3. Quality in Measurement and Testing

Technology and today's global economy depend on reliable measurements and tests that are accepted internationally. As has been explained in Chap. 1, metrology can be considered in categories with different levels of complexity and accuracy.

- Scientific metrology deals with the organization and development of measurement standards and with their maintenance.
- Industrial metrology has to ensure the adequate functioning of measurement instruments used in industry as well as in production and testing processes.
- Legal metrology is concerned with measurements that influence the transparency of economic transactions, health, and safety.

All scientific, industrial, and legal metrological tasks need appropriate *quality* methodologies, which are compiled in this chapter.

3.1 Sampling

Sampling is arguably the most important part of the measurement process. It is usually the case that it is impossible to measure the required quantity, such as concentration, in an entire batch of material. The taking of a sample is therefore the essential first step of nearly all measurements. However, it is commonly agreed that the quality of a measurement can be no better than the quality of the sampling upon which it is based. It follows that the highest level of care and attention paid to the instrumental measurements is ineffectual, if the original sample is of poor quality.

3.1.1 Quality of Sampling

The *traditional approach* to ensuring the quality of sampling is procedural rather than empirical. It relies initially on the selection of a correct sampling protocol for the particular material to be sampled under a particular circumstance. For example the material may be copper metal, and the circumstance could be manufacturers' quality control prior to sale. In general, such a protocol may be specified by a regulatory body, or recommended in an international standard or by a trade organization. The second step is to train the personnel who are to take the samples (i. e., the samplers) in the correct application of the protocol. No sampling protocol can be completely unambiguous in its wording, so uniformity of interpretation relies on the samplers being educated, not just in how to interpret the words, but also in an appreciation of the rationale behind the protocol and how it can be adapted to the changing circumstances that will arise in the real world, without invalidating the protocol. This step is clearly related to the management of sampling by organizations, which is often separated from the management of the instrumental measurements, even though they are both inextricably linked to the overall quality of the measurement. The fundamental basis of the traditional approach to assuring sampling quality is to assume that the correct application of a correct sampling protocol will give a representative sample, by definition.

An *alternative approach* to assuring sampling quality is to estimate the quality of sampling empirically. This is analogous to the approach that is routinely taken to instrumental measurement, where as well as specifying a protocol, there is an initial validation and ongoing quality control to monitor the quality of the measurements actually achieved. The key parameter of quality for instrumental measurements is now widely recognized to be the uncertainty of each measurement. This concept will be discussed in detail later (Sect. 3.4), but informally this uncertainty of measurement can be defined as the range within which the true value lies, for the quantity subject to measurement, with a stated level of probability. If the quantity subject to measurement (the measurand) is defined in terms of the batch of material (the sampling target), rather than merely in the sample delivered to the laboratory, then measurement uncertainty includes that arising from primary sampling. Given that sampling is the first step in the measurement process, then the uncertainty of the measurement will also arise in this first step, as well as in all of the other steps, such as the sampling preparation and the instrumental determination.

The key measure of sampling quality is therefore this sampling uncertainty, which includes contributions not just from the random errors often associated with sampling variance [3.1] but also from any systematic errors that have been introduced by sampling bias. Rather than assuming the bias is zero when the protocol is correct, it is more prudent to aim to include any bias in the estimate of sampling uncertainty. Such bias may often be unsuspected, and arise from a marginally incorrect application of a nominally correct protocol. This is equivalent to abandoning the assumption that samples are representative, but replacing it with a measurement result that has an associated estimate of uncertainty which includes errors arising from the sampling process.

Selection of the most appropriate sampling protocol is still a crucial issue in this alternative approach. It is possible, however, to select and monitor the appropriateness of a sampling protocol, by knowing the uncertainty of measurement that it generates. A judgement can then be made on the fitness for purpose (FFP) of the measurements, and hence the various components of the measurement process including the sampling, by comparing the uncertainty against the target value indicated by the FFP criterion. Two such FFP criteria are discussed below.

Two approaches have been proposed for the estimation of uncertainty from sampling [3.2]. The first or *bottom-up* approach requires the identification of all of the individual components of the uncertainty, the separate estimation of the contribution that each component makes, and then summation across all of the components [3.3]. Initial feasibility studies suggest that the use of sampling theory to predict all of the components will be impractical for all but a few sampling systems, where the material is particulate in nature and the system conforms to a model in which the particle size/shape and analyte concentration are simple, constant, and homogeneously distributed. One recent application successfully mixes theoretical and empirical estimation techniques [3.4]. The second, more practical and pragmatic approach is entirely empirical, and has been called *top-down* estimation of uncertainty [3.5].

Four methods have been described for the empirical estimation of uncertainty of measurement, including that from primary sampling [3.6]. These methods can be applied to any sampling protocol for the sampling of any medium for any quantity, if the general principles are followed. The simplest of these methods (#1) is called the *duplicate method*. At its simplest, a small proportion of the measurements are made in duplicate. This is not just a duplicate analysis (i. e., determination of the quantity), made on one sample, but made on a fresh primary sample, from the same sampling target as the original sample, using a fresh interpretation of the same sampling protocol (Fig. 3.1a). The ambiguities in the protocol, and the heterogeneity of the material, are therefore reflected in the difference between the duplicate measurements (and samples). Only 10% ($n \geq 8$) of the samples need to be duplicated to give a sufficiently reliable estimate of the overall uncertainty [3.7]. If the separate sources of the uncertainty need to be quantified, then extra duplication can be inserted into the experimental design, either in the determination of quantity (Fig. 3.1b) or in other steps, such as the physical preparation of the sample (Fig. 3.1d). This duplication can either be on just one sample duplicate (in an *unbalanced* design, Fig. 3.1b), or on both of the samples duplicated (in a *balanced* design, Fig. 3.1c).

The uncertainty of the measurement, and its components if required, can be estimated using the statistical technique called analysis of variance (ANOVA). The frequency distribution of measurements, such as analyte concentration, often deviate from the normal distribution that is assumed by classical ANOVA. Because of this, special procedures are required to accommodate outlying values, such as robust ANOVA [3.8]. This method

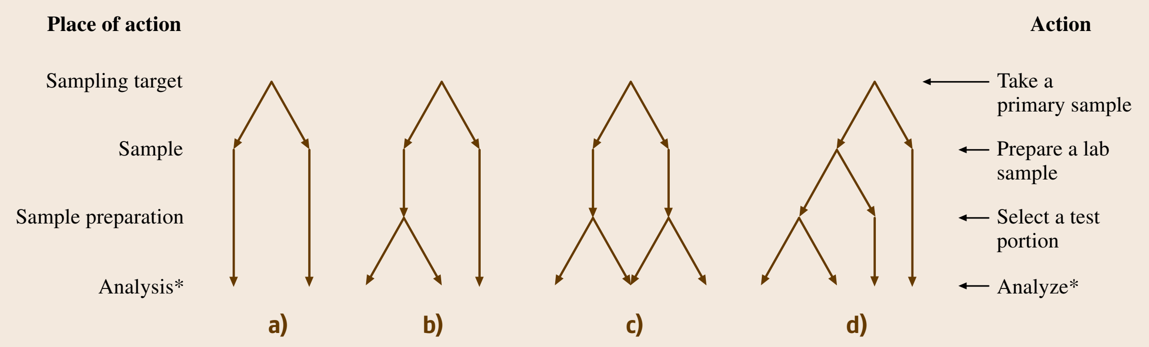


Fig. 3.1a–d Experimental designs for the estimation of measurement uncertainty by the duplicate method. The simplest and cheapest option **(a)** has single analyses on duplicate samples taken on around 10% ($n \geq 8$) of the sampling targets, and only provides an estimate of the random component of the overall measurement uncertainty. If the contribution from the analytical determination is required separately from that from the sampling, duplication of analysis is required on either one **(b)** or both **(c)** of the sample duplicates. If the contribution from the physical sample preparation is required to be separated from the sampling, as well as from that from the analysis, then duplicate preparations also have to be made **(d)**. (*Analysis and Analyze can more generally be described as the determination of the measurand)

has successfully been applied to the estimation of uncertainty for measurements on soils, groundwater, animal feed, and food materials [3.2]. Its weakness is that it ignores the contribution of systematic errors (from sampling or analytical determination) to the measurement uncertainty. Estimates of analytical bias, made with certified reference materials, can be added to estimates from this method. Systematic errors caused by a particular sampling protocol can be detected by use of a different method (#2) in which different sampling protocols are applied by a single sampler. Systematic errors caused by the sampler can also be incorporated into the estimate of measurement uncertainty by the use of a more elaborate method (#3) in which several samplers apply the same protocol. This is equivalent to holding a *collaborative trial in sampling* (CTS). The most reliable estimate of measurement uncertainty caused by sampling uses the most expensive method (#4), in which several samplers each apply whichever protocol they consider most appropriate for the stated objective. This incorporates possible systematic errors from the samplers and the measurement protocols, together with all of the random errors. It is in effect a sampling proficiency test (SPT), if the number of samplers is at least eight [3.6].

Evidence from applications of these four empirical methods suggests that small-scale heterogeneity is often the main factor limiting the uncertainty. In this case, methods that concentrate on repeatability, even with just one sampler and one protocol as in the duplicate method (#1), are good enough to give an acceptable approximation of the sampling uncertainty. Proficiency test measurements have also been used in *top-down* estimation of uncertainty of analytical measurements [3.9]. They do have the added advantage that the participants are scored for the proximity of their measurement value to the true value of the quantity subject to measurement. This true value can be estimated either by consensus of the measured values, or by artificial spiking with a known quantity of analyte [3.10]. The score from such SPTs could also be used for both ongoing assessment and accreditation of samplers [3.11]. These are all new approaches that can be applied to improving the quality of sampling that is actually achieved.

3.1.2 Judging Whether Strategies of Measurement and Sampling Are Appropriate

Once methods are in place for the estimation of uncertainty, the selection and implementation of a correct protocol become less crucial. Nevertheless an appropriate protocol is essential to achieve fitness for purpose. The FFP criterion may however vary, depending on the circumstances. There are cases for example where a relative expanded uncertainty of 80% of the measured value can be shown to be fit for certain purposes. One example is using in situ measurements of lead concentration to identify any area requiring remediation in a contaminated land investigation. The contrast between the concentration in the contaminated and in the uncontaminated areas can be several orders of magnitude, and so uncertainty within one order (i. e., 80%) does not result in errors in classification of the land. A similar situation applies when using laser-ablation inductively coupled plasma for the determination of silver to differentiate between particles of anode copper from widely different sources. The Ag concentration can differ by several orders of magnitude, so again a large measurement uncertainty (e.g., 70%) can be acceptable. One mathematical way of expressing this FFP criterion is that the measurement uncertainty should not contribute more than 20% to the total variance over samples from a set of similar targets [3.8]. A second FFP criterion also includes financial considerations, and aims to set an optimal level of uncertainty that minimizes financial loss. This loss arises not just from the cost of the sampling and the determination, but also from the financial losses that may arise from incorrect decisions caused by the uncertainty [3.12]. The approach has been successfully applied to the sampling of both contaminated soil [3.13] and food materials [3.14].

3.1.3 Options for the Design of Sampling

There are three basic approaches to the design/selection of a sampling protocol for any quantity (measurand) in any material. The first option is to select a previously specified protocol. These exist for most of the material/quantity combinations considered in Chap. 4 of this handbook. This approach is favored by regulators, who expect that the specification and application of a standard protocol will automatically deliver comparability of results between samplers. It is also used as a defense in legal cases to support the contention that measurements will be reliable if a standard protocol has been applied. The rationale of a standard protocol is to specify the procedure to the point where the sampler needs to make no subjective judgements. In this case the sampler would appear not to require any grasp of the rationale behind the design of the protocol, but merely the ability to implement the instructions given. However, experimental video monitoring of samplers implementing specified protocols suggests that individual samplers often do extemporize, especially when events occur that were unforeseen or unconsidered by the writers of the protocols. This would suggest that samplers therefore need to appreciate the rationale behind the design, in order to make appropriate decisions on implementing the protocol. This relates to the general requirement for improved training and motivation of samplers discussed below.

The second option is to use a theoretical model to design the required sampling protocol. Sampling theory has produced a series of increasingly complex theoretical models, recently reviewed [3.15], that are usually aimed at predicting the sampling mass required to produce a given level of variance in the required measurement result. All such models depend on several assumptions about the system that is being modeled. The model of *Gy* [3.1], for example, assumes that the material is particulate, that the particles in the batch can be classified according to volume and type of material, and that the analyte concentration in a contaminated particle and its density do not vary between particles. It was also assumed that the volume of each particle in the batch is given by a constant factor multiplied by the cube of the particle diameter. The models also all require large amounts of information about the system, such as particle diameters, shape factors, size range, liberation, and composition. The cost of obtaining all of this information can be very high, but the model also assumes that these parameters will not vary in space or time. These assumptions may not be justified for many systems in which the material to be sampled is highly complex, heterogeneous, and variable. This limits the real applicability of this approach for many materials. These models do have a more generally useful role, however, in facilitating the prediction of how uncertainty from sampling can be changed, if required, as discussed below.

The third option for designing a sampling protocol is to adapt an existing method in the light of site-specific information, and monitor its effectiveness empirically. There are several factors that require consideration in this adaptation.

Clearly identifying the *objective of the sampling* is the key factor that helps in the design of the most appropriate sampling protocol. For example, it may be that the acceptance of a material is based upon the best estimate of the *mean* concentration of some analyte in a batch. Alternatively, it may be the *maximum* concentration, within some specified mass, that is the basis for acceptance or rejection. Protocols that aim at low uncertainty in estimation of the mean value are often inappropriate for reliable detection of the maximum value.

A *desk-based review* of all of the relevant information about the sampling target, and findings from similar targets, can make the protocol design much more cost effective. For example, the history of a contaminated land site can suggest the most likely contaminants and their probable spatial distribution within the site. This information can justify using *judgemental sampling* in which the highest sampling density is concentrated in the area of highest predicted probability. This approach does however, have the weakness that it may be self-fulfilling, by missing contamination in areas that were unsuspected.

The actual *mode of sampling* varies greatly therefore, depending not just on the physical nature of the materials, but also on the expected heterogeneity in both the spatial and temporal dimension. Some protocols are designed to be random (or nonjudgemental) in their selection of samples, which in theory creates the least bias in the characterization of the measurand. There are various different options for the design of random sampling, such as stratified random sampling, where the target is subdivided into regular units before the exact location of the sampling is determined using randomly selected coordinates. In a situation where judgemental sampling is employed, as described above, the objective is not to get a representative picture of the sampling target. Another example would be in an investigation of the cause of defects in a metallurgical process, where it may be better to select items within a batch by their aberrant visual appearance, or contaminant concentration, rather than at random.

There may also be a question of the most appropriate *medium to sample*. The answer may seem obvious, but consider the objective of detecting which of several freight containers holds nuts that are contaminated with mycotoxins. Rather than sampling the nuts themselves, it may be much more cost effective to sample the atmosphere in each container for the spores released by the fungi that make the mycotoxin. Similarly in contaminated land investigation, if the objective is to assess potential exposure of humans to cadmium at an allotment site, it may be most effective to sample the vegetables that take up the cadmium rather than the soil.

