on performance with certified RMs, (5) within-laboratory precision with certified RMs, (6) assessment of the technical merit of the analytical procedure, (7) number of subsamples analysed compared to that in other laboratories. The usual analysis of variance (ANOVA) and variance component calculations can be carried out on the final dataset. It is this author's view that, in certification, painstaking care is needed in the selection of the co-operating laboratories and analytical methods and the main effort is in the generation of an excellent, tight dataset with small systematic errors and uncertainties, which is then subjected to minor mathematical manipulations to arrive at final property measures [39].

Step 22. Calculation of concentration values and associated uncertainties

This is final step in certification. To avoid what has been denoted 'confusion as to the meaning of the uncertainties that are attached to the concentration values for trace elements in biological materials' and to avoid 'statements that cannot be interpreted in a meaningful quantitative or statistical way', guidelines for evaluating and expressing uncertainty should be consulted [38]. The reference value can be computed as the mean of equally weighed individual laboratory means. The associated SD can be calculated from the three variance components representing within unit (σ_w^2), among unit (σ_u^2) and among laboratory/method (σ_L^2) variation according to the following equation:

$$SD = (\sigma_w^{\,2} + \sigma_u^{\,2} + \sigma_L^{\,2})^{1/2} \qquad (1)$$

where each σ indicates the estimates of the associated variance component obtained from a type I (hierarchal) variance component analysis. The SD is the basis for calculation of a 95% confidence interval or uncertainty interval for a future single observation.

Step 23. Reporting of results and information

A document should be prepared for each QCM developed. The documentation is akin to a Certificate or Report of Analysis issued by the certifying agency or a Report of Investigation whose sole authority is the author for CRMs. Critically important information should be included to define and describe the material, describe its preparation and characterization, list numerical values for properties together with the associated uncertainties (as well as their definitions), stipulate minimum weight to be taken for analysis, indicate conditions of storage and include other details necessary for the analyst to correctly and fully utilize the material. BCR, ISO and other guidelines for contents of certificates should be consulted [25, 39].

Step 24. Publication of protocol followed

In-depth treatment of all aspects of the development procedures should be available in accessible technical or scientific publications for information to the in-house analyst/user and to interested outside parties. These documents or supplementary published material may also contain a listing of all individual data, final values, methods and analysts.

3.3.2. Approaches to the characterization/certification of reference materials

All steps in RM/QCM development require appropriate and critical care in their execution and pose varying degrees of difficulty. The task of chemical characterization (denoted certification when applied to RMs by RM developers), defined as the assignment of concentration data which approaches as closely as possible the "true value", together with uncertainty limits, is one of the most, if not the most, demanding challenges in RM/QCM development. Strictly, certification implies the reliable assignment of a value to a property of a material by a legally-mandated, standards-producing, national or international agency. It encompasses selection of analytes, suitable analytical methodologies, analysts and the certification protocol. Importantly, it relies on an accuracy-based measurement system, that "*... produces precise numerical values of the property under test or analysis that are free of, or corrected for, all known systematic errors, and are also related to the "true value" of the property under test or analysis*" [40]. The ideas and philosophy behind certification can be adopted to some extent to the characterization of QCMs.

Legally, the certification process indicates that a certified RM carries with it the full weight and legal authority of the legally mandated national or international organization. Scientifically, *certification deals with the establishment of "true values"*, with the provisions that [15, 40]: (1) systematic errors in the measurement process leading to certification are always investigated, but it should be realized that advances in the state of the art may uncover additional systematic errors that were unsuspected at the time of the original work; therefore, a cautious, conservative estimate of residual and unknown systematic error is the rule, and this should always be reflected in the final stated uncertainty; (2) every material is inherently unstable and property values will change with time; and (3) certified values are only valid when the RM is used in the manner for which it is intended and with all stated precautions followed by the user. The following approaches to the characterization/certification of RMs will be of use to QCM development and can serve as benchmarks for selecting and developing specific approaches for in-house QCM development. Consideration of such multimethod approaches is especially pertinent to this scenario as in absence of RMs for quality control, reliance on methods takes on significance.

3.3.3. General principles of certification

There are two broad types of approaches for the characterization of RMs: (1) statistical and (2) measurement. The statistical approach relies on the in-depth application of statistical calculations to a body of, often widely scattered and discordant, analytical results obtained from diverse exercises. The measurement-based approach emphasizes laboratory measurement aspects and various analytical measurement possibilities to generate a coherent dataset, followed by necessary (minimal) calculations. The measurement-based approach is the one preferred by the author and focused on here.

The key characteristic of an RM (QCM) is that the properties of interest are measured and certified on the basis of accuracy. Several different philosophies have been utilized in the quest for the best estimate of the true value: a reliable and unassailable numerical value of the concentration of the chemical constituent, under constraints of economics, state of the art analytical technologies, availability of methods, analyst competence, availability of analysts and product end use requirement. The basic requirement for producing reliable data is appropriate methodology, adequately calibrated and properly used.

It is generally accepted that a property can be certified when the value is confirmed by several analysts/laboratories working independently using either one definitive method, or more likely, two or

more methods of appropriate and equivalent accuracy. These approaches require development of a statistical plan for sampling and measurement, selection of reliable methodology of known and demonstrable accuracy, maintenance of statistical control of the measurement process, and quality assessment of the data by concurrent measurements of suitable RMs. These measurements require that the systematic and random errors of the procedures used to determine the particular constituent be sufficiently well known to state the concentration of the analyte within a required uncertainty level.

A meaningful measurement process must yield numerical values that are (1) specific, reflecting only the property under test, (2) precise, and (3) free of systematic error (or bias) within the agreed on or practical limits required for the end use; the resulting numerical value can be equated to the 'true value'. This composite description of the major certification approaches followed by the many agencies and individuals involved in RM development is from a synopsis and a detailed write-up by the author [29, 30]. There are advantages and disadvantages to each certification approach. The most important consideration in any scheme is that systematic errors inherent in the methods are well characterized and minimized to the greatest extent possible. Systematic errors, method precision, material variability and material stability must all be understood and taken into account when deriving the uncertainty statements for certified properties.

3.3.4. Classification of characterization/certification schemes

All of the major RM characterization/certification approaches, for total concentrations of constituents, can be classified in one of four categories. A fifth approach deals with method-specific schemes in which characterization is by a defined method giving a method-specific assessed property value. The following is a classification based on the author's interpretation and adaptation of descriptions of RM certification procedures in the literature [29, 30].

(1) Definitive method — one organization: *a single definitive method used by a single organization, of the highest reputational quality, preferably applied in replicate by two or more highly skilled analysts, in more than one separate laboratories, working totally independently, preferably using different experimental facilities, with equipment and expertise to ensure traceability to the SI system. An accurately characterized, independently different, backup method, independently applied, is employed to provide additional assurance that the data are correct.*

The term “definitive method” is applied to an analytical or measurement method that has a valid and well described theoretical foundation, is based on sound theoretical principles, (“first principles”), and has been experimentally demonstrated to have negligible systematic errors and a high level of precision. While a technique may be conceptually definitive, a complete method based on such a technique must be properly applied and demonstrated to deserve such a status for each individual application. A definitive method is one in which all major significant parameters have been related by a direct chain of evidence to the base or derived units of the SI. The property in question is either directly measured in terms of base units of measurement or indirectly related to the base units through physical or chemical theory expressed in exact mathematical equations. The written protocol indicates how each of these critical parameters in the measurement process has been controlled, how traceability to the base units has been accomplished, and what the bounds are to the limits of systematic error and thus uncertainty. Such methods, applied with high reliability, give 'true values' and provide the fundamental basis for accuracy in chemical analysis. Examples of definitive methods are: isotope dilution mass spectrometry, gravimetry (including fire assay analysis), coulometry and calorimetry.

Limitations of time, technical skills, specialized equipment and resources preclude the widespread use of the definitive methods. Furthermore, most analytical methods cannot be classified as definitive methods, usually because there is no straightforward theory relating all the experimental variables to the final result (e.g. the common techniques of atomic absorption and emission spectroscopies), or because effects, including matrix effects, are too complex to handle by theory. The certification of an RM by one measurement method requires a method of high scientific status,

laboratories of the highest quality and skilled analysts. The method must be sufficiently accurate to stand alone and reported results must have negligible systematic errors relative to end use requirements of the data. The acceptance of an RM certified in this way depends on the user community's confidence in the ability of the certifying agency to carry out the definitive method.

Independence of analysts and analyses in one organization is a fundamental question. It is important to have, even for the most reliable methods, more than one analyst/laboratory involved to avoid possible analyst/laboratory specific biases; certification by a single laboratory, without confirmation by another laboratory or method is risky. Measurement by a single definitive method is usually performed by two or more analysts working independently to minimize possible biases. Frequently, an accurately characterized backup method is employed to corroborate the data. Some agencies feel that a certification campaign should not be based on a single measurement procedure and therefore do not normally certify values on the basis of a definitive method applied in one laboratory.

(2) Independent reference methods — one organization: *Two or more independent reference methods, each based on an entirely conceptually different principle of measurement, independent in theory and experimental procedure, applied in replicate, within a single organization, of the highest reputational quality, by two or more expert analysts, working independently. The methods used can, naturally, include definitive methods. The results should be corroborated by a third or additional, independently different, accurately characterized, well established, thoroughly validated, definitive, reference or other methods.*

A reference method is defined as a method of known and proven accuracy, thoroughly validated and experimentally demonstrated to have negligible systematic errors and a high level of precision. Its development involves removing the principal systematic errors of the measurement process, reducing them to tolerable levels, or when actual physical elimination is impossible, applying correction factors. The meaning of the term "independent" is that the basic theoretical and experimental principles on which one method rests, are entirely different from the principles of the other method(s).

Reference methods are generally arrived at by consensus and fairly extensive testing by a number of laboratories. For example, the flame atomic absorption method for Ca in serum developed under the leadership of NIST [41] was established after several interlaboratory comparison exercises. The results were evaluated after each exercise and the procedure was changed as necessary. After five exercises, it was felt that the state of the art had been reached, with the reference method being capable of measuring Ca in serum with an accuracy of ±2% of the true value determined by IDMS (note that attainment of high accuracy and precision is not only a matter of the method but is a function of both the method and analyst expertise).

Since definitive methods are often unavailable, the multimethod approach is more frequently used in certification. A necessary condition for the certification of constituent concentrations is that determinations must be made by at least two independent, complementary, valid, reliable methods to avoid systematic errors associated with any one particular method or technique. Such measurements must agree within reasonable limits to permit certification. If significant discrepancies among analytical results from the different methods occur, additional work is carried out to reconcile them; otherwise the property values cannot be certified. Every effort should be made to use methods based on more than one principle of measurement (three independent principles being desirable) and to engage trained and experienced analysts. The independent reference methods approach is based on the rationale that the likelihood of two independent methods being biased by the same amount and in the same direction is small. Therefore, when the results from two, three or more independent reference methods agree, one can have a high degree of assurance that they are likely to be accurate. The philosophy of basing RM results on determinations by at least two independent methods of analysis or on determinations by a definitive method is often referred to as 'the National Bureau of Standards approach'. It may be noted that NBS originally relied largely on approach 1, using isotope dilution mass spectrometry.

Concern on the independence of analysts in one organization still remains a fundamental question. Again it is important to have more than one analyst/laboratory involved to avoid possible analyst-specific and laboratory-specific biases. Characterization should be corroborated by additional methods or by laboratories in order to provide additional assurance that the data are correct.

(3) Independent reference and validated methods by selected expert analysts — multiple organizations and laboratories: *Two or preferably three or more independent reference and/or validated methods, each based on an entirely conceptually different principle of measurement, independent in theory and experimental procedure, applied in replicate, by selected expert analysts, of high reputational quality and recognized competence working independently in an ad hoc network of laboratories participating in the collaborative interlaboratory characterization campaign under very carefully prescribed and controlled conditions. The methods used can, naturally, include definitive methods. All analytical methods are well characterized and established, thoroughly validated, of acceptable demonstrated accuracy and uncertainty, and the exercise is planned to incorporate widely different methods, based on different physical or chemical principles. The analysts, are carefully selected, on the basis of their established capabilities for the consistent production of precise and accurate results, reputation, expertise and experience in the specific field of analysis, familiarity with the matrix investigated, appreciation for the RM development concept, and a sense of healthy scepticism, and participate* ***, on invitation****, in the analytical campaign.*

For the wide variety of materials and constituents in RMs, reference methods and definitive methods as well as in-house, single organization competencies are often not available. Thus the certifying agency cannot utilize certification approaches 1 or 2 but must resort to this approach relying on independent analysts and laboratories, using different (validated) methods. In this, a combination of definitive and reference methodologies applied by a single organization (approaches 1 and 2) is augmented by input from external analysts. This characterization philosophy is a variation of the two or more independent and reliable method approaches and can be briefly denoted as the *expert analyst - different independent method* approach. This characterization strategy is viable as long as the selection process selects analytical chemists with the requisite expertise (specific type of measurements and materials) and proven track records of performance, using definitive, reference or validated methods of analysis. Technical discussions with all participants before and after the exercise, as practiced by BCR, is beneficial.

The general premise behind this concept of certification by interlaboratory measurement is based on at least two assumptions: (a) There exists a population of laboratories that is equally capable in determining the characteristics of the RM to provide results with acceptable accuracy; (b) the differences between individual results, both within- and between-laboratories, are statistical in nature regardless of the causes (i.e. variation in measurement procedures, personnel, equipment, etc.). Each laboratory mean is considered to be an unbiased estimate of the characteristic of the material. The interlaboratory comparison mode has been widely used by national and other laboratories for the certification of RMs. The following guidelines, enunciated by NIST are instructive. According to NIST [40], this is a mode that must be used with the greatest restraint and under very carefully prescribed and controlled conditions. At that agency, this approach is used only when the following circumstances apply: (1) The RM under study is in a technical area that is well established and one where many good, reliable methods exist, (2) each of the laboratories in the network are of very high quality and are known to produce very reliable results, (3) each laboratory agrees to the conditions set forth by NIST, (4) NIST controls the experimental design and evaluation of data, (5) a previously issued RM, having similar properties to the RM candidate is used by each laboratory as an internal quality control check. When these conditions are met and maintained, this mode may be used with assurance to produce RMs of high accuracy and integrity. It must be assured that a wide range of reliable independent methods is covered with an absolute minimum of two but preferably three or more; it is furthermore beneficial and advocated that each method is used by at least two but preferably three independent analysts/laboratories. With entirely different basic principles used for the analysis, possible interferences or other systematic errors can reasonably be expected to be

different. Each analyst should use well established method(s), which demonstrate adequate performance in terms of trueness (no significant bias) and reproducibility (standard deviation consistent and explicable on the basis of random experimental errors). If definitive and reference methods exist for the particular constituent/matrix combination, they should be targeted for inclusion in the repertoire of methodologies sought for certification.

Although sophisticated methods may constitute the core methods for certification it is useful to include good, well executed routine methods. In order to further minimize systematic error, a conscious purposeful attempt should be made to get methods and procedures with wide-ranging and different sample preparation steps, including no decomposition as in instrumental neutron activation analysis and particle induced X ray emission spectrometry.

An overriding criterion regarding selection of methods and laboratories is the reputation, expertise and dedication of the scientist, analytical chemist, analyst or technologist conducting and responsible for the analyses. **The choice is made on the basis of the analyst, not the laboratory**, although the laboratory should have an acceptable reputation and environment conducive to good work. The laboratories, collaborating in the analytical campaign should be carefully selected, without political, regional, administrative or other constraints on the basis of reputation, experience in the specific field of analysis, familiarity with the matrix to be investigated and the availability of the required analytical technique. Participation is by invitation.

It is not necessary that participating laboratories be formally recognized, accredited or certified. Measurement of the property of interest should be completed by, or under the supervision of a technically competent manager qualified either in terms of suitable academic qualifications or relevant work experience. The participating laboratory should consider the analysis as a very special one, to be performed with special attention and all possible care, and not have it performed as part of its regular routine.

(4) Volunteer analysts, various methods — multiple organizations and laboratories: *A "round robin" exercise with the participation of volunteer analysts in many laboratories, volunteering freely to participate, or chosen either fairly completely at random or selectively according to some selection criteria, based on political, regional, administrative or other constraints, which may or may not be based on expertise or competence, due to an obligation to involve laboratories from a defined population of countries, regions or other groupings. Analytical methods used are varied, generally self-selected, and include reference, validated, non-validated, routine, as well as definitive methods, and this interlaboratory characterization exercise is carried out without imposition of prescribed conditions and controls. More in-depth statistical treatment is needed to deal with the wide range, and likely, discordant nature, of analytical results received. Calculations and reassessment of data reported in the literature, to arrive at estimates of concentrations and uncertainties, can be considered a component of this approach.*

It is not always possible to have certification based on analyses done in-house or by selected laboratories according to strategies 1, 2 or 3 defined above. One must then resort to the last, and in the author's opinion, least preferred, mode of method-independent characterization, based on analysts, freely volunteering to participate, or selected without necessarily solid regard for expertise or competence, utilizing various methods. This approach, based on volunteer analysts and various methods in multiple organizations and laboratories, represents a round-robin type approach. Since no controls have been imposed upon the investigators, the limitations of such an approach and the data there from must be recognized. It must be appreciated that no mathematical processing can prove the validity of a concentration value derived from a mass of widely scattered data, the typical outcome of an exercise involving contributing analysts of varied backgrounds. Excellent insight into the problems associated with this approach has been provided by several experienced practitioners [34–37, 42–45]. Ingamells (1978) [46], in fact, suggested that the "round robin collaborative analysis" approach was a waste of time and effort and proposed instead a strategy involving only two mutually independent analysts, working in different laboratories and presumably using mutually independent methods. Parr

[44] felt that one of the criticisms that can be levelled against this type of certification procedure is that the participating laboratories are self-selected and some may have very little experience. He postulated that there could be considerable improvement in the analytical results if data were accepted only from experienced laboratories (e.g. approach 3 above). While some improvements in the confidence intervals associated with the certified values can sometimes be achieved in this way, the problem (of scattered data) certainly cannot be made to disappear; selection of laboratories can only be made on rather subjective grounds by the person responsible for certification. Abbey [34–37] has carried out many interesting statistical manipulations and calculations of literature-reported data for standard rocks, recalculating published recommended elemental concentration values. He clearly and forcefully observes that "Given a highly incoherent set of results for the determination of each constituent of a proposed reference sample, the originator is faced with the difficult problem of estimating the 'true' concentration. No known test can prove the validity of a concentration value derived from a mass of incoherent data" [37]. These observations apply equally well to all analytical endeavours.

(5) Method-specific; characterization by a specific, validated method by selected expert or experienced analysts — multiple organizations and laboratories: *One specified analytical method applied in replicate, by selected expert or experienced analysts, of high reputational quality and recognized competence working independently in an ad hoc network of laboratories, participating under carefully prescribed and controlled conditions, giving a method-specific assessed property value. The analysts are selected, on the basis of their established reputation, expertise and experience with the method and the specific field of analysis and familiarity with the matrix investigated, and participate,* ***on invitation****, in the analytical collaborative interlaboratory characterization campaign.*

In a few instances, RMs are certified for the value of a constituent or property that is method-dependent because existing technology or technical or scientific applications require this. In analytical chemistry, examples of this are the Kjeldahl technique for nitrogen, EPA mandated and other extraction procedures for leacheable toxic constituents, extraction procedures for soil nutrients and toxicants in agronomy and soil science, and various enzyme-based or enzyme measuring methods in clinical science. In such cases, demonstration of statistical control of the measurement process and agreement of results by independent analysts are the requirements for certification. As usually viewed by the RM scientist, the philosophy of certification rests on the concept of application of independent methodology to generate concordant results leading to one reliable value for the property. Such values are thus method-independent. Extractable and other concentrations generated by specific procedures are method-dependent, an idea which has to be rationalized with the fundamental method-independent concept in RM certification.

3.4. SUMMARY

The European Commission [25] prepared a detailed and most useful guide for the production and certification of RMs. The following is a summary of items relating to certification:

The participants: The range of participants should, whenever possible, be chosen in such a manner that widely different methods (based on different physical or chemical principles) can be used. The number of participants (recommended 15) should be sufficient to allow meaningful statistical processing of the results. When the laboratories feel the need for a CRM, either because the available calibrants are not comparable and a primary calibrant appears necessary for traceability, or because a reliable certified control material is needed but not available, then it is recommended that these laboratories do not plan a certification project entirely on their own, but that they involve laboratories having a background in traceability.

Quality and traceability: It is not required that participating laboratories are formally recognized, accredited or certified, provided that quality and traceability requirements are met.

The methods: It is advised that, prior to the certification measurements, the participants discuss their methods so that all participants have confidence in each other's methods and there is a good level of agreement between laboratories. As it is preferred to certify on the basis of the agreement between different methods applied in different laboratories, a proposal should include, where relevant and possible, a group of laboratories offering a range of widely different measurement methods. Each laboratory should use well established method(s), with which it can demonstrate adequate performance in terms of trueness and in terms of reproducibility.

Evaluation of results: Evaluation of the results consists of: (1) technical scrutiny of the consistency and of the quality of the data; acceptance on technical (not statistical) grounds of data to be used to calculate the certified value and its uncertainty; (2) calculation (using the appropriate statistical techniques) of the certified value and its uncertainty. The approach includes technical discussion of the results among all co-operators, rejection of outliers, statistical evaluation, and calculation of the certified value and uncertainties.

3.5. CONCLUSIONS

The ultimate goal of the chemical certification strategy is to arrive at reliable, unassailable assigned numerical values for the concentrations of chemical constituents in RMs. All certification approaches comprise several components and the careful and critical implementation accorded the exercise by the RM developer and coordinator of the characterization campaign. The foregoing represents the author's interpretation and composite presentation of the different approaches followed by several of the major RM producers based on information available in the literature. It is this author's view and contention that, in certification painstaking care is needed in the selection of the co-operating analysts, laboratories and analytical methods and the main effort is in the generation of an excellent, tight dataset which is then subjected to minor mathematical manipulations to arrive at final property measures. The analytical determinations should be made with extra care to yield small uncertainties in the results. Throughout the overall task of RM development, there is an overwhelming requirement for a critical approach by critical analytical and measurement scientists and the involvement of national or international government agencies.

Judging from the emphasis placed on analyst expertise and experience, it is clear that the author considers the role of the analyst to be of paramount importance in the certification exercise; good analysis and a good analyst go hand in hand. It is an opinion that it matters little as to **how** the analysis was performed as to **who** did it (Abbey, personal communication); all that is required for proper analysis is the selection of a good analyst. Analyst training, experience, familiarity with the problem on hand, skill, attitude, motivation and judgement are necessary prerequisites with which satisfactory solution of analytical problems is possible.

Reliability and confidence in the stated characteristics of the developed QCMs is a basic critical criterion for their use for quality control. Applications of methods of chemical analysis are fraught with many sources of error and countless opportunities exist for the introduction of error into the final results. Measurement systems must therefore be operated under a complete, regularly-applied quality assurance programme if results are to be meaningful. Characterization philosophy rests on the concept of independent methodology, the application of theoretically and experimentally different measurement techniques and procedures to generate method-independent concordant results. The developer should be aware of the need for, and possible shortage of, highly competent analysts required for characterization work, the difficulties of good work at trace and ultra trace levels and methodological deficiencies for specific measurands. Throughout the overall task of development there is a requirement for a critical approach by critical analytical and measurement scientists in order to produce top QCMs.

3.6. EXAMPLE: RECOMMENDED SAMPLE PROCEDURE OF ASSIGNMENT OF REFERENCE CONCENTRATION VALUES TO QCM USING MULTIPLE METHODS OF ANALYSIS

The following sample (example) recommended procedure for the calculation of reference concentration values to the developed QCM is based on the author's experience with the development of food-related RMs [23, 47–51].

3.6.1. Analytical method selection and testing

The lead, initiating laboratory selects an analytical method (designated as the principal method) appropriate for the intended analysis. This method is typically a core method, routinely and frequently in use in the laboratory by trained, knowledgeable analysts (technicians, chemists, research scientists). The method is taken to be validated to the satisfaction of the laboratory personnel. This indicates that it complies with one, several or more of the following criteria providing suitably precise and accurate data satisfying laboratory requirements for the analysis in question. At the top of the list of methods to be selected are those defined as **definitive methods**; as discussed previously, however, such methods will most likely not be available or useable by the analyst. The method chosen will then most typically be denoted a **reference method**, and will be either (in approximate decreasing order of confidence) an (1) Official Method subjected to collaborative study (e.g. those from AOAC INTERNATIONAL, and other agencies using interlaboratory collaborative studies for method development and verification), (2) an Official Method or recommended method not subjected to collaborative study but essentially accepted on the basis of reported performance and committee deliberations (e.g. ISO, Recommended and Standardized Methods of Analysis, etc.), (3) another reference method developed by a laboratory and verified by peer verification, interlaboratory intercomparisons, etc. A method from such sources is considered to be a validated, reference method. Methods defined as **field methods**, denoting a method used in applications requiring large numbers of individual measurements made on a routine basis, should not be contemplated for this QCM development activity.

This method, denoted as **principal method (A)**, is verified by the analyst for suitability to the analyte/matrix in question, for performance with respect to accuracy and precision. Precision (within laboratory) is determined by conducting 6–10 replicate analyses on several different masses covering the desired range; precision estimates may also be obtained from interlaboratory collaborative studies, information from proficiency and peer verification studies conducted to establish accuracy (nos. 4, 5, 6 under accuracy establishment). Accuracy is estimated (determined) using (1) different masses of material (perhaps a 10-fold range); (2) recovery of added (spiked, fortified) analyte utilizing a limited range of added analyte quantities and one or more material masses (see chapter 3.7); (3) comparison with a second independently different method applied by the analyst, a colleague in the same institution or by an outside collaborator; (4) interlaboratory collaborative studies involving the participation of several outside analysts (if feasible and deemed essential); (5) information from proficiency tests; (6) information from peer verification studies.

A second independently different, validated, reference analytical method is selected by the lead, initiating analyst according to the above criteria and validated as described above. If feasible and desired, a third and additional independently different, validated, reference analytical methods are selected by the lead analyst for use by the lead analyst or colleagues in the same institution or by outside collaborators and similarly validated.

If truly independent methods are not conveniently available, alternative pseudo-independent methods may be devised by considering incorporating variations to the **principal method** such as different solvents, other columns and detectors in gas and liquid chromatographic methods; different wavelengths in atomic absorption and atomic emission spectrometry; different atomic masses in inductively coupled atomic mass spectrometry; internal QC in instrumental neutron activation analysis.

In a simplified application of two or more methods to the chemical characterization of the QCM, the material is analysed by the principal method performing 6–10 replicate analyses. A mean concentration and standard deviation are computed. The material is then (or concurrently) analysed by method B 6–10 times and a mean concentration and standard deviation are again computed. Method means and standard deviations are compared using a t-test and F-test, respectively. Should tests indicate no significant differences of means and variances between the two methods, (or if different but deemed to be acceptable for pooling) means can be averaged and variances can be pooled to calculate an overall mean and variance (uncertainty). If a third (and additional) method is used, similar comparisons of means and variances among all methods (paired calculations) are carried out, and if not significantly different (or if different but deemed to be acceptable for pooling) means can be averaged and variances can be pooled over all three (or more) methods to calculate an overall mean and variance (uncertainty).

If mean and variance results produced by the principal method and method B (and others) are significantly different (or deemed to be so), reasons for the differences are investigated. Sources of differences or errors to be considered can be: blank, material inhomogeneity, uncorrected biases due to background or interferences, calibration (instrument, standards, curve), performance of analyst. As a last resort the method(s) are completely re-evaluated as described above.

Compliance with several prerequisites must be established prior to conducting the QCM characterization exercise. 1) An appropriate analytical method must be applied, by appropriately qualified and trained personnel in a suitable physical (equipment, materials, reagents and laboratory conditions necessary for the proper execution of the method) and administrative (understanding of and support for appropriate data quality by managers) environment. The role of the analyst is of direct paramount importance; analyst training, experience, familiarity with the problem on hand, skill, attitude, motivation and judgement are necessary prerequisites with which satisfactory solution of analytical problems is possible. 2) Suitable quality control/quality assurance procedures should be routinely in use and the need for appropriately reliable analytical information must be recognized. The analytical system must be in a state of statistical control. The method under test should usually give a precision with the RM and other homogeneous materials equal to or better than the uncertainty reported for the RM in the certificate.

3.6.2. Specific example (potassium in Wheat Gluten RM NIST RM 8418)

A specific example of the recommended multi-method procedure discussed above is presented here using results from a real life case from the author's interlaboratory characterization campaign, the multi-method interlaboratory determination of potassium in Wheat Gluten which ended up as NIST RM 8418 [48, 50, 51]. Table V presents a complete listing of all the concentration values received arranged by (1) laboratory number, (2) method code B, (3) material unit number, (4) mass and (5) concentration giving tabulation in ascending order of laboratory number, method, unit number, mass and concentration. More figures for mass and concentration, than warranted by significance, were usually retained in such tabulations but the appropriate number of significant figures, however, was considered in calculation and presentation of final results for the RM.

Observation number (Obs. No.) is a line number for the data as recorded here for each element/RM combination, and has no other significance. Laboratory number (Lab. No.) is the unique code number assigned to each participating laboratory. Unit no. refers to the sequential number assigned to the unit (bottle) in the bottle filling operation and randomly removed and set aside for physical and chemical characterization and the interlaboratory characterization campaigns. Original units were typically submitted for analysis and each laboratory thus generally received a set of uniquely numbered units. Analytical methods used were:

ADFAAS (A01) (denoted on Fig. 5 as AAS01): acid digestion flame atomic absorption spectrometry.

ADFAES (B01) (not accepted in final dataset; not in Fig. 5): acid digestion flame atomic emission spectrometry.

ADICPAES (B02) (denoted on Fig. 5 as AES02): acid digestion inductively coupled plasma atomic emission spectrometry.

CVADICPAES (B03) (denoted on Fig. 5 as AES03): closed vessel acid digestion inductively coupled plasma atomic emission spectrometry.

DAICPAES (B04) (not accepted in final dataset; not in Fig. 5): dry ashing inductively coupled plasma atomic emission spectrometry.

INAA (D01) (denoted on Fig. 5 as NAA01): instrumental neutron activation analysis.