The *specification of the sampling target* needs to be clear. Is it a whole batch, or a whole site of soil, or just the top 1 m of the soil? This relates to the objective of the sampling, but also to the site-specific information (e.g., there is bedrock at 0.5 m) and logistical constraints.

The next key question to address is the *number of samples* required (n). This may be specified in an accepted sampling protocol, but should really depend on the objective of the investigation. Cost–benefit analysis can be applied to this question, especially if the objective is the mean concentration at a specified confidence interval. In that case, and assuming a normal distribution of the variable, the Student t-distribution can be used to calculate the required value of n. A closely related question is whether *composite samples* should be taken, and if so, what is the required number of increments (i). This approach can be used to reduce the uncertainty of measurement caused by the sampling. According to the theory of Gy, taking an i-fold composite sample should reduce the main source of the uncertainty by $\sqrt{i}$, compared with the uncertainty for a single sample with the same mass as one of the increments. Not only do the increments increase the sample mass, but they also improve the sample's ability to represent the sampling target. If, however, the objective is to identify maximum rather than mean values, then a different approach is needed for calculating the number of samples required. This has been addressed for contaminated land by calculating the probability of hitting an idealized hot-spot [3.16].

The *quantity of sample* to be taken (e.g., mass or volume) is another closely related consideration in the design of a specified protocol. The mass may be specified by existing practise and regulation, or calculated from sampling theory such as that of Gy. Although the calculation of the mass from first principles is problematic for many types of sample, as already discussed, the theory is useful in calculating the factor by which to change the sample mass to achieve a specified target for uncertainty. If the mass of the sample is increased by some factor, then the sampling variance should reduce by the same factor, as discussed above for increments. The mass required for measurement is often smaller than that required to give an acceptable degree of representativeness (and uncertainty). In this case, a larger sample must be taken initially and then reduced in mass, without introducing bias. This *comminution of samples*, or reduction in grain size by grinding, is a common method for reducing the uncertainty introduced by this subsampling procedure. This can, however, have unwanted side-effects in changing the measurand. One example is the loss of certain analytes during the grinding, either by volatilization (e.g., mercury) or by decomposition (e.g., most organic compounds).

The *size of the particles* in the original sampling target that should constitute the sample needs consideration. Traditional wisdom may suggest that a representative sample of the whole sampling target is required. However, sampling all particle sizes in the same proportions that they occur in the sampling target may not be possible. This could be due to limitations in the sampling equipment, which may exclude the largest particles (e.g., pebbles in soil samples). A representative sample may not even be desirable, as in the case where only the small particles in soil ($< 100\,\mu\text{m}$) form the main route of human exposure to lead by hand-tomouth activity. The objectives of the investigation may require therefore that a specific size fraction be selected.

Contamination of samples is probable during many of these techniques of sampling processing. It is often easily done, irreversible in its effect, and hard to detect. It may arise from other materials at the sampling site (e.g., topsoil contaminating subsoil) or from processing equipment (e.g., cadmium plating) or from the remains of previous samples left in the equipment. The traditional approach is to minimize the risk of contamination occurring by careful drafting of the protocol, but a more rigorous approach is to include additional procedures that can detect any contamination that has occurred (e.g., using an SPT).

Once a sample has been taken, the protocol needs to describe how to *preserve the sample*, without changing the quantity subject to measurement. For some measurands the quantity begins to change almost immediately after sampling (e.g., the redox potential of groundwater), and in situ measurement is the most reliable way of avoiding the change. For other measurands specific actions are required to prevent change. For example, acidification of water, after filtration, can prevent adsorption of many analyte ions onto the surfaces of a sample container.

The final, and perhaps most important factor to consider in designing a sampling protocol is the logistical *organization of the samples* within the investigation. Attention to detail in the unique numbering and clear description of samples can avoid ambiguity and irreversible errors. This improves the quality of the investigation by reducing the risk of gross errors. Moreover, it is often essential for legal traceability to establish an unbroken chain of custody for every sample. This forms part of the broader quality assurance of the sampling procedure.

There is no such thing as either a perfect sample or a perfect measurement. It is better, therefore, to estimate the uncertainty of measurements from all sources, including the primary sampling. The uncertainty should not just be estimated in an initial method validation, but also monitored routinely for every batch using a sampling and analytical quality control scheme (SAQCS). This allows the investigator to judge whether each batch of measurements are FFP, rather than to assume that they are because some standard procedure was nominally adhered to. It also enables the investigator to propagate the uncertainty value through all subsequent calculations to allow the uncertainty on the interpretation of the measurements to be expressed. This approach allows for the imperfections in the measurement methods and the humans who implement them, and also for the heterogeneity of the real world.

3.2 Traceability of Measurements

3.2.1 Introduction

Clients of laboratories will expect that results are correct and comparable. It is further anticipated that complete results and values produced include an estimated uncertainty. A comparison between different results or between results achieved and given specifications can only be done correctly if the measurement uncertainty of the results is taken into account.

To achieve comparable results, the traceability of the measurement results to SI units through an unbroken chain of comparisons, all having stated uncertainties, is fundamental (Sect. 2.6 *Traceability of Measurements*). Among others, due to the strong request from the International Laboratory Accreditation Cooperation (ILAC) several years ago, the International Committee for Weights and Measures (CIPM), which is the governing board of the International Bureau of Weights and Measures (BIPM), has realized under the scope of the Metre Convention the CIPM mutual recognition arrangement (MRA) on the mutual recognition of national measurement standards and of calibration and measurement certificates issued by the national metrology institutes, under the scope of the Metre Convention. Details of this MRA can be found in Chap. 2 *Metrology Principles and Organization* Sect. 2.7 or at http://www1.bipm.org/en/convention/mra/.

The range of national measurement standards and best measurement capabilities needed to support the calibration and testing infrastructure in an economy or region can normally be derived from the websites of the respective national metrology institute or from the website of the BIPM. Traceability to these national measurement standards through an unbroken chain of comparisons is an important means to achieve accuracy and comparability of measurement results.

Access to suitable national measurement standards may be more complicated in those economies where the national measurement institute does not yet provide national measurement standards recognized under the BIPM MRA. It is further to be noted that an unbroken chain of comparisons to national standards in various fields such as the chemical and biological sciences is much more complex and often not available, as appropriate standards are lacking. The establishment of standards in these fields is still the subject of intense scientific and technical activities, and reference procedures and (certified) reference materials needed must still be defined. As of today, in these fields there are few reference materials that can be traced back to SI units available on the market. This means that other tools should also be applied to assure at least comparability of measurement results, such as, e.g., participation in suitable proficiency testing programs or the use of reference materials provided by reliable and competent reference material producers.

3.2.2 Terminology

According to the *International Vocabulary of Metrology – Basic and General Concepts and Associated Terms (VIM 2008)* [3.17], the following definitions apply.

Primary Measurement Standard

Measurement standard established using a primary reference measurement procedure, or created as an artifact, chosen by convention.

International Measurement Standard

Measurement standard recognized by signatories to an international agreement and intended to serve worldwide.

National Measurement Standard, National Standard

Measurement standard recognized by national authority to serve in a state or economy as the basis for assigning quantity values to other measurement standards for the kind of quantity concerned.

Reference Measurement Standard, Reference Standard

Measurement standard designated for the calibration of other measurement standards for quantities of a given kind in a given organization or at a given location.

Working Standard

Measurement standard that is used routinely to calibrate or verify measuring instruments or measuring systems.

Note that a working standard is usually calibrated against a reference standard. Working standards may also at the same time be reference standards. This is particularly the case for working standards directly calibrated against the standards of a national standards laboratory.

3.2.3 Traceability of Measurement Results to SI Units

The formal definition of traceability is given in Chap. 2, Sect. 2.6 as: the property of a measurement result relating the result to a stated metrological reference through an unbroken chain of calibrations or comparisons, each contributing to the stated uncertainty. This chain is also called the traceability chain. It must, as defined, end at the respective primary standard.

The uncertainty of measurement for each step in the traceability chain must be calculated or estimated according to agreed methods and must be stated so that an overall uncertainty for the whole chain may be calculated or estimated. The calculation of uncertainty is officially given in the *Guide to the Expression of Uncertainty in Measurement* (GUM) [3.18]. The ILAC and regional organizations of accreditation bodies (see under peer and third-party assessment) provide application documents derived from the GUM, providing instructive examples. These documents are available on their websites.

Competent testing laboratories, e.g., those accredited by accreditation bodies that are members of the ILAC MRA, can demonstrate that calibration of equipment that makes a significant contribution to the uncertainty and hence the measurement results generated by that equipment are traceable to the international system of units (SI units) wherever this is technically possible.

In cases where traceability to the SI units is not (yet) possible, laboratories use other means to assure at least comparability of their results. Such means are, e.g., the use of certified reference materials, provided by a reliable and competent producer, or they assure at least comparability by participating in interlaboratory comparisons provided by a competent and reliable provider. See also Sects. 3.6 and 3.7 on *Interlaboratory Comparisons and Proficiency Testing* and *Reference Materials*, respectively.

The Traceability Chain

National Metrology Institutes. In most cases the national metrology institutes maintain the *national standards* that are the sources of traceability for the quantity of interest. The national metrology institutes ensure the comparability of these standards through an international system of key comparisons, as explained in detail in Chap. 2, Sect. 2.7.

If a national metrology institute has an infrastructure to realize a given primary standard itself, this national standard is identical to or directly traceable to that *primary standard*. If the institute does not have such an infrastructure, it will ensure that its national standard is traceable to a primary standard maintained in another country's institute. Under http://kcdb.bipm.org/AppendixC/default.asp, the calibration and measurement capabilities (CMCs) declared by national metrology institutes are shown.

Calibration Laboratories. For *calibration laboratories* accredited according to the ISO/International Electrotechnical Commission (IEC) standard ISO/IEC 17025, accreditation is granted for specified calibrations with a defined calibration capability that can (but not necessarily must) be achieved with a specified measuring instrument, reference or working standard.

The *calibration capability* is defined as the smallest uncertainty of measurement that a laboratory can achieve within its scope of accreditation, when performing more or less routine calibrations of nearly ideal measurement standards intended to realize, conserve or reproduce a unit of that quantity or one or more of its values, or when performing more or less routine calibrations of nearly ideal measuring instruments designed for the measurement of that quantity.

Most of the accredited laboratories provide calibrations for customers (e.g., for organizations that do not have their own calibration facilities with a suitable measurement capability or for testing laboratories) on request. If the service of such an accredited calibration laboratory is taken into account, it must be assured that its scope of accreditation fits the needs of the customer. Accreditation bodies are obliged to provide a list of accredited laboratories with a detailed technical description of their scope of accreditation. http://www.ilac.org/ provides a list of the accreditation bodies which are members of the ILAC MRA.

If a customer is using a nonaccredited calibration laboratory or if the scope of accreditation of a particular calibration laboratory does not fully cover a specific calibration required, the customer of that laboratory must ensure that

- the tractability chain as described above is maintained correctly,
- there is a concept to estimate the overall measurement uncertainty in place and applied correctly,
- the staff is thoroughly trained to perform the activities within their responsibilities,
- clear and valid procedures are available to perform the required calibrations,
- a system to deal with errors is applied, and the calibration operations include statistical process control such as, e.g., the use of control charts.

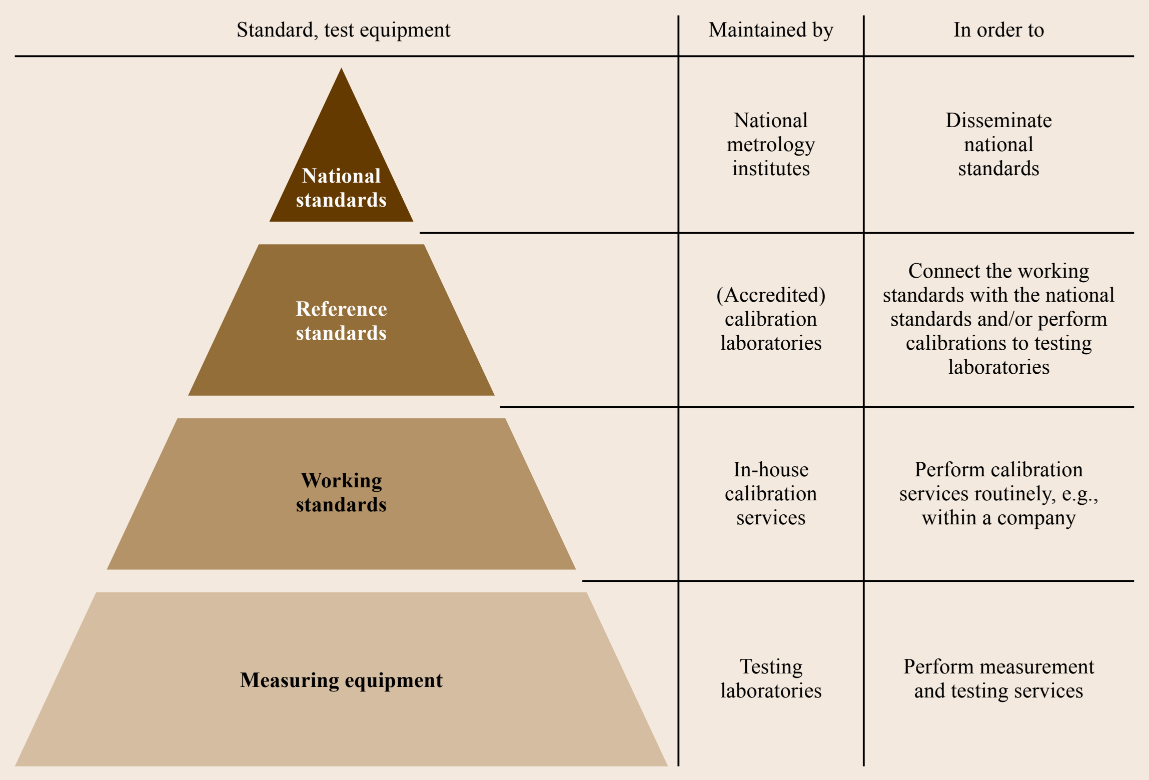
In-House Calibration Laboratories (Factory Calibration Laboratories). Frequently, calibration services are provided by in-house calibration laboratories which regularly calibrate the measuring and test equipment used in a company, e.g., in a production facility, against its reference standards that are traceable to an accredited calibration laboratory or a national metrology institute.

An in-house calibration system normally assures that all measuring and test equipment used within a company is calibrated regularly against working standards, calibrated by an accredited calibration laboratory. In-house calibrations must fit into the internal applications in such a way that the results obtained with the measuring and test equipment are accurate and reliable. This means that for in-house calibration the following elements should be considered as well.

- The uncertainty contribution of the in-house calibration should be known and taken into account if statements of compliance, e.g., internal criteria for measuring instruments, are made.
- The staff should be trained to perform the calibrations required correctly.
- Clear and valid procedures should be available also for in-house calibrations.
- A system to deal with errors should be applied (e.g., in the frame of an overall quality management system), and the calibration operations should include a statistical process control (e.g., the use of control charts).

To assure correct operation of the measuring and test equipment, a concept for the maintenance of that equipment should be in place. Aspects to be considered when establishing calibration intervals are given in Sect. 3.5.

The Hierarchy of Standards. *The hierarchy of standards* and a resulting metrological organizational structure for tracing measurement and test results within a company to national standards are shown in Fig. 3.2.