PIXE (E01) (denoted on Fig. 5 as XRE01): particle-induced X ray emission spectrometry.

XRF (E02) (not accepted in final dataset; not in Fig. 5): X ray fluorescence spectrometry.

The aim of this work was to arrive at **total** analyte concentrations, i.e. the sum of all species of the elemental analyte in all phases of the RM. Analysts were requested to take appropriate measures to measure such contents and thus all concentration results submitted were deemed to represent measures of total elemental content. Concentration data were reported either as means (of several instrumental readings) or individual values, and the reported mean or calculated mean value is recorded here; each concentration datum in this table represents one analysis. Concentration results were generally received on a dry matter basis with moisture content determined by the analyst using the prescribed procedure for determining moisture contents on eight separate 2 g portions by drying at 85 C for 4.0 h in an air convection oven. When submitted results were not reported on a dry matter basis or were reported corrected for moisture determined by another procedure, results were converted or corrected, by the initiating analyst, to the appropriate dry matter basis using mean moisture values determined in the initiating analyst's laboratory, by the prescribed procedure, on random units of material retained by the initiating for analysis in his laboratory.

Concentrations identified by an asterisk (*) were deemed outliers and removed in the various stages of data treatment to yield a final dataset for calculation of best estimate and informational concentrations. The remaining data correspond to those accepted for final calculations and are plotted in Figs 4 (d) and 5 (c).

Prior to final calculations of concentration values, the analytical concentration results were carefully inspected using technical, statistical (variation and bias) and judgement criteria to remove aberrant, outlying or non-representative data. The procedure for dealing with outliers followed three steps: (1) deletion of obviously erroneous, aberrant or outlying data, (2) inspection of concentration versus laboratory number plots and deletion of all data for an element/material with excessive within-laboratory variation or systematic errors (bias) relative to data from the other laboratories; confirmation of rejection by noting performance with certified RMs (3) repetition of (2) and rejection of additional individual outliers or entire sets from a laboratory when their retention had a serious impact on final uncertainty (spread of accepted results).

Dataset reduction (outlier rejection) criteria included the following considerations: (1) poor within-laboratory precision (among subsample precision revealed by the range of data and the magnitude of within-laboratory variance) compared that of other laboratories, (2) poor within-subsample precision (within-laboratory instrumental precision provided by the analyst or computed from replicate data) compared with similar parameters of other laboratories, (3) laboratory systematic error judged by deviation of laboratory mean from overall mean, (4) accuracy, based on performance with certified RMs, (5) within-laboratory precision with certified RMs, (6) assessment of the

technical merit of the analytical procedure, (7) number of subsamples analysed compared to that in other laboratories. Generally, an all or none principle was applied — either all the data from a laboratory for an element/material were retained or all were rejected. Occasionally, however, only limited selected data were eliminated when these clearly deviated from the rest of the data.

TABLE V. INDIVIDUAL CONCENTRATION VALUES (MG/KG, DRY MATTER BASIS) RECEIVED IN THE INTERLABORATORY CHARACTERIZATION CAMPAIGN ARRANGED BY LABORATORY NUMBER AND MATERIAL UNIT NUMBER WITHIN LABORATORIES FOR POTASSIUM IN WHEAT GLUTEN REFERENCE MATERIAL NIST RM 8418[A]

Obs. No.[b]	Lab No.[b]	Method Codes[c] Code A	Code B	Generic	Mass[d] g	Unit No.[e]	Concn[f] mg/kg
		[Total number of results received: 67; accepted: 42 (Reference file: FIL18417)]					
01	2	ADICPAES	B02	AES	0.20	32	475.0
02	2	ADICPAES	B02	AES	0.20	32	501.0
03	2	ADICPAES	B02	AES	0.20	1362	460.0
04	2	ADICPAES	B02	AES	0.20	1362	464.0
05	14	DAICPAES	B04	AES		516	3110.0*
06	14	DAICPAES	B04	AES		516	3220.0*
07	14	DAICPAES	B04	AES		1257	3430.0*
08	14	DAICPAES	B04	AES		1257	4080.0*
09	15	INAA	D01	NAA	0.50	841	520.0
10	15	INAA	D01	NAA	0.50	887	500.0
11	15	INAA	D01	NAA	0.50	1041	450.0
12	15	INAA	D01	NAA	0.50	1163	520.0
13	15	INAA	D01	NAA	0.50	1163	840.0*
14	25	ADICPAES	B02	AES	1.00	725	452.0
15	25	ADICPAES	B02	AES	1.00	725	453.0
16	25	ADICPAES	B02	AES	1.00	725	482.0
17	25	ADICPAES	B02	AES	1.00	859	476.0
18	25	ADICPAES	B02	AES	1.00	859	477.0
19	25	ADICPAES	B02	AES	1.00	1333	452.0
20	25	ADICPAES	B02	AES	1.00	1333	455.0
21	25	ADICPAES	B02	AES	1.00	1420	484.0
22	25	ADICPAES	B02	AES	1.00	1420	492.0
23	33	ADFAAS	A01	AAS	0.50	194	436.0*
24	33	ADFAAS	A01	AAS	0.50	194	449.0*
25	33	ADFAAS	A01	AAS	2.00	194	452.1
26	33	ADFAAS	A01	AAS	0.50	554	430.0*
27	33	ADFAAS	A01	AAS	0.50	554	454.0*
28	33	ADFAAS	A01	AAS	2.00	554	453.5
29	33	ADFAAS	A01	AAS	0.50	613	426.0*
30	33	ADFAAS	A01	AAS	0.50	613	452.0*
31	33	ADFAAS	A01	AAS	2.00	613	455.6
32	33	ADFAAS	A01	AAS	0.50	1281	459.0
33	33	ADFAAS	A01	AAS	0.50	1281	464.0
34	33	ADFAAS	A01	AAS	2.00	1281	456.1
35	47	PIXE	E01	XRE	0.30	43	424.0*
36	47	PIXE	E01	XRE	0.30	43	449.0
37	47	PIXE	E01	XRE	0.30	1362	442.0

38	47	PIXE	E01	XRE	0.30	1362	446.0
39	51	ADFAES	B01	AES	0.90	68	521.0*
40	51	ADFAES	B01	AES	0.90	68	531.0*
41	51	ADFAES	B01	AES	0.90	68	553.0*
42	51	ADFAES	B01	AES	0.90	68	678.0*
43	51	ADFAES	B01	AES	0.90	117	486.0*
44	51	ADFAES	B01	AES	0.90	117	513.0*
45	51	ADFAES	B01	AES	0.90	1242	607.0*
46	51	ADFAES	B01	AES	0.90	1242	634.0*
47	51	ADFAES	B01	AES	0.90	1242	683.0*
48	51	ADFAES	B01	AES	0.90	1242	684.0*
49	51	ADFAES	B01	AES	0.90	1319	523.0*
50	51	ADFAES	B01	AES	0.90	1319	525.0*
51	56	ADFAAS	A01	AAS	0.20	253	439.0
52	56	ADFAAS	A01	AAS	0.20	253	446.0
53	56	ADFAAS	A01	AAS	0.20	375	437.0
54	56	ADFAAS	A01	AAS	0.20	375	446.0
55	56	ADFAAS	A01	AAS	0.20	998	447.0
56	56	ADFAAS	A01	AAS	0.20	998	453.0
57	56	ADFAAS	A01	AAS	0.20	1419	452.0
58	56	ADFAAS	A01	AAS	0.20	1419	456.0
59	59	CVADICPAES	B03	AES	0.10	54	488.0
60	59	CVADICPAES	B03	AES	0.10	54	490.0
61	59	CVADICPAES	B03	AES	0.10	54	503.0
62	59	CVADICPAES	B03	AES	0.10	54	523.0
63	59	CVADICPAES	B03	AES	0.10	680	516.0
64	59	CVADICPAES	B03	AES	0.10	680	528.0
65	59	CVADICPAES	B03	AES	0.10	680	530.0
66	59	CVADICPAES	B03	AES	0.10	680	538.0
67	82	XRF	E02	XRE	0.01	362	397.0*

[a] This is a complete listing of all the combined concentration values received in the interlaboratory characterization campaigns arranged by (1) laboratory number, (2) method code B (3) material unit number within methods/laboratories, (4) mass within unit number and (5) finally by concentration (ascending order of laboratory number, method, unit number, mass, concentration). More figures for mass and concentration than warranted by significance have usually been retained in this presentation; the appropriate number of significant figures, however, was considered in calculation and presentation of final results for the RM.

[b] Observation number (Obs. No.) is simply a line number for the data as recorded here for each element/RM combination, and has no other significance. Laboratory number (Lab. No.) is the unique code number assigned to each participating laboratory.

[c] Analytical methods used are: ADFAAS (A01): acid digestion flame atomic absorption spectrometry; ADFAES (B01): acid digestion flame atomic emission spectrometry; ADICPAES (B02): acid digestion inductively coupled plasma atomic emission spectrometry; CVADICPAES (B03): closed vessel acid digestion inductively coupled plasma atomic emission spectrometry; DAICPAES (B04): dry ashing inductively coupled plasma atomic emission spectrometry; INAA (D01): instrumental neutron activation analysis; PIXE (E01): particle-induced X ray emission spectrometry; XRF (E02): X ray fluorescence spectrometry.

[d] Mass refers to nominal subsample masses, reported by the analysts, taken for chemical analysis.

[e] Unit no. refers to the sequential number assigned to the unit (bottle) in the bottle filling operation and randomly removed and set aside for physical and chemical characterization and the interlaboratory characterization campaigns. Original units were typically submitted for analysis and each laboratory thus generally received a set of uniquely numbered units.

[f] The aim of this work was to arrive at **total** analyte concentrations, i.e. the sum of all species of the elemental analyte in all phases of the RM. Analysts were requested to take appropriate measures to measure such contents and thus all concentration results submitted were deemed to represent measures of total elemental content. Concentration data were reported either as means (of several instrumental readings) or individual values, and the reported mean or calculated mean value is recorded here; each concentration datum in this table represents one analysis.

Results for potassium in Wheat Gluten are plotted versus laboratory number in Fig. 4 showing progression upon successive dataset reductions due to removal of outliers. The top Fig. 4(a) (file FIL18417) depicts a plot of the 67 original results received; not all are visible on the scale of the plot due to hidden values. Deletion of four deemed erroneous values leads to the second Fig. 4(b) (file FIC18417). Removal of additional outliers upon inspection of Fig. 4(b) and consideration of data quality according to guidelines enumerated above leads to 3rd Fig. 4(c) (file FIN18417) and to a final dataset of 42 values (25 outliers rejected) depicted in the bottom Fig. 4(d) (file FNN18417). This final dataset contains results used to arrive at a recommended concentration value and associated uncertainty depicted by the point at the extreme right hand side in Fig. 4(d). The error bar depicts an assigned uncertainty representing a 95% confidence interval for a single future determination.

Fig. 5 presents the identical data plotted in a different way, pooled by method, a depiction more relevant to the exercise at hand: comparing method performances and combining data over methods to arrive at a recommended value. In this figure, mean concentrations (final dataset after elimination of outliers, file FNN18417) are grouped by method used for their generation and plotted against laboratory number. Variants of the same generic method are separated by dashed vertical lines, whereas solid vertical lines indicate demarcation between fundamentally different methods. Thus, Fig. 5 shows the following breakdown of accepted mean results: 2 by method AAS01, 2 by AES02, 1 by AES03, 1 by NAA01 and 1 by XRE01, for a generic subtotal of 2 means by AAS, 3 by AES, 1 by NAA and 1 by XRE for an overall total of 7 means, each generated from an average of 6 separate analyses. It is evident that accepted results for potassium in Wheat Gluten RM were generated by a total of five variants of four fundamentally different methods, AAS: atomic absorption spectrometry, AES: atomic emission spectrometry, NAA: neutron activation analysis and XRE: particle-induced X ray emission spectrometry. Vertical lines about mean points depict ranges of results permitting a rough visual intercomparison of method/laboratory performance. The horizontal dashed line, as well as the point on the right labelled REF, is the mean over all methods/laboratories calculated as described previously, and the error bar on the reference value depict, as before, an assigned uncertainty representing a 95% confidence interval for a single future determination.

Based on visual intercomparison of means, considering ranges, it appears that data from all methods/laboratories generally overlap the overall mean value (reference value) as well as each other; that means they do not differ significantly from each other (statistical tests do however indicate differences). However, visual inspection of estimations/indications of standard deviations, presented as ranges of accepted results in these figures, does suggest some differences. Simple statistical t-tests and F-tests can be applied, based on means and variances, as presented in Table VI for the K/Wheat Gluten example.

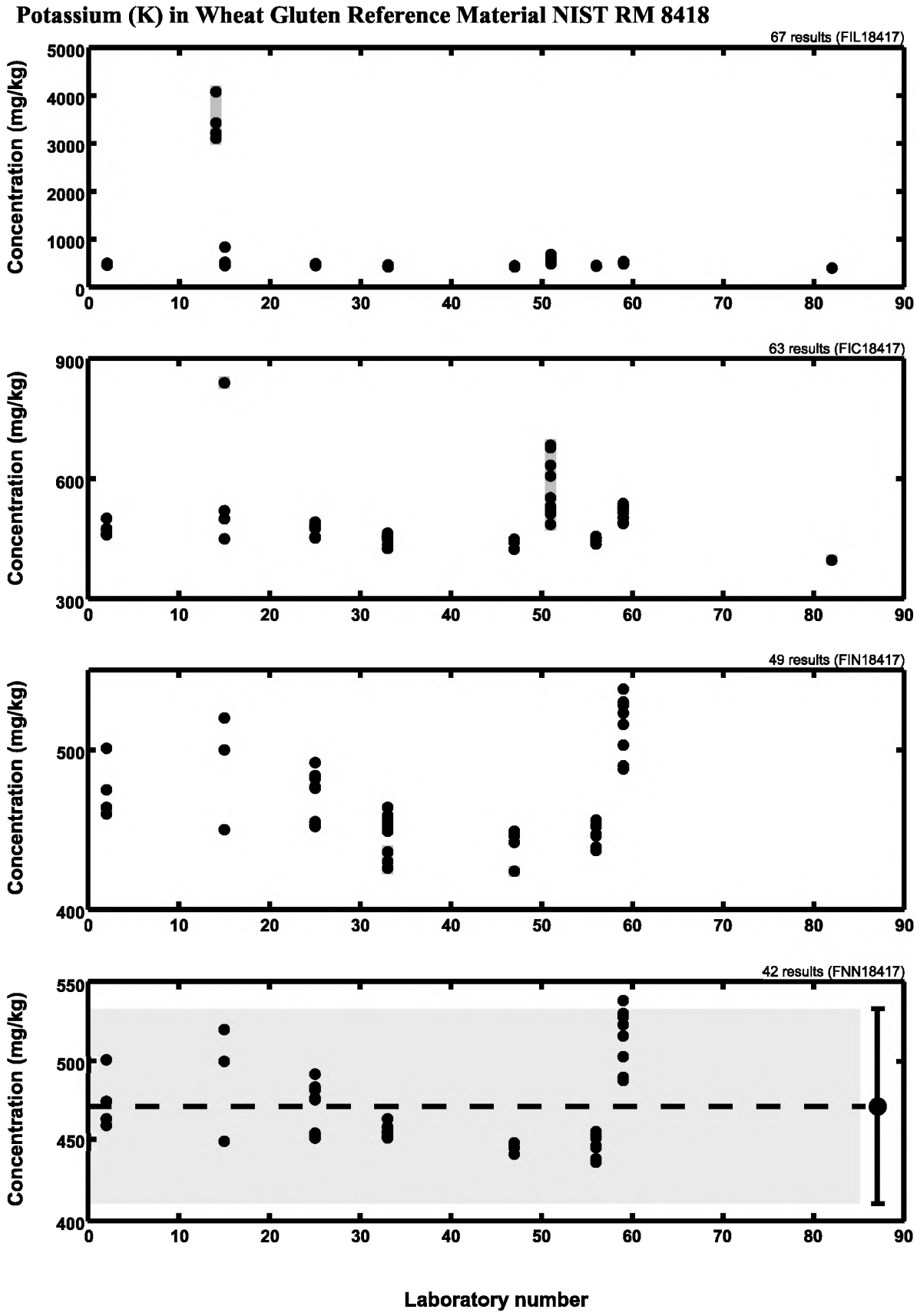

FIG. 4. Data reduction of results from potassium determination in Wheat Gluten versus laboratory number of participants in the RM characterization (original data, successive removal of outliers).

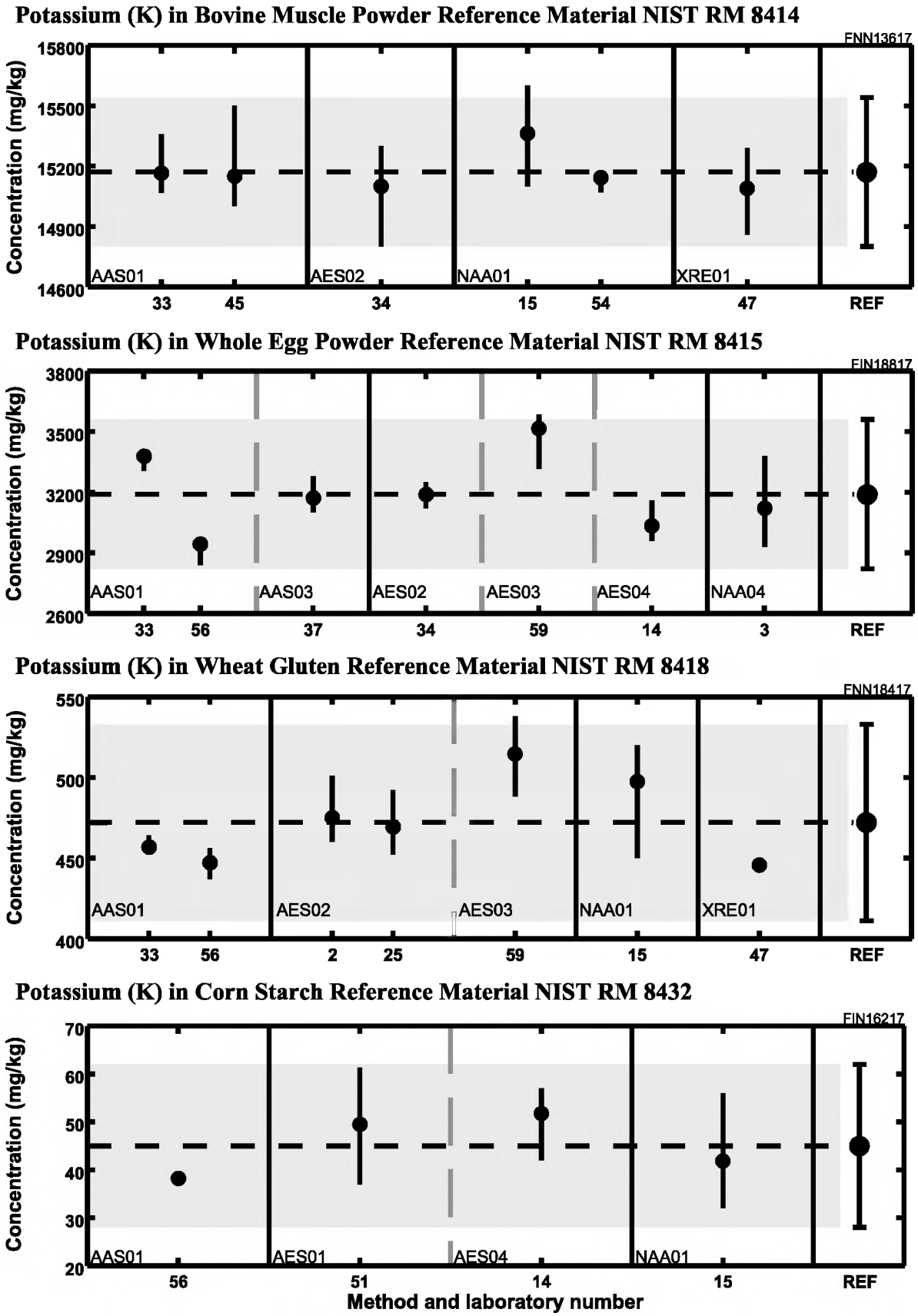

FIG. 5. Data reduction of results from potassium determination in various RMs versus laboratory number of participants in the RM characterization (grouped by method).

TABLE VI. STATISTICAL PARAMETERS FOR THE DIFFERENT METHODS (ACCEPTED VALUES) MEASURING POTASSIUM IN WHEAT GLUTEN REFERENCE MATERIAL ARRANGED BY METHOD

Lab. No.	Method Generic	Code B	N[a]	Mean mg/kg	Variance[b]	Std. Dev.[b]
33	AAS	A01	6	456.78	18.309667	4.27898
56	AAS	A01	8	447.00	44.000000	6.63325
2	AES	B02	4	475.00	340.66667	18.45716
25	AES	B02	9	469.22	258.19444	16.06843
59	AES	B03	8	514.50	354.85714	18.83765
15	NAA	D01	4	497.50	1091.66667	33.04038
47	XRE	E01	3	445.67	12.33333	3.51189
REF:[c]			42	472.2	946.0060	30.76

[a] Number of analyses carried out by the laboratory using the method.
[b] Variances and standard deviations are simple within-laboratory calculations.
[c] The reference value REF is an equally-weighted mean of results from the individual methods/laboratories; the associated standard deviation is calculated from the three variance components for this dataset from: $SD = (\sigma_w^2 + \sigma_u^2 + \sigma_L^2)^{1/2}$, where the three variance components are - within-unit variance (σ_w^2):106.03033; among-unit variance (σ_u^2): 205.915144; among-laboratory/method variance (σ_L^2): 634.060559.

3.6.3. Simple paired t-test calculations

Paired t-test calculations may be performed, pair-wise, on all possible combinations of the seven individual means to ascertain statistical agreement according to the following equation for paired t-tests:

$$t = X_1 - X_2 / s\,(1/n_1 + 1/n_2)^{1/2} \quad (2)$$

$$s = [(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2]/(n_1 + n_2 - 2) \quad (3)$$

where

X_1 and s_1 are the mean concentration and standard deviation, respectively, found by method/laboratory specified in the first member of the pair,

X_2 and s_2 are the mean concentration and standard deviation, respectively, found by method/laboratory specified in the second member of the pair,

and where the assumption is that $s_1 = s_2$ (i.e. variances are assumed to be homogeneous).

By way of example, calculated paired t-test results are given here for selected data from the four methods: AAS (AAS01 lab 33), AES (AES03 lab 59), NAA (NAA01 lab 15) and XRE (XRE01 lab 47), giving six possible pair-wise comparisons. Comparison of data within the following pairs, 33/59, 33/15, 33/47, 59/15, 59/47 and 15/47 gave respective calculated t's of 7.295 (S), 3.075 (S), 3.856 (S), 1.157 (NS), 6.090 (S) and 2.642 (S). Comparison with predicted t's at the respective degrees of freedom at the usual test level of $p = 0.05$ suggests non-significant (NS) and significant (S) differences as indicated.

Similarly, F-tests using variances (V) for the identical six possible pairs of data gave the following F ratios: V59/V33 = 19.38 (S); V15/V33 = 59.62 (S); V33/V47 = 1.48 (NS); V15/V59 =

3.08 (NS); V59/V47 = 28.77 (S); and V15/V47 = 88.51 (S). Comparison with predicted F ratios at the respective degrees of freedom for p = 0.05 suggests non-significant (NS) and significant (S) differences as indicated. Thus, for these six methods/laboratories, statistical calculations indicate some differences among means and variances.

Although actual statistical calculations do indicate some differences among means and inhomogeneity of variances, in the RM project from which the K/Wheat Gluten example is taken, such data were combined and utilized for calculation of reference concentration values and associated uncertainties. It is up to the initiating analyst to decide on pooling criteria and whether strict adherence to non-significant t-tests and F-tests need be maintained. The variances discussed so far are simply those generated within laboratories, i.e. within-laboratory variances, including within-unit and among-unit variances; it is surmised that consideration and addition of real errors (e.g. based on between-laboratory deviations) to those indicated in Fig. 2(c) would indicate more universal agreement among means and uncertainties. Plots in Fig. 2(a), 2(b) and 2(d) may be consulted for additional examples of visual comparisons and indications of the kind of data accepted for generation of reference concentration values in that RM characterization project.

In that project, such individual t-test and F-test calculations were not performed but the data tests were subjected to ANOVA calculations to extract the three variance components. Some statistical ANOVA calculations for the accepted set of analytical results are presented in Table VI. The following results were generated by detailed ANOVA calculations: overall mean (reference value): 472.2 mg/kg; within-unit variance (σ_w^2): 106.03033; among-unit variance (σ_u^2): 205.915144; among-laboratory/method variance (σ_L^2): 634.060559. The reference value was computed as the arithmetic average of equally weighed individual laboratory means. The associated SD was calculated from the three variance components according to the equation:

$$SD = (\sigma_w^2 + \sigma_u^2 + \sigma_L^2)^{1/2} \qquad (4)$$

$$SD = (106.03033 + 205.915144 + 634.060559)^{1/2}$$

$$SD = 30.76$$

where each σ indicates the estimates of the associated variance component obtained from a type I (hierarchal) variance component analysis.

Thus the recommended value and associated uncertainty (rounded to appropriate numbers of significant figures) is 472 ± 61 mg/kg, (mean ± 95% confidence limits) where the latter was calculated from the SD using the appropriate t value (usually 2) based on the degrees of freedom. It is worth pointing out that the confidence limits in this work, as in that of other RM producers, is based on **standard deviation** and **not standard error**, as by IRMM (BCR). Using standard deviation leads to more realistic estimates of uncertainties and is related to standard deviations estimated for a **single** future determination on the RM. Use of standard errors provides an unrealistically tight, narrow uncertainty estimate. Table IV is an example of a Table from a Report of Investigation [50] listing reference concentrations of constituent elements in RM Wheat Gluten WG 184 (NIST RM 8418). The following statement used by the author in reports of analysis for co-operatively produced AAFC/NIST RMs [52] may be taken as a guide for the reporting target values for in-house developed QCMs:

Reference values, weight percentage or mg/kg (ppm), presented in Reports of Investigation provided with these RMs are based on the dry material, and are equally weighed means of results from generally at least two, but typically several, different analytical methods applied by analysts in different laboratories. Uncertainties are estimates expressed either as 95% confidence intervals or occasionally as intervals based on ranges of accepted results for a single future determination based on a sample weight of at least 0.5 g. These uncertainties, based on among-method and laboratory/among-unit and within-unit estimates of variances, include measures of analytical method and laboratory imprecisions as well as biases and material inhomogeneity.

TABLE VII. REFERENCE CONCENTRATIONS OF CONSTITUENT ELEMENTS IN WHEAT GLUTEN REFERENCE MATERIAL WG 184 (NIST RM 8418)

Major Constituents (weight %)

Element	*Content and uncertainty (a)*	*Methods(b)*
Nitrogen	14.68 ± 0.26	I01 I02 J01 J02
Sulphur	0.845 ± 0.085	B02 B03 F04 J02 M02
Chlorine	0.362 ± 0.022	D01 F02 K01 K02
Phosphorus	0.219 ± 0.015	B02 B03 F01 F02 M01
Sodium	0.142 ± 0.011	A01 B01 B02 D01
Minor and Trace Constituents (mg/kg)		
Magnesium	510 ± 47	A01 B02 B03 D01
Potassium	472 ± 61	A01 B02 B03 D01 E01
Calcium	369 ± 35	A01 B02 B03 D01 E02
Iron	54.3 ± 6.8	A01 B02 B03 D01 D03 E01 E02
Zinc	53.8 ± 3.7	A01 B02 B03 D03 E01
Manganese	14.3 ± 0.8	A01 B02 B04 D01 E01 E02
Aluminium	10.8 ± 3.0	A05 B02 B03 D01
Copper	5.94 ± 0.72	A01 A05 B02 C03 C06 E01 H01
Selenium	2.58 ± 0.19	B02 C01 C04 D01 D03 G01
Strontium	1.71 ± 0.26	B02 B03 C03 E01
Barium	1.53 ± 0.26	B02 B03 C03
Molybdenum	0.76 ± 0.09	B02 C03 C06 D01 D03 F01 H06
Nickel	0.13 ± 0.04	A16 H01
Lead	0.10 ± 0.05	A05 A16 C03 H01
Cadmium	0.064 ± 0.022	A04 A05 A16 C03 D03 H01
Iodine	0.060 ± 0.013	D03 D05 D06 F02 H03
Chromium	0.053 ± 0.013	A12 C05 D03
Cobalt	0.010 ± 0.006	A16 D01 H01
Mercury	0.0019 ± 0.0006	A10 D03

(a) Best estimate values, weight percent or mg/kg (ppm), are based on the dry material, dried according to instructions in this report and are equally-weighted means of results from generally at least two, but typically several, different analytical methods applied by analysts in different laboratories. Uncertainties are estimates expressed either as a 95% confidence interval or occasionally (Co, S, Se) as an interval based on the range of accepted results for a single future determination based on a sample weight of at least 0.5g. These uncertainties, based on among-method and laboratory, among-unit and within-unit estimates of variances, include measures of analytical method and laboratory imprecisions and biases and material inhomogeneity.
(b) Analytical method codes and descriptions are provided above.

3.7. THE TECHNIQUE OF RECOVERY FOR METHOD VERIFICATION

Recovery testing is one important component in the arsenal of the analyst for method verification and validation. This is especially valuable in the absence of certified RMs and other QC possibilities. The following discussion, based on the report of Dabeka and Ihnat [53] is presented here in some depth as an aid to the analyst.

Many parameters impinge on the philosophy and practice of the determination and use of recovery factors. This section focuses on (1) general considerations for the use of recovery materials, (2) estimation of recovery based on added analytes, (3) calculation of recovery factors and associated

uncertainties, (4) the impact of method and laboratory bias/systematic error on recovery and (5) the application of recovery factors. Although the ideas herein relate to recovery considerations in general, the emphasis is on determination of inorganic constituents and further only on total elemental concentrations.