Standard, test equipment	Maintained by	In order to
National standards	National metrology institutes	Disseminate national standards
Reference standards	(Accredited) calibration laboratories	Connect the working standards with the national standards and/or perform calibrations to testing laboratories
Working standards	In-house calibration services	Perform calibration services routinely, e.g., within a company
Measuring equipment	Testing laboratories	Perform measurement and testing services

Fig. 3.2 The calibration hierarchy

Equipment used by testing and calibration laboratories that has a significant effect on the reliability and uncertainty of measurement should be calibrated using standards connected to the national standards with a known uncertainty.

Alternative Solutions

Accreditation bodies which are members of the ILAC MRA require accredited laboratories to ensure traceability of their calibration and test results. Accredited laboratories also know the contribution of the uncertainty derived through the traceability chain to their calibration and test results.

Where such traceability is not (yet) possible, laboratories should at least assure comparability of their results by alternative methods. This can be done either through the use of appropriate reference materials (RM) or by participating regularly in appropriate proficiency tests (PT) or interlaboratory comparisons. *Appropriate* means that the RM producers or the PT providers are competent or at least recognized in the respective sector.

3.2.4 Calibration of Measuring and Testing Devices

The VIM 2008 gives the following definition for calibration:

Definition

Operation that, under specified conditions, in a first step, establishes a relation between the quantity values with measurement uncertainties provided by measurement standards and corresponding indications with associated measurement uncertainties and, in a second step, uses this information to establish a relation for obtaining a measurement result from an indication.

The operation of calibration and its two steps is described in Sect. 3.4.2 with an example from dimensional metrology (Fig. 3.10).

It is common and important that testing laboratories regularly maintain and control their testing instruments, measuring systems, and reference and working standards. Laboratories working according to the ISO/IEC 17025 standard as well as manufacturers working according to, e.g., the ISO 9001 series of standards maintain and calibrate their measuring instruments, and reference and working standards regularly according to well-defined procedures.

Clause 5.5.2 of the ISO/IEC 17025 standard requires that:

> *Calibration programmes shall be established for key quantities or values of the instruments where these properties have a significant effect on the results.*
>
> *Whenever practicable, all equipment under the control of the laboratory and requiring calibration shall be labeled, coded, or otherwise identified to indicate the status of calibration, including the data when last calibrated and the date or expiration criteria when recalibration is due. (Clause 5.5.8)*

Clause 7.6 of ISO 9001:2000 requires that:

> *Where necessary to ensure valid results, measuring equipment shall be calibrated or verified at specified intervals, or prior to use, against measurement standards traceable to international or national measurement standards.*

In the frame of the calibration programs of their measuring instruments, and reference and working standards, laboratories will have to define the time that should be permitted between successive calibrations (recalibrations) of the used measurement instruments, and reference or working standards in order to

- confirm that there has not been any deviation of the measuring instrument that could introduce doubt about the results delivered in the elapsed period,
- assure that the difference between a reference value and the value obtained using a measuring instrument is within acceptable limits, also taking into account the uncertainties of both values,
- assure that the uncertainty that can be achieved with the measuring instrument is within expected limits.

A large number of factors can influence the time interval to be defined between calibrations and should be taken into account by the laboratory. The most important factors are usually

- the information provided by the manufacturer,
- the frequency of use and the conditions under which the instrument is used,
- the risk of the measuring instrument drifting out of the accepted tolerance,
- consequences which may arise from inaccurate measurements (e.g., failure costs in the production line or aspects of legal liability),
- the cost of necessary corrective actions in case of drifting away from the accepted tolerances,
- environmental conditions such as, e.g., climatic conditions, vibration, ionizing radiation, etc.,
- trend data obtained, e.g., from previous calibration records or the use of control charts,
- recorded history of maintenance and servicing,
- uncertainty of measurement required or declared by the laboratory.

These examples show the importance of establishing a concept for the maintenance of the testing instruments and measuring systems. In the frame of such a concept the definition of the calibration intervals is one important aspect to consider. To optimize the calibration intervals, available statistical results, e.g., from the use of control charts, from participation in interlaboratory comparisons or from reviewing own records should be used.

3.2.5 The Increasing Importance of Metrological Traceability

An increasing awareness of the need for metrological underpinning of measurements can be noticed at least in the past years. Several factors may be the reason for this process, including

- the importance of quality management systems,
- requirements by governments or trading partners for producers to establish certified quality management systems and for calibration and testing activities to be accredited,
- aspects of legal reliability.

In a lot of areas it is highly important that measurement results, e.g., produced by testing laboratories, can be compared with other results produced by other parties at another time and quite often using different methods. This can only be achieved if measurements are based on equivalent physical realizations of units. Traceability of results and reference values to primary standards is a fundamental issue in competent laboratory operation today.

3.3 Statistical Evaluation of Results

Statistics are used for a variety of purposes in measurement science, including mathematical modeling and prediction for calibration and method development, method validation, uncertainty estimation, quality control and assurance, and summarizing and presenting results. This section provides an introduction to the main statistical techniques applied in measurement science. A knowledge of the basic descriptive statistics (mean, median, standard deviation, variance, quantiles) is assumed.

3.3.1 Fundamental Concepts

Measurement Theory and Statistics

The traditional application of statistics to quantitative measurement follows a set of basic assumptions related to ordinary statistics

1. That a given measurand has a value – the value of the measurand – which is unknown and (in general) unknowable by the measurement scientist. This is generally assumed (for univariate quantitative measurements) to be a single value for the purpose of statistical treatment. In statistical standards, this is the *true value*.
2. That each measurement provides an *estimate* of the value of the measurand, formed from an *observation* or set of observations.
3. That an observation is the sum of the measurand value and an *error*.

Assumption 3 can be expressed as one of the simplest statistical models

$$x_i = \mu + e_i ,$$

in which x_i is the i-th observation, μ is the measurand value, and e_i is the error in the particular observation.

The error itself is usually considered to be a sum of several contributions from different sources or with different behavior. The most common partition of error is into two parts: one which is constant for the duration of a set of experiments (the *systematic* error) and another, the *random* error, which is assumed to arise by random selection from some distribution. Other partitioning is possible; for example, collaborative study uses a statistical model based on a systematic contribution (method bias), a term which is constant for a particular laboratory (the laboratory component of bias) but randomly distributed among laboratories, and a residual error for each observation. Linear calibration assumes that observations are the sum of a term that varies linearly and systematically with measurand value and a random term; least-squares regression is one way of characterizing the behavior of the systematic part of this model.

The importance of this approach is that, while the value of the measurand may be unknown, studying the distribution of the observations allows inferences to be drawn about the probable value of the measurand. Statistical theory describes and interrelates the behaviors of different distributions, and this provides quantitative tools for describing the probability of particular observations given certain assumptions. Inferences can be drawn about the value of the measurand by asking what range of measurand values could reasonably lead to the observations found. This provides a range of values that can reasonably be attributed to the measurand. Informed readers will note that this is the phrase used in the definition of uncertainty of measurement, which is discussed further below.

This philosophy forms the basis of many of the routine statistical methods applied in measurement, is well established with strong theoretical foundations, and has stood the test of time well. This chapter will accordingly rely heavily on the relevant concepts. It is, however, important to be aware that it has limitations. The basic assumption of a point value for the measurand may be inappropriate for some situations. The approach does not deal well with the accumulation of information from a variety of different sources. Perhaps most importantly, real-world data rarely follow theoretical distributions very closely, and it can be misleading to take inference too far, and particularly to infer very small probabilities or very high levels of confidence. Furthermore, other theoretical viewpoints can be taken and can provide different insights into, for example, the development of confidence in a value as data from different experiments are accumulated, and the treatment of estimates based on judgement instead of experiment.

Distributions

Figure 3.3 shows a typical measurement data set from a method validation exercise. The tabulated data shows a range of values. Plotting the data in histogram form shows that observations tend to cluster near the center of the data set. The histogram is one possible graphical representation of the *distribution* of the data.

If the experiment is repeated, a visibly different data distribution is usually observed. However, as the

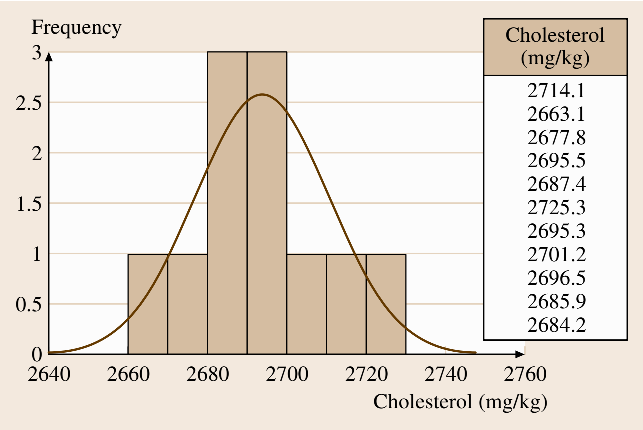


Fig. 3.3 Typical measurement data. Data from 11 replicate analyses of a certified reference material with a certified value of 2747±90 mg/kg cholesterol. The *curve* is a normal distribution with mean and standard deviation calculated from the data, with vertical scaling adjusted for comparability with the histogram

number of observations in an experiment increases, the distribution becomes more consistent from experiment to experiment, tending towards some underlying form. This underlying form is sometimes called the *parent* distribution. In Fig. 3.3, the smooth curve is a plot of a possible parent distribution, in this case, a *normal distribution* with a mean and standard deviation estimated from the data.

There are several important features of the parent distribution shown in Fig. 3.3. First, it can be represented by a mathematical equation – a distribution function – with a relatively small number of parameters. For the normal distribution, the parameters are the mean and population standard deviation. Knowing that the parent distribution is normal, it is possible to summarize a large number of observations simply by giving the mean and standard deviation. This allows large sets of observations to be summarized in terms of the distribution type and the relevant parameters. Second, the distribution can be used predictively to make statements about the likelihood of further observations; in Fig. 3.3, for example, the curve indicates that observations in the region of 2750–2760 mg kg^{-1} will occur only rarely. The distribution is accordingly important in both describing data and in drawing inferences from the data.

Distributions of Measurement Data. Measurement data can often be expected to follow a normal distribution, and in considering statistical tests for ordinary cases, this will be the assumed distribution. However, some other distributions are important in particular circumstances. Table 3.1 lists some common distributions, whose general shape is shown in Fig. 3.4. The most important features of each are

- The normal distribution is described by two independent parameters: the mean and standard deviation. The mean can take any value, and the standard

Table 3.1 Common distributions in measurement data

Distribution	Density function	Mean	Expected variance	Remarks
Normal	$\frac{1}{\sigma\sqrt{2\pi}}\exp\left(\frac{(x-\mu)^2}{2\sigma^2}\right)$	μ	σ^2	Arises naturally from the summation of many small random errors from any distribution
Lognormal	$\frac{1}{\sigma\sqrt{2\pi}}\exp\left(\frac{(\ln(x)-\mu)^2}{2\sigma^2}\right)$	$\exp\left(\mu+\frac{\sigma^2}{2}\right)$	$\exp\left(2\mu+\sigma^2\right)\left[\exp(\sigma^2)-1\right]$	Arises naturally from the product of many terms with random errors. Approximates to normal for small standard deviation
Poisson	$\lambda^{x\exp(-\lambda)/x!}$	λ	λ	Distribution of events occuring in an interval; important for radiation counting. Approximates to normality for large λ
Binomial	$\binom{n}{x}p^x(1-p)^{(n-x)}$	np	$np(1-p)$	Distribution of x, the number of successes in n trials with probability of success p. Common in counting at low to moderate levels, such as microbial counts; also relevant in situations dominated by particulate sampling
Contaminated normal	Various			Contaminated normal is the most common assumption given the presence of a small proportion of aberrant results. The *correct* data follow a normal distribution; aberrant results follow a different, usually much broader, distribution

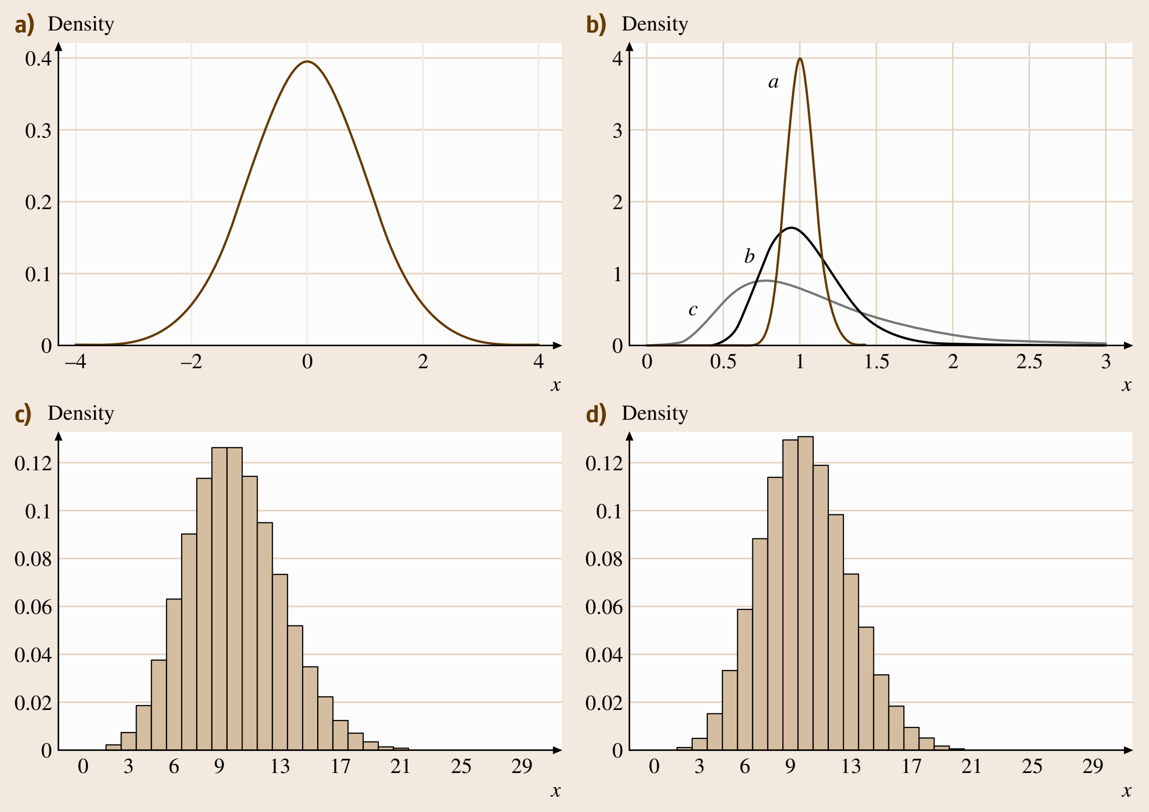


Fig. 3.4a–d Measurement data distributions. Figure 3.4 shows the *probability density function* for each distribution, not the probability; the area under each curve, or sum of discrete values, is equal to 1. Unlike probability, the probability density at a point x can be higher than 1. **(a)** The standard normal distribution (mean $=0$, standard deviation $=1.0$). **(b)** Lognormal distributions; mean on log scale: 0, standard deviation on log scale $= a$: 0.1, b: 0.25, c: 0.5. **(c)** Poisson distribution: lambda $=10$. **(d)** Binomial distribution: 100 trials, p(success) $=0.1$. Note that this provides the same mean as **(c)**

deviation any nonnegative value. The distribution is symmetric about the mean, and although the density falls off sharply, it is actually infinite in extent. The normal distribution arises naturally from the additive combination of many effects, even, according to the central limit theorem, when those effects do not themselves arise from a normal distribution. (This has an important consequence for means; errors in the mean of even three or four observations can often be taken to be normally distributed even where the parent distribution is not.) Furthermore, since small effects generally behave approximately additively, a very wide range of measurement systems show approximately normally distributed error.

- The lognormal distribution is closely related to the normal distribution; the logarithms of values from a lognormal distribution are normally distributed. It most commonly arises when errors combine multiplicatively, instead of additively. The lognormal distribution itself is generally asymmetric, with positive skew. However, as shown in the figure, the shape depends on the ratio of standard deviation to mean, and approaches that of a normal distribution as the standard deviation becomes small compared with the mean. The simplest method of handling lognormally distributed data is to take logarithms and treat the logged data as arising from a normal distribution. As the standard deviation becomes small relative to the mean, the lognormal distribution tends towards the normal distribution.
- The Poisson and binomial distributions describe counts, and accordingly are *discrete* distributions;

they have nonzero density only for integer values of the variable. The Poisson distribution is applicable to cases such as radiation counting; the binomial distribution is most appropriate for systems dominated by sampling, such as the number of defective parts in a batch, the number of microbes in a fixed volume or the number of contaminated particles in a sample from an inhomogeneous mixture. In the limit of large counts, the binomial distribution tends to the normal distribution; for small probability, it tends to the Poisson distribution. Similarly, the Poisson distribution tends towards normality for small probability and large counts. Thus, the Poisson distribution is often a convenient approximation to the binomial, and as counts increase, the normal distribution can be used to approximate either.

Distributions Derived from the Normal Distribution. Before leaving the topic of distributions, it is important to be aware that other distributions are important in analyzing measurement data with normally distributed error. The most important for this discussion are

- the t-distribution, which describes the distribution of the means of small samples taken from a normal distribution. The t-distribution is routinely used for checking a method for significant bias or for comparing observations with limits,
- the chi-squared distribution, which describes inter alia the distribution of estimates of variance. Specifically, the variable $(n-1)s^2/\sigma^2$ has a chi-squared distribution with $\nu = n-1$ degrees of freedom. The chi-squared distribution is asymmetric with mean ν and variance 2ν,
- the F-distribution, which describes the distribution of ratios of variances. This is important in comparing the spread of two different data sets, and is extensively used in analysis of variance as well as being useful for comparing the precision of alternative methods of measurement.

Probability and Significance

Given a particular distribution, it is possible to make predictions of the probability that observations will fall within a particular range. For example, in a normal distribution, the fraction of observations falling, by chance, within two standard deviations of the mean value is very close to 95%. This equates to the probability of an observation occurring in that interval. Similarly, the probability of an observation falling more than 1.65 standard deviations above the mean value is close to 5%. These proportions can be calculated directly from the area under the curves shown in Fig. 3.4, and are available in tabular form, from statistical software and from most ordinary spreadsheet software.

Knowledge of the probability of a particular observation allows some statement about the significance of an observation. Observations with high probability of chance occurrence are not regarded as particularly significant; conversely, observations with a low probability of occurring by chance are taken as significant. Notice that an observation can only be allocated a probability if there is some assumption or hypothesis about the true state of affairs. For example, if it is asserted that the concentration of a contaminant is below some regulatory limit, it is meaningful to consider how likely a particular observation would be *given this hypothesis*. In the absence of any hypothesis, no observation is more likely than any other. This process of forming a hypothesis and then assessing the probability of a particular observation *given the hypothesis* is the basis of significance testing, and will be discussed in detail below.

3.3.2 Calculations and Software

Statistical treatment of data generally involves calculations, and often repetitive calculation. Frequently, too, best practise involves methods that are simply not practical manually, or require numerical solutions. Suitable software is therefore essential. Purpose-designed software for statistics and experimental design is widely available, including some free and open-source packages whose reliability challenges the best commercial software. Some such packages are listed in Sect. 3.12 at the end of this chapter. Many of the tests and graphical methods described in this short introduction are also routinely available in general-purpose spreadsheet packages. Given the wide availability of software and the practical difficulties of implementing accurate numerical software, calculations will not generally be described in detail. Readers should consult existing texts or software for further details if required.

However, it remains important that the software used is reliable. This is particularly true of some of the most popular business spreadsheet packages, which have proven notoriously inaccurate or unstable on even moderately ill-conditioned data sets. Any mathematical software used in a measurement laboratory should therefore be checked using typical measurement data to ensure that the numerical accuracy is sufficient. It may additionally be useful to test software using more extreme test sets; some such sets are freely available (Sect. 3.12).

3.3.3 Statistical Methods

Graphical Methods

Graphical methods refer to the range of graphs or plots that are used to present and assess data visually. Some have already been presented; the histogram in Fig. 3.3 is an example. Graphical methods are easy to implement with a variety of software and allow a measurement scientist to identify anomalies, such as outlying data points or groups, departures from assumed distributions or models, and unexpected trends, quickly and with minimal calculation. A complete discussion of graphical methods is beyond the scope of this chapter, but some of the most useful, with typical applications, are presented below. Their use is strongly recommended in routine data analysis.

Figure 3.5 illustrates some basic plots appropriate for reviewing simple one-dimensional data sets. Dot plots and strip charts are useful for reviewing small data sets. Both give a good indication of possible outliers and unusual clustering. Overinterpretation should be avoided; it is useful to gain experience by reviewing plots from random normal samples, which will quickly indicate the typical extent of apparent anomalies in small samples. Strip charts are simpler to generate (plot the data as the x variable with a constant value of y), but overlap can obscure clustering for even modest sets. The stacked dot plot, if available, is applicable to larger sets. Histograms become more appropriate as the number of data points increases. Box plots, or box-and-whisker plots (named for the lines extending from the rectangular box) are useful for summarizing the general shape and extent of data, and are particularly useful for grouped data. For example, the range of data from replicate measurements on several different test items can be reviewed very easily using a box plot. Box plots can represent several descriptive statistics, including, for example, a mean and confidence interval. However, they are most commonly based on quantiles. Traditionally,

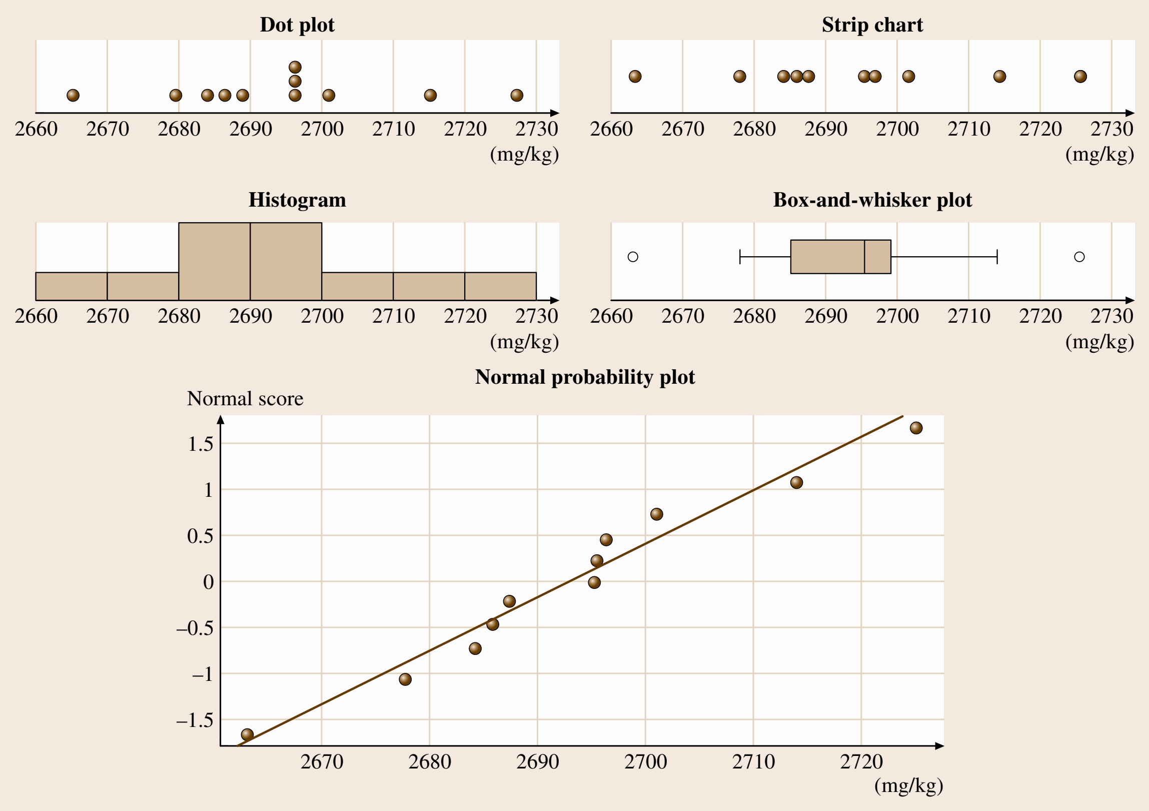


Fig. 3.5 Plots for simple data set review

the box extends from the first to the third quartile (that is, it contains the central 50% of the data points). The median is marked as a dividing line or other marker inside the box. The *whiskers* traditionally extend to the most distant data point within 1.5 times the interquartile range of the ends of the box. For a normal distribution, this would correspond to approximately the mean ±2.7 standard deviations. Since this is just beyond the 99% confidence interval, more extreme points are likely to be outliers, and are therefore generally shown as individual points on the plot. Finally, a normal probability plot shows the distribution of the data plotted against the expected distribution assuming normality. In a normally distributed data set, points fall close to the diagonal line. Substantial deviations, particularly at either end of the plot, indicate nonnormality.

The most common graphical method for two-dimensional measurement data (such as measurand level/instrument response pairs) is a scatter plot, in which points are plotted on a two-dimensional space with dimensions corresponding to the dimensions of the data set. Scatter plots are most useful in reviewing data for linear regression, and the topic will accordingly be returned to below.

Planning of Experiments

Most measurements represent straightforward application of a measuring device or method to a test item. However, many experiments are intended to test for the presence or absence of some specific treatment effect – such as the effect of changing a measurement method or adjusting a manufacturing method. For example, one might wish to assess whether a reduction in preconditioning time had an effect on measurement results. In these cases, it is important that the experiment measures the intended effect, and not some external nuisance effect. For example, measurement systems often show significant changes from day to day or operator to operator. To continue the preconditioning example, if test items for short preconditioning were obtained by one operator and for long preconditioning by a different operator, operator effects might be misinterpreted as a significant conditioning effect. Ensuring that nuisance parameters do not interfere with the result of an experiment is one of the aims of good experimental design.

A second, but often equally important aim is to minimize the cost of an experiment. For example, a naïve experiment to investigate six possible effects might investigate each individually, using, say, three replicate measurements at each level for each effect: a total of 36 measurements. Careful experimental designs which vary all parameters simultaneously can, using the right statistical methods, reduce this to 16 or even 8 measurements and still achieve acceptable power.

Experimental design is a substantial topic, and a range of reference texts and software are available. Some of the basic principles of good design are, however, summarized below.

1. *Arrange experiments for cancelation*: the most precise and accurate measurements seek to cancel out sources of bias. For example, null-point methods, in which a reference and test item are compared directly by adjusting an instrument to give a zero reading, are very effective in removing bias due to residual current flow in an instrument. Simultaneous measurement of test item and calibrant reduces calibration differences; examples include the use of internal standards in chemical measurement, and the use of comparator instruments in gage block calibration. Difference and ratio experiments also tend to reduce the effects of bias; it is therefore often better to study differences or ratios of responses obtained under identical conditions than to compare absolute measurements.
2. *Control if you can; randomize if you cannot*: a good experimenter will identify the main sources of bias and control them. For example, if temperature is an issue, temperature should be controlled as far as possible. If direct control is impossible, the statistical analysis should include the nuisance parameter. *Blocking* – systematic allocation of test items to different strata – can also help reduce bias. For example, in a 2 day experiment, ensuring that every type of test item is measured an equal number of times on each day will allow statistical analysis to remove the between-day effect. Where an effect is known but cannot be controlled, and also to guard against unknown systematic effects, randomization should be used. For example, measurements should always be made in random order within blocks as far as possible (although the order should be recorded to allow trends to be identified), and test items should be assigned randomly to treatments.
3. *Plan for replication or obtaining independent uncertainty estimates*: without knowledge of the precision available, and more generally of the uncertainty, the experiment cannot be interpreted. Statistical tests all rely on comparison of an effect with some estimate of the uncertainty of the effect, usually based on observed precision. Thus, exper-

iments should always include some replication to allow precision to be estimated, or provide for additional information of the uncertainty.

4. *Design for statistical analysis*: *To consult a statistician after an experiment is finished is often merely to ask him to conduct a post-mortem examination. He can perhaps say what the experiment died of.* (R. A. Fisher, Presidential Address to the First Indian Statistical Congress, 1938). An experiment should always be planned with a specific method of statistical analysis in mind. Otherwise, despite the considerable range of tools available, there is too high a risk that no statistical analysis will be applicable. One particular issue in this context is that of *balance*. Many experiments test several parameters simultaneously. If more data are obtained on some combinations than others, it may be impossible to separate the different effects. This applies particularly to two-way or higher-order analysis of variance, in which interaction terms are not generally interpretable with unbalanced designs. Imbalance can be tolerated in some types of analysis, but not in all.

Significance Testing

General Principles. Because measurement results vary, there is always some doubt about whether an observed difference arises from chance variation or from an underlying, real difference. Significance testing allows the scientist to make reliable objective judgements on the basis of data gathered from experiments, while protecting against overinterpretation based on chance differences.

A significance test starts with some hypothesis about a true value or values, and then determines whether the observations – which may or may not appear to contradict the hypothesis – could reasonably arise by chance if the hypothesis were correct. Significance tests therefore involve the following general steps.