3.7.1. General considerations for use of recovery materials

3.7.1.1. Sources of error in analytical methods

The general lack of agreement among analytical results from different analysts and laboratories arises from numerous factors influencing the validity and reliability of the final numerical results. These factors can be broadly categorized as presampling, sampling, sample manipulation and measurement. Other important considerations such as contamination control, data quality control and the analyst capability transcend the above boundaries. A tabular summary is presented in Table VIII [54]. It is no wonder that such an extensive collection of potential pitfalls seriously impacts on data quality and typically imparts to it substantial questions of validity. Tests with recovery materials can monitor and control, to a good extent, the performance of the collection of laboratory procedures subsequent to the point of introduction of the material. Errors arising from activities occurring prior to this point of introduction, such as sampling, preservation, storage and presampling considerations are generally impossible to monitor by use of recovery tests.

3.7.1.2. Procedures for use of recovery materials

The choice of recovery material, which generally can be either a spiked sample or a native analyte/natural matrix RM, is limited in this scenario to spiking. Samples of the QCM and spiked QCM samples can be physically adjacent to each other (alternate) in the analytical sequence or for logistic and convenience considerations, the spiked samples may be placed at either the beginning or end of the batch.

Selection of recovery level for monitoring recovery should consider the expected concentration of the analyte determined and the possible effect of concentration on recovery. As a first choice, the level should be at or near the level actually present in the sample for most appropriate monitoring of method performance, particularly when dealing with highly homogeneous materials. Alternatively, the spike level can be a certain multiple of the limit of quantitation, limit of determination or limit of reporting.

3.7.1.3. Multi-analyte determinations

The determination of a recovery factor for one analyte, and its application to monitoring/ correction of other analytes, assuming it to be constant and applicable to all analytes determined, is unsuitable and strongly discouraged. This is equally true whether dealing with different or related organic analytes or elements. Although there will be a semblance of similarity in chemical behaviour among related analytes throughout the various chemical reactions constituting the method, behavioural differences can be significant. Error types and magnitudes can be quite specific to each analyte. The various errors in sample collection and manipulation and measurement impact differently on recovery of analytes from the sample bulk matrix. It is to be expected that collection techniques, sample storage and transportation, reduction of the gross sample to laboratory sample, sample manipulation, analyte volatilization during decomposition, incomplete extraction/separation, analyte retention by solid residue from incomplete destruction of matrix, alteration of oxidation state during decomposition and extraction, contamination, calibration, matrix effect management, selection of proper analytical technique, correction for physical, chemical and background interferences, starting calibrant material purity and composition and specific calculation details can all impinge differently on each analyte of interest. Thus when multielement/multianalyte determinations are conducted it is vital to determine an individual recovery factor for each analyte.

TABLE VIII. GENERAL SOURCES OF ERROR IN ANALYSIS OF BIOLOGICAL MATERIALS[a]

PRESAMPLING
BIOLOGICAL VARIATION
- Genetic predisposition
- Long term physiological influences (age, sex, geographical location, environment, diet, pregnancy, lactation)
- Short term physiological influences (circadian rhythms, recent meals, posture, stress)
- Seasonal changes (physiologic and climatic)
POSTMORTEM CHANGES
- Cell swelling, imbibition and autolysis
INTRINSIC ERRORS
- Medication, hemolysis, subclinical conditions, medical restrictions

SAMPLING
- Identification of population
- Sampling model and plan
- Representative and proper collection
- Storage and transportation
- Reduction of gross sample to laboratory sample
- Alteration of oxidation state
- Contamination from and losses to devices and containers

SAMPLE TREATMENT
DECOMPOSITION
- Volumetric ware verification, calibration, technique of use
- Drying and/or moisture determination
- Volatilization losses in dry ashing or wet decomposition
- Incomplete destruction of matrix, recovery and analysis of insoluble residue
- Alteration of oxidation state during decomposition
- Contamination from ashing aids and acids
- Contamination from and losses to decomposition vessels
- Dilution schemes

CHEMICAL SEPARATION
- Alteration of oxidation state
- Incomplete separation/extraction
- Contamination from reagents
- Contamination from and losses to vessels
ALL STEPS
- Reagent blanks (procedural and standard), non identity of method performance with pure reagents and actual sample
- Laboratory environment

MEASUREMENT
- Selection of proper analytical technique
- Instrument optimization, performance characteristics and utilization
- Physical and chemical interferences
- Background correction
- Calibrants (starting material purity and composition, preparation techniques of stock and working calibrants, verification, dilution schemes)
- Calibration solutions (single analyte, composite, matrix matching)
- Calibration technique (Calibration curve, bracketing for high precision)

DATA HANDLING AND INTERPRETATION
- Recording, data entry and calculation
- Calibration curve fitting and calculation techniques
- Interpretation and evaluation (controls, statistical treatment, data presentation basis)

OVERALL
- Analyst, specialist
- Data quality control (accuracy verification by recovery testing and performance with appropriate reference materials)

[a]From [54].

3.7.1.4. Preliminary requirements

In order to produce valid analytical data and to properly and cost effectively make use of recovery, it is essential that compliance with several prerequisites be established; the principal ones are correct analytical method and quality control.

An appropriate analytical method must be applied to the task on hand, by appropriately qualified and trained personnel in a suitable physical and administrative environment. Suitable physical environment refers to the equipment, materials, reagents and laboratory conditions necessary for the proper execution of the method; suitable administrative environment includes understanding of and support for appropriate data quality by the analyst's supervisor and all other managers. The role of the analyst is of direct paramount importance; good analysis and good analyst go hand in hand. Analyst training, experience, familiarity with the problem on hand, skill, attitude, motivation and judgement are necessary for satisfactory solution of analytical problems.

Suitable quality control/quality assurance procedures should be routinely in use and the need for appropriately reliable analytical information must be recognized. The analytical system must be in a state of statistical control, that is, operating optimally and consistently generating acceptable data.

When dealing with the determination of total concentrations of elements, that is, the sum of all the element concentrations in all material (sample) phases and molecular species, it must be ascertained that the method is in fact measuring all of the element. The sample decomposition procedure must bring into solution all of the material with no grains or insoluble fraction left behind. In addition, the element must be in the correct oxidation state required by the various chemical reactions constituting the procedure.

3.7.2. Determination of recovery based on added analyte

3.7.2.1. Applicability of the spiking approach

Frequently determination of recovery is based on the addition of the analyte being sought to a sample of material being analysed. In this approach, a known quantity of pure analyte is introduced at some stage in the analysis process, the sample/analyte combination is carried through the analysis and comparison of results with the baseline value determined for the sample gives an estimate of recovery. The nature of the added analyte, selected from available elements or compounds used for calibrant preparation, is not necessarily identical to or representative of the nature and form of the native analyte occurring in the natural material being analysed. This consideration is true for both organic and inorganic constituents. Thus, in principle, recovery estimated in this manner is not strictly accurate and should be regarded as solely an estimate. With organic analytes, reliance on spiking with analyte(s) of interest is, at times, the sole alternative for recovery determination.

Dependence of measured recovery on the nature of the element is not expected to be significant when sample destruction is complete, the material has been quantitatively brought into solution and all of the element is in solution and available for reaction and detection as required. It is also not expected to matter for certain instances of incomplete decomposition when the characteristics of undissolved residues are known. For example, if it is acceptable that Pb retained by silicate or siliceous residue of plant foodstuffs is not recovered, detected and measured by the applicable test method, then recovery of added Pb will be a reasonable measure of recovery. However, generally, the nature of the element should be taken as an important consideration when incomplete digestion is encountered. In instances of incomplete destruction of matrix, the time from spiking to analysis may also be an important factor.

3.7.2.2. Procedures for recovery determinations based on added analyte

Consideration of determination of recovery for each matrix type encountered in analysis or only on selected matrices representative of the materials scheduled for analysis depends on the criticality of the analyses. A preliminary semi-quantitative analysis of selected samples would be advantageous to establish at least an estimate of matrix composition to facilitate selection of the typical, representative and most suitable materials for spiking from among the lot to be analysed. This usually may not be feasible unless the analyst has access to high throughput multielement analytical techniques and the analyst must decide whether it is worthwhile to devote such additional effort to cover a wider range of material.

Introduction of analyte can be *via* a given volume of a solution of chosen concentration or with a known mass of dry analyte or compound. Solution addition is substantially a more feasible and convenient technique on account of the generally very small quantities of analyte required. Manipulative constraints in weighing and dealing with minute or microscopic amounts of solid analyte or compounds thereof, preclude sufficient accuracy with this approach.

The nature of the recovery solution(s) used for spiking is also dependent on the criticality of the analysis and the analyst's judgement and convenience. As a first choice, the recovery solution should be prepared, independently of the calibrants, from the pure solid element/compound weighed from the supplier's container, or from the concentrated stock solution from a reputable government agency or commercial supplier. This approach will ensure a more independent and accurate determination of recovery. A second choice is separate preparation of the recovery solution, at the required concentration, from the same single element or composite stock solution used for preparation of calibrants. Yet a third suggestion is simply use of a solution identical to one of those used for calibration, i.e. one of the higher concentration calibration solutions. For multielement analyses, it would be efficient and cost effective to use one aliquot of a suitable multielement recovery solution containing the multiple elements at appropriate concentrations. Addition of spike in the analytical scheme should be at the earliest opportunity.

Spiking levels should be selected to represent the expected level of analyte to be measured or a certain multiple of the enforcement/quantitation/determination / reporting limit. To incur acceptable error in the recovery factor, the ratio of added analyte to analyte present in the material should be several-fold the concentration of naturally occurring analyte with the actual ratio depending on circumstances. On the other hand, from the point of view of possible differences in analyte/matrix interactions at different, non-natural ratios, spiking levels should be closer to native content. Due consideration should be given to ensuring that the response of the spiked sample falls on the calibration scale to permit adherence to identical conditions of dilution, calibration etc. as for actual sample. How much one can deviate from this again depends on the specific circumstances; however, the greater the ratio the greater the certainty of recovery factor measurement.

It may be deemed important to carry out a sufficient number of repeat measurements at each concentration level in order to get a good estimate of the uncertainty, a parameter we believe essential in estimating recovery. Repeats are especially necessary when ratios of spike to native or total level

are small or unfavourable due to existing native levels leading to high uncertainty in the recovery factors. Assessment of amount of analyte (native level) already present in the sample prior to conduction of recovery tests is absolutely vital as proper definition of spike recovery refers to recovery of the **added** quantity. The rate of incorporation of spiked samples is at the discretion of the analyst and could range from less than one spiked sample per 100 samples (1/100) to more than 1/10 depending on the nature of the work and data quality requirements. In large routine analysis operations, where many similar samples are analysed concurrently in a batch or run, one suitable spiked recovery sample will suffice to monitor the performance of the method for quite a number of samples.

3.7.3. Calculation of recovery

3.7.3.1. Calculation of recovery values

The recovery factor is a quantitative representation of the proportion of added or endogenous analyte recovered and measured by the specified overall method and defines one performance aspect, perhaps the most important one, of the method. The factor is expressed as the ratio of analyte recovered to analyte added or known to be present and is given as a numerical value with an associated uncertainty. Ideally it has a numerical value of 1.000 but in reality will deviate positively or negatively from unity and is to be reported, as determined, as a number greater or less than 1. Measurements and calculations are separately conducted for added or native analyte and the two are not mixed. For example, when recovery of added analyte is determined, correction is made for the background content of native analyte so that the recovery factor reflects performance solely with added analyte. The practice of calculating 'blended recoveries' where the issue of spike and native recovery is confounded is not advocated.

Generally, two independent determinations of concentration constitute the basis for the calculation of the recovery factor. For calculations of added analyte (spiking), the first measurement is the analysis of the sample without added analyte to establish the baseline value. The second analysis is of the sample with the spike. The factor is calculated from:

$$F = [C_{b+s} - C_b] / C_s \qquad (5)$$

where

F is the recovery factor,
C_{b+s} is the concentration of the analyte determined in the spiked sample,
C_b is the concentration of the analyte in the sample (baseline content),
C_s is the theoretically expected concentration of the added analyte, due to the spike, in the sample.

This determination and calculation indicate the performance of the method with respect to the added analyte only and yields a value for the recovery of the added analyte independent of the performance of the method with native analyte.

3.7.3.2. Estimating uncertainty of recovery values

In line with good analytical and measurement practice, where an estimate of error should always be reported with every datum, the recovery factor should be presented with an associated uncertainty. Uncertainty is defined as any appropriate combination of precision and bias/systematic errors, in all the required determinations or known values, giving an estimate or indication of the overall possible error of the recovery factor. Uncertainty calculations can be based on laboratory — determined, laboratory — estimated or known precision and bias/systematic errors. Appropriate statistical calculations are applied in calculating uncertainties associated with the recovery factors. Since precision is dependent on concentration (as aptly presented by Margosis, et al. [55] who established a relation (Horwitz curve) between precision and concentration) and, to a first

approximation, independent of method as well as other parameters, it must be determined at the various concentration levels encountered for most reliable estimation of uncertainties of recovery factors.

In the case of spike recoveries from samples, repetitive determinations of the background (C_b) and spiked concentrations (C_{b+s}) will provide precision (standard deviation) estimates for the two required determinations which will lead to precision error estimates for the recovery factor. Precision and systematic error in C_s, the theoretically expected concentration of the added analyte, is expected to be small, and, for all intents and purposes, except the most rigorous endeavour, may be ignored. Error in C_s comes basically from (a) reliability of spike solution preparation and (b) reliability of spike addition technique when introducing it into the sample. Error in C_b depends on material homogeneity and the analytical method while error in C_{b+s} depends on material homogeneity, analytical method and error in C_s. Propagation of error formulas may be utilized to calculate the transmission of accumulated error to the final recovery factor. Much more mathematically complex uncertainty computations will have to be resorted to should the concentrations be considered as dependent variables, instead of independent variables. Identical computations are followed for the similar cases of recovery of added analyte from Reference, in-house or proficiency materials, where background and spiked concentrations are measured by repetitive determinations on the control and spiked control materials. If only determination of C_{b+s} is necessary and C_b is taken from the certificate of analysis, then precision of C_{b+s} and uncertainty in the actual, certified level from the certificate of analysis are the two uncertainty components.

As an example for the case of spike recovery, if we assume the following values and errors (say standard deviations) for each of the variables used in calculating F from the equation: $C_b = 2 \pm 0.2$ mg/kg (RSD = 10 %), $C_{b+s} = 12 \pm 0.4$ mg/kg (3.2%), $C_s = 10 \pm 0$ mg/kg (0 %), the recovery factor, $F = 1.000 \pm 0.045$ (4.5 %). In this case it is assumed that the baseline level is determined to ± 10 %, and the higher level spiked sample to ± 3.3 %; the error in the theoretically expected concentration of the added spike is taken to be negligible and assigned a value of 0 %. Further, for demonstration purposes F is taken to be unity.

3.7.4. Differentiation between recovery and bias/systematic error

One of the major misconceptions of less experienced analysts is that good recovery means good accuracy. Recovery studies are the most frequently used method of sample result or method validation. They do not reflect analytical accuracy, however, because they only evaluate recovery of the analyte added to the sample and tell us nothing about the amount of analyte present in the sample. That is, they give no indication about the accuracy of the unspiked sample signal. Thus, recovery can be 100%, yet analytical results can be biased and in error by orders of magnitude.

Situations causing such errors are (a) contamination of samples but not blanks, (b) contamination of blanks but not samples, (c) presence of uncorrected background in atomic absorption spectrometry contributing to a portion or to all of the analyte signal, or (d) an invalid baseline in chromatography or stripping voltammetry. For example, when foods are dry ashed for lead analysis in quartz or Pyrex vessels, lead present in the surface of the interior of the vessel can be leached into the sample ash by the aggressive nature of some of the ash components. Because the blank has no sample ash present, the measured reagent blank will be artificially low, and sample concentrations will be biased toward higher concentrations.

The ways to evaluate the presence of any of the above errors are to include appropriate RMs with baseline-levels of analyte (refer to blank mention above), or to analyse the samples using a completely independent method of analysis. The former is infrequently used because of availability and cost of good control materials and even when RMs are used, analysts prefer to choose those with higher concentrations because the quality control results "look better". The use of an independent method of analysis is usually impractical due to productivity demands on the analyst.

A sample weight test, which overcomes the above limitations with little additional time or cost to the analyst, has been described recently [56]. The test involves analysing two different weights of the same sample. One weight should be at least twice that of the other. The number of replicates determined at each sample weight depends on the critical nature of the sample, the homogeneity of the sample and the precision of the method.

If the same analyte concentration is obtained for the two different sample weights, then the result can be considered accurate. If different concentrations are obtained for the two sample weights, then the results should not be reported and a cause for the discrepancy should be sought. The sample weight test is particularly sensitive at low concentrations, and can reveal most of the method bias problems mentioned above. Thus, it is complementary to recovery studies as an evaluation of accuracy. The test should be applied to all samples of a critical nature (sample result validation) and to all test samples when a method is being validated.

Also relevant is the question of differences in the nature and extent of interactions of analyte with sample matrix referred to in Section 3.2. Complex kinetically- and thermodynamically-driven interactions can occur including intra-particle diffusion, physical and chemical binding, precipitation and other phenomena making the analyte unavailable to subsequent detection. The resolution of this issue would be easier if one had more understanding of such interactions.

3.7.5. Application of recovery factors

Recovery study results should be used with caution. As a general rule, recovery is only used to assess the performance of the method with a particular material. If a numerical value differing from unity is obtained for the recovery factor, a discrepancy is deemed to exist between the measured and correct concentration value indicating the analytical method/procedure not to be operating well. Should it be ascertained that an unacceptable error exists, a correction should not generally be applied. Instead, diagnostic steps should be taken to identify sources of unacceptable error or imprecision and remedial action should be taken to eliminate or at least minimize such errors in the method. Recovery factors generally should not be used to adjust the results to correct for recovery. The recovery factor, measured throughout the various stages of method fine tuning, development and application, serves to track method performance during development with the goal of arriving at a method with quantitative or sufficiently acceptable performance. Having stated that, there are circumstances when it is valid to use recovery studies to adjust sample concentration results for losses or enhancement by application of the recovery factor, and we feel that such adjustment is sometimes justified. For example, if a method is well defined and used by an experienced analyst, it is known whether or not the method has a bias. If the method bias is directly proportional to recovery, and this is known without a doubt, then it is valid, when high accuracy is needed, to use the recovery obtained in careful spiking studies to adjust the analytical results and thereby correct for recovery. When such corrections are made, however, it should be realized that the operation is defined as internal standardization rather than a recovery study, and the quotation of the recovery study results as part of quality control is invalid.

4. SCENARIO 3: CASE OF UNSTABLE ANALYTES AND/OR UNSTABLE MATRIX

Present situation with organic food contaminants such as Pesticide residues, veterinary drugs and mycotoxins

4.1. ANALYTES AND MATRICES

The case of unstable analytes in unstable matrices is typically represented by organic analytes in organic matrices. In this respect, analysis of food constituents and/or traces of organic compounds (in food for example) are targeted and are considered as typical case.

Being limited by the IAEA scope of activities, we would review mainly the trace organic analysis for unstable matrices, as is the case with the majority of foodstuff.

Although this field of analysis constitutes a significant part of analytical chemistry, it is noticeable that there is a lack of (C)RMs on the market that boosts, in conclusion, the need for in-house QCMs. Identical questions or issues are facing the preparation of (C)RMs and in-house QCMs for pesticide and veterinary drug residues as well as mycotoxin analysis.

The absence of these materials is not related to unawareness in the matter. The explanation for this fact is related to a number of factors, some of which are listed below.

4.1.1. Number of possible combinations of analytes and sample matrices

Just in the case of pesticide residues, over 1000 pesticide active ingredients are currently in use. The Food and Feed Crops of the USA lists over 400 crops that represent a large part of a healthy diet.

Besides that, currently the CCPR[3] programme includes about 3000 MRLs[4]. Table IX gives some examples of Codex MRLs (Codex MRLs are recommended on the basis of appropriate residue data obtained mainly from supervised field trials carried out in accordance with 'good agricultural practices'. MRLs represent levels that are toxicologically acceptable).

TABLE IX: COMMODITIES AND NUMBER OF SUBSTANCES FOR WHICH CODEX MRLS ARE ESTABLISHED

Commodity	# of MRLs
Potato	58
Citrus	41
Wheat	36
Cabbage	37
Apple	33
Banana	25

We can have a similar scenario with veterinary drugs also when it comes to their residues in the totality of an animal or poultry. As for mycotoxins the existence of only tolerable or recommendation levels along with the smaller number of mycotoxins analysed in one sample makes the matter a bit simpler.

4.1.2. The method of analysis and its uncertainty

It is recommended to use reference methods as they provide unequivocal identification and/or quantification of analytes. The "reference method status" is only valid if the method is implemented under an appropriate QA regime.

Knowing the number of combinations plus the mandatory and critical monitoring programmes conducted in several areas, e.g. food, environment and illicit drugs, etc., the application of multi-residue procedures is the only feasible option for regulatory analysts.

They are used in pesticide residues, veterinary drugs and mycotoxin analysis as well.

[3] Codex Committee on Pesticide Residues.
[4] Maximum residue limits in food are FAO/WHO Codex Alimentarius International Standards.

Multi-residues procedures are able to detect several analytes in one operation. In pesticide residue analysis, certain multi-residue methods can detect up to 250 residues in a food commodity. In this case, the instrumental determination e.g. with GC or HPLC requires an analysis time of 40–60 minutes for one sample if not longer.

Additionally to the multi-residue methods, special individual methods exist for specific analytes.

But when it comes to trace analysis, whether applying a reference method or not, being multi-residue method or an individual one:

The relative uncertainty of repeated analysis alone is 10–35% (strict metrologists consider such methods not quantitative, for which the criterion is ≤10%) [57].

The analyst, in general, should keep in mind that the inter-laboratory CV for the repeated analysis of a Reference or fortified material, under reproducibility conditions should not exceed the level calculated by the Horwitz Equation:

$$CV = 2^{(1 - 0.5 \log C)} \quad (6)$$

where

C is the concentration of the pesticide as a decimal fraction (1mg/kg = 10^{-6}).

Usually ¾ of this CV, expressed in Equation 6, can be expected within a laboratory.

Table X exposes agreed criteria for validated method in the case of veterinary drugs and pesticide residue analysis [58]. (Refer to the definition of these two terms *Repeat.* and *Reprod.* in [64]).

Table XI divides the overall uncertainty in the case of pesticide residues in relative uncertainties. We notice, in particular, the importance of the sampling uncertainty compared to the uncertainty of other components.

TABLE X. WITHIN LABORATORY METHOD VALIDATION CRITERIA FOR ANALYSIS OF PESTICIDE RESIDUES AND VETERINARY DRUGS[(a)]

Concentration	Repeatability		Reproducibility		Trueness[(b)],
	CV_A%[(c)]	CV_L%[(d)]	CV_A%	CV_L%	Range of mean % recovery
≤1 µg/kg	35	36	53	54	50–120
> 1 µg/kg ≤ 0.01 mg/kg	30	32	45	46	60–120
> 0.01 mg/kg ≤ 0.1 mg/kg	20	22	32	34	70–120
> 0.1 mg/kg ≤ 1 mg/kg	15	18	23	25	70–110
> 1 mg/kg	10	14	16	19	70–110

(a) With multi-residue methods, there may be certain analytes where these quantitative performance criteria cannot be strictly met. The acceptability of data produced under these conditions will depend on the purpose of the analyses, e.g. when checking for MRL compliance the indicated criteria should be fulfilled as far as technically possible, while any data well below the MRL may be acceptable with the higher uncertainty.

(b) These recovery ranges are appropriate for multi-residue methods. Stricter criteria may be necessary for some purposes, e.g. methods for single analytes or veterinary drug residues.

(c) CV_A: CV for analysis excluding sample processing. The parameter can be estimated from tests performed with reference materials or analytical portions spiked before extraction. A reference material prepared in the laboratory may be used in the absence of a certified reference material.

(d) CV_L: Overall CV of a laboratory result, allowing up to 10% variability of sample processing.

TABLE XI. QUANTIFICATION OF SOURCES OF UNCERTAINTY IN PRACTICE

Typical relative uncertainties %					
Sampling	Sample processing	Analysis (CV_A)			
			Extraction	Cleanup	GLC
30–40	5–56	16–53	≥ 1.5–3	5–10	8–15

TABLE XII. EXAMPLE FOR REPRESENTATIVE COMMODITIES/SAMPLES FOR VALIDATION OF ANALYTICAL PROCEDURES FOR PESTICIDE RESIDUES

Group	Common properties	Commodity group	Representative species
		Plant products	
I.	High water and chlorophyll content	Leafy vegetables	spinach or lettuce
		Brassica leafy vegetables	broccoli, cabbage, kale
		Legume vegetables	green beans
II.	High water and low or no chlorophyll content	Pome fruits	apple, pear
		Stone fruits	peach, cherry
		Berries	strawberry
		Small fruits	grape,
		Fruiting vegetables	tomato, bell pepper, melon mushroom
		Root vegetables	potato, carrot, parsley
III.	High acid content	Citrus fruits	orange, lemon
IV.	High sugar content		raisins, dates
V.	High oil or fat	Oil seeds	avocado, sunflower seed,
		Nuts	walnut, pecan nut, pistachios
VI.	Dry materials	Cereals	wheat, rice or maize grains
		Cereal products	wheat bran, wheat floor
Commodities requiring individual test			e.g. garlic, hops, tea, spices, cranberry
Products of animal origin			
		Meats	Cattle meat, chicken meat
		Edible offals	Liver, kidney
		Fat	Fat of meat
		Milk	Cow milk
		Eggs	Chicken egg
		Fish	Bivalves, sea fish, fresh water fish, etc

For additional information, always in the case of PRA[5]: (a) the combined CVA of repeatability ranges from 9.6 to 18%; (b) the combined CVR (R for reproducibility) for residues above 0.01 mg/kg ranges from 33 to 70%.

4.2. QA/QC MEASURES

Under these conditions, it is obvious that the implementation of a QA/QC system is a strong requisite, along with a close and regular monitoring of the important components. This situation will engender a considerable use or consumption of the (C)RMs or QCMs.

The major components of this implementation follow.

[5] Pesticide residue analysis.

4.2.1. Method validation

In order to validate the MRM[6] to be used, working with representative commodities and representative analytes is a good option to extrapolate a great deal of analytes/matrices combinations.

The FAO/WHO CAC[7] [59] organized food commodities in 33 groups, this classification can be reduced to fewer as presented in the Guidelines for Single-Laboratory Validation of Methods for Trace-Level Organic Chemicals [58], see Table XII.

In terms of analytes, especially if the number is very considerable and a grouping/classification is possible, the use of representative compounds for the QA/QC measures or the preparation of QCM should be explored.

This concept is applied for PRA and can be applicable in other analytical areas. Some representative compounds are shown in Table XIII for pesticides.

For veterinary drugs also, some practical guidance is given for the selection of appropriate matrix for testing, as shown in Table XIV.

TABLE XIII. SUMMARY OF PHYSICO-CHEMICAL PROPERTIES OF SELECTED REPRESENTATIVE COMPOUNDS

Active ingredient	Water solubility		LogPow[8] at	Vapor pressure	Hydrolysis	
	mg/l	°C	pH and/or °C	mPa at °C	DT_{50} [day]	pH; °C
DDE-p,p'	0.065	24				
Permethrin	0.2	30	6.1 at 20°C	0.045 at 25°C	>720	4, 50
Endosulfan [a]	0.32	22	4.74 at pH 5	0.83 at 25[ao]C		
Chlorothalonil	0.81	-	2.89	0.076 at 25°C		
Chlorpyrifos	1.4	20	4.7	2.7 at 25°C	In water, 1.5	8, 25
Lindane	7.3	25		5.6 at 20°C	191	7, 22
Iprodione	13	20	3 at pH 3 & 5°C	0.0005 at 25°C	1 to7	7, -
Dimethoate	23.3	20 (pH 5)	0.704	1.1 at 25°C	12	9, -
Azinphos-methyl	28	20	2.96	0.18 at 20°C	87	4, 22
Diazinon	60	20	3.3	12 at 25°C	0.49 185	3.1, 20 7.4, 20
Progargite	632	25	3.73	0.006 at 25°C	800	7, -
Methamidophos	200,0	20	-0.8 at 20°C	2.3 at 20°C	657	4, 22

Note (a): 2:1 mixture of α and β isomers.

[6] Multi-residue method.
[7] Codex Alimentarius Commission.
[8] Partition coefficient between n-octanol and water (as the log value).

TABLE XIV. APPROPRIATE TEST MATRIX FOR EXAMINATION OF RESIDUES OF VETERINARY DRUGS IN FOOD

Species/commodity for method validation	Usual target tissue or matrix for method validation	
	Water-soluble	Fat soluble
Ruminant (e.g. cattle, sheep)	Liver or kidney, muscle	Fat, muscle
Non-ruminant (e.g. pig)	Liver or kidney, muscle	Fat, muscle
Poultry (e.g. chicken, turkey)*	Liver, muscle	Fat, or muscle with adhering skin in normal proportions
Fish	Muscle with adhering skin in normal proportions	Muscle with adhering skin in normal proportions
Shellfish/crustacean (e.g. prawn)	Muscle	Muscle
Milk (usually cows' milk)	Whole milk	Whole milk
Honey	Honey	Honey

In the validation process, use of the representative commodities and representative analytes concept would help to establish the basic characteristics of an analytical method.

At a later stage, appropriate internal QC measures shall be implemented. Moreover, refinement of the performance characteristics during the regular use of the method is an important internal QA measure.

Coming back to our scenario on the trace organic contaminants, applied methods are usually either validated and/or qualified as standard, official and reference methodologies. This condition fulfils an important QA criterion.

However, as the extent of validation is limited, use of CRMs is very pertinent.

In the case of repeated analyses of a CRM, the experimentally determined mean content should not deviate from the certified value more than ±10 %.

When no such (C)RMs are available, it is acceptable that the trueness of measurements is assessed through recovery of additions of known amounts of the analyte to the unknown or preferably blank samples. Attention is drawn to the fact that the added analyte is not chemically bound in the real matrix and therefore results obtained by this approach have lesser validity than those achieved through the use of natural matrix (C)RMs

Unfortunately, in this case the available CRMs do not match the need or do not exist at all. Indeed, with respect to unstable matrices, such as foodstuffs, the matrix is often powdered and dried to ensure homogeneity and stability, creating finally a matrix mismatch.

Nevertheless, occasional check with the available (C)RMs (even if not matching), at or near either the maximum residue limit or decision limit, would deliver another proof for the goodness of the implemented QA/QC system.

4.2.2. Quality control

In order to compensate for the absence of (C)RMs and/or QCMs, the following basic QC steps are to be performed regularly and/or carried out simultaneously for each batch of test samples analysed:

- System suitability testing of the used instrument(s): to be performed first place
- Control charting for the principle parameters of the applied technique, e.g. recovery, yield, instrument performance and response
- Statistically sound acceptance criteria and evaluation of results

- Distribution of the reference substance (standard) over the whole analytical batch for the calibration
- Checking the calibration: use of two parameters: the coefficient of correlation (r) and the SD of (relative) residuals
- Analysis of blank samples to check interference: included in each analytical batch. Reagent blanks can be optional
- Recovery testing: included in each analytical batch
- Surrogate recovery testing[9], if blank samples are not available
- Reference and fortified material containing known amounts of the analyte(s) as well as blank material should be treated in the same way as the test samples
- Qualitative and quantitative confirmation of measurand or residues detected above an action or acceptable level: each result
- Repeated analysis of positive samples: e.g. one in each analytical batch
- Internal checks with blind samples in order to avoid the analysts' bias
- Finally, some guidelines even give the recommended order for injecting the extracts into the analytical instrument, e.g. reagent blank, negative control sample, sample(s) being confirmed, negative control sample and finally positive control sample [60].

4.2.3. Additional quality control measures

Added to the points mentioned above, robustness or ruggedness of the analytical methods under different realistic conditions is a useful means to check the fitness for the purpose of the applied method. The variable conditions might be different sample weight, extraction solvent, detectors, wavelength, operators, etc.

With respect to the peculiar points of the instrument calibration [61] and the QCM preparation, use of a substance[10] with certified purity as RM, is a must. As for recovery studies and spiking/fortification (whether for blanks or surrogates), they can be made using secondary standards[11] if the first choice is not possible and if secondary ones were certified against primary standards.

After viewing the essential internal QC measures, we can draw a limit between the advantages and drawbacks of usage of QCMs. Some of these are listed below:

Advantages:

(1) Provides consistent information on the performance of the method when applied for different samples of small numbers (as it is the best compromise. In case of large number of similar samples, a real matrix match RM is preferable);
(2) May be sufficient for QC of screening runs under certain conditions;
(3) Enables the application of control charts;
(4) Enables comparability between laboratories using the same materials.

Drawbacks:

(1) Not specific for the analyte/commodity combination analysed;
(2) Recovery studies with the tested combination are required for quantitative confirmation;
(3) Does not reveal, in general, information on the efficiency of sample processing (repeated analyses of test portions are still required).

Bearing these limitations in mind, we will present in the following section the relevant and feasible means of sample and analyte preservation leading to an appropriate and improved integrity of QCMs.

---

[9] For more information about surrogates, check the 'Harmonized Guidelines for the Use of recovery Information in Analytical Measurement (technical report)' by M. Thompson, S. Ellison, A. Fajgelj, P. Willets and R. Wood, IUPAC, Pure Appl. Chem. **71** (1999) 337–348.
[10] In the sense employed by the ISO Guide 11095.
[11] With purity less than 95% or technical grade substance.

4.3. PREPARATION OF QCMS

Since Section 3 of this report deals with the development of QCMs, we will shed light on the important aspect of stability of QCMs that is divided into two:

- Stability of the matrix;
- Stability of the analyte(s).

In case natural matrix (C)RMs are not available and the preparation of QCM is carried out, special measures for stabilizing either the matrix and/or the analytes should be undertaken.

A basic recommendation for ensuring the stability of any sample, which applies to QCM also, is the storage at –20°C. It is even a prerequisite for QCMs as their stability might strongly affect the ruggedness of the analytical technique if not considered.

Another basic rule is the transport of samples from and to laboratories that must be carried on ice blocks, preferably. In some laboratories storage is even made at –75°C. This latter option must be considered if it is cost effective, affordable and imperative.

In a study made on tetracycline residues over a three month storage period, no difference was detected between the storage at –20 °C and –75°C [62].

On the other hand, a third basic recommendation in the case of natural matrix QCMs, is to check the effect of sample preparation that can significantly affect the concentration of anlaytes [63] (ref. D- Case study) from the very beginning. For instance, grinding the meat before storage allows to keep the penicillin residues near the higher level of concentration, but when bulk meat is stored frozen, systematically a decrease of concentration could be detected [64].

A non-exhaustive list of examples for stabilization of matrix and analytes are given below and developed later in the text.

4.3.1. Matrix

- Freezing (storage and processing under reduced temperature)
- Lyophilization [65, 66]; (freeze drying)
- γ irradiation; doses up to 10–25 kGy were reported in the literature
- Storage of digests or extracts at low temperatures
- Microwaving [67] or heating [68] without cooking in order to denature the matrix enzymes and decrease their degradation potential on pesticide residues
- Treating with preservatives, such as sodium azide (NaN_3) at 0.02%, thimerosal ($C_8H_9HgO_2SNa$) at 0.01%, and additives used in home made preparations, based on, for example, benzoic acid at 0.4%.

4.3.2. Analytes

- Storage at –20°C or below
- Addition of a keeper substance
- Trapping on column, e.g. SPE column.

An important work on the stability of both analytes and matrices can be found in the literature. This work aimed to ensure that generated data are valid and the measurands remain accurately quantifiable from the time of sampling to analysis, whatever the sample or the QCM (in our case) is submitted to.

4.3.3. General guidance for stabilization

Various experiments were done on the stabilization of analytes, mainly for organic residues, in water matrices. These researches converged towards the preservation of analytes, which can be linked to the preparation of QCM for unstable analytes.

Organic analytes are subject to degradation by different modes: biological, chemical and physical. Some examples of these paths are hydrolysis, photolysis, oxidation, etc. The temperature and the matrix or the solvent in which the analytes of interest are contained constitute a major factor for their stability.

4.3.1.1. Trapping or binding

As we mentioned earlier, one of the advantages of QCMs is to enable comparability between laboratories using the same materials. For instance, in the case of water analysis, intercomparison runs were organized [69] based on the use of loaded solid phase cartridges or columns with analytes.

The trapping of analytes revealed to be more efficient, in terms of stability, than the biologically inhibited water. Moreover, it is a practical method for transporting analytes and or samples. The latter fact would ensure better comparability between laboratories, in case they are used in different locations.

It was found that trapping avoids or reduces degradation by preventing the breakdown of sorbed hydrocarbons by bacteria. Yet the loaded cartridges, namely graphitized carbon black ones [70] and C_{18} solid-phase extraction pre-columns [71–73] are preferably kept at –20°C. This type of stabilization, under the described conditions, permitted good recovery of the tested analytes.

4.3.1.2. Freezing: (storage at reduced temperature)

Freezing has been applied to a variety of matrices from food to soil, passing by water and organic extracts in different forms of these matrices. Different modes of storage are applied in analytical laboratories, with respect to pure analytical standards and their solutions. In some laboratories, these are stored at –18°C in order to extend their shelf life. One study proved that considerable number of pure substances related to pesticides is stable in a freezer at –18°C up to 15 years [74]. As for the corresponding stock solutions, prepared at a 1mg/ml level in toluene, they were stable for three years at the same temperature. Likewise, mycotoxins solutions are better stored at –18°C [75].

4.3.1.3. An alternative to freezing is freeze drying

A group of researchers tested the freeze drying of water (drinking and Milli-Q) spiked with pesticides in order to prepare samples for an interlaboratory comparison [76, 77] and consequently avoid the risk of hydrolysis, for instance.

To summarize, there is no best technique for stabilizing a matrix or an analyte. They vary according to the nature of the compound or the tested food matrix.

Other options, explored below, exist for laboratories in order to prepare positive stabilized samples.

4.3.4. Analytical samples and portions

An interesting way to have a matrix-matched QCM, yet under limited stability conditions, is to inject the analytes of interest inside the studied matrix (e.g. oranges and peaches) and freeze it. This practice, explored by a Greek group [78], is very practical for two reasons:

- To have a perfect matrix matched QCM
- To study the effect of sample processing, in terms of stability and degree of homogeneity, along the goodness of the analytical procedure.

However, the preparation, storage (at –20°C) and use of analytical portions[12] are preferred for many reasons also, if:

- The homogeneity uncertainty of the analytical sample is affecting significantly the overall analytical uncertainty. This component disappears in case of spiked analytical portions;
- The analytical sample[13] is not used in its totality;
- There is a risk of segregation in the stored analytical sample [79].

This aspect of storage in analytical portions was even well emphasized early for PRA [80].

4.3.2.1. Extracts

Preservation of analytes for an extended period of time can be accomplished through storage in dry organic extracts, at low temperatures, i.e. ca. 4°C. Some studies showed that the stability of most pesticides, like synthetic pyrethroids, organochlorines, some organophosphates, in extracts is comparable to their stability in pure solvents [81].

Selection of the solvent and the concentration level for the standards or extracts conditioning is crucial for ensuring their stability during storage. Beside the fact that exposure to light and the possibility of existence of active sites on the glass containers, solutions of analytes in *n*-alkanes for pesticides and acetonitrile/water for mycotoxins, for instance, would not have the same stability compared to, for example, in methanolic solutions [82, 75].

In case of proven stability, the cleaned organic extracts can be used as matrix matched calibrants or standards for matrix sensitive instruments.

Cleaned organic extracts (free of analytes) can be spiked with compounds of interest, in this case. An important aspect though, is that organic extracts should be free from any trace of water and preferably stored in sealed brown glass ampoules.

An interesting means to extend the stability of extracts is to concentrate those or even to evaporate them in the presence of a "keeper" substance. The role of these substances, being characterized with a high boiling point, is to condense and dissolve, e.g. the pesticide residues, hence avoiding their losses, or to complex them in order to improve their stability. Usually a small miscible amount is used up to 10% of the analytes' solutions or extracts [83].

Some of the tested keepers are: ethylene and propylene glycol, glycine, stearic acid, white oil, $HgCl_2$ and dodecane.

Another means to keep some specific analytes is via acidification with HCL, diluted sulphuric acid or a monochloroacetic buffer to prevent the degradation of volatile compounds and pesticides, e.g. herbicides and carbamates [76]. Maintaining a pH <3, in case of water analysis for carbamates residues, is necessary to inhibit both chemical and biological degradation [84].

As mentioned in the part dealing with the QA/QC measures, statistically based approaches are recommended for planning and interpreting the results. Therefore we devoted the following section to the important statistical aspects related to usage of QCMs.

---

[12] Analytical portion or test portion: a representative quantity of material removed from the analytical sample, of proper size for measurement of the analyte concentration.

[13] Analytical sample: the material prepared for analysis from the laboratory sample (a representative quantity of material removed from the bulk sample) by separation of the portion of the product to be analysed and then by mixing, grinding, fine chopping, etc., for the removal of analytical portions with minimal sampling error. Extended terminology on samples can be found in the FAO publications, e.g. Joint FAO/WHO Food Standards Programme, Codex Alimentarius Commission Vols 2A & 2B, Pesticide Residues in Food, Methods of Analysis and Sampling. 2nd Ed (2000).

4.4. STATISTICAL ASPECTS

4.4.1. Accuracy and precision

Before the end of the 1960s, the time of the adaptation of the GC technique to PRA, scientists used colorimetry (despite the problematic aspect of its use) TLC, spectrophotometry or cholinesterase tests in e.g. PRA [85].

The results dating from that period, mentioned:

- Surprising possibilities of detecting degradation rates of few percent! [86, 87][14], even if the amounts used were 100 times higher than the ones used presently.
- On the contrary, very high degradation rates and in some cases contradictory results [88, 89] vs. [90]. (The case of parathion in frozen storage conditions).

Under these conditions, it was difficult to draw any conclusion or to generalize when it comes to stability of pesticide residues.

In the event of a QCM preparation, special care has to be taken to ensure a minimum uncertainty accompanying the QCM agreed or consensus value. Therefore, it is preferable to have this value based on the analysis of this material in several laboratories, as the uncertainty is inversely proportional to n (n being the number of laboratories that participated in the analysis).

The QCM uncertainty value should not be much higher than the uncertainty of the applied analytical procedure in individual laboratories.

In general, as seen above, in the case of trace analysis, analytical variations are more important than variation due to residue degradation [91]. Therefore, for an accurate stability testing of organic analytes, we can increase accuracy of the analytical method by:

- Analysing a statistically sound number of replicates
- Extending the stability testing period.

When short stability testing period are set, unless it is the case of accelerated tests, the important variability inherent to trace analysis, makes it hard for an analyst to quantify degradation. Therefore, stability studies under these conditions would rather denote tendencies for degradation as even recovery studies could give sometimes very low results. This fact is understandable as long as the stability and the analytical uncertainties are overlapping.

To illustrate this difficult issue, we will make use of the performance limits (warning and action limits) set for the applied analytical procedure [58].

$$\text{Warning limits} = Qa \pm (2\ CV_{typ}\ Qa)\ \&\ \text{Action limits} = Qa \pm (3\ CV_{typ}\ Qa) \qquad (6)$$

Where Qa is the average of recoveries at all fortification levels and for all analytes, from which we can calculate a typical CV.

From that point, control chart rules (Westgard) can be applied [92], see Sections 2.10 and 6.9 for more information.

From real laboratory data, we have tried to calculate the previously mentioned limits. We obtained the following interesting results, presented in the table below.

[14] The same author, in 1972, did not detect any degradation of dimethoate in the same matrices studied in 1969.

TABLE XV. ACCEPTABILITY LIMITS OF ANALYTICAL RECOVERIES

Qa (in %)	CV_{typ}	Warning Limits (in %)		Action Limits (in %)	
		Higher	Lower	Higher	Lower
90.5	0.1725	121.7	59.3	137.3	43.7
94.7	0.204	133.3	56.1	152.6	36.7

Under these conditions, one can conclude that the stability studies are rather challenging. Hence the importance of keeping analytical variation of residue data to a minimum.

In parallel, we can make use of the critical range (2.8 x SD).

In this case, the difference between (C_{max}): the highest value and the lowest (C_{min}), of a number of replicates, must be lower than the critical range as defined in equation 3.

$$C_{max} - C_{min} = 2.8 \times CV_{Ltyp}Q \quad (7)$$

CV_L represent the intra-laboratory reproducibility including the sample processing uncertainty.

The use of the analytical instrument response's variation, to detect non-stability, requires rather robust instrumentation that is characterized with a very low uncertainty.

On the other hand, an extended stability study would enable the analyst to give a mathematical model to the degradation pattern characterized by a specific degradation rate with confidence intervals given to the model. The interesting aspect of these models is the possibility of predicting shelf lives.

4.5. STATISTICAL INTERPRETATION OF DATA

This subject is divided into several parts:

Recovery: we have to know what is the significance of the difference between our recovery rates based on spiking blank samples and the theoretical 100% recovery.

Bias: we have to know what is the significance of the difference between the certified value of a (C)RM and our in-house related values.

The two previous parts are independent from the following third part.

Stability: In case of in-house QCM, we have to know what is the significance of the difference between the replicate analysis at two different periods of time, i.e. stability testing period.

In general, the number of independent replicate measurements, in analytical laboratories, does not exceed 2 and rarely 3. This number might serve to answer the first question but not the second or the third.

With variations, reaching 20 even 30% (table 2), statistical tests, like the *t*, *F* tests or the "least significant difference", are not capable of detecting significant differences indicating some instability.

More powerful tools should be used, allowing for significance testing taking into account the errors of types I and II (α: false negative and β: false positive, respectively) and therefore consolidating our conclusions.

To interpret the third case, there is one statistical test giving the number of replicate analysis in order to detect a difference between 2 means, i.e. the difference between the average results of samples subject of the stability study (μ_1 and μ_2) [93].

This test is the *t*-test for the sample size (= the number of replicates to analyse).

To apply this test:

- The analytical procedure should be characterized with a typical acceptable standard deviation very close to the true standard deviation, therefore valid for the analysis at time 1 and time 2 ($\sigma_1 = \sigma_2$).
- The number de replicates measurement allowing detecting a deviation > σ, should be chosen based on the value Δ calculated from equation 4.

The ratio of the desired difference and σ gives a value Δ that, at certain probability levels for the α and β errors, indicates the required number of replicate measurement. This number is taken from the correspondent statistical table.

$$\Delta = \frac{|\mu_1 - \mu_2|}{\sigma} \qquad (8)$$

Considering a $\Delta = 1.1$ with $\alpha = \beta = 0.05$, the number of replicates would be 23 for each study period, so 23 independent analytical portions. The smaller the σ , the higher the Δ and, therefore, the smaller the number of replicates.

The Harmonized Guidelines [58] propose the analysis of minimum 5 replicates at each stability study period. In this case, an extended stability study period is required in order to built the mathematical model already mentioned.

In this following section we will present a case study related to the sample processing uncertainty, an important aspect of the preparation of natural matrix QCMs.

4.6. CASE STUDY

Within the programme of the FAO/IAEA Training and Reference Centre, one of the major research axes at the Agrochemicals Unit is the estimation of uncertainties due to sample processing and instability.

These are as well dominant aspects in the step-wise preparation of (C)RMs or in-house QCM.

4.6.1. Sample processing

The effect of sample processing was explored in two cases with tomato samples:

A surface treated primary analytical samples (not homogenized) with a radioactive pesticide solution at its MRL level.

Fortification of a secondary analytical sample (already homogenized) with a similar pesticide solution.

In the first case, we made use of the sampling constant (K_S) concept to evaluate the degree of homogeneity, by taking replicate analytical portions of different sample size [94], according to Fig. 6.

On the other hand, several processing techniques and equipment were also tested. (e.g. after freezing the sample, with and without addition of dry ice, double processing).

In the second case, a one-way ANOVA was used to evaluate the homogeneity, again by taking replicate analytical portions of different sample size.

In both cases the best estimate of the sample processing uncertainty was 4.1 and 5% for the fortified homogenized sample and the surface treated sample subjected to a double processing, respectively.

The stability study was conducted under common laboratory conditions including sample processing and storage. Surprisingly, sample processing had a pronounced effect on the concentration — up to 50% loss of analyte — of some pesticide residues [63]. Therefore, this check should be always undertaken.

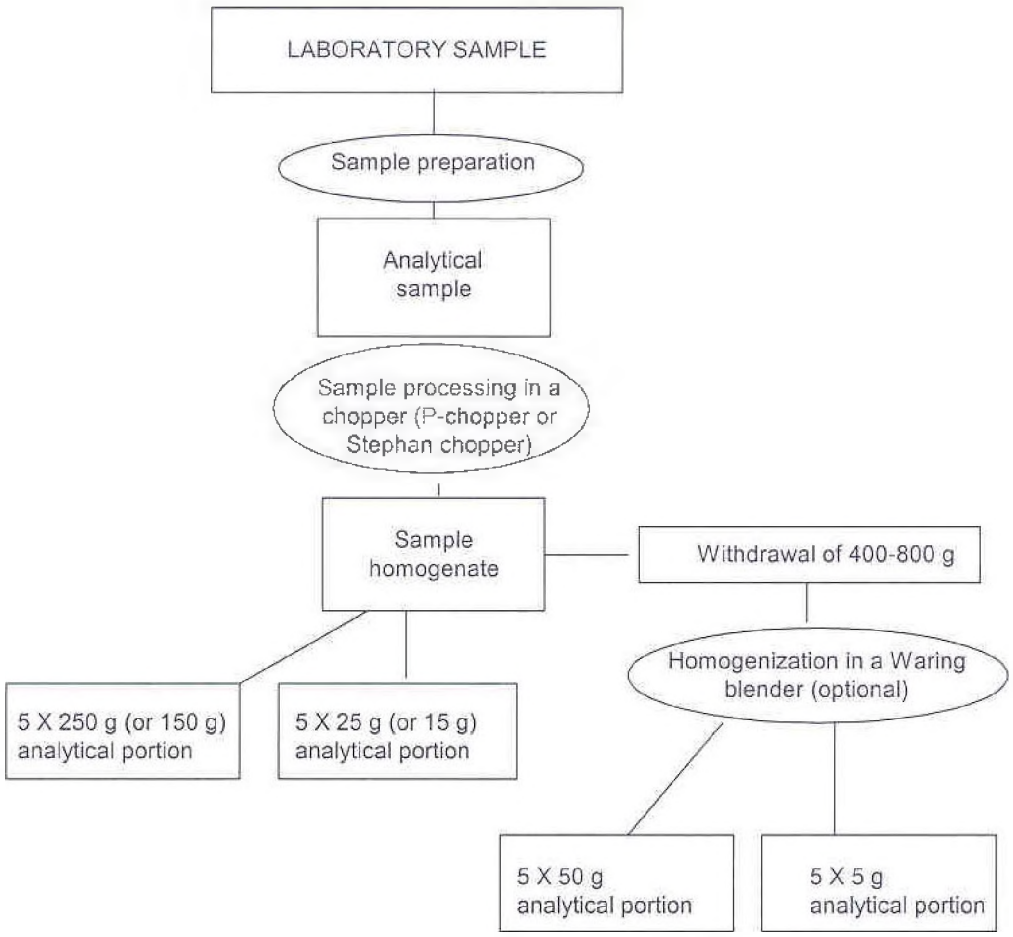

FIG. 6. Sampling plan scheme of analytical portions.

We concluded also that the storage of analytical portions is far better than the storage of sub-analytical portion. Indeed, a second contribution to uncertainty related to the in-homogeneity of the portions would be added to the in-homogeneity one of the composite sample, and therefore amplifying the overall uncertainty. In our research case, after storage of sub-analytical samples, the CV of the in-homogeneity passes from 5.3 to 18.8% [79].

Storage stability at –18°C is a preferable condition for maintaining the integrity of the residues (pesticides, mycotoxins, etc.). This holds, however, for a limited period only, as even at this temperature residues might undergo degradation to a certain degree.

Consequently, the determination of the decline profile and/or a shelf life estimate is primordial for a good use of the in-house QCM.

4.7. CONCLUSION

The need for (C)RMs in analytical chemistry is a must for all its application, this is enforced by more and more demanding requirements for establishing a Quality system in analytical laboratories. Added to that the strong recommendations for analytical methods validation.

It is clear that dealing with unstable analytes as well as unstable matrices is a challenging task for an analyst, yet some options are given in order to improve the stability for both components. These options deal with modifying the matrix (freezing, lyophilizing, heating, etc.), which would change the analyte's environment and, at the same time, improve its stability, as it was mentioned for the penicillin residues in meat. In all cases this stability shall be monitored and established preferably with an uncertainty for a better use of the related QCMs.

The implementation and use of a validated method or at least characterized with acceptable performance criteria is a must for the characterization of a QCM. Once this is made, its stability shall be tested. It can be done at best with the use of statistical tools in the planning stage as well as the interpretation of the results.

5. BASIC STATISTICAL TOOLS FOR THE ANALYTICAL CHEMIST

5.1. INTRODUCTION

The analytical chemist faces a number of requirements to do his job properly and, most importantly, to have their results recognized and accepted not only by the scientific community but also by traders, health authorities, governmental institutions, regulatory agencies, environmental organizations just to mention some of the users of analytical data. In the past, the analytical chemist was concerned with the analysis of samples received from a customer who was interested in either a qualitative or a quantitative answer to his problem. Common questions were whether a given analyte was present in the sample and, if yes, in how much quantity. The analytical chemist used then an analytical method obtained from the open scientific literature or established by an organization, either national or international, as a reference or standard method, for the analysis of the sample. The result, usually the average of duplicate analyses, was reported as a number in the appropriate units.

Today, national and international acceptance of analytical data is more demanding. It is requested that, prior to the routine use of a given analytical method, it has been demonstrate that it is useful for its intended purpose and been properly *validated*. It is also essential to demonstrate the *traceability* of the results and, most important, to estimate their *uncertainty*. On the other hand, most field laboratories needs to have a record of the determination of given analytes in frequently analysed samples. This information is kept in the form of *control charts*, which are important for demonstrating the performance of the laboratory (analytical method) in the long term.

To comply with these demands, the analytical chemist needs to be familiar with some statistical procedures to demonstrate the validity of the results. These procedures are simple and, in most cases, straightforward. The availability of computers and statistical software allows the execution of these tasks faster and straightforward. Even more, there are several computer programmes dedicated specifically to cope with matters such as method validation, traceability and for the evaluation of *collaborative tests*. The information contained in this paper does not pretend to cover all aspects of statistics or the mathematical basis of them. It is intended to complement the information written in the other chapters of this TECDOC and to give the analytical chemist a more practical view of the applications of statistics to his day to day work. If the reader is interested in a deeper look to the matters presented here or to others related to statistics, his attention is drawn to many excellent books and papers available, some of which are indicated in the bibliography.

5.2. WHAT IS STATISTICS?

Statistics is an area of science concerned with the design of experiments or sampling procedures, the analysis of data and the making of inferences about a *population* of measurements from information contained in a *sample*.

*A **population** is the set representing all measurements of interest to the sample collector.*

*A **sample** is a subset of measurements selected from the population of interest.*

Statistics helps in studying various inferential procedures, in looking for the best predictor or decision making process for a given situation. Even more important, it provides information concerning the goodness of an inferential procedure. When predicting, it is important to know something about the error in such prediction. If a decision is taken, what is the chance that our decision is incorrect. Our built-in individual prediction and decision making systems do not provide immediate answers to these important questions and could be evaluated only by observation over a long period. In contrast, statistical procedures do provide answers to these questions.

To make an inference about the population from which the sample is drawn, it is essential the inspection of the observed data and, second, the selection of the appropriate statistical procedure. For the purpose of this contribution, we will suppose that the numerical data, or any type of observation, was obtained through controlled experimentation or data collection. Furthermore, we will include additional data that might come from various methods of experimentation giving varying amounts of information. Hence, essential to statistical problems is the design of the experiment, or sampling procedure, which must enable the gathering of a maximum amount of information for a given condition. This aspect of a statistical problem may be less important when data collection is easily done. However, in many data collecting situations where it is impossible to repeat poorly conducted experiments or where the data are costly, the design of the experiment or sampling procedure assumes a very important role.

To summarize, a statistical problem involves the:

(1) design of the experiment or, sampling procedure,
(2) collection and analysis of data, and
(3) making of inferences about the population based upon information in the sample.

It is extremely important to note that the steps in the solution of a statistical problem are sequential; that is, you must plan how you will collect the data before you can collect and analyse it. And all these operations must precede the final step, making inferences about the population based on information contained in the sample. These steps, designing the experiment or sampling procedure, can be, and often are, omitted. The experimenter may plan the data collection in a manner that intuitively seems reasonable or logical but which may be an extremely poor plan from a statistical point of view. The resulting data may be difficult or impossible to analyse, may contain little or no pertinent information or, inadvertently, the sample might not be **representative of the population** of interest. This means that if the experimenter is not knowledgeable in the statistical design of experiments and (or) sample surveys, he should consult an applied statistician for the appropriate design *before* the data are collected.

5.3. ESSENTIAL CONCEPTS

There are a number of concepts that have to be known to the analyst. Several of them may be known to the reader. However, they are presented here in order to harmonize the bases for the following sections.

In our daily work in the laboratory or even in the normal life, we always talk about *variables* to indicate parameters, which may have different values. Variables are things that we measure, control, or manipulate in research. There are two types of variables: *independent variables* and *dependent variables*. The terms dependent and independent variable apply mostly to experimental research where some variables are manipulated, and in this sense they are "independent" from the initial reaction patterns, features, intentions, etc. of the subjects. Some other variables are expected to be "dependent" on the manipulation or experimental conditions. Regardless of their type, two or more variables are related if in a sample of observations, the values of those variables are distributed in a consistent manner. In other words, variables are related if their values systematically correspond to each other for these observations.

When applying statistics to our results, often we refer that a given result or a relation between variables is *statistically significant*. The statistical significance of a result is an estimation of the "trueness" of such results. More technically, the value of the so-called *p-value*, the indicator of the significance, represents a decreasing index of the reliability of a result. Specifically, the *p-value* represents the probability of error that is involved in accepting our observed result as valid. For example, a p-value of 0.05 (i.e., 1/20) indicates that there is a 5% probability that the relation between the variables found in our sample is due to chance. In many areas of research a *p-value* of 0.05 is usually considered the limit for an acceptable error level and, for the purpose of our applications, this will be the case.

The significance of a relation between variables depends on the size of the sample. If there are very few observations, then there are also respectively few possible combinations of the values of the variables, and thus the probability of obtaining by chance a combination of those values, indicative of a strong relation, is relatively high. Therefore, in many procedures in analytical chemistry, it is important to carry out as many experiments as possible to strength the bases for establishing a significant relation for the parameters of interest.

An important point, which is often not carefully studied when analysing a set of results, is the determination of the type of distribution that such data follows. One usually assumes that the data follows a *normal distribution*. However, this is not always the case. There are several distributions which can our data fit into: normal, Poisson, log, log-normal, semi-log, etc.

5.4. MEASURES OF THE CENTRAL TENDENCY AND THE DISPERSION OF THE DATA

Normal distribution is important because many statistical tests are applicable and the inference made from them is valid only, if the data follows such distribution. The exact shape of the normal distribution, graphically represented by the well known "bell curve", is defined by a function, which has only two parameters: *mean* and *standard deviation*.

*The **arithmetic mean** of a set of n measurements x_1, x_2, x_3, ..., x_n, is equal to the sum of the measurements divided by n:*

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad (9)$$

*The **variance** of a set of measurements x_1, x_2, x_3, ..., x_n, is the average of the square of deviation of measurements about their mean*:

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n} \qquad (10)$$

*The **standard deviation** of a set of n measurements x_1, x_2, x_3, ..., x_n, is equal to the positive square root of the variance:*

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}} \qquad (11)$$

A characteristic property of the normal distribution is that 68% of all of its observations fall within a range of ∀1 standard deviation from the mean, and ∀2 standard deviations include 95% of the data.

Problems may occur or wrong conclusions are made when a test based on the normal distribution is applied to a set of data, which does not follow this type of distribution. In such situations there are two alternatives to solve the problem. First, we can use some alternative *non-parametric test* or the so-called "distribution-free test". However, such tests are less powerful and the conclusions they would provide may not be definitive. Alternatively, in many cases one can still use the normal distribution-based test if the size of the sample is large enough. As the sample size increases, the shape of the sampling distribution approaches to a normal shape, even if the distribution of the variable in question is not normal.