1. *State the question clearly, in terms of a* null hypothesis *and an* alternate hypothesis: in most significance testing, the null hypothesis is that there is no effect of interest. The alternate is always an alternative state of affairs such that the two hypotheses are mutually exclusive and that the combined probability of one or the other is equal to 1; that is, that no other situation is relevant. For example, a common null hypothesis about a difference between two values is: *there is no difference between the true values* ($\mu_1 = \mu_2$). The relevant alternate is that *there* is *a difference between the true values* ($\mu_1 \neq \mu_2$). The two are mutually exclusive (they cannot both be true simultaneously) and it is certain that one of them is true, so the combined probability is exactly 1.0. The importance of the hypotheses is that different initial hypotheses lead to different estimates of the probability of a contradictory observation. For example, if it is hypothesized that the (true) value of the measurand is exactly equal to some reference value, there is some probability (usually equal) of contradictory observations both above and below the reference value. If, on the other hand, it is hypothesized that the true value is less than or equal to the reference value, the situation changes. If the true value may be anywhere below or equal to the reference value, it is less likely that observations above the reference value will occur, because of the reduced chance of such observations from true values very far below the reference value. This change in probability of observations on one side or another must be reflected either in the choice of critical value, or in the method of calculation of the probability.
2. *Select an appropriate test*: different questions require different tests; so do different distribution assumptions. Table 3.2 provides a summary of the tests appropriate for a range of common situations. Each test dictates the method of calculating a value called the *test statistic* from the data.
3. *Calculate the test statistic*: in software, the test statistic is usually calculated automatically, based on the test chosen.
4. *Choose a significance level*: the significance level is the probability at which chance is deemed sufficiently unlikely to justify rejection of the null hypothesis. It is usually the measurement scientist's responsibility to choose the level of significance appropriately. For most common tests on measurement results, the significance level is set at 0.05,

Table 3.2 Common significance tests for normally distribution data. The following symbols are used: α is the desired significance level (usually 0.05); μ is the (true) value of the measurand; σ is the population standard deviation for the population described by μ (*not* that calculated from the data). $\overline{a}$ is the observed mean; s is the standard deviation of the data used to calculate $\overline{x}$; n is the number of data points. x_0 is the reference value; x_U, x_L are the upper and lower limits of a range. $\mu_1, \mu_2, \overline{x_1}, \overline{x_2}, s_1, s_2, n_1, n_2$ are the corresponding values for each of two sets of data to be compared ▶

Test objective	Test name	Test statistic	Remarks
Tests on a single observed mean $\overline{x}$ against a reference value or range			
Test for significant difference from the reference value x_0	Student t-test	$\lvert x_0-\overline{x}\rvert\,/(s/\sqrt{n})$	Hypothesis $(\mu=x_0)$ against alternate $(x_0\neq\mu)$. Use a table of two-tailed critical values
Test for $\overline{x}$ significantly exceeding an upper limit x_0	Student t-test	$(\overline{x}-x_0)/(s/\sqrt{n})$	Hypothesis $(\mu_1=\mu_2)$ against alternate $(\mu_1\neg\mu_2)$. Use a table of *one*-tailed critical values. Note that the sign of $x_0-\overline{x}$ is retained
Test for $\overline{x}$ falling significantly below a lower limit x_0	Student t-test	$(x_0-\overline{x})/(s/\sqrt{n})$	
Test for $\overline{x}$ falling significantly outside a range $[x_L, x_U]$	Student t-test	$\max\begin{bmatrix}(x_L-\overline{x})/(s/\sqrt{n})\\(\overline{x}-x_U)/(s/\sqrt{n})\end{bmatrix}$	Hypothesis: $x_L\leq\mu\leq x_U$ against the alternate $\mu<x_L$, $x_U<\mu$. Use a table of *one*-tailed critical values. This test assumes that the range is large compared with s, but $(x_U-x_L)>s$ gives adequate accuracy at the 5% significance level
Tests for significant difference between two means			
(a) With equal variance	Equal-variance t-test	$\lvert\overline{x}_1-\overline{x}_2\rvert$	Hypothesis $\mu_1=\mu_2$ against alternate $(\mu_1\neq\mu_2)$
(b) With significantly different variance	Unequal-variance t-test		Use a table of two-tailed critical values. For equal variance, take degrees of freedom equal to n_1+n_2-2. For unequal variance, take degrees of freedom equal to $\dfrac{\left(s_1^2/n_1+s_2^2/n_2\right)^2}{\left[(n_1-1)/(s_1/n_1)^2+(n_2-1)(s_2/n_2)^2\right]}$
For testing the hypothesis $\mu_1>\mu_2$ against the alternative $\mu_1\leq\mu_2$, where μ_1 is the *expected* larger mean (not necessarily the larger observed mean), calculate the test statistic using $(\overline{x}_1-\overline{x}_2)$ instead of $\lvert\overline{x}_1-\overline{x}_2\rvert$ and use a one-tailed critical value			
Test n paired values for significant difference (constant variance)	Paired t-test	$\lvert\overline{d}\rvert\,/\left(s_d/\sqrt{n}\right)$, where $\overline{d}=\dfrac{1}{n}\sum_i x_{1,i}-x_{2,i}$ and $s_d=\dfrac{1}{n-1}\sum_i\left(x_{1,i}-x_{2,i}\right)^2$	Hypothesis $\mu_d=0$ against alternate $\mu_d\neq0$. The sets must consist of pairs of measurements, such as measurements on the same test items by two different methods
Tests for standard deviations			
Test an observed standard deviation against a reference or required value σ_0	i) Chi-squared test	i) $(n-1)s^2/\sigma_0$	i) Compare $(n-1)s^2/\sigma_0$ with critical values for the chi-squared distribution with $n-1$ degrees of freedom
	ii) F-test	ii) s^2/σ_0	ii) Compare s^2/σ_0 with critical values for F for $(n-1)$ and infinite degrees of freedom For a test of $\sigma\leq\sigma_0$ against $\sigma>\sigma_0$, use the upper one-tailed critical value of chi-squared or F for probability α. To test $\sigma=\sigma_0$ against $\sigma\neq\sigma_0$, use two-tailed limits for chi-sqared or compare max $(s^2/\sigma_0, \sigma_0/s^2)$ against the upper one-tailed value for F for probability $\alpha/2$
Test for a significant difference between two observed standard deviations	F-test	$s_{\max}^2/s_{\min}^2$	Hypothesis: $\sigma_1=\sigma_2$ against $\sigma_1\neq\sigma_2$. $s_{\max}$ is the larger observed standard deviation. Use the upper *one*-tailed critical value for F for a probability $\alpha/2$ using n_1-1, n_2-1 degrees of freedom
Test for one observed standard deviations s_1 significantly exceeding another (s_2)	F-test	s_1^2/s_2^2	Hypothesis: $\sigma_1\leq\sigma_2$ against $\sigma_1>\sigma_2$. Use the upper *one*-tailed critical value for F for a probability α using n_1-1, n_2-1 degrees of freedom
Test for homogeneity of variance among several groups of data	Levene's test	N/A	Levene's test is most simply estimated as a one-way analysis of variance performed on absolute values of group residuals, that is, $\lvert x_{ij}-\hat{x}_j\rvert$, where $\hat{x}_j$ is an estimate of the population mean of group j; $\hat{x}_j$ is usually the median, but the mean or another robust value can be used

or 5%. For stringent tests, 1% significance or less may be appropriate. The term *level of confidence* is an alternative expression of the same quantity; for example, the 5% level of significance is equal to the 95% level of confidence. Mathematically, the significance level is the probability of incorrectly rejecting the null hypothesis given a particular critical value for a test statistic (see below). Thus, one chooses the critical value to provide a suitable significance level.

5. *Calculate the degrees of freedom for the test*: the distribution of error often depends not only on the number of observations n, but on the number of degrees of freedom ν (Greek letter nu). ν is usually equal to the number of observations minus the number of parameters estimated from the data: $n-1$ for a simple mean value, for example. For experiments involving many parameters or many distinct groups, the number of degrees of freedom may be very different from the number of observations. The number of degrees of freedom is usually calculated automatically in software.
6. *Obtain a critical value*: critical values are obtained from tables for the relevant distribution, or from software. Statistical software usually calculates the critical value automatically given the level of significance.
7. *Compare the test statistic with the critical value or examine the calculated probability (p-value).* Traditionally, the test is completed by comparing the calculated value of the test statistic with the critical value determined from tables or software. Usually (but not always) a calculated value higher than the critical value denotes significance at the chosen level of significance. In software, it is generally more convenient to examine the calculated probability of the observed test statistic, or p-value, which is usually part of the output. The p-value is always between 0 and 1; small values indicate a low probability of chance occurrence. Thus, if the p-value is below the chosen level of significance, the result of the test is significant and the null hypothesis is rejected.

Significance Tests for Specific Circumstances. Table 3.2 provides a summary of the most common significance tests used in measurement for normally distributed data. The calculations for the relevant test statistics are included, although most are calculated automatically by software.

Interpretation of Significance Test Results. While a significance test provides information on whether an observed difference could arise by chance, it is important to remember that statistical significance does not necessarily equate to practical importance. Given sufficient data, very small differences can be detected. It does not follow that such small differences are important. For example, given good precision, a measured mean 2% away from a reference value may be statistically significant. If the measurement requirement is to determine a value within 10%, however, the 2% bias has little practical importance.

The other chief limitation of significance testing is that a lack of statistical significance cannot prove the absence of an effect. It should be interpreted only as an indication that the experiment failed to provide sufficient evidence to conclude that there was an effect. At best, statistical insignificance shows only that the effect is not large compared with the experimental precision available. Where many experiments fail to find a significant effect, of course, it becomes increasingly safe to conclude that there is none.

Effect of Nonconstant Standard Deviation. Significance tests on means assume that the standard deviation is a good estimate of the population standard deviation and that it is constant with μ. This assumption breaks down, for example, if the standard deviation is approximately proportional to μ, a common observation in many fields of measurement (including analytical chemistry and radiological counting, although the latter would use intervals based on the Poisson distribution). In conducting a significance test in such circumstances, the test should be based on the best estimate of the standard deviation at the hypothesized value of μ, and not that at the value $\bar{x}$. To take a specific example, in calculating whether a measured value significantly exceeds a limit, the test should be based on the standard deviation at the limit, not at the observed value.

Fortunately, this is only a problem when the standard deviation depends very strongly on μ in the range of interest and where the standard deviation is large compared with the mean to be tested. For $s/\bar{x}$ less than about 0.1, for example, it is rarely important.

Confidence Intervals

Statistical Basis of Confidence Intervals. A confidence interval is an interval within which a statistic (such as a mean or a single observation) would be expected to be observed with a specified probability.

Significance tests are closely related to the idea of confidence intervals. Consider a test for significant difference between an observed mean $\bar{x}$ (taken from n values with standard deviation s) against a hypothesized measurand value μ. Using a t-test, the difference is considered significant at the level of confidence $1-\alpha$ if

$$\frac{|\bar{x}-\mu|}{s/\sqrt{n}} > t_{\alpha,\nu,2} ,$$

where $t_{\alpha,\nu,2}$ is the two-tailed critical value of Student's t at a level of significance α. The condition for an *insignificant* difference is therefore

$$\frac{|\bar{x}-\mu|}{s/\sqrt{n}} \le t_{\alpha,\nu,2} .$$

Rearranging gives $|\bar{x}-\mu| \le t_{\alpha,\nu,2} s/\sqrt{n}$, or equivalently, $-t_{\alpha,\nu,2} s/\sqrt{n} \le \bar{x}-\mu \le t_{\alpha,\nu,2} s/\sqrt{n}$. Adding $\bar{x}$ and adjusting signs and inequalities accordingly gives

$$\bar{x} - t_{\alpha,\nu,2} s \big/ \sqrt{n} \le \mu \le \bar{x} + t_{\alpha,\nu,2} s \big/ \sqrt{n} .$$

This interval is called the $1-\alpha$ *confidence interval* for μ. Any value of μ within this interval would be considered consistent with $\bar{x}$ under a t-test at significance level α.

Strictly, this confidence interval cannot be interpreted in terms of the probability that μ is within the interval $\bar{x} \pm t_{\alpha,\nu,2} s/\sqrt{n}$. It is, rather, that, in a long succession of similar experiments, a proportion $100(1-\alpha)\%$ of the calculated confidence intervals would be expected to contain the true mean μ. However, because the significance level α is chosen to ensure that this proportion is reasonably high, a confidence interval does give an indication of the range of values that can reasonably be attributed to the measurand, based on the statistical information available so far. (It will be seen later that other information may alter the range of values we may attribute to the measurand.)

For most practical purposes, the confidence interval is quoted at the 95% level of confidence. The value of t for 95% confidence is approximately 2.0 for large degrees of freedom; it is accordingly common to use the range $\bar{x} \pm 2s/\sqrt{n}$ as an approximate 95% confidence interval for the value of the measurand.

Note that, while the confidence interval is in this instance symmetrical about the measured mean value, this is by no means always the case. Confidence intervals based on Poisson distributions are markedly asymmetric, as are those for variances. Asymmetric confidence intervals can also be expected when the standard deviation varies strongly with μ, as noted above in relation to significance tests.

Before leaving the topic of confidence intervals, it is worth noting that the use of confidence intervals is not limited to mean values. Essentially any estimated parameter estimate has a confidence interval. It is often simpler to compare some hypothesized value of the parameter with the confidence interval than to carry out a significance test. For example, a simple test for significance of an intercept in linear regression (below) is to see whether the confidence interval for the intercept includes zero. If it does, the intercept is not statistically significant.

Analysis of Variance

Introduction to ANOVA. Analysis of variance (ANOVA) is a general tool for analyzing data grouped by some factor or factors of interest, such as laboratory, operator or temperature. ANOVA allows decisions on which factors are contributing significantly to the overall dispersion of the data. It can also provide a direct measure of the dispersion due to each factor.

Factors can be qualitative or quantitative. For example, replicate data from different laboratories are grouped by the qualitative factor *laboratory*. This single-factor data would require one-way analysis of variance. In an experiment to examine time and temperature effects on a reaction, the data are grouped by both time and temperature. Two factors require two-way analysis of variance. Each factor treated by ANOVA must take two or more values, or *levels*. A combination of factor levels is termed a *cell*, since it forms a cell in a table of data grouped by factor levels. Table 3.3 shows an example of data grouped by time and temperature. There are two factors (time and temperature), and each has three levels (distinct values). Each cell (that is, each time/temperature combination) holds two observations.

The calculations for ANOVA are best done using software. Software can automate the traditional manual calculation, or can use more general methods. For example, simple grouped data with equal numbers of replicates within each cell are relatively simple to ana-

Table 3.3 Example data for two-way ANOVA

Time (min)	Temperature (K)		
	298	315	330
10	6.4	11.9	13.5
10	8.4	4.8	16.7
12	7.8	10.6	17.6
12	10.1	11.9	14.8
9	1.5	8.1	13.2
9	3.9	7.6	15.6

lyze using summation and sums of squares. Where there are different numbers of replicates per cell (referred to as an unbalanced design), ANOVA is better carried out by linear modeling software. Indeed, this is often the default method in current statistical software packages. Fortunately, the output is generally similar whatever the process used. This section accordingly discusses the interpretation of output from ANOVA software, rather than the process itself.

One-Way ANOVA. One-way ANOVA operates on the assumption that there are two sources of variance in the data: an effect that causes the true mean values of groups to differ, and another that causes data within each group to disperse. In terms of a statistical model, the i-th observation in the j-th group, x_{ij}, is given by

$$x_{ij} = \mu + \delta j + \varepsilon_{ij} \; ,$$

where δ and ε are usually assumed to be normally distributed with mean 0 and standard deviations σ_b and σ_w, respectively. The subscripts "b" and "w" refer to the between-group effect and the within-group effect, respectively. A typical ANOVA table for one-way ANOVA is shown in Table 3.4 (The data analyzed are shown, to three figures only, in Table 3.5). The important features are

- The row labels, *Between groups* and *Within groups*, refer to the estimated contributions from each of the two effects in the model. The *Total* row refers to the total dispersion of the data.
- The columns "SS" and "df" are the sum of squares (actually, the sum of squared deviations from the relevant mean value) and the degrees of freedom for each effect. Notice that the total sum of squares and degrees of freedom are equal to the sum of those in the rows above; this is a general feature of ANOVA, and in fact the between-group SS and df can be calculated from the other two rows.
- The "MS" column refers to a quantity called the mean square for each effect. Calculated by dividing the sum of squares by the degrees of freedom, it can be shown that each mean square is an estimated variance. The between-group mean square ($\mathrm{MS_b}$) estimates $n_w\sigma_b^2 + \sigma_w^2$ (where n_w is the number of values in each group); the within-group mean square ($\mathrm{MS_w}$) estimates the within-group variance σ_w^2.