In rigour, therefore, it is required that the first step in a statistical analysis should be to examine if the data to be analysed follow a normal distribution. There are several statistical tests, which can be used to determine whether the distribution of the data is normal. One of these parameters is the *kurtosis*. The *kurtosis coefficient* is an indication of how flat or steep the distribution of the data is compared to a normal distribution. For a normal distribution, the kurtosis coefficient is zero. When the coefficient is less than zero, the "bell curve" is flat with short tails. When the coefficient is greater than zero, the curve either is very steep at the centre or has relatively long tails.

A second parameter is the *skewness*, which is used to measure the symmetry or shape of the data. A skewness of zero suggests that the data are symmetrically distributed. Positive values of skewness indicate that the upper tail of the "bell curve" is longer than the lower tail; negative values indicate that the lower tail is longer.

If the *kurtosis* and the *skewness* have values between ∀2, the data follow a normal distribution.

Another statistical parameter used quite extensively when reporting results from the analyses of a number of samples, is the *confidence interval* of the mean. A *confidence interval* for a mean specifies a range of values within which the unknown population parameter, in this case the mean, may lie. These intervals may be calculated by, for example, a producer who wishes to estimate his mean daily output; a medical researcher who wishes to estimate the mean response by patients to a new drug; etc. The width of the *confidence interval* gives us some idea about how uncertain we are about the unknown population parameter, in this case the mean. A very wide interval may indicate that more data should be collected before anything very definite can be said about the parameter.

We calculate these intervals for different *confidence levels*, depending on how precisely we want to be. We interpret an interval calculated at a 95% level as, we are 95% confident that the interval contains the true population mean. We could also say that 95% of all confidence intervals formed in this manner (from different samples of the population) will include the true population mean.

In general, the confidence interval for the mean can be calculated using:

Assuming that the distribution of the data is normal, we can define the confidence interval of the mean with a 95% confidence level, as:

$$\mu = \bar{x} \pm z\frac{\sigma}{\sqrt{n}} \qquad (12)$$

where z (coefficients of area under the normal curve) takes different values according to the ***degrees of freedom*** and the confidence level. Thus, for a 95% confidence level, z is equal to 1.96 and for a 99.7% confidence level z takes the value of 2.97. Usually, to facilitate calculations z takes the value of 2 for a 95% confidence level.

As the sample size gets smaller, the uncertainty introduced by using *s* (the standard deviation) increases. To allow for this, the equation applied to calculate the confidence interval is modified to:

$$\mu = \bar{x} \pm t\frac{s}{\sqrt{n}} \qquad (13)$$

where *t* corresponds to the *distribution of Student's t,* which is used for a small number of data following a *normal distribution.*

Table XVI includes data with results for the determination of zinc in a candidate reference material for chemical analysis; these data will be used to illustrate several applications of statistical tests to analytical results.

As an example, we will calculate the parameters explained so far using the data in Table XVI. We will assume, for the calculation of the confidence interval a 95% confidence level. Therefore, applying the equations shown above, we find,

TABLE XVI. MASS FRACTION OF ZN IN A CANDIDATE REFERENCE MATERIAL AS DETERMINED BY SEVERAL ANALYTICAL TECHNIQUES

Number of data	Analytical technique	Number of measurements	Mass fraction of Zn (mg/kg)
1	A	6	32.8
2	B	5	32.8
3	A	6	33.5
4	B	6	33.7
5	C	6	34.4
6	C	6	34.6
7	D	6	34.7
8	C	6	34.9
9	C	1	34.9
10	E	4	36.2
11	F	6	36.4
12	C	6	36.7
13	B	6	36.8
14	A	6	37.4
15	G	6	37.9
16	B	6	38.2
17	C	6	40.8
18	B	6	41.0
19	C	2	41.2
20	D	6	41.4

TABLE XVII. STATISTICAL PARAMETERS DESCRIBING THE DATA SET PRESENTED IN TABLE XVI

Statistical parameter	Value
Count	20
Average	36.5
Variance	7.9
Standard deviation	2.8
Range	8.6
Skewness	0.554916
Kurtosis	0.817833
Confidence interval for the mean	36.5 ∀ 1.3 [35.2–37.8]

Observe the values for the skewness and kurtosis, which can be used to determine whether the sample comes from a normal distribution. As mentioned, values of these statistics outside the range of -2 to +2 indicate significant departures from normality, which would tend to invalidate any statistical test regarding the standard deviation. In this case, both the skewness and the kurtosis have value within the range expected for data from a normal distribution.

The interpretation for the confidence interval is that, in repeated sampling, this interval will contain the true mean of the population from which the data come 95.0% of the time. In practical terms, we can state with 95.0% confidence that the true mean of the data is somewhere between 35.2 and 37.8. It is assumed that the population from which the sample comes can be represented by a normal distribution.

5.5. RELATIONSHIP BETWEEN TWO SETS OF DATA

In analytical chemistry it is essential to validate a given analytical method to determine its applicability, reproducibility, repeatability and the accuracy of the data obtained. The analyst should establish some basis to prove that the method is working for its intent use. Normally, the amount of data is rather small and the so-called ***Student t distribution*** should be used.

Large sample methods for making inferences concerning a population are not common in normal research projects or in the routine work of field or service laboratories. Cost, available time, and other factors limit the size of the sample that may be acquired. When this occurs, the large sample procedures described before are inadequate and other tests and estimation procedures must be employed. We will now study several small samples, and inferential procedures that are closely related to the large sample methods already presented. Specifically, we shall consider methods for estimating and testing hypotheses concerning population means, the difference between two means, a population variance, and a comparison of two population variances. These aspects are closely related to modern procedures in analytical chemistry such as method validation, comparison of experimental results to certified values in RMs and the determination of differences in the results when using two or more analytical method. Of particular importance, and highly related to the main topic of this document is the determination of important parameters to be determined in a RM or a QCM such as the homogeneity of the material as regards a given property or analyte mass fraction.

5.6. HYPOTHESIS TESTING

Setting up and testing hypotheses is an essential part of statistical inference. In order to formulate such a test, usually some theory has been put forward, either because it is assumed to be true or because it is to be used as a basis for argument, but has not been proved, for example, claiming that a new analytical method is better than the current one.

For each problem, the question is simplified into two competing claims/hypotheses between which one has to select: the so-called *null hypothesis*, denoted by H_o, against the *alternative hypothesis*, denoted by H_1. These two competing hypotheses are not, however, treated on an equal basis, but special consideration is given to the null hypothesis. We have two common situations:

(1) The experiment has been carried out in an attempt to disprove or reject a particular hypothesis, the null hypothesis, thus we give that one priority so it cannot be rejected unless the evidence against it is sufficiently strong.
(2) If one of the two hypotheses is "simpler" we give it priority so that a more "complicated"' theory is not adopted unless there is sufficient evidence against the simpler one.

The hypotheses are often statements about population parameters like expected value and variance. For example, H_o might be that the expected value of the concentration of ozone in the atmosphere of a village is not different from that in another town situated closely.

The outcome of a hypothesis test is "reject H_o"or "do not reject H_o".

The null hypothesis H_o represents a theory that has been put forward, either because it is believed to be true or because it is to be used as a basis for argument, but has not been proved. For example, in the development of a new analytical method, the null hypothesis might be that the new method is no better, on average, than the currently used. We would write H_o: there is no difference between the two methods on average.

We give special consideration to the null hypothesis. This is due to the fact that the null hypothesis relates to the statement being tested, whereas the alternative hypothesis relates to the statement to be accepted if / when the null is rejected.

The final conclusion once the test has been carried out is always given in terms of the null hypothesis. We either "reject H_o in favour of H_1" or "do not reject H_o". One never "rejects H_1", or even "accept H_1".

If we conclude "do not reject H_o", this does not necessarily mean that the null hypothesis is true, it only suggests that there is not sufficient evidence against H_o in favour of H_1; rejecting the null hypothesis then, suggests that the alternative hypothesis may be true.

5.6.1. Alternative hypothesis

The alternative hypothesis, H_1, is a statement of what a statistical hypothesis test is set up to establish. For example, if a new analytical method is being tested, the alternative hypothesis might be that the new method is different, on average, compared to the current one. We would write H_1: the two methods give different results, on average. The alternative hypothesis might also be that the new method is better (i.e. more accurate), on average, than the current one. In this case we would write H_1: the new method is better (i.e. more accurate) than the current one, on average.

In hypothesis testing, there are two types of errors one can make: *Type I* and *Type II* errors.

5.6.1.1. Type I error

In a hypothesis test, a *Type I* error occurs when the null hypothesis is rejected when it is in fact true; that is, H_o is wrongly rejected. A *Type I* error would occur if we concluded that the two analytical methods produced different results when in fact there was no difference between them.

A *Type I* error is often considered to be more serious, and therefore more important to avoid, than a *Type II* error. The hypothesis test procedure is therefore adjusted so that there is a guaranteed 'low' probability of rejecting the null hypothesis wrongly; this probability is never zero.

5.6.1.2. Type II error

In a hypothesis test, a *Type II* error occurs when the null hypothesis H_o, is not rejected when it is in fact false. For example, when developing a new analytical method, the null hypothesis might be that the new method is no better (e.g.., more accurate), on average, than the current one; that is H_o: there is no difference between the average of the two methods. A *Type II* error would occur if it was concluded that the two methods produce comparable results, that is, there is no difference between the average of the two methods, when in fact they produced different ones. A *Type II* error is frequently due to sample sizes being too small.

If we do not reject the null hypothesis, it may still be false (a *Type II* error) as the sample may not be big enough to identify the falseness of the null hypothesis (especially if the truth is very close to the hypothesis). For any given set of data, *Type I* and *Type II* errors are inversely related; the smaller the risk of one, the higher the risk of the other.

5.7. STUDENT'S *t* DISTRIBUTION

We introduce our topic by considering the following problem. An experiment has been carried out to evaluate a new analytical method to determine arsenic in seafood as part of the control for export products. The maximum allowed amount of As in the commodity is 0.5 mg/kg. Six independent determinations were made with the following results: 0.46, 0.61, 0.52, 0.48, 0.57, and 0.54 mg/kg. Do the six measurements present sufficient evidence to indicate that the average mass fraction exceeds the 0.5 mg/kg?

The distribution of

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}} \qquad (14)$$

for samples drawn from a normally distributed population was discovered by W.S. Gosset and published (1908) under the pen name of Student. He referred to the quantity under study as *t* and it has ever since been known as Student's t. We omit the complicated mathematical expression for the density function for *t* but describe some of its characteristics.

The distribution of the test statistic

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}} \qquad (15)$$

in repeated sampling is, like z, bell-shaped and perfectly symmetrical, about t=0. Unlike z, it is much more variable, tailing rapidly out to the right and left, a phenomenon that may readily be explained. The variability of *z* in repeated sampling is due solely to $\bar{x}$, the other quantities appearing in *z (n* and Φ) are non-random. On the other hand, the variability of *t* is contributed by two random quantities, $\bar{x}$ and *s*, which can be shown to be independent of one another. Thus when $\bar{x}$ is very large, *s* may be very small, and vice versa. As a result, *t* will be more variable than *z* in repeated sampling. Finally, as we might assume, the variability of *t* decreases as n increases because the estimate of *s*, will be based upon more and more a larger set of sample. When *n* is infinitely large, the *t* and *z* distributions will be identical. Thus, Gosset discovered that the distribution of *t* depended upon the sample size, *n*.

The divisor of the sum of squares of deviations (n-1), which appears in the formula for *s2* is called the number of degrees of freedom associated with *s2*. The origin of the term *"degrees of freedom"* is linked to the statistical theory underlying the probability distribution of *s2*. One may say that the test statistic *t* is based upon a sample of n measurements or that it possesses (n -1) degrees of freedom.

The critical values of t, which separate the rejection and acceptance regions for the statistical test are presented in Table XV. The tabulated value $t_{\forall}$, records the value of t in such a way that an area $\forall$ lies to its right. The degrees of freedom associated with s_2, d.f., are shown in the first and last columns of Table II, and the $t_{\forall}$, corresponding to various values of $\forall$, appear in the top row. Thus, if we wish to find the value of t, in such a way that 5% of the area lies to its right, we would use the column marked $t_{0.05}$. The critical value of t, for our example, is found in the $t_{0.05}$ column opposite to d.f. = $(n - 1) = (6 - 1) = 5$, is $t = 2.015$. Thus, we would reject H_o: = 0.5 when t>2.015.

The reason for choosing n = 30 as dividing line between large and small samples is apparent. For $n = 30$ (d.f. = 29), the critical value of $t_{0.05} = 1.699$ is numerically quite close to $z_{0.05} = 1.645$. For a two-tailed test based upon n = 30 measurements and $\forall = 0.05$, we would place 0.025 in each tail of the t distribution and reject H_o: : = :$_o$ when $t > 2.045$ or $t < -2.045$. Note that this is very close to the $z_{0.025} = 1.96$ employed in the z test.

It is important to note that the Student's t and corresponding tabulated critical values are based upon the assumption that the sampled population possesses a *normal probability distribution*, This indeed is a very restrictive assumption because, in many sampling situations, the properties of the population will be completely unknown and may well be non-normal (non-parametric). If this were to seriously effect the distribution of the t statistic, the application of the t test would be very limited. Fortunately, this point is of little consequence, as it can be shown that the distribution of the t statistic is relatively stable for populations not normally distributed, but possesses a bell-shaped probability distribution. This property of the t statistic and the common occurrence of bell-shaped distributions of data in nature, enhance the value of Student's t for use in statistical inference.

One would note that $\bar{x}$ and *s2* must be independent (in a probabilistic sense) in order that the quantity below (Equation 16) exhibit a t distribution in repeated sampling. As mentioned previously, this requirement will automatically be satisfied when the sample has been randomly drawn from a normal population.

$$\frac{\bar{x} - \mu}{s / \sqrt{n}} \qquad (16)$$

Having discussed the origin of Student's t and the tabulated critical values, Table XV, we now return to the problem of making an inference about the mean mass fraction of As in our seafood based upon our $n = 6$ measurements.

The statistical test of a hypothesis concerning a population mean may be stated as follows:

Test of a hypothesis concerning a population mean: H_o: : = :$_o$

Alternative hypothesis, H_1: specified by the experimenter depending upon the alternative values he wishes to detect.

Test statistic:

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}} \qquad (17)$$

To apply this test to the data, we must first calculate the sample mean, $\bar{x}$, and its standard deviation, s. This latter quantity is calculated using the formula explained before.

Mean $\bar{x} = 0.53$ (18)

Standard deviation, $s = 0.0559$ (19)

Remember that we wish to test the null hypothesis that the mean mass fraction of As is not significantly different from 0.5 mg/kg, against the alterative hypothesis that it is greater than 0.5. Then the elements of the test as defined above are:

$$H_o: := 0.5. \quad (20)$$

Test statistic:

$$t = (0.53 - 0.5)\ \%6 / 0{,}0559 = 1.31 \quad (21)$$

The rejection region for the H_o, for 0.05 and $(n - 1)$ (6 - 1) = 5 degrees of freedom is $t > 2.015$. The calculated value of the test statistic does not fall in the rejection region. Therefore, we do not reject H_o. This implies that the data do not present sufficient evidence to indicate that the mean mass fractions of As in the sample do not exceed 0.5 mg/kg.

Performing hypothesis tests, one found that there are to two types of approaches to the problem, depending on how it is presented: a one-sided test and a two-sided test.

5.7.1. One sided test

A one sided test is a statistical hypothesis test in which the values for which we can reject the null hypothesis, H_o, are located entirely in one tail of the probability distribution. In other words, the critical region for a one-sided test is the set of values smaller than the critical value of the test, or the set of values greater than the critical value of the test. A one sided test is also referred to as a one-tailed test of significance.

5.7.2. Two sided test

A two sided test is a statistical hypothesis test in which the values for which we can reject the null hypothesis, H_o, are located in both tails of the probability distribution. In other words, the critical region for a two-sided test is the set of values smaller than a first critical value of the test and the set of values greater than a second critical value of the test. A two sided test is also referred to as a two-tailed test of significance.

The choice between a one-sided test and a two-sided test is determined by the purpose of the investigation.

As an example, let us suppose we want to test a manufacturer claim that there is, on average, 50 matches in a box. We could set up the following hypotheses

$$H_o: := 50 \text{ against } H_1; : < 50 \text{ or } H_1;. : > 50 \quad (22)$$

Either of these two alternative hypotheses would lead to a one-sided test. Presumably, we would want to test the null hypothesis against the first alternative hypothesis since it would be useful to know if there is likely to be less than 50 matches, on average, in a box (no one would complain if they get the correct number of matches in a box or more).

Another alternative hypothesis could be tested against the same null hypothesis, leading this time to a two-sided test:

$$H_o: := 50 \text{ against } H_1: : \quad 50 \quad (23)$$

That is, nothing specific can be said about the average number of matches in a box; only that, if we could reject the null hypothesis in our test, we would know that the average number of matches in a box is likely to be less than or greater than 50.

Hypothesis testing can also be performed to establish the properties of a one given sample or to relate or compare two samples.

5.7.3. One sample t-test

A one sample t-test is a hypothesis test for answering questions about the mean where the data is a random sample of independent observations from a normally distributed population.

The null hypothesis for the one sample t-test is: H_o: : = :$_o$ (where :$_o$ known)

That is, the sample has been drawn from a population of a given mean and unknown variance (which therefore has to be estimated from the sample).

This null hypothesis, H_o is tested against one of the following alternative hypotheses, depending on the question posed:

- H1: : 0
- H1: : > :$_o$
- H1: : < :$_o$

5.7.4. Two sample t-test

A two sample t-test is a hypothesis test for answering questions about the mean where the data are collected from two random samples of independent observations, each from a normally distributed population.

When carrying out a two sample t-test, it is usual to assume that the variances for the two populations are equal, that is:

$$\Phi_1^2 = \Phi_2^2 \qquad (24)$$

The null hypothesis for the two samples t-test is:

$$H_o: :_1 = :_2 \qquad (25)$$

That is, the two samples have both been drawn from the same population.

This null hypothesis is tested against one of the following alternative hypotheses, depending on the question to be answered.

- H_1: : 0
- H_1: : > :$_o$
- H_1: : < :$_o$

To illustrate the above mentioned concepts here are some examples:

5.7.5. Testing the mean against a given value

This is the case when validating an analytical method or when comparing the results from a routine analytical method with the value established for that analyte in a RM or a QCM.

We will calculate *t* using the equation

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}} \qquad (26)$$

where $\bar{x}$ is the mean of your data, is the given (certified or reference) value, n is the number of measurements and s is the standard deviation of your data. Let's suppose that we have analysed an RM for Cu and found the following results: 10.5, 11.0, 10.0, 10.8 and 10.4 mg/kg. The mass fraction in the RM is 10.0 mg/kg. Our question would be whether there is evidence, at the 95% confidence level, of any significant difference between the mean and the reference value.

The procedure would be to calculate *t* from the above equation, then compare this value with the tabulated *t* value (at the chosen confidence level) and for (n-1) degrees of freedom. If the t_{calc} is lower that the *t* tabulated, we can accept the H_o, thus, there is no significant difference between both values.

H0: are the results coming from a population with mean = 10.00 mg/kg?

$=\bar{x} = 10.00$

$\bar{x} = 10.54$ mg/kg, s = 0.385, $t_{calc} = 2.105$

$t_{(tab,0.05,4)} = 2.78$, so $t_{calc} < t_{(tab,0.05,4)}$; therefore, there is no significant difference between the mean and the reference value.

5.7.6. Testing two means

In this example we will suppose that two samples have been analysed by the same method. We can test if the means are significantly different by a t-test

We will assume that the standard deviations of each set are not significantly different. Thus, H_o is that there is no significant difference between the means, i.e., the difference between the means should be zero

$$:_1 = :_2 \qquad (27)$$

Proceed, as follows:

(1) Calculate mean and standard deviation of each set
(2) Calculate pooled standard deviation using the following equation

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{(n_1 + n_2 - 2)} \qquad (28)$$

(3) Calculate t

$$t = \frac{\left|\bar{x}_1 - \bar{x}_2\right|}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad (29)$$

(4) Compare with $t_{(0.05,\,(n1+n2)\,-\,2)}$.

As an example, let's supposed that two replicates from one sample have been analysed by two methods and we want to test if the means are significantly different by a t-test:

With method A: mean = 28.0,

standard deviation = 0.3,

n = 10

With method B: mean = 26.25,

standard deviation = 0.23,

n = 10

The results are the following:

Pooled standard deviation: $s_p^2 = (9 \times 0.32 + 9 \times 0.232)/18$

$= 0.0715$

$s_p = 0.267$

$$t = \frac{(28.0 - 26.25)}{0.267\sqrt{\frac{1}{10} + \frac{1}{10}}} \qquad (30)$$

$t = 14.7$

$t_{(0.05, 18)} = 2.1$

Therefore, since $t_{calc} > t_{tab}$ the null hypothesis is rejected which means that the results from the two methods are significantly different

5.8. ANALYSIS OF VARIANCE

In general, the purpose of the analysis of variance (ANOVA) is to test for significant differences between means. If we are only comparing two means, then ANOVA will give the same results as the *t test* if we are comparing two different groups of cases or observations as already seen.

Why the name analysis of variance? It may seem odd to you that a procedure that compares means is called analysis of variance. However, this name is derived from the fact that in order to test for statistical significance between means, we are actually comparing (i.e., analysing) variances.

To understand better this statistical procedure we will explain some concepts and use a rather simple example and then we will go into more real analytical situations.

Why the name analysis of variance? It may seem funny to you that a procedure that compares means is called analysis of variance. However, this name is derived from the fact that in order to test for statistical significance between means, we are actually comparing (i.e., analysing) variances. ANOVA is based on the fact that variances can be divided up, that is, partitioned. Remember that the variance is computed as the sum of squared deviations from the overall mean, divided by n-1 (sample size minus one). Thus, given a certain n, the variance is a function of the sums of (deviation) squares, or SS for short. Partitioning of variance works as follows. Consider the following data set:

TABLE XVIII. DATA SET FOR EVALUATION OF THE VARIANCE OF RESULTS

	Group 1	Group 2
Data 1	2	6
Data 2	3	7
Data 3	1	5
Mean	2	6
Sum of squares (SS)	2	2
Overall mean	4	
Total sum of squares	28	

Note that the means for the two groups are quite different (2 and 6, respectively). The sums of squares within each group are equal to 2. Adding them together, we get 4. If we now repeat these computations, ignoring group membership – that is, if we compute the total SS based on the overall mean, we get the number 28. In other words, computing the variance (sums of squares) based on the within-group variability yields a much smaller estimate of variance than computing it based on the total variability (the overall mean). The reason for this, as in the above example, is of course the large difference between means, and it is this difference that accounts for the difference in the SS. In fact, if we were to perform an ANOVA on the above data, we would get the following result:

TABLE XIX. RESULTS OF ANOVA OF THE DATASET FROM TABLE XVIII

	Main effects				
	SS	df	MS	F	p
Effect	240	1	240	240	8
Error	40	4	10		

As one can see in the above table, the total SS (28) was partitioned into the SS due to within-group variability (2+2=4) and variability due to differences between means (28-(2+2)=24).

The within-group variability (SS), usually referred to as the ***error variance***, denotes the fact that we cannot readily explain or account for it in the current design. On the other hand, the SS effect is due to the differences in means between the groups.

Many statistical tests represent ratios of explained to unexplained variability. ANOVA is a good example of this. Here, we base this test on a comparison of the variances due to the between-groups variability (called Mean Square Effect) with the within-group variability (called Mean Square Error). Under the null hypothesis (that there are no mean differences between groups in the population), we would still expect some minor random fluctuation in the means for the two groups when taking small samples (as in our example). Therefore, under the null hypothesis, the variance estimated based on within-group variability should be about the same as the variance due to between-groups variability. We can compare those two estimates of variance via the F test, which tests whether the ratio of the two variance estimates is significantly greater than 1. In our example above, that test is highly significant, and we would in fact conclude that the means for the two groups are significantly different from each other.

In summary, the purpose of the ANOVA is to test differences in means (for groups or variables) for statistical significance. This is accomplished by analysing the variance, that is, by partitioning the total variance into the component that is due to true random error (i.e., within-group SS) and the

components that are due to differences between means. These latter variance components are then tested for statistical significance, and, if significant, we reject the null hypothesis of no differences between means, and accept the alternative hypothesis that the means (in the population) are different from each other.

From an analytical point of view, ANOVA can help us to answer questions such as which of many variables are important for a method or, if a linear relationship between variables is significant. For a round robin analysis between several laboratories, what is the inter-laboratory precision (reproducibility) and what the intra-laboratory precision (repeatability) or whether the inter-laboratory precision is significantly greater than the intra-laboratory precision. For our particular interest, and the subject of this TECDOC, ANOVA can help us to determine the degree of homogeneity of a RM or a QCM prepared in our laboratory. We will illustrate this case with a practical example. Before that, the other aspects should be explained.

Normally, one sets up a table with the values, which are going to be evaluated, in the form described below: different variables of interest in columns while replicates in rows.

TABLE XX. EXAMPLE OF A DATA TABLE FOR STATISTICAL EVALUATION

		Variables (k)		
		A	B	C
Replicates (n)	1	$X_{1,1}$	$X_{1,2}$	$X_{1,j}$
	2	$X_{2,1}$	$X_{2,2}$	$X_{2,j}$
	3	$X_{i,1}$	$X_{i,2}$	$X_{i,j}$

For the reader who wants to explore the basic mathematics and equations used for ANOVA he is referred to the list of books and other documents mentioned in the bibliography. Since there is a wide selection of tools for calculating statistics parameters, such as computer software, spreadsheets and electronic calculators, we will deal directly with practical examples.

The result of an ANOVA is presented in the form of Table XXI:

where SSc is the sum of squares due to the factor studied, also known as the treatment sum of squares, the heterogeneity sum of squares, or the between-column sum of squares and can be represented by the following equation:

$$SS_c = \sum_j n_j \frac{(x_{i,j} - \bar{x})}{n_j}^2 = \sum_j n_j(\bar{x}_j - \bar{x}) \qquad (31)$$

TABLE XXI. RESULTS OF ANOVA OF THE DATA FROM TABLE XX

Source	Sum of squares	Degrees of freedom	Mean squares	Expected mean squares
Between variables	SS_C	k-1	SS_C / (k-1)	$\Phi^2 + n_j\Phi_c^2$
Within variables	SS_R	N-k	SS_R / (N-k)	Φ^2
Total	SS_T	N-1		

In the same way, SST, the total sum of squares or the corrected sum of squares, can be represented by the equation:

$$SS_T = \sum_i \sum_j (x_{i,j} - \bar{x})^2 \tag{32}$$

and SS_R, the residual sum of squares or the within column sum of squares is (SS_T - SS_C).

Finally, to decide if there is a significant difference in the variance between columns (variables and samples), one has to use a one-tailed F-test, known also as the Fisher test. To calculate the F(0.05, k-1, N-k), the following equation is used:

$$F = \frac{s_2 + n_j s_c^2}{s^2} \tag{33}$$

and this value is compared to the critical value in the tables. The Fisher F distribution is used to compare ***variances*** for two sets of data with standard deviations s_1 and s_2. The F distribution is at a given probability level (e.g. 0.05 = 95%), and at the relevant number of degrees of freedom (ni – 1) for numerator and denominator. In this case, always the larger s2 goes in the numerator.

Let us suppose the following case: two mollusc samples from different harbours are analysed to determine their content of Fe. The results are the following:

TABLE XXII. FE CONCENTRATIONS DETERMINED IN MOLLUSC SAMPLES FROM TWO DIFFERENT HARBOURS

Sample A	Sample B
49	44
44	57
70	34
50	48
58	50

The question here is whether the mass fraction of Fe in the sample A is significantly different from that in sample B.

After calculating the respective parameters, the ANOVA Table will have this data:

TABLE XXIII. RESULTS OF ANOVA OF THE DATA PRESENTED IN TABLE XXII

Source	Sum of squares	Degrees of freedom	Mean squares	Expected mean squares
Between variables	144.4	1	144.4	$\Phi^2 + 5\Phi_c^2$
Within variables	700	8	87.5	Φ^2
TOTAL	844.4	9		

Then,

$F = (144.40 / 87.50) = 1.65$

From the table, $F(0.05, 1,.8) = 5.3$, thus, $F < F_{table}$, therefore the difference is not significant at 95% probability.

As mentioned earlier, there are several computer softwares that can perform an ANOVA and give the respective results. As an example, we present here such outputs. It is important to notice that, in this case, the ANOVA Table also informs the ***p-value,*** which gives us additional information regarding the relationship between the means.

TABLE XXIV. ANALYSIS OF VARIANCE

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Between groups	144.4	1	144.4	1.65	0.2349
Within groups	700.0	8	87.5		
Total (Corr.)	844.4	9			

As mentioned, the F ratio, which in this case equals 1.65, is a ratio of the between-group estimate to the within-group estimate. Since the ***p-value*** of the F test is greater than or equal to 0.05, there is not a statistically significant difference between the mean Fe mass fractions from one sample to another at the 95.0% confidence level.

This particular statistics is normally used to determine the homogeneity of a sample, a material candidate to RM or to QCM for chemical analysis. In this case, a well-established procedure during the preparation of those materials includes the fractionation of the bulk sample or material into units of a reasonable amount of mass. Usually, such units contain about 20–25 grams of the material. It is necessary to prove that the properties being measured (i.e., the mass fraction or concentration) of one (or several) analyte(s) is not significantly different between the units or within the units for a given amount of mass. The use of ANOVA will give the answer to this question.

Suppose we have a material candidate to RM or QCM that has to be tested for homogeneity for Cd. From the total of units available we select 17 and analyse them, taking 6 replicates from each unit. The results of the analysis are shown in the following table (values are in mg/kg).