It follows that, if the between-group contribution were zero, the two mean squares should be the same, while if there were a real between-group effect, the between-group mean square would be larger than the within-group mean square. This allows a test for significance, specifically, a one-sided F-test. The table accordingly gives the calculated value for F ($= \mathrm{MS_b}/\mathrm{MS_w}$), the relevant critical value for F using the degrees of freedom shown, and the p-value, that is, the probability that $F \geq F_{\text{calc}}$ given the null hypothesis. In this table, the p-value is approximately 0.08, so in this instance, it is concluded that the difference is not statistically significant. By implication, the instruments under study show no significant differences.

Finally, one-way ANOVA is often used for interlaboratory data to calculate repeatability and reproducibility for a method or process. Under interlaboratory conditions, repeatability standard deviation s_r is simply $\sqrt{\mathrm{MS_w}}$.

The reproducibility standard deviation s_R is given by

$$s_R = \sqrt{\frac{\mathrm{MS_b} + (n_w - 1)\mathrm{MS_w}}{n_w}} \; .$$

Two-Way ANOVA. Two-way ANOVA is interpreted in a broadly similar manner. Each effect is allocated a row in an ANOVA table, and each main effect (that is, the effect of each factor) can be tested against the within-group term (often called the residual, or error, term in higher-order ANOVA tables). There is, however, one additional feature found in higher-order ANOVA tables: the presence of one or more *interaction terms*.

By way of example, Table 3.6 shows the two-way ANOVA table for the data in Table 3.3. Notice the *Interaction* row (in some software, this would be labeled *Time:Temperature* to denote which interaction it referred to). The presence of this row is best understood

Table 3.4 One-way ANOVA. Analysis of variance table

Source of variation	SS	df	MS	F	P-value	F_{crit}
Between groups	8.85	3	2.95	3.19	0.084	4.07
Within groups	7.41	8	0.93			
Total	16.26	11				

Table 3.5 One-way ANOVA. Data analyzed

Instrument			
A	***B***	***C***	***D***
58.58	59.89	60.76	61.80
60.15	61.02	60.78	60.60
59.65	61.40	62.90	62.50

by reference to a new statistical model

$$x_{ijk} = \mu + A_j + B_k + \mathrm{AB}_{jk} + \varepsilon_{ijk} \,.$$

Assume for the moment that the factor A relates to the columns in Table 3.3, and the factor B to the rows. This model says that each level j of factor A shifts all results in column j by an amount A_j, and each level k of factor B shifts all values in row k by an amount B_j. This alone would mean that the effect of factor A is independent of the level of factor B. Indeed it is perfectly possible to analyze the data using the statistical model $x_{ijk} = \mu + A_j + B_k + \varepsilon_{ijk}$ to determine these main effects – even without replication; this is the basis of so-called two-way ANOVA without replication. However, it is possible that the effects of A and B are not independent; perhaps the effect of factor A depends on the level of B. In a chemical reaction, this is not unusual; the effect of time on reaction yield is generally dependent on the temperature, and vice versa. The term AB_{jk} in the above model allows for this, by associating a possible additional effect with every combination of factor levels A and B. This is the interaction term, and is the term referred to by the *Interaction* row in Table 3.6. If it is significant with respect to the within-group, or error, term, this indicates that the effects of the two main factors are not independent.

In general, in an analysis of data on measurement systems, it is safe to assume that the levels of the factors A and B are chosen from a larger possible population. This situation is analyzed, in two-way ANOVA, as a *random-effects model*. Interpretation of the ANOVA table in this situation proceeds as follows.

1. Compare the interaction term with the within-group term.
2. If the interaction term is *not* significant, the main effects can be compared directly with the within-group term, as usually calculated in most ANOVA tables. In this situation, greater power can be obtained by pooling the within-group and interaction term, by adding the sums of squares and the degrees of freedom values, and calculating a new mean square from the new combined sum of squares and degrees of freedom. In Table 3.6, for example, the new mean square would be 4.7, and (more importantly) the degrees of freedom for the pooled effect would be 13, instead of 9. The resulting p-values for the main effects drop to 0.029 and 3×10^{-5} as a result. With statistical software, it is simpler to repeat the analysis omitting the interaction term, which gives the same results.
3. If the interaction term *is* significant, it should be concluded that, even if the main effects are not statistically significant in isolation, their combined effect is statistically significant. Furthermore, the effects are not independent of one another. For example, high temperature and long times might increase yield more than simply raising the temperature or extending the time in isolation. Second, compare the main effects with the interaction term (using an F-test on the mean squares) to establish whether each main effect has a statistically significant additional influence – that is, in addition to its effect in combination – on the results.

The analysis proceeds differently where both factors are *fixed effects*, that is, not drawn from a larger population. In such cases, all effects are compared directly with the within-group term.

Higher-order ANOVA models can be constructed using statistical software. It is perfectly possible to analyze simultaneously for any number of effects and all their interactions, given sufficient replication. However,

Table 3.6 Two-way ANOVA table

Source of variation	**SS**	**df**	**MS**	***F***	***P*-value**	F_{crit}
Time	44.6	2	22.3	4.4	0.047	4.26
Temperature	246.5	2	123.2	24.1	0.0002	4.26
Interaction	15.4	4	3.8	0.8	0.58	3.63
Within	46.0	9	5.1			
Total	352.5	17	154.5			

in two-way and higher-order ANOVA, some cautionary notes are important.

Assumptions in ANOVA. ANOVA (as presented above) assumes normality, and also assumes that the within-group variances arise from the same population. Departures from normality are not generally critical; most of the mean squares are related to sums of squares of group means, and as noted above, means tend to be normally distributed even where the parent distribution is nonnormal. However, severe outliers can have serious effects; a single severe outlier can inflate the within-group mean square drastically and thereby obscure significant main effects. Outliers can also lead to spurious significance – particularly for interaction terms – by moving individual group means. Careful inspection to detect outliers is accordingly important. Graphical methods, such as box plots, are ideal for this purpose, though other methods are commonly applied (see *Outlier detection* below).

The assumption of equal variance (homoscedasticity) is often more important in ANOVA than that of normality. Count data, for example, manifest a variance related to the mean count. This can cause seriously misleading interpretation. The general approach in such cases is to transform the data to give constant variance (not necessarily normality) for the transformed data. For example, Poisson-distributed count data, for which the variance is expected to be equal to the mean value, should be transformed by taking the square root of each value before analysis; this provides data that satisfies the assumption of homoscedasticity to a reasonable approximation.

Effect of Unbalanced Design. Two-way ANOVA usually assumes that the design is balanced, that is, all cells are populated and all contain equal numbers of observations. If this is not the case, the order that terms appear in the model becomes important, and changing the order can affect the apparent significance. Furthermore, the mean squares no longer estimate isolated effects, and comparisons no longer test useful hypotheses.

Advanced statistical software can address this issue to an extent, using various modified sums of squares (usually referred to as type II, III etc.). In practise, even these are not always sufficient. A more general approach is to proceed by constructing a linear model containing all the effects, then comparing the residual mean square with that for models constructed by omitting each main effect (or interaction term) in turn. Significant differences in the residual mean square indicate a significant effect, independently of the order of specification.

Least-Squares Linear Regression

Principles of Least-Squares Regression. Linear regression estimates the coefficients α_i of a model of the general form

$$Y = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \cdots + \alpha_n X_n ,$$

where, most generally, each variable X is a *basis function*, that is, some function of a measured variable. Thus, the term covers both multiple regression, in which each X may be a different quantity, and polynomial regression, in which successive basis functions X are increasing powers of the independent variable (e.g., x, x^2 etc.). Other forms are, of course, possible. These all fall into the class of linear regression because they are linear in the coefficients α_i, not because they are linear in the variable X. However, the most common use of linear regression in measurement is to estimate the coefficients in the simple model

$$Y = \alpha_0 + \alpha_1 X ,$$

and this simplest form – the form usually implied by the unqualified term *linear regression* – is the subject of this section.

The coefficients for the linear model above can be estimated using a surprisingly wide range of procedures, including robust procedures, which are resistant to the effects of outliers, and nonparametric methods, which make no distribution assumptions. In practise, by far the most common is simple least-squares linear regression, which provides the minimum-variance unbiased estimate of the coefficients when all errors are in the dependent variable Y and the error in Y is normally distributed. The statistical model for this situation is

$$y_i = \alpha_0 + \alpha_1 x_i + \varepsilon_i ,$$

where ε_i is the usual error term and α_i are the true values of the coefficients, with estimated values a_i. The coefficients are estimated by finding the values that minimize the sum of squares

$$\sum_i w_i \left[y_i - (a_0 + a_1 x_i) \right]^2 ,$$

where the w_i are weights chosen appropriately for the variance associated with each point y_i. Most simple regression software sets the weights equal to 1, implicitly assuming equal variance for all y_i. Another common procedure (rarely available in spreadsheet implementations) is to set $w_i = 1/s_i^2$, where s_i is the standard deviation at y_i; this *inverse variance weighting* is the correct weighting where the standard deviation varies significantly across the y_i.

The calculations are well documented elsewhere and, as usual, will be assumed to be carried out by software. The remainder of this section accordingly discusses the planning and interpretation of linear regression in measurement applications.

Planning for Linear Regression. Most applications of linear regression for measurement relate to the construction of a calibration curve (actually a straight line). The instrument response for a number of reference values is obtained, and the calculated coefficients a_i used to estimate the measurand value from signal responses on test items. There are two stages to this process. At the validation stage, the linearity of the response is checked. This generally requires sufficient power to detect departures from linearity and to investigate the dependence of precision on response. For routine measurement, it is sufficient to reestablish the calibration line for current circumstances; this generally requires sufficient uncertainty and some protection against erroneous readings or reference material preparation.

In the first, validation, study, a minimum of five levels, approximately equally spaced across the range of interest, are recommended. Replication is vital if a dependence of precision on response is likely; at least three replicates are usually required. Higher numbers of both levels and replication provide more power.

At the routine calibration stage, if the linearity is very well known over the range of interest and the intercept demonstrably insignificant, single-point calibration is feasible; two-point calibration may also be feasible if the intercept is nonzero. However, since there is then no possibility of checking either the internal consistency of the fit, or the quality of the fit, suitable quality control checks are essential in such cases. To provide additional checks, it is often useful to run a minimum of four to five levels; this allows checks for outlying values and for unsuspected nonlinearity. Of course, for extended calibration ranges, with less well-known linearity, it will be valuable to add further points. In the following discussion, it will be assumed that at least five levels are included.

Interpreting Regression Statistics. The first, and perhaps most important, check on the data is to inspect the fitted line visually, and wherever possible to check a residual plot. For unweighted regression (i. e., where $w_i = 1.0$) the residual plot is simply a scatter plot of the values $y_i - (a_0 + a_1 x_i)$ against x_i. Where weighted regression is used, it is more useful to plot the weighted residuals $w_i[y_i - (a_0 + a_1 x_i)]$. Figure 3.6 shows an example, including the fitted line and data (Fig. 3.6a) and the residual plot (Fig. 3.6b). The residual plot clearly provides a much more detailed picture of the dispersion around the line. It should be inspected for evidence of curvature, outlying points, and unexpected changes in precision. In Fig. 3.6, for example, there is no evidence of curvature, though there might be a high outlying point at $x_i = 1$.

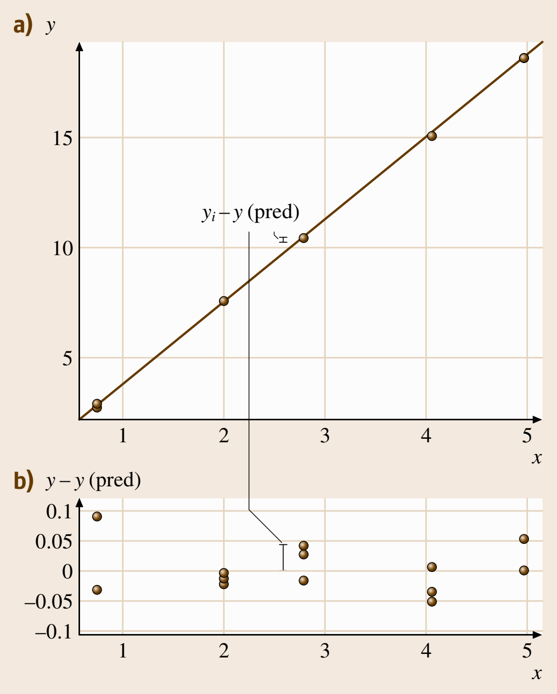


Fig. 3.6a,b Linear regression

Regression statistics include the correlation coefficient r (or r^2) and a derived correlation coefficient r^2 (adjusted), plus the regression parameters a_i and (usually) their standard errors, confidence intervals, and a p-value for each based on a t-test for difference compared with the null hypothesis of zero for each.

The regression coefficient is always in the range -1 to 1. Values nearer zero indicate a lack of linear relationship (not necessarily a lack of *any* relationship); values near 1 or -1 indicate a strong linear relationship. The regression coefficient will always be high when the data are clustered at the ends of the plot, which is why it is good practise to space points approximately evenly. Note that r and r^2 approach 1 as the number of degrees of freedom approaches zero, which can lead to overinterpretation. The adjusted r^2 value protects against this, as it decreases as the number of degrees of freedom reduces.

The regression parameters and their standard errors should be examined. Usually, in calibration, the intercept a_0 is of interest; if it is insignificant (judged by a high p-value, or a confidence interval including zero) it may reasonably be omitted in routine calibration. The slope a_1 should always be highly significant in any practical calibration. If a p-value is given for the regression as a whole, this indicates, again, whether there is a significant linear relationship; this is usually well known in calibration, though it is important in exploratory analysis (for example, when investigating a possible effect on results).

Prediction from Linear Calibration. If the regression statistics and residual plot are satisfactory, the curve can be used for prediction. Usually, this involves estimating a value x_0 from an observation y_0. This will, for many measurements, require some estimate of the uncertainty associated with prediction of a measurand value x from an observation y. Prediction uncertainties are, unfortunately, rarely available from regression software. The relevant expression is therefore given below.

$$s_{x_0} = \frac{s_{(y/x)}}{a_1}\left(w_0 + \frac{1}{n} + \frac{(y_0 - \bar{y}_\mathrm{w})^2}{a_1^2 \sum\left(w_i x_i^2 - n\bar{x}_\mathrm{w}^2\right)}\right)^{1/2} .$$

s_{x_0} is the *standard error of prediction* for a value x_0 predicted from an observation y_0; $s_{(y/x)}$ is the (weighted) residual standard deviation for the regression; $\bar{y}_\mathrm{w}$ and $\bar{x}_\mathrm{w}$ are the weighted means of the x and y data used in the calibration; n is the number of (x, y) pairs used; w_0 is a weighting factor for the observation y_0; if y_0 is a mean of n_0 observations, w_0 is $1/n_0$ if the calibration used unweighted regression, or is calculated as for the original data if weighting is used; s_{x_0} is the uncertainty arising from the calibration and precision of observation of y_0 in a predicted value x_0.