TABLE XXV. RESULTS OF CD ANALYSIS IN A CANDIDATE RM IN μG/G

Unit/ measurement	U1	U2	U3	U4	U5	U6	U7	U8	U9	U10	U11	U12	U13	U14	U15	U16	U17
1	0.59	0.48	0.53	0.54	0.54	0.47	0.45	0.5	0.54	0.49	0.58	0.49	0.46	0.59	0.5	0.5	0.48
2	0.47	0.58	0.42	0.56	0.46	0.55	0.46	0.49	0.56	0.56	0.57	0.54	0.49	0.47	0.57	0.53	0.54
3	0.53	0.47	0.5	0.51	0.57	0.55	0.57	0.54	0.53	0.54	0.58	0.49	0.46	0.46	0.54	0.49	0.46
4	0.51	0.49	0.48	0.43	0.45	0.48	0.44	0.45	0.42	0.49	0.47	0.49	0.42	0.51	0.42	0.43	0.48
5	0.44	0.54	0.46	0.44	0.51	0.5	0.51	0.46	0.47	0.46	0.48	0.49	0.49	0.48	0.47	0.47	0.49
6	0.46	0.42	0.49	0.49	0.47	0.44	0.48	0.49	0.47	0.48	0.5	0.5	0.54	0.48	0.46	0.49	0.55

Applying ANOVA, the ANOVA Table obtained is the following:

TABLE XXVI. ANALYSIS OF VARIANCE

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Between groups	0.0133706	16	0.000835662	0.41	0.9770
Within groups	0.174117	85	0.00204843		
Total (Corr.)	0.187487	101			

The F ratio, which in this case equals 0.41, is a ratio of the between-group estimate to the within-group estimate. Since the p-value of the F test is greater than or equal to 0.05, there is not a statistically significant difference between the means from one unit to another at the 95.0% confidence level. Therefore, one can conclude that Cd is homogeneous in the material when using a minimum amount of material as established by the producer.

If there is a significant difference between the means as detected by ANOVA, it is not possible to determine, with this information, which is (are) the mean(s) that differ from the rest. Additional tests are then necessary to identify these results.

5.9. QUALITY CONTROL CHARTS

In all production processes, it is necessary to monitor the extent to which the products meet specifications. In the most general terms, there are two opponents to product quality: (1) deviations from target specifications, and (2) excessive variability around target specifications. During the earlier stages of developing the production process, designed experiments are often used to optimize these two quality characteristics. The methods provided in Quality Control are on-line or in-process quality control procedures to monitor an ongoing production process.

The general approach to on-line quality control is straightforward: We simply extract samples of a certain size from the ongoing production process. We then produce line charts of the variability in those samples, and consider their closeness to target specifications. If a trend emerges in those lines, or if samples fall outside pre-specified limits, then we declare the process to be out of control and take action to find the cause of the problem. These types of charts are sometimes also referred to as Shewhart control charts (named after W.A. Shewhart who is generally credited for being the first to introduce these methods).

This procedure has been extended and applied to analytical chemistry and the control of the "production" of data in the laboratory. The principle is the same as described before, but instead of taking samples from the production process, one plots the results of the determination of a given analyte in a specific sample. This practice helps the analytical chemist to determine whether there arise unexpected problems with his analytical procedure and to detect the presence of systematic errors. Result outside the predetermined warning or action limits imply immediate review of the complete methodology and correction for any problem found.

In the chart shown above, a Shewhart or X chart, the horizontal axis represents the results obtained when analysing a given sample at time intervals. The vertical axis represents the content (individual or mean mass fraction or concentration) of the analyte of interest. A typical chart includes four additional horizontal lines to represent the upper and lower warning limits (UWL, LWL, respectively) and the upper and lower action limits (UAL, LAL, respectively).

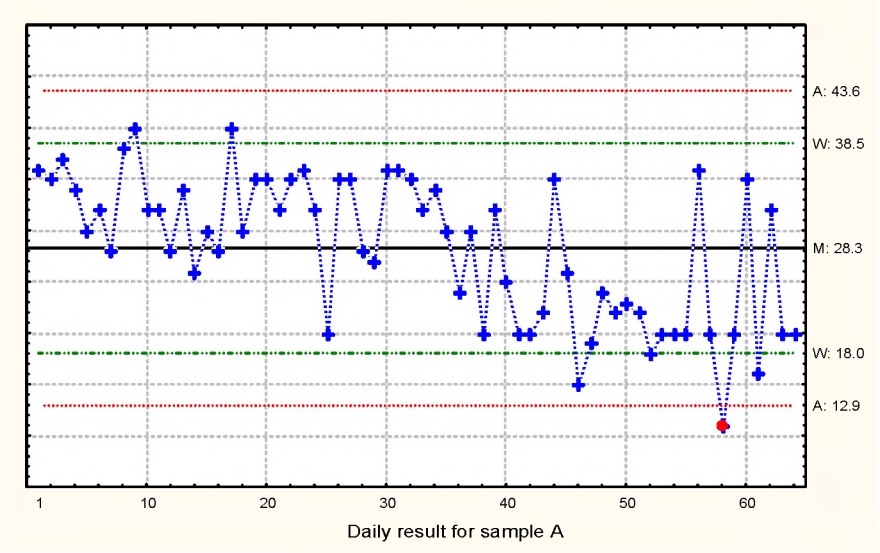

FIG. 7. Example of a Shewhart or X quality control.

Typically, the individual points in the chart, representing the results for the analyte, are connected by a line. If this line moves outside the upper or lower control limits or exhibits systematic patterns across consecutive samples, then a quality problem may potentially exist. Even though one could arbitrarily determine when to declare a process out of control (that is, outside the UWL-LWL range), it is common practice to set this limits at a 2s (two standard deviations) from the central line. The UAL and LAL are set at 3s (three standard deviations) from the mean (central) line.

Results falling outside the control limits are only one indication of a measurement out of control. Another indication occurs when the values fall into some sort of pattern over time. That is, an analysis in control should result in random errors about the centre line; non-random errors indicate that assignable-cause variability may exist. There are a variety of rules to use when looking for non-randomness, four of which are given here.

(1) Two of three observations in a row beyond two sigma
(2) Eight consecutive observations above or eight consecutive observations below the centre line
(3) Seven observations in a row up or down
(4) Four out of five beyond one sigma.

The Shewhart chart provides a way of monitoring the results from the analysis, but it does not monitor the variability of it. Sometimes the chart will indicate that the analysis appears lo be under control, but the variability of it is not in control. More variability means the analysis is not under control because of assignable causes. Because Shewhart charts are designed to monitor the process and not the process variability, an additional control measure is necessary.

In most quality control applications, variability is measured using the range of the items in each sample. Recall that the range is the difference between the highest and lowest values in a sample. The use of the range to measure variability in quality control is partly statistical and partly historical. The statistical part stems from an advantage in the estimation process, especially for small sample sizes. The historic part is a result of the fact that, when statistical process control first originated, it was much easier for QC employees to calculate (and understand) a range rather than a standard deviation.

Small ranges suggest a small variation from results to results. That is, the analysis output is similar from item to item. A large range indicates sample items that tend to differ from one another. Thus, small values for the range of an analysis are desirable, as long as the process is under control.

Variability is monitored with a range chart, which is abbreviated as R chart. An R chart provides a plot, through time, of the range of the observations at a point in time. Even for a process where the Shewhart chart appears to be in control, an R chart may indicate that a process is not in control. As with means, a process in control will result in ranges that fall in a random pattern, within three-sigma limits. Thus, points outside the three-sigma limits and non-random points on an R chart indicate an analysis that appears to be out of control.

The patterns to look for in an R chart are the same as those in a Shewhart chart, except one no longer counts the number of observations outside one or two sigmas (because the R values cannot be assumed to be normally distributed).

Conceptually, an R chart is similar to a Shewhart chart. The average range ($\bar{R}$) is the centreline. This is the average of all the ranges. The estimated standard deviation of the range is designated as R, The three-sigma limits are then calculated and designated as UCLR, and LCLR. It may help to think of these control limits as follows:

$$UCLR = R + 3\,R$$

$$LCLR = R - 3\,R$$

5.10. COMPUTERS, SOFTWARE AND STATISTICS

Computers are essential to many daily activities. The analytical laboratory is not an exception and, normally, there is one such machine available. Computers can be used for many things and, of course, for calculations, plotting, drawing, word processing and, not to say, games.

Computers can carry out large number of calculations in few seconds. They can process massive amounts of data and do with them whatever the operator wants. In statistics, they are very helpful since they can perform many useful calculations provided they are loaded with the appropriate software.

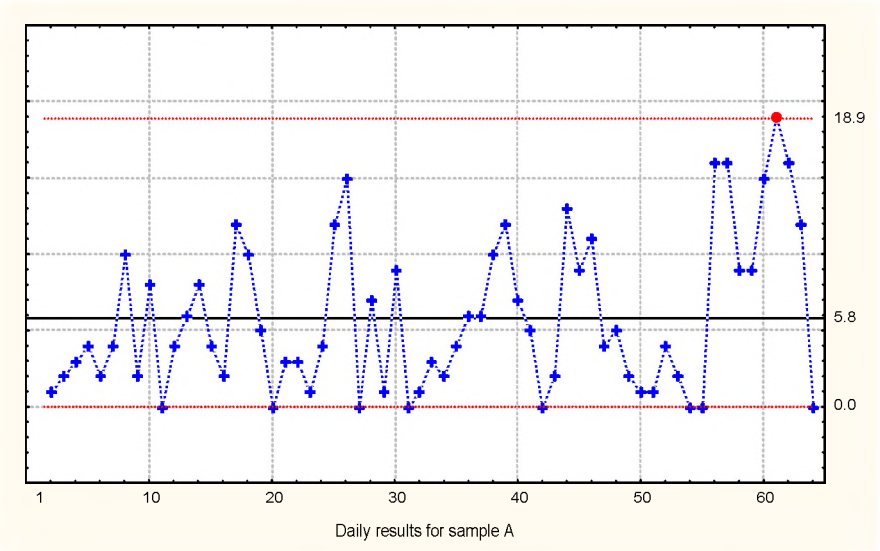

FIG. 8. Example of an R (range) quality control chart.

There are essentially two types of software for the analytical chemist to deal with matters like the one treated in this contribution. One group, although limited in number, is formed by programmes dedicated specifically for purposes such as method validation, curve calibrations, proficiency testing, control charts, etc. Most of these programmes are commercially available from laboratories or institutions devoted to quality control and collaborative tests and their prices are relatively high, in the range of several thousand US dollars.

The second group constitutes programmes that are mathematically (statistically) oriented. This type of software normally includes a number of mathematical (statistical) procedures, and allows the calculation of many statistical parameters. These programmes are also available at a relative high cost and sold by modules, depending on the interest of the customer. They have a "standard" module that permits calculation of the most common parameters. However, other applications, such as experimental design, quality control, advance regression, multivariate methods, time series, etc. come in separate modules, each one has a separate price. Despite the cost of these softwares, it is worthwhile to have at least one of them since they can provide almost all statistical information needed for the analytical laboratory and for data processing and evaluation.

A good, although limited, alternative is the so-called spreadsheet software. The most popular of them are included in software packages oriented to office applications. Present versions of these spreadsheets allow the determination of several very useful statistical parameters for the analytical chemist: descriptive statistics, single and two factors ANOVA, correlation, covariance, regression, F, t and z tests, among others.

5.11. CONCLUSIONS

Present requirements for the analytical laboratory are far more demanding than some years ago when it was enough to submit a number as the result of an analysis. Today, international guides and regulatory issues request not only a figure from the laboratory but additional information. It is necessary to demonstrate the validity and applicability of the analytical method being used for the intended purpose, to give proofs of quality control and quality assurances procedures and to give an estimation of the uncertainty associated to the measurement. All this information needs the use of appropriate statistical tools.

Statistics is used for the design of sampling strategies, to determine a calibration curve for our instruments, to find consensus values when performing intercomparison round robins or collaborative tests, to estimate the dispersion of a series of measurements, to determine if a given process is under statistical control, to demonstrate the traceability of the measurements, to compare data from different samples or from different analytical methods, and, most important nowadays, to estimate the uncertainty associated to the measurements.

To correctly apply statistical procedures, however, there are a few considerations to be taken into account: Answers to questions such as — do we know what we want to do?, do we know how to do them?, do we have enough information about (the distribution) of our data?, and can we interpret correctly the output of these calculations? — have to be clearly established before proceeding with the final and definitive calculations and drawing conclusions.

REFERENCES

[1] INTERNATIONAL ORGANIZATION FOR STANDARDIZATION, ISO 8402:1994, Quality Management and Quality Assurance - Vocabulary, Geneva (1994).

[2] INTERNATIONAL ORGANIZATION FOR STANDARDIZATION, ISO/IEC 17025:1999 (E), General Requirements for Competence of Testing and Calibration Laboratories (formerly ISO Guide 25), Geneva (1999).

[3] IUPAC/ISO/AOAC, Harmonized Guidelines for Internal Quality Control in Analytical Chemistry Laboratory, prepared for publication by R. Wood and M. Thompson, Pure & Appl. Chem. **67(4)** (1995) 649–666.

[4] INTERNATIONAL ORGANIZATION FOR STANDARDIZATION, ISO Guide 30, Terms and Definitions used in connection with Reference Materials, 2nd Edition, Geneva, Switzerland (1992).

[5] INTERNATIONAL ORGANIZATION FOR STANDARDIZATION, ISO Guide 33, Uses of Certified Reference Materials, 2nd Edition, Geneva, Switzerland (2000).

[6] STEGER, H.F., "Uses of matrix reference materials", The Use of Matrix Reference Materials in Environmental Analytical Process, the Royal Society of Chemistry Special Publication No. 238 (FAJGELJ, A., PARKANY, M., Eds.) Cambridge (1999).

[7] BIPM, IEC, IFCC, ISO, IUPAC, IUPAP, OIML, Guide to the Expression of Uncertainty in Measurement (GUM), Geneva (1995).

[8] THE ROYAL SOCIETY OF CHEMISTRY, The Use of Matrix Reference Materials in Environmental Analytical Process, (FAJGELJ, A., PARKANY, M., Eds.), Special Publication No. 238, Cambridge (1999).

[9] WILEY-VCH, Reference Materials for Chemical Analysis, (STOEPPLER, M., WOLF, V., JENKS, P.J., Eds.) Wienheim (2001).

[10] EURACHEM, The Fitness for Purpose of Analytical Methods - A Laboratory Guide to Method Validation and Related Topics, LGS Teddington, available for free download from the Internet address: http://www.eurachem.bam.de/ (1998).

[11] EURACHEM, Quantifying Uncertainty in Analytical Measurement, EURACHEM/CITAC Guide, 2nd Edition, EURACHEM QUAM:200.P1, available for free download from the Internet address: http://www.eurachem.bam.de/ .

[12] AMERICAN SOCIETY FOR TESTING AND MATERIALS, ASTM Manual on Presentation of Data and Control Chart Analysis, 6th Edition, Manual Series: MNL 7, Baltimore (1991).

[13] IHNAT, M., "Biological and related reference materials for determination of elements", Quantitative Trace Analysis of Biological Materials, (MCKENZIE, H.A., SMYTHE, L.E., Eds.), Elsevier, Amsterdam (1988) Appendix 1, 739–760.

[14] IHNAT, M., "Biological reference materials for quality control", Quantitative Trace Analysis of Biological Materials, (MCKENZIE, H.A., SMYTHE, L.E., Eds.), Elsevier, Amsterdam (1988) Chapter 19, 331–351.

[15] CALI, J.P., et al., The Role of Standard Reference Materials in Measurement Systems, National Bureau of Standards Monograph 148, Washington, D.C. (1975) 51 pp.

[16] HUNTOON, R.D., "Standard reference materials and meaningful measurements, an overview", Standard Reference Materials and Meaningful Measurements, (SEWARD, R.W., Ed.), National Bureau of Standards Spec. Publ. 408, Washington, DC (1975) 4–56.

[17] URIANO, G.A., GRAVATT, C.C., "The role of reference materials and reference methods in chemical analysis", CRC Crit. Rev. Anal. Chem. **6** (1977) 361–411.

[18] IHNAT, M., "Proposals for the use of reference materials and for the development of in-house QCMs for food analysis", (Proc. Consultants Meeting on Proper Use of Reference and Control Materials, IAEA, Vienna, Austria, 2001) 1–31.

[19] NATIONAL INSTITUTE FOR ENVIRONMENTAL STUDIES, Preparation, Analysis and Certification of Pepperbush Standard Reference Material, (OKAMOTO, K., Ed.) Research Report No. 18, Tsukuba, Ibaraki, Japan; also individual information sheets on other NIES materials (1980).

[20] ROSSBACH, M., SCHLADOT, J.D., OSTAPCZUK P., Specimen Banking, Environmental Monitoring and Modern Analytical Approaches, Springer-Verlag, Berlin (1992) 242 pp.
[21] QUEVAUVILLER, Ph., MAIER, E., KRAMER, K.J.M., Production of Certified Reference Materials for Pollutants in Environmental Matrices, European Commission Report EUR 18157, European Commission, Brussels and CCF Academic Press, Tarbes (1998) 251 pp.
[22] WILEY-VCH, Reference Materials for Chemical Analysis - Certification, Availability and Proper Usage, (STOEPPLER, M., WOLF, W.R., JENKS, P.J., Eds.), Weinheim (2001) 297 pp.
[23] IHNAT, M., "Development of a new series of agricultural/food reference materials for analytical quality control of elemental determinations", J. AOAC INTL. **77** [E 3700] (1994) 1605–1627.
[24] IHNAT, M., "Twenty five years of reference material activity at Agriculture and Agri-Food Canada", Fresenius J. Anal. Chem. **370** (2001) 279–285.
[25] EUROPEAN COMMISSION, STANDARDS, MEASUREMENT AND TESTING PROGRAMME (BCR), Guidelines for the Production and Certification of BCR Reference Materials, Doc. BCR/48/93, Brussels (1994) 54 pp.
[26] INTERNATIONAL ORGANIZATION FOR STANDARDIZATION, Quality System Guidelines for the Production of Reference Materials, Guide 34-1996, Geneva (1996).
[27] INTERNATIONAL ORGANIZATION FOR STANDARDIZATION, Certification of Reference Materials - General and Statistical Principles, Guide 35-1989, Geneva (1989).
[28] IHNAT, M., "Selection and preparation of relevant reference materials for agricultural purposes", Specimen Banking - Environmental Monitoring and Modern Analytical Approaches, (ROSSBACH, M., SCHLADOT, J.D., OSTAPCZUK, P., Eds), Chapter 4.1, Springer-Verlag, Berlin (1992) 57–73.
[29] IHNAT, M., "A synopsis of different approaches to the certification of reference materials", Fresenius J. Anal. Chem. **360** (1998) 308–311.
[30] IHNAT, M., "Certification philosophy of RM producers", Reference Materials for Chemical Analysis - Certification, Availability and Proper Usage, (STOEPPLER, M., WOLF, W.R., JENKS, P.J., Eds.), Wiley-VCH, Weinheim, Section 3.1, Chp. 3 (2001) 49–60.
[31] IHNAT, M., Performance of neutron activation analytical methods in an international interlaboratory reference material characterization campaign, J. Radioanal. Nucl. Chem. **245** (2000) 73–80.
[32] IHNAT, M., et al., "Certification of elements", Reference Materials for Chemical Analysis - Certification, Availability and Proper Usage, (STOEPPLER, M., WOLF, W.R., JENKS, P.J., Eds.), Wiley-VCH, Weinheim, Section 3.2, Chapter 3 (2001) 60–75.
[33] INTERNATIONAL ORGANIZATION FOR STANDARDIZATION, VIM, International Vocabulary of Basic and General Terms in Metrology, 2nd Ed., Geneva.
[34] ABBEY, S., US Geological Survey standards - a critical study of published analytical data, Can. Spectrosc. **15** (1970) 10–16.
[35] ABBEY, S., "Standard samples: how standard are they?" Geostand. Newsl., **1** (1977) 39–45.
[36] ABBEY, S., "Studies in standard samples of silicate rocks and minerals", 1969–1982, Pap.-Geol. Surv., Can. Paper (1983) 83–15.
[37] ABBEY, S., "Studies in 'standard samples' for use in the general analysis of silicate rocks and minerals" Part 6:1979 Edition of Usable Values, Pap.-Geol. Surv. Can. Paper 80–14 (1980).
[38] TAYLOR, B.N., KUYATT, C.E., Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results, NIST Tech. Note 1297, 1994 Edition, National Institute of Standards and Technology, Gaithersburg, MD (1994).
[39] INTERNATIONAL ORGANIZATION FOR STANDARDIZATION, Contents of Certificates of Reference Materials, ISO Guide 31-1981, Geneva (1981).
[40] CALI, J.P., REED, W.P., "The role of the National Bureau of Standards, standard reference materials in accurate trace analysis", Accuracy in Trace Analysis: Sampling, Sample Handling and Analysis (LAFLEUR, Ed.), NBS Special Publication 422, Vol. **1**, National Bureau of Standards, Washington, D.C. (1976) 41–63.

[41] CALI, J.P., et al., "A referee method for the determination of calcium in serum standard reference materials", NBS Spec. Publ. 260-36, National Bureau of Standards, Washington, D.C. (1972).
[42] FLANAGAN, F.J., "Reference samples for the earth sciences", Geochem. Cosmochim. Acta **38** (1974) 1731–1744.
[43] GLADNEY, E.S., et al., "Elemental concentrations in the United States Geological Survey's geochemical exploration reference samples - a review", Anal. Chem. **51** (1979) 1557–1569.
[44] PAAR, R.M., "The reliability of trace element analysis as revealed by analytical reference materials". Trace Element Analytical Chemistry in Medicine and Biology, (BRATTER, P., SCHRAMEL, P., Eds.), Walter de Gruyter & Co., New York (1980) 631–655.
[45] VON LEHMDEN, D.J., JUNGERS, R.H., LEE, R.E., JR., "Determination of trace elements in coal, fly ash, fuel oil, and gasoline - a preliminary comparison of selected analytical techniques", Anal. Chem. **46** (1974) 239–345.
[46] INGAMELLS, C.O., "Standard reference materials in geoexploration and extractive metallurgy research", Geoanalysis '78 Symposium, Ottawa, 1978, from ABBEY, S., "Studies in standard samples of silicate rocks and minerals", 1969–1982, Pap.-Geol. Surv. Can. Paper 83–15 (1983).
[47] IHNAT, M., WOLYNETZ, M.S., "An interlaboratory characterization (certification) campaign to establish the elemental composition of a new series of agricultural/food reference materials", Fresenius' J. Anal. Chem. **348** (1994) 452–458.
[48] IHNAT, M., "Characterization (certification) of three wheat flours and a wheat gluten reference material (NIST RM 8436, 8437, 8438 and 8418) for essential and toxic major, minor and trace element constituents", Fresenius' J. Anal. Chem. **348** (1994) 468–473.
[49] IHNAT, M., Biological Reference Materials for Quality Control of Elemental Composition Analytical Data. J. Radioanal. Nucl. Chem. **245** (2000) 65–72.
[50] IHNAT, M., Report of Investigation, Reference Material WG 184 (NIST RM 8418) Wheat Gluten, Centre for Land and Biological Resources Research, Agriculture Canada, Ottawa, ON. (1993) 15 pp; Report of Investigation, Reference Material 8418 Wheat Gluten, National Institute of Standards and Technology, Gaithersburg, MD (1994, 1999).
[51] IHNAT, M., Technical Report on the Development of Agricultural/Food Reference Materials, National Bureau of Standards, Gathresburg, MD (in preparation).
[52] IHNAT, M., Report of Investigation, Reference Material 8418 Wheat Gluten, also other such rep+orts on Reference Materials 8412, 8413, 8414, 8415, 8416, 8432, 8433, 8435, 8436, 8437 and 8438, with updates, National Institute of Standards and Technology, Gaithersburg, MD (1993).
[53] DABEKA, R.W., IHNAT, M., Considerations in the estimation of recovery in inorganic analysis, (PARKANY, M., Ed.), The Use of Recovery Factors in Trace Analysis, (Proc. Seventh International Harmonization Symposium on a Protocol for Recovery Factors, 1996 Sept. 4–5, Orlando, FL), Special Publication No. 194, Royal Society of Chemistry, Cambridge, UK (1996) 5–23.
[54] IHNAT, M., Pick a Number - Analytical Data Reliability and Biological Reference Materials. Sci. Total Environ. **71** (1988) 85–103.
[55] MARGOSIS, M., HORWITZ, W., ALBERT, R., J. Assoc. Offic. Anal. Chem. **71** (1988) 619.
[56] DABEKA, R.W., HAYWARD, S., Quality Assurance for Analytical Laboratories (PARKANY, M., Ed.), Royal Society of Chemistry, London (1993) 67.
[57] HORWITZ, W., KAMPS, L.R., BOYER, K.W., "Quality assurance in the analysis of foods for trace constituents", J. Asso. of Anal. Chem., **63**, No. 6 (1980).
[58] ROYAL SOCIETY OF CHEMISTRY, Guidelines for Single-Laboratory Validation of Methods for Trace-Level Organic Chemicals, (FAJGELJ, A., AMBRUS, A., Eds.), AOAC/FAO/ IAEA/IUPAC Principles and Practices of Method Validation (2000).
[59] FAO/WHO Codex Alimentarius Commission, Pesticide Residues in Food Sampling and Analysis Methods, 2nd Edition (2000).
[60] EUROPEAN COMMISSION, Directorate General for General Health and Consumer Protection, "Guidance document on residue analytical methods" (2000) 16 pp.

[61] INTERNATIONAL ORGANIZATION FOR STANDARDIZATION, ISO/DIS 11095, Linear Calibration using Reference Materials, Geneva (1993).
[62] PODHORNIAK, L.V., LEAKE, S., SCHENCK, F.J., "Stability of tetracycline antibiotics in raw milk under laboratory storage conditions", J. Food Prot. **62** (1999) 547–8.
[63] EL-BIDAOUI, et al., Stability of Pesticide Residues during Sample Processing and Storage. FAO/IAEA/IUPAC/AOAC Principles and Practices of Method Validation, (FAJGELJ, A., AMBRUS, A., Eds.), Royal Society of Chemistry (2000).
[64] VERDON, E., et al., "Stability of penicillin antibiotic residues in meat during storage", J. of Chromatography A, 882 (2000) 135–143.
[65] ZEISLER, R., et al., "The preparation of a cabbage candidate reference material to be certified for residues of agrochemicals", Fresenius J. Anal. Chem. **345** (1993) 202–206.
[66] IHNAT, M., "Development of a new series of agricultural/food reference materials for analytical quality control of elemental determinations", J. of AOAC International **77** No. 6 (1994).
[67] HILL, A.R.C., HARRIS, C.A., WARBURTON, A.G., Effects Of Sample Processing on Pesticide Residues in Fruit and Vegetables, AOAC/FAO/IAEA/IUPAC Principles and Practices of Methods Validation, (FAJGELJ, A., AMBRUS, A., Eds.), Royal Society of Chemistry (2000).
[68] HEISE, S., WEBER, H., ADLER, L., "Reasons for the decomposition of the fungicide thiram during preparation of fruit and vegetable samples and consequences for residue analysis", Fresenius J. Anal. Chem. **366** (2000) 851–856.
[69] LACORTE, S., EHRESMANN, N., BARCELO, D., "Stability of organophosphorus pesticides on disposable solid-phase extraction precolumns", Environ. Sci. Technol. **29** (1995) 2834–2841.
[70] SABIK, H., JEANNOT, R., "Stability of organophosphorus insecticides on graphitized carbon black extraction cartridges used for large volumes of surface water", J. of Chromatography A **879** (2000) 73–82.
[71] LISKA, I., BILIKOVA, K, "Stability of polar pesticides on disposable solid-phase extraction precolumns", J. of Chromatography A **795** (1998) 61–69.
[72] JOHNSON, W.G., LAVY, T.L., SENSEMAN, S.A., "Stability of selected pesticides on solid extraction disks", J. Env. Qual. **23** (1994) 1027–1031.
[73] SENSEMAN, S.A., et al., "Stability of various pesticides on membranous solid-phase extraction media", Environ. Sci. Technol. **27** (1993) 516–519.
[74] DE KOK, A., et al., "The use of validated standard mixtures in pesticide multi-residue analysis", Poster 8E-013, 9th International Congress of Pesticide Chemistry (IUPAC), London (1998).
[75] VISCONTI, A., et al., "Stability of fumonisins (FB1 and FB2) in solution", Food Addit. Contam. **11**:427–31.
[76] FERRER, I., BARCELO, D., "Determination and stability of pesticides in freeze-dried water samples by automated on-line solid-phase extraction followed by liquid chromatography with diode-array detection", Journal of Chromatography A, **737** (1996) 93–99.
[77] MARTIN-ESTEBAN, A., et al., "The preparation of certified reference material of polar pesticides in freeze-dried water (CRM 606)", Fresenius J. Anal. Chem. **363** (1999) 632–640.
[78] ATHANASOPOULOS, P., KYRIAKIDIS, N.B., GEORGITSANAKOU, I., "Effect of storage temperature on the degradation of dimethoate fortified oranges and peach juices", J. Agric. Food Chem. **48** (2000) 4896–4899.
[79] EL-BIDAOUI, M., AMBRUS, A., POULIQUEN, I., "Stability of pesticide residues during storage, Poster, 2nd International Symposium of the MGPR, Valencia, Spain (2001).
[80] EGLI, H., "Storage stability of pesticide residues", J. Agric. Food Chem. **30** (1982) 861–866.
[81] KOCOUREK, V., HAJSLOVA, J., HOLADOVA, K., POUTSKA, J., "Stability of pesticides in plant extracts used as calibrants in the gas chromatographic analysis of residues", J. of Chromatography A, **800** (1998) 297–304.