Outlier Detection

Identifying Outliers. Measurement data frequently contain a proportion of extreme values arising from procedural errors or, sometimes, unusual test items. It is, however, often difficult to distinguish erroneous values from chance variations, which can also give rise to occasional extreme values. Outlier detection methods help to distinguish between chance occurrence as part of the normal population of data, and values that cannot reasonably arise from random variability.

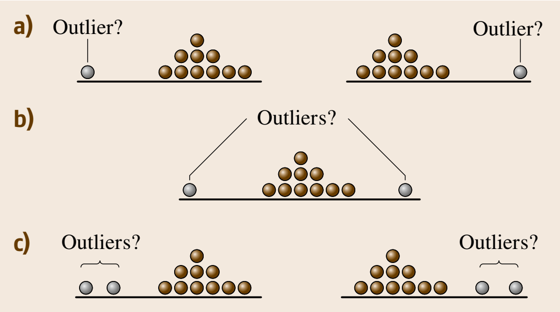


Fig. 3.7a–c Possible outliers in data sets

Graphical methods are effective in identifying possible outliers for follow-up. Dot plots make extreme values very obvious, though most sets have at least some apparent extreme values. Box-and-whisker plots provide an additional quantitative check; any single point beyond the whisker ends is unlikely to arise by chance in a small to medium data set. Graphical methods are usually adequate for the principal purpose of identifying data points which require closer inspection, to identify possible procedural errors. However, if critical decisions (including rejection – see below) are to be taken, or to protect against unwarranted follow-up work, graphical inspection should always be supported by statistical tests. A variety of tests are available; the most useful for measurement work are listed in Table 3.7. Grubb's tests are generally convenient (given the correct tables); they allow tests for single outliers in an otherwise normally distributed data set (Fig. 3.7a), and for simultaneous outlying pairs of extreme values (Fig. 3.7b, c), which would otherwise cause outlier tests to fail. Cochran's test is effective in identifying outlying variances, an important problem if data are to be subjected to analysis of variance or (sometimes) in quality control.

Successive application of outlier tests is permitted; it is not unusual to find that one exceptionally extreme value is accompanied by another, less extreme value. This simply involves testing the remainder of the data set after discovering an outlier.

Action on Detecting Outliers. A statistical outlier is only *unlikely* to arise by chance. In general, this is a signal to investigate and correct the cause of the problem. As a general rule, outliers should not be removed from the data set simply because of the result of a statistical test. However, many statistical procedures are seriously undermined by erroneous values, and long experience suggests that human error is the most common cause

Table 3.7 Tests for outliers in normally distributed data. The following assume an ordered set of data $x_1 \dots x_n$. Tables of critical values for the following can be found in ISO 5725:1995 part 2, among other sources. Symbols otherwise follow those in Table 3.2

Test objective	Test name	Test statistic	Remarks
Test for a single outlier in an otherwise normal distribution	i) Dixon's test	$n = \;\; 3 \dots \;\; 7:$ $(x_n - x_{n-1})/(x_n - x_1)$ $n = \;\; 8 \dots 10:$ $(x_n - x_{n-1})/(x_n - x_2)$ $n = 10 \dots 30:$ $(x_n - x_{n-2})/(x_n - x_3)$	The test statistics vary with the number of data points. Only the test statistic for a high outlier is shown; to calculate the test statistic for a low outlier, renumber the data in descending order. Critical values must be found from tables of Dixon's test statistic if not available in software
	ii) Grubb's test 1	$(x_n - \overline{x})/s$ (high outlier) $(\overline{x} - x_1)/s$ (low outlier)	Grubb's test is simpler than Dixon's test if using software, although critical values must again be found from tables if not available in software
Test for two outliers on opposite sides of an otherwise normal distribution	Grubb's test 2	$1 - \left(\frac{(n-3)[s(x_3 \dots x_n)]^2}{(n-1)s^2} \right)$	$s(x_3 \dots x_n)$ is the standard deviation for the data excluding the two suspected outliers. The test can be performed on data in both ascending and descending order to detect paired outliers at each end. Critical values must use the appropriate tables
Test for two outliers on the same side of an otherwise normal distribution	Grubb's test 3	$(x_n - x_1)/s$	Use tables for Grubb's test 3
Test for a single high variance in l groups of data	Cochran's test	$C_{\overline{n}} = \frac{(s^2)_{\max}}{\sum\limits_{i=1,l} s_i^2}$	$\overline{n} = \frac{1}{l} \sum\limits_{i=1,l} n_i$

of extreme outliers. This experience has given rise to some general rules which are often used in processing, for example, interlaboratory data.

1. Test at the 95% and the 99% confidence level.
2. All outliers should be investigated and any errors corrected.
3. Outliers significant at the 99% level may be rejected unless there is a technical reason to retain them.
4. Outliers significant only at the 95% level should be rejected only if there is an additional, technical reason to do so.
5. Successive testing and rejection is permitted, but not to the extent of rejecting a large proportion of the data.

This procedure leads to results which are not unduly biased by rejection of chance extreme values, but are relatively insensitive to outliers at the frequency commonly encountered in measurement work. Note, however, that this objective can be attained without outlier testing by using robust statistics where appropriate; this is the subject of the next section.

Finally, it is important to remember that an outlier is only outlying in relation to some prior expectation. The tests in Table 3.7 assume underlying normality. If the data were Poisson distributed, for example, too many high values would be rejected as inconsistent with a normal distribution. It is generally unsafe to reject, or even test for, outliers unless the underlying distribution is known.

Robust Statistics

Introduction. Instead of rejecting outliers, robust statistics uses methods which are less strongly affected by extreme values. A simple example of a robust estimate of a population mean is the median, which is essentially unaffected by the exact value of extreme points. For example, the median of the data set (1, 2, 3, 4, 6) is identical to that of (1, 2, 3, 4, 60). The median, however, is substantially more variable than the mean when the data are not outlier-contaminated. A variety of estimators have accordingly been developed that retain a useful degree of resistance to outliers without unduly affecting performance on normally distributed data. A short summary of the main estimators for means and standard deviations is given below. Robust methods also exist for analysis of variance, linear regression, and other modeling and estimation approaches.

Robust Estimators for Population Means. The median, as noted above, is a relatively robust estimator, widely available in software. It is very resistant to extreme values; up to half the data may go to infinity without affecting the median value. Another simple robust estimate is the so-called trimmed mean: the mean of the data set with two or more of the most extreme values removed. Both suffer from increases in variability for normally distributed data, the trimmed mean less so.

The mean suffers from outliers in part because it is a least-squares estimate, which effectively gives values a weight related to the square of their distance from the mean (that is, the *loss function* is quadratic). A general improvement can be obtained using methods which use a modified loss function. Huber (see Sect. 3.12 *Further Reading*) suggested a number of such estimators, which allocate a weight proportional to squared distance up to some multiple c of the estimated standard deviation $\hat{s}$ for the set, and thereafter a weight proportional to distance. Such estimators are called M-estimators, as they follow from maximum-likelihood considerations. In Huber's proposal, the algorithm used is to replace each value x_i in a data set with z_i, where

$$z_i = \begin{vmatrix} x_i & \text{if} \quad \hat{X} - c \times \hat{s} < x_i < \hat{X} + c \times \hat{s} \\ \hat{X} \pm c \times \hat{s} & \text{otherwise} \end{vmatrix} ,$$

and recalculate the mean $\hat{X}$, applying the process iteratively until the result converges. A suitable one-dimensional search algorithm may be faster. The estimated standard deviation is usually determined using a separate robust estimator, or (in Huber's proposal 2) iteratively, together with the mean. Another well-known approach is to use Tukey's biweight as the loss function; this also reduces the weight of extreme observations (to zero, for very extreme values).

Robust Estimators of Standard Deviation. Two common robust estimates of standard deviation are based on rank order statistics, such as the median. The first, the median absolute deviation (MAD), calculates the median of absolute deviations from the estimated mean value $\hat{x}$, that is, median $(|x_i - \hat{x}|)$. This value is not directly comparable to the standard deviation in the case of normally distributed data; to obtain an estimate of the standard deviation, a modification known as MAD_e should be used. This is calculated as MAD/0.6745. Another common estimate is based on the interquartile range (IQR) of a set of data; a normal distribution has standard deviation IQR/1.349. The IQR method is slightly more variable than the MAD_e method, but is usually easier to implement, as quartiles are frequently available in software. Huber's proposal 2 (above) generates a robust estimate of standard deviation as part of the procedure; this estimate is expected to be identical to the usual standard deviation for normally distributed data. ISO 5725 provides an alternative iterative procedure for a robust standard deviation independently of the mean.

Using Robust Estimators. Robust estimators can be thought of as providing good estimates of the parameters for the good data in an outlier-contaminated set. They are appropriate when

- The data are expected to be normally distributed. Here, robust statistics give answers very close to ordinary statistics.
- The data are expected to be normally distributed, but contaminated with occasional spurious values, which are regarded as unrepresentative or erroneous. Here, robust estimators are less affected by occasional extreme values and their use is recommended. Examples include setting up quality control (QC) charts from real historical data with occasional errors, and interpreting interlaboratory study data with occasional problem observations.

Robust estimators are not recommended where

- The data are expected to follow nonnormal distributions, such as binomial, Poisson, chi-squared, etc. These generate extreme values with reasonable likelihood, and robust estimates based on assumptions of underlying normality are not appropriate.
- Statistics that represent the whole data distribution (including extreme values, outliers, and errors) are required.

3.3.4 Statistics for Quality Control

Principles

Quality control applies statistical concepts to monitor processes, including measurement processes, and detect significant departures from normal operation. The general approach to statistical quality control for a measurement process is

1. regularly measure one or more typical test items (control materials),
2. establish the mean and standard deviation of the values obtained over time (ignoring any erroneous results),
3. use these parameters to set up warning and action criteria.

The criteria can include checks on stability of the mean value and, where measurements on the control material are replicated, on the precision of the process. It is also possible to seek evidence of emerging trends in the data, which might warn of impending or actual problems with the process.

The criteria can be in the form of, for example, permitted ranges for duplicate measurements, or a range within which the mean value for the control material must fall. Perhaps the most generally useful implementation, however, is in the form of a control chart. The following section therefore describes a simple control chart for monitoring measurement processes.

There is an extensive literature on statistical process control and control charting in particular, including a wide range of methods. Some useful references are included in Sect. 3.12 *Further Reading*.

Control Charting

A control chart is a graphical means of monitoring a measurement process, using observations plotted in a time-ordered sequence. Several varieties are in common use, including cusum charts (sensitive to sustained small bias) and range charts, which control precision. The type described here is based on a Shewhart mean chart. To construct the chart

- Obtain the mean $\bar{x}$ and standard deviation s of at least 20 observations (averages if replication is used) on a control material. Robust estimates are recommended for this purpose, but at least ensure that no erroneous or aberrant results are included in this preliminary data.
- Draw a chart with date as the x-axis, and a y-axis covering the range approximately $\bar{x} \pm 4s$.
- Draw the mean as a horizontal line on the chart. Add two *warning limits* as horizontal lines at $\bar{x} \pm 2s$, and two further *action limits* at $\bar{x} \pm 3s$. These limits are approximate. Exact limits for specific probabilities are provided in, for example, ISO 8258:1991 *Shewhart control charts*.

As further data points are accumulated, plot each new point on the chart. An example of such a chart is shown in Fig. 3.8.

Interpreting Control Chart Data

Two rules follow immediately from the action and warning limits marked on the chart.

- A point outside the action limits is very unlikely to arise by chance; the process should be regarded as out of control and the reason investigated and corrected.
- A point between the warning and action limits could happen occasionally by chance (about 4–5% of the time). Unless there is additional evidence of loss of control, no action follows. It may be prudent to remeasure the control material.

Other rules follow from unlikely sequences of observations. For example, two points outside the warning limits – whether on one side or alternate sides – is very unlikely and should be treated as actionable. A string of seven or more points above, or below, the mean – whether within the warning limits or not – is unlikely and may indicate developing bias (some recommendations consider ten such successive points as actionable). Sets of such rules are available in most textbooks on statistical process control.

Action on Control Chart *Action* Conditions

In general, actionable conditions indicate a need for corrective action. However, it is prudent to check that the control material measurement is valid before undertaking expensive investigations or halting a process. Taking a second control measurement is therefore advised, particularly for warning conditions. However, it is not sensible to continue taking control measurements until one falls back inside the limits. A single remeasurement is sufficient for confirmation of the out-of-control condition.

If the results of the second check do not confirm the first, it is sensible to ask how best to use the duplicate data in coming to a final decision. For example, should one act on the second observation? Or perhaps take the mean of the two results? Strictly, the correct answer requires consideration of the precision of the means of duplicate measurements taken over the appropriate time interval. If this is available, the appropriate limits can

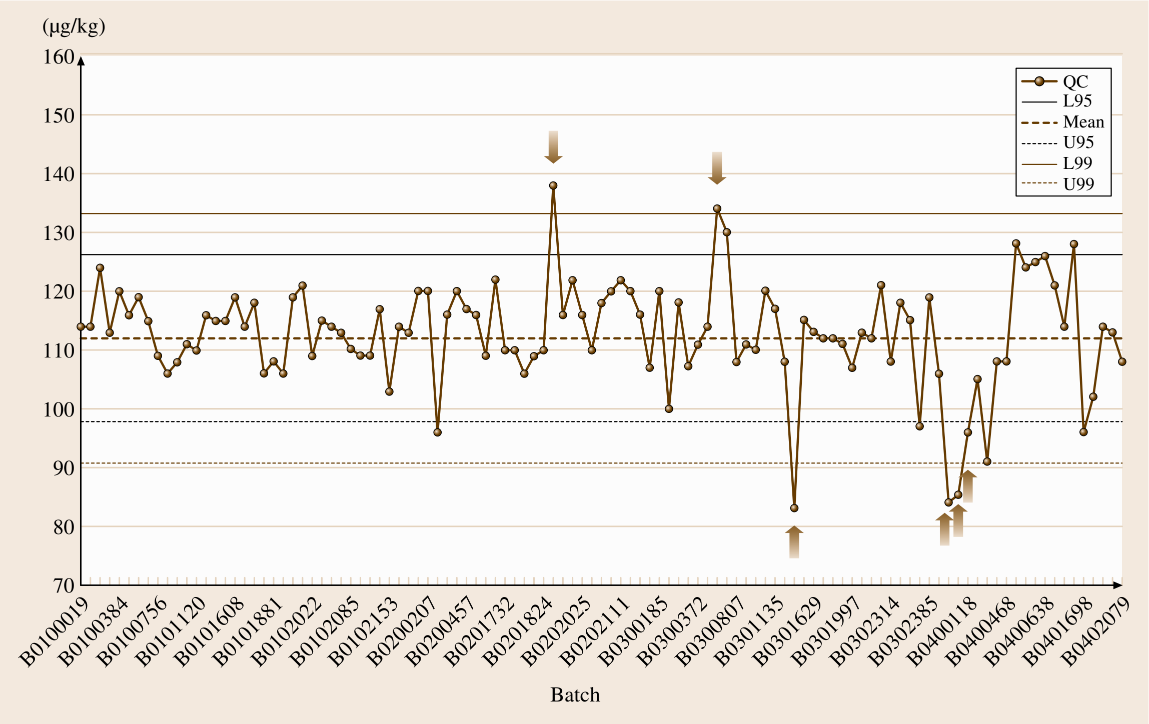


Fig. 3.8 QC chart example. The figure shows successive QC measurements on a reference material certified for lead content. There is evidence of loss of control at *points marked by arrows*

be calculated from the relevant standard deviation. If not, the following procedure is suggested: First, check whether the difference between the two observations is consistent with the usual operating precision (the results should be within approximately 2.8s of one another). If so, take the mean of the two, and compare this with new limits calculated as $\bar{x} \pm 2s/\sqrt{2}$ and $\bar{x} \pm 3s/\sqrt{2}$ (this is conservative, in that it assumes complete independence of successive QC measurements; it errs on the side of action). If the two results do not agree within the expected precision, the cause requires investigation and correction in any case.