[82] BERNAL, J.L., DEL NOZAL, M.J., JIMENEZ, J.J., "Influence of solvent and storage conditions on the stability of acaricide standard stock solutions", J. of Chromatography A, **765** (1997) 109–114.
[83] TERENIUS, O., AKERBLOM, M., "Evaporated extracts of samples for pesticide residue analysis simplifies transport from remote places", Bull Env. Cont Tox. (1996).
[84] MUNCH, D.J., FREBLS, C.P., "Analyte stability conducted during the National Pesticide Survey", Environ. Sci. Technol. **26** (1992) 921–925.
[85] BECKMAN, H., BRUCE, R., MACDOUGALL, D., Analytical Methods for Pesticides, Plant Growth Regulators and Food Additives, Vol. **I**, Chapter 8, Spectrophotometric Methods, (ZWEIG. G., Ed.) (1963).
[86] GEISSBÜHLER, H., HASELBACH, C., "On the behaviour of DDVP upon storage and processing of insecticide-treated cereals", Report to CIBA Ltd, Batelle Memorial Institute, Geneva (1963).
[87] GUNTHER, F.A., "Insecticide residues in California citrus fruits and products", Residue Reviews **28** 1 (1969).
[88] KOIVISTOINEN, P., ROINE, P., "Occurrence and disappearance of parathion and malathion residues in vegetables and fruits", Maataloustieteellinin Aikakauskirja **31** 294 (1959), from Chem. Abstr. **54** 15751d (1960).
[89] LAMB, F.C., et. al., "Removal of DDT, parathion and carbaryl from spinach by commercial and home preparative methods", J. Agr. Food Chem. **16** 967 (1968).
[90] BECKMAN, H., THORNBURG, W., "Effect of frozen storage on parathion residues", J. Food Sci. **30** 656 (1965).
[91] LINSINGER, T.P., et al., "Homogeneity and stability of reference materials", Accred. Qual. Assur. **6** (2001) 20–25.
[92] THOMPSON, M., WOOD, R., Harmonized Guidelines for Internal Quality Control in Analytical Chemistry Laboratories (Technical report), Pure & Appl. Chem., Vol. **67** No. 4, IUPAC (1995) 649–666.
[93] BEYER, W.H.. CRC Handbook of Tables for Probability and Statistics, CRC Press, Boca Raton, Florida (1996).
[94] MAESTRONI, B., et al., Testing the Efficiency and Uncertainty of Sample Processing Using ^{14}C-Labelled Chlorpyrifos, Parts I&II, (FAJGELJ, A., AMBRUS, V., Eds.) AOAC/FAO/IAEA/ IUPAC Principles and Practices of Methods Validation, Royal Society of Chemistry (2000).
[95] INTERNATIONAL ORGANIZATION FOR STANDARDIZATION, ISO Guide 30, Terms and Definitions used in connection with Reference Materials, 2nd Edition, Geneva, Switzerland (1992).
[96] STEGER, H.F., "Uses of matrix reference materials", The Use of Matrix Reference Materials in Environmental Analytical Process, Special Publication No. 238, (FAJGELJ, A., PARKANY, M., Eds.), the Royal Society of Chemistry, Cambridge (1999).
[97] IUPAC/ISO/AOAC Harmonized Guidelines for Internal Quality Control in Analytical Chemistry Laboratory, prepared for publication by R. Wood and M. Thompson, Pure & Appl. Chem. **67** (4) (1995) 649–666.
[98] INTERNATIONAL ORGANIZATION FOR STANDARDIZATION, ISO 8402:1994, Quality - Vocabulary, Geneva, Switzerland (1994).
[99] IUPAC, Compendium of Analytical Terminology - IUPAC Orange Book, Blackwell Scientific (1998).
[100] EURACHEM, The Fitness for Purpose of analytical Methods – A Laboratory Guide to Method Validation and Related Topics (1998).
[101] INTERNATIONAL ORGANIZATION FOR STANDARDIZATION, et al., International Vocabulary of Basic and General Terms in Metrology (VIM), Geneva, Switzerland (1993).
[102] INTERNATIONAL ORGANIZATION FOR STANDARDIZATION, et al., Guide to the Expression of Uncertainty in Measurement, Geneva, Switzerland (1995).

BIBLIOGRAPHY

ANDERSON, T.W., An Introduction to Multivariate Statistical Analysis, 2nd Edition, Wiley, New York, (1984).

AOAC INTERNATIONAL, ASSOCIATION OF ANALYTICAL COMMUNITIES (Gaithersburg, MD)

Quality Assurance Principles for Analytical Laboratories, (GARFIELD, F.M., KLESTA, E., HIRSCH, J., Eds.) (2000).

Statistical Manual of the AOAC, (YOUDEN, W.J., STEINER, E.H., Eds.) (1987).

Use of Statistics to Develop and Evaluate Analytical Methods, (WERNIMONT, G.T., Author, SPENDLEY, W., Ed.) (1987).

BAYNE, C.K., RUBIN, I.B., Practical Experimental Designs and Optimization Methods for Chemists, Deerfield Beach, FL, VCH Publishers (1986).

BOX, G.E.P., HUNTER, W.G., HUNTER, S.J., Statistics for experimenters: An Introduction to Design, Data Analysis and Model Building, Wiley, New York (1978).

CENTRO DE INVESTIGACIONES ENERGÉTICAS, MADIOAMBIENTALES Y TECNOLÓGICAS, Garantía de Calidad y Control de Calidad en Química Analítica, Serie Ponencias, Madrid, España (1999).

CONSUMING INDUSTRIES TRADE ACTION COALITION (CITAC) (Washington, D.C.)

Quality Assurance for Research and Development and Non-routine Analysis (1998).

Quantifying Uncertainty in Analytical Measurement (2000).

DEMING, S.N., MORGAN, S.L., Experimental Design: A Chemometric Approach, 2nd Edition, Elsevier Science Publishers B.V., Amsterdam, The Netherlands (1993).

DIXON, W.J., MASSEY, F.J., Introduction to Statistical Analysis, 4th Edition, McGraw-Hill, New York (1983).

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION (Geneva, Switzerland)

Accuracy (trueness and precision) of Measurement Methods and Results, ISO-5725 Parts 1–6 (1994).

Guide To The Expression Of Uncertainty In Measurement (1993).

Statistics – Vocabulary and Symbols, ISO-3534 (1993).

Uses of Certified Reference Materials, Revised ISO Guide 33 (2000).

LANGLEY, R., Practical Statistics – Simply Explained, Pan Books, London (1979).

MENDENHALL, W., Introduction to Probability and Statistics, Wadsworth Publishing Co., Belmont (1975).

MILLER, J.C., MILLER, J.N., Statistics for Analytical Chemistry, John Wiley and Sons, New York (1988).

QUEVAUVILLER, PH., MAIER, E.A., Interlaboratory Studies and Certified Reference Materials for Environmental Analysis, The BCR Approach, Elsevier, Amsterdam (1999) 558 pp.

STOEPPLER, M., WOLF, W.R., JENKS, P.J., Reference Materials for Chemical Analysis, Certification, Availability, and Proper Usage, Wiley-VCH, Weinheim (2001).

TAYLOR, J.K., Standard Reference Materials: Handbook for SRM Users, NIST Special Publications 260-100, National Institute of Standards and Technology, Gaithersbury, MD (1993).

DEFINITIONS

Analytical portion or test portion: A representative quantity of material removed from the analytical sample, of proper size for measurement of the analyte concentration.

Bias: Difference between the expectation of the test result and an accepted reference value.

Certification produces precise numerical values of the property under test or analysis that are free of, or corrected for, all known systematic errors, and are also related to the "true value" of the property under test or analysis. Certification deals with the establishment of "true values", with the provisions that (1) systematic errors in the measurement process leading to certification are always investigated, but it should be realized that advances in the state of the art may uncover additional systematic errors that were unsuspected at the time of the original work; therefore, a cautious, conservative estimate of residual and unknown systematic error is the rule, and this should always be reflected in the final stated uncertainty; (2) every material is inherently unstable and property values will change with time; and (3) certified values are only valid when the reference material is used in the manner for which it is intended and with all stated precautions followed by the user. It is generally accepted that a property can be certified when the value is confirmed by several analysts/laboratories working independently using either one definitive method, or more likely, two or more methods of appropriate and equivalent accuracy.

Certified reference material: Reference material accompanied by a certificate, one or more of whose property values are certified by a procedure which establishes traceability to an accurate realization to the unit in which the property values are expressed, and for which each certified value is accompanied by an uncertainty at a stated level of confidence [4]. Certified reference materials are generally prepared in batches for which the property values are determined within the stated uncertainty limits by measurements on samples representative for the whole batch. All certified reference materials lie within the definition of 'measurement standards' or 'etalons' given in the International Vocabulary of Basic and General Terms in Metrology.

Characterization: For reference materials and quality control materials, is the determination of one or more physical, chemical, biological or technological property values that are relevant to its intended end use.

Definitive method of chemical analysis is one that has a valid and well-described theoretical foundation, has been experimentally evaluated to lead to negligible systematic errors and a high level of precision. Definitive methods provide the fundamental basis for accuracy in chemical analysis.

Fitness for purpose: degree to which data produced by a measurement process enables a user to make technically and administratively correct decision for a stated purpose.

Homogeneity: Condition of being of uniform structure or composition with respect to one or more specified properties. A reference material or quality control material is said to be homogeneous with respect to a specified property if the property value, as determined by tests on samples of specified size, is found to lie within the specified uncertainty limits, the samples being taken either from different supply units (bottles, packages, etc.) or from a single supply unit. (Adopted from [4]).

Internal quality control: set of procedures undertaken by laboratory staff for the continuous monitoring of operation and the results of measurements in order to decide whether results are reliable enough to be released [99].

Matrix (or compositional) reference materials: A "natural" substance more representative of laboratory samples that has been chemically characterized for one or more elements, constituents, etc. with a known uncertainty. (*Note:* This is not a standardized definition).

Matrix (or compositional) RMs: A "natural" substance more representative of laboratory samples that has been chemically characterized for one or more elements, constituents, etc. with a known uncertainty. (*Note:* This is not a standardized definition) [97].

Method validation: The process of establishing the performance characteristics and limitations of a

method and the identification of the influences that 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 of verifying that a method is fit for a purpose, i.e. for solving a particular analytical problem [100].

Precision: Closeness of the agreement between independent test results obtained under prescribed conditions.

Quality assurance comprises all those planned and systematic actions undertaken by the organization necessary to provide adequate confidence that a product or service will satisfy given requirements for quality [1]. In other words, quality assurance describes the overall measures that a laboratory uses to ensure the quality of its operations.

Quality control material: Material used for the purposes of internal quality control and subjected to the same part of the same measurement procedure as that used for test materials [97]

Quality control: Operational techniques and activities that are used to fulfil requirements for quality [98].

Quality, according to the general ISO definition, is the totality of characteristics of an entity that bear on its ability to satisfy stated and implied needs.

Reference material: Material or substance one or more of whose property values are sufficiently homogeneous and well established to be used for the calibration of an apparatus, the assessment of a measurement method, or for assigning values to materials. A reference material may be in the form of a pure or mixed gas, liquid or solid. Examples are water for the calibration of viscometers, sapphire as a heat-capacity calibrant in calorimetry, and solutions used for calibration in chemical analysis.

Reference material: Material or substance one or more of whose property values are sufficiently homogeneous and well established to be used for the calibration of an apparatus, the assessment of a measurement method, or for assigning values to materials. A reference material may be in the form of a pure or mixed gas, liquid or solid. Examples are water for the calibration of viscometers, sapphire as a heat-capacity calibrant in calorimetry, and solutions used for calibration in chemical analysis.

Reference method is a method of proven and demonstrated accuracy established by direct comparison with a definitive method or with a primary reference material.

Routine analysis: A type of chemical analysis in which the analytical problem will have been encountered before. A suitable validated method for solving the problem would exists and may be in frequent use. The degree of associated staff training, calibration and quality control used with the method will depend on sample throughput.

Stability: Ability of a RM or QCM, when stored under specified conditions, to maintain a stated property value within specified limits for a specified period of time.

Traceability: Property of the result of a measurement or the value of a standard (including calibrants, CRMs and QCMs, etc.) whereby it can be related to stated references, usually national or international standards, through an unbroken chain of comparisons all having stated uncertainties [101].

(1) The concept is often expressed by the adjective **traceable**.
(2) The unbroken chain of comparisons is called a **traceability chain**.

Trueness: Closeness of the agreement between the average value obtained from a large series of test results and an accepted value.

Uncertainty of measurement: Parameter associated with the result of measurement that characterizes the dispersion of the values that could reasonably be attributed to the measurand.

(1) The parameter may be, for example, a standard deviation (or a given multiple of it), or the half-width of an interval having stated level of confidence.
(2) Uncertainty of measurement comprises, in general, many components (uncertainty sources).

Some of these components may be evaluated from the statistical distribution of the results of series of measurements and can be characterized by experimental standard deviations. The other components, which can also be characterized by standard deviations, are evaluated from assumed probability distributions based on experience or other information.

(3) It is understood that the result of measurement is the best estimate of the value of the measurand, and that all components of uncertainty, including those arising from systematic effects, such as components associated with corrections and reference standards, contribute to the dispersion.

Validated method: Analytical method which has undergone full or partial method validation, assuring that this method is fit for a specific purpose. *Note*: in this text the meaning of a **well established method** is similar as for validated method. The main difference arises from the frequency in which the methods are applied and for the type of analytical task. It is assumed that validated methods in a laboratory are being applied for routine analyses [102].

ABBREVIATIONS

ANOVA	analysis of variance
AOAC	Association of Analytical Communities
ASTM	American Society for Testing Materials
CRM	certified reference materials
CUSUM	cumulative sum (control chart)
ISO	International Organization for Standardization
IUPAC	International Union of Pure and Applied Chemistry
QA	quality assurance
QC	quality control
QCM	quality control materials
RM	reference materials
SS	sum of squares
VIM	International Vocabulary of Basic and General Terms in Metrology

Annex

HARMONIZED GUIDELINES FOR INTERNAL QUALITY CONTROL IN ANALYTICAL CHEMISTRY LABORATORIES

Guidelines used by the: International Union of Pure and Applied Chemistry, International Organization for Standardization, AOAC International

PREFACE

ISO, IUPAC and AOAC INTERNATIONAL have cooperated to produce agreed protocols on the "Design-, Conduct and Interpretation of Collaborative Studies" [1] and on the "Proficiency Testing of- (Chemical) Analytical Laboratories" [2] . The Working Group that produced these protocols has prepared a further protocol on the internal quality control of data produced in analytical laboratories. The document was discussed at the Fifth International Symposium on the Harmonisation of Quality Assurance Systems in Chemical Analysis, sponsored by IUPAC/ISO/AOAC INTERNATIONAL and held in Washington D.C. in July, 1993, and finalised at a Working Group Meeting held in Delft in May 1994.

1 INTRODUCTION

1.1 Basic concepts

This document sets out guidelines for the implementation of internal quality control (IQC) in analytical laboratories. IQC is one of a number of concerted measures that analytical chemists can take to ensure that the data produced in the laboratory are fit for their intended purpose. In practice, fitness for purpose is determined by a comparison of the accuracy achieved in a laboratory at a given time with a required level of accuracy.

Internal quality control therefore comprises the routine practical procedures that enable the analytical chemist to accept a result or group of results as fit for purpose, or reject the results and repeat the analysis. As such, IQC is an important determinant of the quality of analytical data, and is recognised as such by accreditation agencies.

Internal quality control is undertaken by the inclusion of particular reference materials, here called "control materials", into the analytical sequence and by duplicate analysis. The control materials should, wherever possible, be representative of the test materials under consideration in respect of matrix composition, the state of physical preparation and the concentration range of the analyte. As the control materials are treated in exactly the same way as the test materials, they are regarded as surrogates that can be used to characterise the performance of the analytical system, both at a specific time and over longer intervals.

Internal quality control is a final check of the correct execution of all of the procedures (including calibration) that are prescribed in the analytical protocol and all of the other quality assurance measures that underlie good analytical practice. IQC is therefore necessarily retrospective. It is also required to be as far as possible independent of the analytical protocol, especially the calibration, that it is designed to test.

Ideally both the control materials and those used to create the calibration should be traceable to appropriate certified reference materials or a recognised empirical reference method. When

this is not possible, control materials should be traceable at least to a material of guaranteed purity or other well characterised material. However, the two paths of traceability must not become coincident at too late a stage in the analytical process.

For instance, if control materials and calibration standards were prepared from a single stock solution of analyte, IQC would not detect any inaccuracy stemming from the incorrect preparation of the stock solution.

In a typical analytical situation several, or perhaps many, similar test materials will be analysed together, and control materials will be included in the group. Often determinations will be duplicated by the analysis of separate test portions of the same material. Such a group of materials is referred to in this document as an analytical "run". (The words "set", "series" and "batch" have also been used as synonyms for "run".) Runs are regarded as being analysed under effectively constant conditions. The batches of reagents, the instrument settings, the analyst, and the laboratory environment will, under ideal conditions, remain unchanged during analysis of a run. Systematic errors should therefore remain constant during a run, as should the values of the parameters that describe random errors. As the monitoring of these errors is of concern, the run is the basic operational unit of IQC.

A run is therefore regarded as being carried out under repeatability conditions, *i.e.*, the random measurement errors are of a magnitude that would be encountered in a "short" period of time. In practice the analysis of a run may occupy sufficient time for small systematic changes to occur. For example, reagents may degrade, instruments may drift, minor adjustments to instrumental settings may be called for, or the laboratory temperature may rise. However, these systematic effects are, for the purposes of IQC, subsumed into the repeatability variations. Sorting the materials making up a run into a randomised order converts the effects of drift into random errors.

1.2 Scope of this document

This document is a harmonisation of IQC procedures that have evolved in various fields of analysis, notably clinical biochemistry, geochemistry and environmental studies, occupational hygiene and food analysis [3–9] . There is much common ground in the procedures from these various fields. However, analytical chemistry comprises an even wider range of activities, and the basic principles of IQC should be able to encompass all of these. The present document provides guidelines that will be applicable in the great majority of instances. This policy necessarily excludes a number of IQC practices that are restricted to individual sectors of the analytical community. In addition in some sectors it is common to combine IQC as defined here with other aspects of quality assurance practice. There is no harm in such combination, but it must remain clear what the essential aspects of IQC are.

In order to achieve a harmonisation and provide basic guidance on IQC, some types of analytical activity have been excluded from this document. Issues specifically excluded are as follows.

(i) *Quality control of sampling.* While it is recognised that the quality of the analytical result can be no better than that of the sample, quality control of sampling is a separate subject and in many areas is not fully developed. Moreover, in many instances analytical laboratories have no control over sampling practice and quality.

(ii) *In-line analysis and continuous monitoring*. In this style of analysis there is no possibly of repeating the measurement, so the concept of IQC as used in this document is inapplicable.
(iii) *Multivariate IQC*. Multivariate methods in IQC are still the subject of research and cannot be regarded as sufficiently established for inclusion here. The current document regards multianalyte data as requiring a series of univariate IQC tests. Caution is necessary in the interpretation of this type of data to avoid inappropriately frequent rejection of data.
(iv) *Statutory and contractual requirements*.
(v) *Quality assurance measures* such as checks on instrumental stability before and during analysis, wavelength calibration, balance calibration, tests on resolution of chromatography columns, and problem diagnostics are not included. For present purposes they are regarded as part of the analytical protocol, and IQC tests their effectiveness together with the other aspects of the methodology.

1.3 Internal quality control and uncertainty

A prerequisite of analytical chemistry is the recognition of "fitness for purpose", the standard of accuracy that is required for an effective use of the analytical data. This standard is arrived at by consideration of the intended uses of the data although it is seldom possible to foresee all of the potential future applications of analytical results. For this reason, in order to prevent inappropriate interpretation, it is important that a statement of the uncertainty should accompany analytical results, or be readily available to those who wish to use the data.

Strictly speaking an analytical result cannot be interpreted unless it is accompanied by knowledge of its associated uncertainty at a stated level of confidence. A simple example demonstrates this principle. Suppose that there is a statutory requirement that a foodstuff must not contain more than 10 $\mu g\ g^{-1}$ of a particular constituent. A manufacturer analyses a batch and obtains a result of 9 $\mu g\ g^{-1}$ for that constituent.
If the uncertainty of the result expressed as a half range (assuming no sampling error) is 0.1 $\mu g\ g^{-1}$ (i.e. the true result falls, with a high probability, within the range 8.9–9.1) then it may be assumed that the legal limit is not exceeded. If, in contrast, the uncertainty is 2 $\mu g\ g^{-1}$ then there is no such assurance. The interpretation and use that may be made of the measurement thus depends on the uncertainty associated with it.

Analytical results should therefore have an associated uncertainty if any definite meaning is to be attached to them or an informed interpretation made. If this requirement cannot be fulfilled, the use to which the data can be put is limited. Moreover, the achievement of the required measurement uncertainty must be tested as a routine procedure, because the quality of data can vary, both in time within a single laboratory and between different laboratories. IQC comprises the process of checking that the required uncertainty is achieved in a run.

2 DEFINITIONS

2.1 International definitions

Quality assurance: all those planned and systematic actions necessary to provide adequate confidence that a product or service, will satisfy given requirements for quality [10].

Trueness: closeness of the agreement between the average value obtained from a large series of test results and an accepted reference value [11].

Precision: closeness of agreement between independent test results obtained under prescribed conditions [12].

Bias: difference between the expectation of the test results and an accepted reference value [11].

Accuracy: closeness of the agreement between the result of a measurement and a true value of the measurand [13].
Note 1. Accuracy is a qualitative concept.
Note 2. The term precision should not be used for accuracy.

Error: result of a measurement minus a true value of the measurand [13].

Repeatability conditions: conditions where independent test results are obtained with the same method on identical test items in the same laboratory by the same operator using the same equipment within short intervals of time [11].

Uncertainty of measurement: parameter, associated with the result of a measurement, that characterises the dispersion of the values that could reasonably be attributed to the measurand [14].
Note 1. The parameter may be, for example, a standard deviation (or a given multiple of it), or the half-width of an interval having a stated level of confidence.
Note 2. Uncertainty of measurement comprises, in general, many components. Some of these components may be evaluated from the statistical distribution of results of a series of measurements and can be characterised by experimental standard deviations. The other components, which can also be characterised by standard deviations, are evaluated from assumed probability distributions based on experience or other information.
Note 3. It is understood that the result of a measurement is the best estimate of the value of a measurand, and that all components of uncertainty, including those arising from systematic effects, such as components associated with corrections and reference standards, contribute to the dispersion.

Traceability: property of the result of a measurement or the value of a standard whereby it can be related to stated references, usually national or international standards, through an unbroken chain of comparisons 0 having stated uncertainties [13].

Reference material: material or substance one of whose property values are sufficiently homogeneous and well established to be used for the calibration of an apparatus, the assessment of a measurement method, or for assigning values to materials [13].

Certified reference material: reference material, accompanied by a certificate, one or more of whose property values are certified by a procedure which establishes its traceability to an accurate realisation of the unit in which the property values are expressed, and for which each certified value is accompanied by an uncertainty at a stated level of confidence [13].

2.2 Definitions of terms specific to this document

Internal quality control: set of procedures undertaken by laboratory staff for the continuous monitoring of operation and the results of measurements in order to decide whether results are reliable enough to be released.

Control material: material used for the purposes of internal quality control and subjected to the same or part of the same measurement procedure as that used for test materials.

Run (analytical run): set of measurements performed under repeatability conditions. 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

Analytical system: range of circumstances that contribute to the quality of analytical data, including equipment, reagents, procedures, test materials, personnel, environment and quality assurance measures.

3 QUALITY ASSURANCE PRACTICES AND INTERNAL QUALITY CONTROL

3.1 Quality assurance

Quality assurance is the essential organisational infrastructure that underlies all reliable analytical measurements. It is concerned with achieving appropriate levels in matters such as staff training and management, adequacy of the laboratory environment, safety, the storage, integrity and identity of samples, record keeping, the maintenance and calibration of instruments, and the use of technically validated and properly documented methods. Failure in any of these areas might undermine vigorous efforts elsewhere to achieve the desired quality of data. In recent years these practices have been codified and formally recognised as essential. However, the prevalence of these favourable circumstances by no means ensures the attainment of appropriate data quality unless IQC is conducted.

3.2 Choice of analytical method

It is important that laboratories restrict their choice of methods to those that have been characterised as suitable for the matrix and analyte of interest. The laboratory must possess documentation describing the performance characteristics of the method, estimated under appropriate conditions.

The use of a method does not in itself guarantee the achievement of its established performance characteristics. There is, for a given method, only the potential to achieve a certain standard of reliability when the method is applied under a particular set of circumstances. It is this collection of circumstances, known as the "analytical system", that is therefore responsible for the accuracy of analytical data. Hence it is important to monitor the analytical system in order to achieve fitness for purpose. This is the aim of the IQC measures undertaken in a laboratory.

3.3 Internal quality control and proficiency tests

Proficiency testing is a periodic assessment of the performance of individual laboratories and groups of laboratories that is achieved by the distribution by an independent testing body of typical materials for unsupervised analysis by the participants [(2)]. Although important, participation in proficiency testing schemes is not a substitute for IQC measures, or vice versa. Proficiency testing schemes can be regarded as a routine, but relatively infrequent, check on analytical errors. Without the support of a well-developed IQC system, the value of participation in a proficiency test is negligible. Probably the main beneficial effect of proficiency tests is that of encouraging participants to install effective quality control systems. It has been shown that laboratories with effective IQC systems performed better in a proficiency testing scheme [(15)].

4 INTERNAL QUALITY CONTROL PROCEDURES

4.4 Introduction

Internal quality control involves the practical steps undertaken to ensure that errors in analytical data are of a magnitude appropriate for the use to which the data will be put. The practice of IQC depends on the use of two strategies, the analysis of reference materials to monitor trueness and statistical control, and duplication to monitor precision.

The basic approach to IQC involves the analysis of control materials alongside the test materials under examination. The outcome of the control analyses forms the basis of a decision regarding the acceptability of the test data. Two key points are worth noting in this context.

(i) The interpretation of control data must be based on documented, objective criteria, and on statistical principles wherever possible.

(ii) The results of control analyses should be viewed primarily as indicators of the performance of the analytical system, and only secondarily as a guide to the errors associated with individual test results. Substantial changes in the apparent accuracy of control determinations can sometimes be taken to imply similar changes to data for contemporary test materials, but correction of analytical data on the basis of this premise is unacceptable.

4.5 General approach – statistical control

The interpretation of the results of IQC analyses depends largely on the concept of statistical control, which corresponds with stability of operation. Statistical control implies that an IQC result x can be interpreted as arising independently and at random from a normal population with mean μ and variance σ^2.

Under these constraints only about 0.3% of results (x) would fall outside the bounds of $\mu \pm 3\sigma$. When such extreme results are encountered they are regarded as being "out of control" and interpreted to mean that the analytical system has started to behave differently. Loss of control therefore implies that the data produced by the system are of unknown accuracy and hence cannot be relied upon. The analytical system therefore requires investigation and remedial action before further analysis is undertaken. Compliance with statistical control can be monitored graphically with Shewhart control charts (see Appendix 1). An equivalent numerical approach, comparing values of $z = (x - \mu)/\sigma$ against appropriate values of the standard normal deviate, is also possible.

4.6 Internal quality control and fitness for purpose.

For the most part, the process of IQC is based on a description in terms of the statistical parameters of an ongoing analytical system in normal operation. Control limits are therefore based on the estimated values of these parameters rather than measures derived from considerations of fitness for purpose. Control limits must be narrower than the requirements of fitness for purpose or the analysis would be futile.

The concept of statistical control is inappropriate, however, when the so-called ad hoc analysis is being undertaken. In ad hoc analysis the test materials may be unfamiliar or rarely encountered, and runs are often made up of only a few such test materials. Under these circumstances there is no statistical basis for the construction of control charts. In such an instance the analytical chemist has to use fitness for purpose criteria, historical data or

consistency with the visual properties of the test material for judging the acceptability of the results obtained.

Either way, agreed methods of establishing quantitative criteria to characterise fitness for purpose would be desirable. Unfortunately, this is one of the less-developed aspects of IQC. In specific application areas guidelines may emerge by consensus. For example, in environmental studies it is usually recognised that relative uncertainties of less than ten percent in the concentration of a trace analyte are rarely of consequence. In food analysis the Horwitz curve[16] is sometimes used as a fitness for purpose criterion. Such criteria have been defined for clinical analysis [17,18] . In some areas of applied geochemistry a systematic approach has given rise to fitness for purpose criteria for sampling and analytical precisions. However, it is not practicable here to give guidelines in these areas, and at present no general principles can be advanced that would allow specific applications to be addressed.

4.7 The nature of errors

Two main categories of analytical error are recognised, namely random errors and systematic errors, which give rise to imprecision and bias respectively. The importance of categorising errors in this way lies in the fact that they have different sources, remedies and consequences for the interpretation of data.

Random errors determine the precision of measurement. They cause random positive and negative deviations of results about the underlying mean value. *Systematic errors* comprise displacement of the mean of many determinations from the true value. For the purposes of IQC two levels of systematic error are worth consideration.
(i) *Persistent bias* affects the analytical system (for a given type of test material) over a long period and affects all data. Such bias, if small in relation to random error, may be identifiable only after the analytical system has been in operation for a long time. It might be regarded as tolerable, provided it is kept within prescribed bounds.

(ii) *The run effect* is exemplified by a deviation of the analytical system during a particular run. This effect, where it is sufficiently large, will be identified by IQC at the time of occurrence as an out of control condition.

The conventional division of errors between the random and the systematic depends on the timescale over which the system is viewed. Run effects of unknown source can be regarded in the long term as the manifestation of a random process. Alternatively, if a shorter term view is taken, the same variation could be seen as a bias-Re change affecting a particular run.

The statistical model used for IQC in this document is as follows[1]. The value of a measurement (x) in a particular run is given by:
x = true value + persistent bias + run effect + random error (+ gross error).
Tle variance of x (σ^2_x) in the absence of gross errors is given by:

$$\sigma^2_x = \sigma^2_0 + \sigma^2_1$$

where

[1] The model could be extended if necessary to include other features of the analytical system.

σ^2_0 = variance of the random error (within run) and
σ^2_1 = variance of the run effect.

The variances of the true value and the persistent bias are both zero. An analytical system in control is fully described by σ^2_0 , σ^2_1 and the value of the persistent bias. Gross errors are implied when the analytical system does not comply with such a description.

5 IQC AND WITHIN-RUN PRECISION

5.8 Precision and duplication

A limited control of within-run precision is achieved by the duplication within a run of measurements made on test materials. The objective is to ensure that the differences between paired results are consistent with or better than the level implied by the value of σ_0 used by a laboratory for IQC purposes[2]. Such a test alerts the user to the possibility of poor within-run precision and provides additional information to help in interpreting control charts. The method is especially useful in ad hoc analysis, where attention is centred on a single run and information obtained from control materials is unlikely to be completely satisfactory.

As a general approach all of the test materials, or a random selection from them, are analysed in duplicate. The absolute differences $|d| = |x_1 - x_2|$ between duplicated analytical results x_1 and x_2 are tested against an upper control limit based on an appropriate value of σ_0. However, if the test materials in the run have a wide range of concentration of analyte, no single value of σ_0 can be assumed [(19)].

Duplicates for IQC must reflect as far as possible the full range of variation present in the run. They must not be analysed as adjacent members of the run, otherwise they will reveal only the smallest possible measure of analytical variability. The best placing of duplicates is at random within each run. Moreover the duplication required for IQC requires the complete and independent analysis (preferably blind) of separate test portions of the test material. A duplication of the instrumental measurement of a single test solution would be ineffective because the variations introduced by the preliminary chemical treatment of the test material would be absent.

5.9 Interpretation of duplicate data

5.2.1 *Narrow concentration range*. In the simplest situation the test materials comprising the run have a small range of analyte concentrations so that a common within-run standard deviation σ_0 can be applied. A value of this parameter must be estimated to provide a control limit. The upper 95% bound of $|d|$ is $2\sqrt{2}\,\sigma_0$ and on average only about three in a thousand results should exceed $3\sqrt{2}\,\sigma_0$

A group of n duplicated results can be interpreted in several ways. For example, the standardised difference
$z_d = d / \sqrt{2}\,\sigma_0$

---

[2] There is no intention here of estimating the standard deviation of repeatability σ, from IQC data or of comparing estimates: there would usually be too few results for a satisfactory outcome. Where such an estimate is needed the formula $s_r = \sqrt{\Sigma d^2 / 2n}$ can be used.

should have a normal distribution with zero mean and unit standard deviation. The sum of a group of n such results would have a standard deviation of $\sqrt{n}$, so only about three runs in a thousand would produce a value of $|\Sigma z_d| > 3\sqrt{n}$. Alternatively a group of n values of z_d from a run can be combined to form Σz^2_d and the result interpreted as a sample from a chi-squared distribution with n degrees of freedom, (χ^2_n) . Some caution is needed in the use of this statistic, however, as it is sensitive to outlying results.

5.2.2 *Wide concentration range*. If the test materials comprising a run have a wide range of analyte concentrations, no common standard of precision (σ_0) can be assumed. In such an instance σ_0 must be expressed as a functional relationship with concentration. The value of concentration for a particular material is taken to be $(x_1 + x_2)/2$, and an appropriate value of σ_0 obtained from the functional relationship, the parameters of which have to be estimated in advance.

6 CONTROL MATERIALS IN IQC

6.10 Introduction

Control materials are characterised substances that are inserted into the run alongside the test materials and subjected to exactly the same treatment. A control material must contain an appropriate concentration of the analyte, and a value of that concentration must be assigned to the material. Control materials act as surrogates for the test materials and must therefore be representative, i.e., they should be subject to the same potential sources of error. To be fully representative, a control material must have the same matrix in terms of bulk composition, including minor constituents that may have a bearing on accuracy. It should also be in a similar physical form, i.e., state of comminution, as the test materials. There are other essential characteristics of a control material. It must be adequately stable over the period of interest. It must be possible to divide the control material into effectively identical portions for analysis. It is often required in large amounts to allow its use over an extended period.

Reference materials in IQC are used in combination with control charts that allow both persistent bias and run effects to be addressed (Appendix 1). Persistent bias is evident as a significant deviation of the centre line from the assigned value. The
variation in the run effect is predictable in terms of a standard deviation when the system is under statistical control, and that standard deviation is used to define action limits and warning limits at appropriate distances from the true value.

6.11 The role of certified reference materials

Certified reference materials (CRM) as defined in Section 2 (i.e., with a statement of uncertainty and traceability), when available and of suitable composition, are ideal control materials in that they can be regarded for traceability purposes as ultimate standards of trueness [20]. In the past CRMs were regarded as being for reference purposes only and not for routine use. A more modem approach is to treat CRMs as consumable and therefore suitable for IQC.

The use of CRMs in this way is, however, subject to a number of constraints.
(i) Despite the constantly increasing range of CRMs available, for the majority of analyses there is no closely matching CRM available.

(ii) Although the cost of CRMs is not prohibitive in relation to the total costs of analysis, it may not be possible for a laboratory with a wide range of activities to stock every relevant kind of reference material.
(iii) The concept of the reference material is not applicable to materials where either the matrix or the analyte is unstable.
(iv) CRMs are not necessarily available in sufficient amounts to provide for IQC use over extended periods.
(v) It must be remembered that not all apparently certified reference materials are of equal quality. Caution is suggested when the information on the certificate is inadequate.

If for any of the above reasons the use of a CRM is not appropriate it falls on individual laboratories or groups of laboratories to prepare their own control materials and assign traceable[3] values of analyte concentration to them. Such a material is sometimes referred to as a "house reference material" (HRM). Suggestions for preparing HRMs are listed in Section 6.3. Not all of the methods described there are applicable to all analytical situations.

6.12 Preparation of control materials

6.3.1 *Assigning a true value by analysis*. In principle a working value can be assigned to a stable reference material simply by careful analysis. However, precautions are necessary to avoid biases in the assigned value. This requires some form of independent check such as may be provided by analysis of the materials in a number of laboratories and, where possible, the use of methods based on different physico-chemical principles. Lack of attention to independent validation of control materials has been shown to be a weakness in IQC systems (15).

One way of establishing a traceable assigned value in a control material is to analyse a run comprising the candidate material and a selection of matching CRMS, with replication and randomisation. This course of action would be appropriate if only limited amounts of CRMs were available. The CRMs must be appropriate in both matrix composition and analyte concentration. The CRMs are used directly to calibrate the analytical procedure for the analysis of the control material. An appropriate analytical method is a prerequisite for this approach. It would be a dangerous approach if, say, a minor and variable fraction of the analyte were extracted for measurement. The uncertainty introduced into the assigned value must also be considered.

6.3.2 *Materials validated in proficiency testing* comprise a valuable source of control materials. Such materials would have been analysed by many laboratories using a variety of methods. In the absence of counter-indications, such as an obvious bias or unusual frequency distribution of results, the consensus of the laboratories could be regarded as a validated assigned value to which a meaningful uncertainty could be attached. (There is a possibility that the consensus could suffer from a bias of consequence,-but this potential is always present in reference values.) There would be a theoretical problem of establishing the traceability of such a value, but that does not detract from the validity of the proposed procedure. The range of such materials available would be limited, but organisers of proficiency tests could ensure a copious supply by preparing batches of material in excess of

[3] Where a CRM is not available traceability only to a reference method or to a batch of a reagent supplied by a manufacturer may be necessary.

the immediate requirements of the round. The normal requirements of stability would have to be demonstrable.

6.3.3 *Assigning a true value by formulation.* In favourable instances a control material can be prepared simply by mixing constituents of known purity in predetermined amounts. For example, this approach would often be satisfactory in instances where the control material is a solution. Problems are often encountered in formulation in producing solid control materials in a satisfactory physical state or in ensuring that the speciation and physical distribution of the analyte in the matrix is realistic. Moreover an adequate mixing of the constituents must be demonstrable.

6.3.4 *Spiked control materials.* "Spiking" is a way of creating a control material in which a value is assigned by a combination of formulation and analysis. This method is feasible when a test material essentially free of the analyte is available. After exhaustive analytical checks to ensure the background level is adequately low, the material is spiked with a known amount of analyte. The reference sample prepared in this way is thus of the same matrix as the test materials to be analysed and of known analyte level — the uncertainty in the assigned concentration is limited only by the possible error in the unspiked determination. However, it may be difficult to ensure that the speciation, binding and physical form-n of the added analyte is the same as that of the native analyte and that the mixing is adequate.

6.3.5 *Recovery checks.* If the use of a reference material is not practicable then a limited check on bias is possible by a test of recovery. This is especially useful when analytes or matrices cannot be stabilised or when ad hoc analysis is executed. A test portion of the test material is spiked with a known amount of the analyte and analysed alongside the original test material. The recovery of the added analyte (known as the "marginal recovery") is the difference between the two measurements divided by the amount that is added. The obvious advantages of recovery checks are that the matrix is representative and the approach is widely applicable — most test materials can be spiked by some means. However, the recovery check suffers from the disadvantage previously noted regarding the speciation, binding and physical distribution of the analyte. Furthermore, the assumption of an equivalent recovery of the analyte added as a spike and of the native analyte may not be valid. However, it can normally be assumed that a poor performance in a recovery check is strongly indicative of a similar or worse performance for the native analyte in the test materials.

Spiking and recovery testing as an IQC method must be distinguished from the method of standard additions, which is a measurement procedure: a single spiking addition cannot be used to fulfil the roles of both measurement and IQC.

6.13 Blank donations

Blank determinations are nearly always an essential part of the analytical process and can conveniently be effected alongside the IQC protocol. The simplest form of blank is the "reagent blank", where the analytical procedure is executed in all respects apart from the addition of the test portion. This kind of blank, in fact, tests more than the purity of the reagents. For example it is capable of detecting contamination of the analytical system originating from any source, e.g., glassware and the atmosphere, and is therefore better described as a "procedural blank". In some instances, better execution of blank determinations is achieved if a simulated test material is employed. The simulant could be an actual test material known to be virtually analyte-free or a surrogate (e.g., ashless filter paper used

instead of plant material). Where it can be contrived, the best type of blank is the "field blank", which is a typical matrix with zero concentration of analyte.

An inconsistent set of blanks in a run suggests sporadic contamination and may add weight to IQC evidence suggesting the rejection of the results. When an analytical protocol prescribes the subtraction of a blank value, the blank value must be subtracted also from the results of the control materials before they are used in IQC.

6.14 Traceability in spiking and recovery checks

Potential problems of the traceability of reagents used for spikes and recovery checks must be guarded against. Under conditions where CRMs are not available, traceability can often be established only to the batch of analyte provided by a manufacturer. In such cases, confirmation of identity and a check on purity must be made before use. A further precaution is that the calibration standards and spike should not be traceable to the same stock solution of analyte or the same analyst. If such a common traceability existed, then the corresponding sources of error would not be detected by the IQC.

7 RECOMMENDATIONS

The following recommendations represent integrated approaches to IQC that are suitable for many types of analysis and applications areas. Managers of laboratory quality systems will have to adapt the recommendations to the demands of their own particular requirements. Such adoption could be implemented, for example, by adjusting the number of duplicates and control material inserted into a run, or by the inclusion of any additional measures favoured in the particular application area. The procedure finally chosen and its accompanying decision rules must be codified in an IQC protocol that is separate from the analytical system protocol.

The practical approach to quality control is determined by the frequency with which the measurement is carried out and the size and nature of each run. The following recommendations are therefore made. The use of control charts and decision rules are covered in Appendix 1.

In each of the following the order in the run in which the various materials are analysed should be randomised if possible. A failure to randomise may result in an underestimation of various components of error.

(i) *Short (e.g., $n<20$) frequent runs of similar materials.* Here the concentration range of the analyte in the run is relatively small, so a common value of standard deviation can be assumed.
Insert a control material at least once per run. Plot either the individual values obtained, or the mean value, on an appropriate control chart. Analyse in duplicate at least half of the test materials, selected at random. Insert at least one blank determination.
(ii) *Longer (e.g., $n>20$) frequent runs of similar materials.* Again a common level of standard deviation is assumed.
Insert the control material at an approximate frequency of one per ten test materials. If the run size is likely to vary from run to run it is easier to standardise on a fixed number of insertions per run and plot the mean value on a control chart of means. Otherwise plot individual values. Analyse in duplicate a minimum of five test materials selected at random. Insert one blank determination per ten test materials.

(iii) *Frequent runs containing similar materials but with a wide range of analyte concentration*. Here we cannot assume that a single value of standard deviation is applicable. Insert control materials in total numbers approximately as recommended above. However, there should be at least two levels of analyte represented, one close to the median level of typical test materials, and the other approximately at the upper or lower decile as appropriate. Enter values for the two control materials on separate control charts. Duplicate a minimum of five test materials, and insert one procedural blank per ten test materials.

(iv) *Ad hoc analysis*. Here the concept of statistical control is not applicable. It is assumed, however, that the materials in the run are of a single type, i.e., sufficiently similar for general conclusions on errors to be made.

Carry out duplicate analysis on all of the test materials. Carry out spiking or recovery tests or use a formulated control material, with an appropriate number of insertions (see above), and with different concentrations of analyte if appropriate. Carry out blank determinations. As no control limits are available, compare the bias and precision with fitness for purpose limits or other established criteria.

8 CONCLUSIONS

Internal quality control is an essential aspect of ensuring that data released from a laboratory are fit for purpose. If properly executed, quality control methods can monitor the various aspects of data quality on a run-by-run basis. In runs where performance falls outside acceptable limits, the data produced can be rejected and, after remedial action on the analytical system, the analysis can be repeated.

It must be stressed, however, that internal quality control is not foolproof even when properly executed. Obviously it is subject to "errors of both kinds", i.e., runs that are in control will occasionally be rejected and runs that are out of control occasionally accepted. Of more importance, IQC cannot usually identify sporadic gross errors or short term disturbances in the analytical system that affect the results for individual test materials. Moreover, inferences based on IQC results are applicable only to test materials that fall within the scope of the analytical method validation. Despite these limitations, which professional experience and diligence can alleviate to a degree, internal quality control is the principal recourse available for ensuring that only data of appropriate quality are released from a laboratory. When properly executed it is very successful.

Finally, it must be appreciated that a perfunctory execution of any quality system will not guarantee the production of data of adequate quality. The correct procedures for feedback, remedial action and staff motivation must also be documented and acted upon. In other words, there must be a genuine commitment to quality within a laboratory for an internal quality control programme to succeed, i.e., the IQC must be part of a total quality management system.

REFERENCES TO THE ANNEX

[1] "Protocol for the Design, Conduct and Interpretation of Method Performance Studies", Edited W Horwitz, Pure Appl Chem, 1988, 60, 855– 864. (Revision in press)

[2] "7th International Harmonised Protocol for the Proficiency Testing of (Chemical) Analytical Laboratories", Edited M Thompson and R Wood, Pure Appl. Chem., 1993, 65, 2123–2144. (Also published in J. AOAC International, 1993, 76, 926–940.

[3] "IFCC approved recommendations on quality control in clinical chemistry. Part 4: internal quality control", J. Clin. Chem Clin. Biochem, 1980, 18, 534–541.

[4] S Z Cekan, S B Sufi and E W Wilson, "Internal quality control for assays of reproductive hormones: Guidelines for laboratories", WHO, Geneva, 1993.

[5] M Thompson, "Control procedures in geochemical analysis", in R J Howarth (Ed), "Statistics and data analysis in geochemical prospecting", Elsevier, Amsterdam, 1983.

[6] M Thompson, "Data quality in applied geochemistry: the requirements and how to achieve them", J. Geochem Explor., 1992, 44, 3–22.

[7] Health and Safety Executive, "Analytical quality in workplace air monitoring", London, 1991.

[8] "A protocol for analytical quality assurance in public analysts' laboratories", Association of Public Analysts, 342 Coleford Road, Sheffield S9 5PH, UK, 1986.

[9] "Method evaluation, quality control, proficiency testing" (AMIQAS PC Program), National Institute of Occupational Health, Denmark, 1993.

[10] ISO 8402:1994, "Quality assurance and quality management – vocabulary".

[11] ISO 3534 -1: 1993 (E/F), "Statistics, vocabulary and symbols - Part 1: Probability and general statistical terms".

[12] ISO Guide 30:1992, "Terms and definitions used in connections with reference materials".

[13] "International vocabulary for basic and general terms in metrology", 2nd Edition, 1993, ISO, Geneva.

[14] "Guide to the expression of uncertainty in measurement", ISO, Geneva, 1993, 15, M Thompson and P J Lowthian, Analyst, 1993, 118, 1495–1500.

[15] W Horwitz, L R Kamps and K W Boyer, J. Assoc. Off. Anal. Chem, 1980, 63, 1344.

[16] D Tonks, Clin. Chem, 1963, 9, 217–223.

[17] G C Fraser, P H Petersen, C Ricos and R Haeckel, "Proposed quality specifications for the imprecision and inaccuracy of analytical systems for clinical chemistry", Eur. J. Clin. Chem. Clin. Biochem., 1992, 30, 311–317.

[18] M Thompson, Analyst, 1988, 113, 1579–1587.

[19] ISO Guide 33: 1989, "Uses of Certified Reference Materials", Geneva.

Appendix

SHEWHART CONTROL CHARTS

1 INTRODUCTION

The theory, construction and interpretation of the Shewhart chart [1] are detailed in numerous texts on process quality control and applied statistics, and in several ISO standards [2–5]. There is a considerable literature on the use of the control chart in clinical chemistry [6, 7]. Westgard and coworkers have formulated multiple rules for the interpretation of such control charts[8], and the power of these results has been studied in detail [9, 10]. In this appendix only simple Shewhart charts are considered.

In IQC a Shewhart control chart is obtained when values of concentration measured on a control material in successive runs are plotted on a vertical axis against the run number on the horizontal axis. If more than one analysis of a particular control material is made in a run, either the individual results x or the mean value — can be used to form a control chart. The chart is completed by horizontal lines derived from the normal distribution $N(\mu, \sigma^2)$ that is taken to describe the random variations in the plotted values. The selected lines for control purposes are μ, $\mu \pm 2\sigma$ and $\mu \pm 3\sigma$. Different values of σ are required for charts of individual values and of means. For a system in statistical control, on average about one in twenty values fall outside the $\mu \pm 2\sigma$ lines, called the "warning limits", and only about three in one thousand fall outside the $\mu \pm 3\sigma$ lines, the "action limits". In practice the estimates x and s of the parameters μ and σ are used to construct the chart. A persistent bias is indicated by a significant difference between x and the assigned value. A control chart showing results from a system in statistical control over 40 runs is shown in Figure A-1.

2 ESTIMATES OF THE PARAMETERS μ AND σ

An analytical system under control exhibits two sources of random variation, the within- run, characterised by variance σ^2_0, and the between-run with variance σ^2_1 .The two variances are typically comparable in magnitude. The standard deviation CF. used in a chart of individual values is given by

$$\sigma_x = (\sigma^2_0 + \sigma^2_1)^{1/2}$$

whereas for a control chart of mean values the standard deviation is given by

$$\sigma_x = (\sigma^2_0 / n + \sigma^2_1)^{1/2}$$

where n is the number of control measurements in a run from which the mean is calculated. The value of n therefore must be constant from run to run, otherwise control limits would be impossible to define. If a fixed number of repeats of a control material per run cannot be guaranteed (e.g., if the run length were variable) then charts of individual values must be used. Furthermore the equation indicates that σ_x or σ_x must be estimated with care. An attempt to base an estimate on repeat values from a single run would result in unduly narrow control limits.

Estimates must therefore include the between-run component of variance. If the use of a particular value of *n* can be assumed at the outset, then σ_x can be estimated directly from the

m means $\bar{x}_i = \sum_{j=1}^{n} x_{ij} / n$, (i = 1,.....,m) of the n repeats in each of m successive runs. Thus the estimate of μ is

$$\bar{x} = \sum_i \bar{x}_i / m,$$

and the estimate of $\sigma_{\bar{x}}$ is

$$s_x = \sqrt{\frac{\sum_i (\bar{x}_i - \bar{x})^2}{m-1}}$$

If the value of n is not predetermined, then separate estimates of σ_0 and σ_1 could be obtained by one-way analysis of variance. If the mean squares within- and between groups are MS_w. and MS_b, respectively, then
σ^2_0 is estimated by MS_w and
σ^2_1 is estimated by $(MS_b - MS_w.)/n$

Often in practice it is necessary to initiate a control chart with data collected from a small number of runs, which may be to a degree unrepresentative, as estimates of standard deviation are very variable unless large numbers of observations are used. Moreover, during the initial period, the occurrence of out-of-control conditions are more than normally likely and will produce outlying values. Such values of would bias x and inflate s beyond its proper value. It is therefore advisable to recalculate x and s after a further "settling down" period. One method of obviating the effects of outliers in the calculation is to reject them after the application of Dixon's Q or Grubbs' [11] test, and then use the classical statistics given above. Alternatively, the methods of robust statistics could be applied to the data [12, 13].

3 THE INTERPRETATION OF CONTROL CHARTS

The following simple rules can be applied to control charts of individual results or of means.

Single control chart. An out-of-control condition in the analytical system is signalled if any of the following occur.

(i) The current plotting value falls outside the action limits.
(ii) The current value and the previous plotting value fall outside the warning limits but within the actions limits.
(iii) Nine successive plotting values fall on the same side of the mean line.

Two control charts. When two different control materials are used in each run, the respective control charts are considered simultaneously. This increases the chance of a type I error (rejection of a sound run) but decreases the chance of a type 2 error (acceptance of a flawed run). An out-of-control condition is indicated if any of the following occur.

(i) At least one of the plotting values falls outside the action limits.
(ii) Both of the plotting values are outside the warning limits.
(iii) The current value and the previous plotting value on the same control chart both fall outside the warning limits.

(iv) Both control charts simultaneously show that four successive plotting values on the same side of the mean line.
(v) One of the charts shows nine successive plotting values falling on the same side of the mean line.

A more thorough treatment of the control chart can be obtained by the application of the full Westgard rules, illustrated in Figure A-2.

The analytical chemist should respond to an out-of-control condition by cessation of analysis pending diagnostic tests and remedial action followed by rejection of the results of the run and reanalysis of the test materials.

REFERENCES TO THE APPENDIX

[1] W A Shewhart, "Economic control of quality in manufactured product", Van Nostrand, New York, 1931.
[2] ISO 8258:1991, "Shewhart control charts".
[3] ISO 7873:1993, "Control charts for arithmetic means with warning limits".
[4] ISO 7870:1993, "Control charts - general guide and introduction".
[5] ISO 7966:1993, "Acceptance control charts".
[6] S Levey and E R Jennings, Am. J. Clin. Pathol, 1950, 20, 1059–1066.
[7] ABJ Nix, R J Rowlands, K W Kemp, D W Wilson and K Griffiths, Stat. Med., 1987,6, 425–440.
[8] J 0 Westgard, P L Barry and M R Hunt, Clin. Chem., 1981, 27, 493–501.
[9] C A Parvin, Clin. Chem. 1992, 38, 358–363.
[10] J Bishop and A B J Nix, Clin. Chem., 1993, 39, 1638–1649.
[11] W Horwitz, Pure Appl Chem., (in press).
[12] Analytical Methods Committee, Analyst, 1989, 114, 1693–1697.
[13] Analytical Methods Committee, Analyst, 1989, 114, 1699–1702.

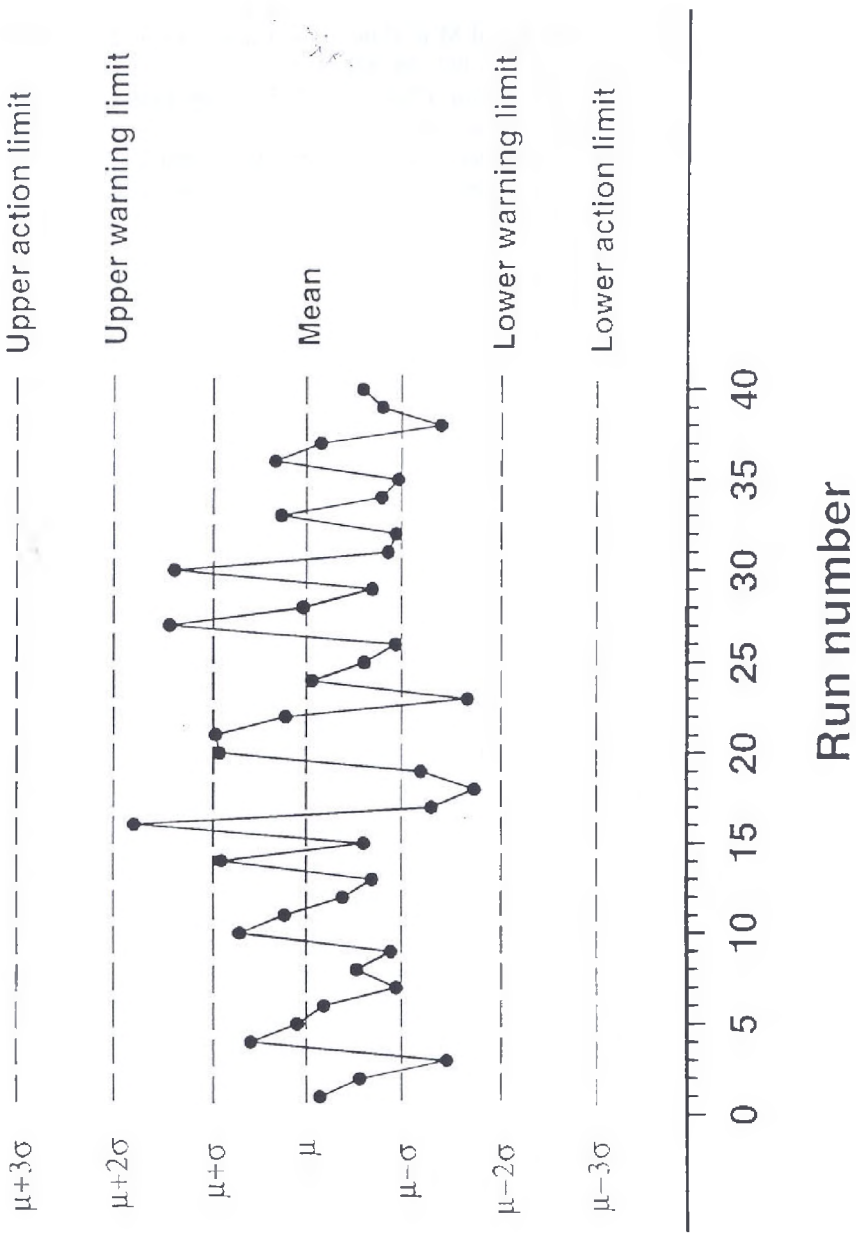

FIG. A-1. Results from a system in statistical control.

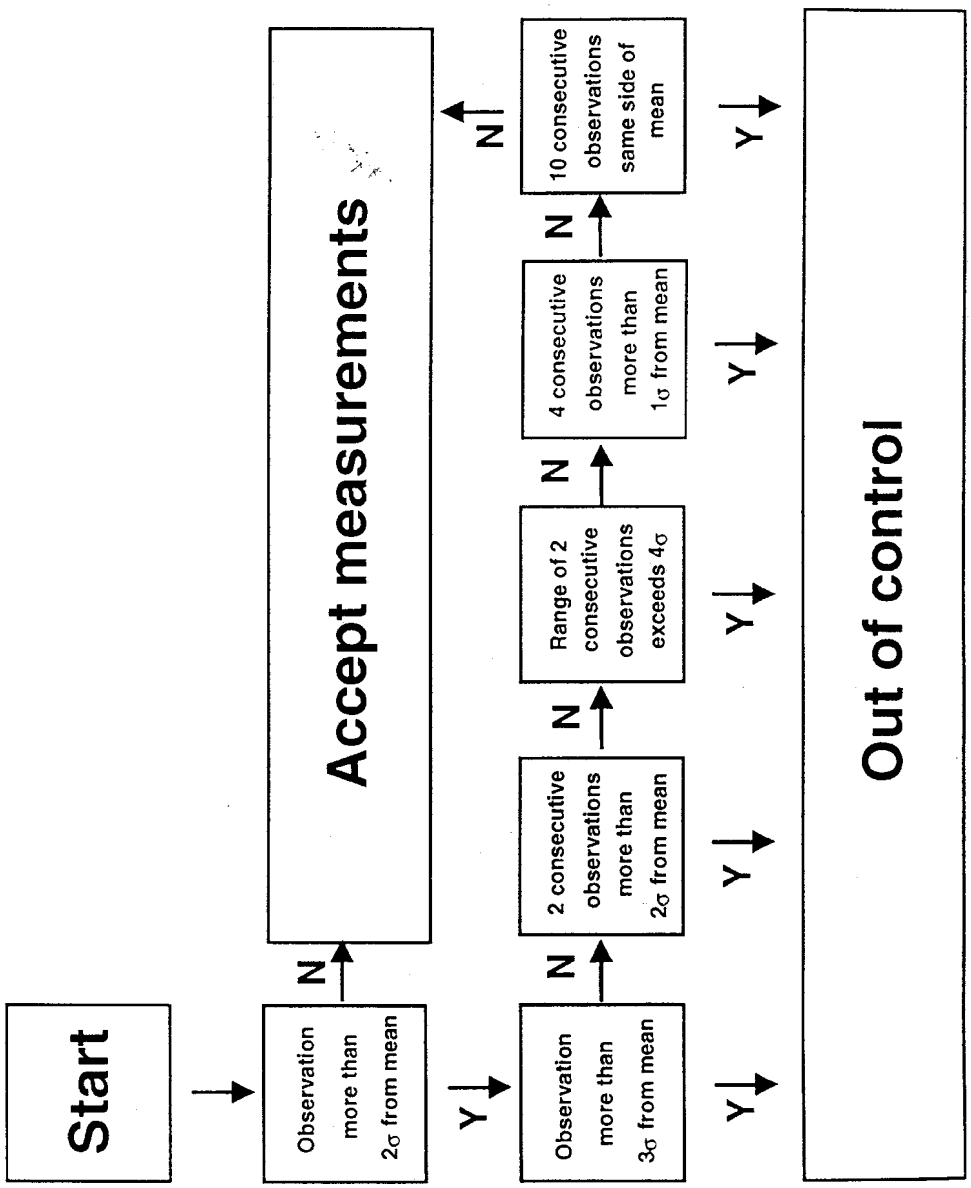

FIG. A-2. The Westgard rules.

CONTRIBUTORS TO DRAFTING AND REVIEW

Ambrus, A.	International Atomic Energy Agency
Cortes Toro, E.	Comisión Chilena de Energía Nuclear, Centro Nuclear "La Reina", Chile
El-Bidaoui, Maha	International Atomic Energy Agency
Fajgelj, A.	International Atomic Energy Agency
Gardiner, P.H.E.	Sheffield Hallam University, United Kingdom
Ihnat, M.	Pacific Agri-Food Research Centre, Canada
Qian Chuanfan	China Agricultural University, China
Rossbach, M.	International Atomic Energy Agency