3.4 Uncertainty and Accuracy of Measurement and Testing

3.4.1 General Principles

In metrology and testing, the result of a measurement should always be expressed as the measured quantity value together with its uncertainty. The uncertainty of measurement is defined as a nonnegative parameter characterizing the dispersion of the quantity values being attributed to a measurand [3.17].

Measurement accuracy, which is the closeness of agreement between a measured quantity value and the true quantity value of a measurand, is a positive formulation for the fact that the measured value is deviating from the true value, which is considered unique and, in practise, unknowable. The deviation between the measured value and the true value or a reference value is called the measurement error.

Since the 1990s there has been a conceptual change from the traditionally applied *error approach* to the *uncertainty approach*.

In the *error approach* it is the aim of a measurement to determine an *estimate of the true value* that is as close as possible to that single true value. In the *uncertainty approach* it is assumed that the information from measurement only permits assignment of an *interval of reasonable values* to the measurand.

The master document, which is acknowledged to apply to all measurement and testing fields and to all types of uncertainties of quantitative results, is the *Guide to the Expression of Uncertainty in Measurement* (GUM) [3.19]. The Joint Committee for Guides in Metrology Working Group 1 (JCGM-WG1), author of the GUM, is producing a complementary series of documents to accompany the GUM.

The *GUM uncertainty philosophy* has already been introduced in Chap. 1, its essential points are

- A measurement quantity X, of which the true value is not known exactly, is considered as a stochastic variable with a probability function. Often it is assumed that this is a normal (*Gaussian*) distribution.
- The result x of a measurement is an estimate of the expectation value $E(X)$ for X.
- The standard uncertainty $u(x)$ of this measured value is equal to the square root of the variance $V(X)$.
- Expectation (quantity value) and variance (standard uncertainty) are estimated either
  - by statistical processing of repeated measurements (*type A uncertainty evaluation*) or
  - by other methods (*type B uncertainty evaluation*).
- The result of a measurement has to be expressed as a quantity value together with its uncertainty, including the unit of the measurand.

The methodology of measurement evaluation and determination of measurement uncertainty are compiled in Fig. 3.9. The statistical evaluation of results has been described in detail in Sect. 3.3.

3.4.2 Practical Example: Accuracy Classes of Measuring Instruments

All measurements of quantity values for single measurands as well as for multiple measurands need to be performed with appropriate measuring instruments, devices for making measurements, alone or in conjunction with one or more supplementary devices.

The quality of measuring instruments is often defined through *limits of errors* as description of the *accuracy*.

Accuracy classes are defined [3.17] as classes of measuring instruments or measuring systems that meet stated metrological requirements that are intended to keep measurement errors or instrumental measurement uncertainties within specified limits under specified operating conditions. An accuracy class is usually denoted by a number or symbol adopted by convention. Analog measuring instruments are divided conventionally into accuracy classes of 0.05, 0.1, 0.2, 0.3, 0.5, 1, 1.5, 2, 2.5, 3, and 5. The accuracy classes p represent the maximum permissible relative measurement error in %. For example an accuracy class of 1.0 indicates that the limits of error – in both directions – should not exceed 1% of the full-scale deflection. In digital instruments, the limit of indication error is ±1 of the least significant unit of the digital indication display.

In measuring instruments with an analog indication, the measured quantity is determined by the position

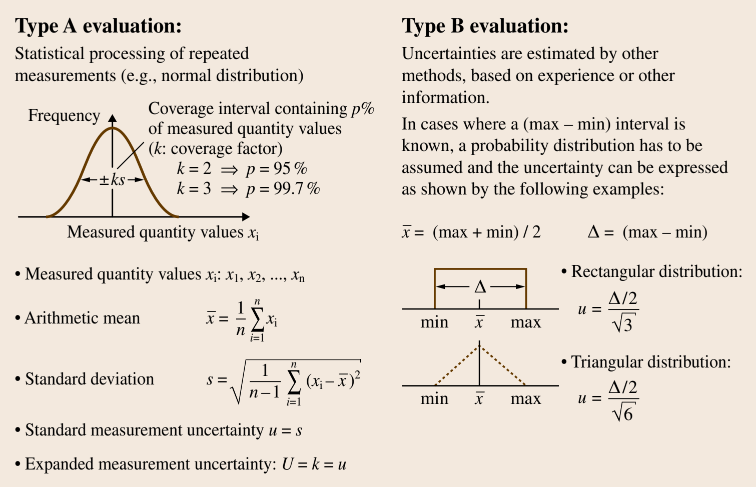
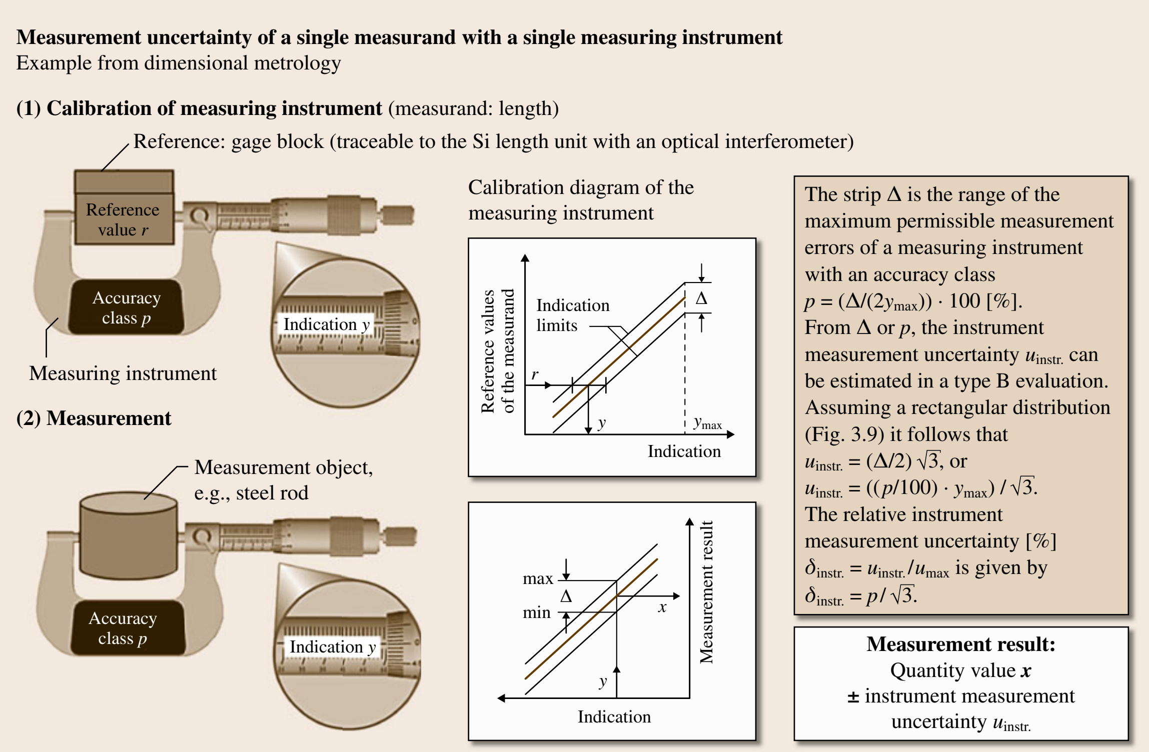

Type A evaluation:
Statistical processing of repeated measurements (e.g., normal distribution)
Frequency
Coverage interval containing p% of measured quantity values (k: coverage factor)
$k = 2 \Rightarrow p = 95\,\%$
$k = 3 \Rightarrow p = 99.7\,\%$
$\pm ks$
Measured quantity values x_i
• Measured quantity values x_i: $x_1, x_2, \ldots, x_n$
• Arithmetic mean $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
• Standard deviation $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$
• Standard measurement uncertainty $u = s$
• Expanded measurement uncertainty: $U = k = u$
Type B evaluation:
Uncertainties are estimated by other methods, based on experience or other information.
In cases where a (max – min) interval is known, a probability distribution has to be assumed and the uncertainty can be expressed as shown by the following examples:
$\bar{x} = (\max + \min)/2$
$\Delta = (\max - \min)$
Δ
min
$\bar{x}$
max
• Rectangular distribution:
$u = \frac{\Delta/2}{\sqrt{3}}$
min
$\bar{x}$
max
• Triangular distribution:
$u = \frac{\Delta/2}{\sqrt{6}}$

Fig. 3.9 Principles of measurement evaluation and determination of uncertainty of measurement for a single measurand x



Measurement uncertainty of a single measurand with a single measuring instrument
Example from dimensional metrology

(1) Calibration of measuring instrument (measurand: length)

Reference: gage block (traceable to the Si length unit with an optical interferometer)
Reference value r
Accuracy class p
Indication y
Measuring instrument
Calibration diagram of the measuring instrument
Reference values of the measurand
Indication limits
Δ
r
y
y_{max}
Indication

The strip Δ is the range of the maximum permissible measurement errors of a measuring instrument with an accuracy class $p = (\Delta/(2y_{max})) \cdot 100$ [%]. From Δ or p, the instrument measurement uncertainty $u_{instr.}$ can be estimated in a type B evaluation. Assuming a rectangular distribution (Fig. 3.9) it follows that $u_{instr.} = (\Delta/2)\sqrt{3}$, or $u_{instr.} = ((p/100) \cdot y_{max}) / \sqrt{3}$. The relative instrument measurement uncertainty [%] $\delta_{instr.} = u_{instr.}/u_{max}$ is given by $\delta_{instr.} = p/\sqrt{3}$.

(2) Measurement

Measurement result:
Quantity value $\boldsymbol{x}$
± instrument measurement
uncertainty $u_{instr.}$

Fig. 3.10 Method for obtaining a measurement result and estimating the instrument measurement uncertainty

of the indicator on the scale. The limits of errors (percentages) are usually given at the full-scale amplitude (maximum value of measurement range). From the accuracy class p, also the instrumental measurement uncertainty u_{instr} can be estimated. In Fig. 3.10, the method for obtaining a measurement result and measurement uncertainty for a single measurand with a single measuring instrument is shown.

As illustrated in Fig. 3.10, a measuring instrument gives as output an *indication*, which has to be related to the quantity value of the measurand through a calibration diagram. A calibration diagram represents the relation between indications of a measuring instrument and a set of reference values of the measurand. At the maximum indication value (maximum measurement range) y_{max} the width Δ of the strip of the calibration diagram is the range of the maximum permissible measurement errors.

From the determination of Δ the accuracy class p in % follows as

$$p = \left(\frac{\Delta}{(2y_{max})}\right) \cdot 100\ [\%]\ .$$

Note that, at indicator amplitudes lower than the maximum y_{max}, the actual relative maximum permissible measurement errors p_{act} for the position y_{act} on the scale need to be determined as

$$p_{act} = p \cdot \left(\frac{y_{max}}{y_{act}}\right)\ .$$

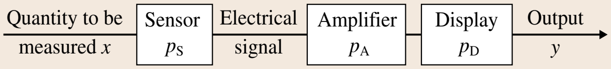
For the estimation of the standard measurement uncertainty it can be considered in an uncertainty estimation of type B that all values in the range between the limits of indications have the same probability – as long as no other information is available. This kind of distribution is called a rectangular distribution (Fig. 3.9). Therefore, the standard uncertainty is equal to

$$u_{instr} = \frac{(\Delta/2)}{\sqrt{3}} = \frac{((p/100) \cdot y_{max})}{\sqrt{3}}\ .$$

Example 3.1: What is the measurement uncertainty of a measurement result obtained by a measurement with an analog voltmeter (accuracy class 2.0) with a maximum amplitude of 380 V, when the indicator is at 220 V?

Measurement uncertainty of a measurement system or measurement chain

Consider a measurement system, consisting in the simplest case of three components, namely a sensor (accuracy class p_S), an amplifier (accuracy class p_A) and a display (accuracy class D_3)



The measurement uncertainty of the system can be estimated by applying the law of the propagation of uncertainties (see Sect. 3.4.3)

$$u_{System}/|y| = \sqrt{(u_S^2/x_S^2 + u_A^2/x_A^2 + u_D^2/x_D^2)},$$

where $u_S/x_S + u_A/x_A$, u_D/x_D, are the relative instrument uncertainties of sensor, amplifier and display, which can be expressed by their accuracy classes as $p_S/\sqrt{3}$, $p_A/\sqrt{3}$, $p_D/\sqrt{3}$.

It follows that $$u_{System}/|y| = \frac{\sqrt{(p_S^2 + p_A^2 + p_D^2)}}{\sqrt{3}}$$

For a measurement system of n components in line, the following formula characterizes the relative *uncertainty budget* of the measuring chain

$$\delta_{chain} = u_{chain}/|y| = \frac{\sqrt{(\Sigma p_i^2)}}{\sqrt{3}} \quad (i = 1 \ldots n)$$

Fig. 3.11 Method for estimating the measurement system uncertainty

Consideration: actual relative maximum permissible measurement errors for 220 V and limits of error expressed in measurement units (V as scale divisions) are

$$p_{220,rel} = 2.0\% \cdot \frac{380\,V}{220\,V} = 3.5\%\ ;$$
$$p_{abs} = 380\,V \cdot \frac{2.0\%}{100\%} = 7.6\,V \quad \text{(limits of error)}$$
$$u_{instr,rel} = \frac{p_{rel}}{\sqrt{3}} = \frac{3.5\%}{\sqrt{3}} = 2.0\% \quad \text{and}$$
$$u_{instr,abs} = \frac{p_{abs}}{\sqrt{3}} = \frac{7.6\,V}{\sqrt{3}} = 4.4\,V\ .$$

It is obvious that the relative standard uncertainties are smallest at y_{max}.

Since a rectangular distribution was assumed, it is not reasonable to apply the coverage factor k, because this approach assumes a Gaussian distribution. Instead, the standard uncertainty u_{instr} should be stated. It normally suffices to report the uncertainties to at most two significant digits – and also to provide information on how it was determined. Finally, the measurement uncertainty allows the experimenter to decide whether the used instrument is appropriate for his/the customer's needs.

Answer: The result of the example could be reported as 220 V ± 4.4 V. The measurement uncertainty of the instrument is reported as a standard uncertainty (coverage factor $k = 1$) and was obtained by type B evaluation only considering the instrument accuracy class.

If instead of a single measuring instrument, a measuring system or a measuring chain is used, consisting in the simplest case of a sensor, an amplifier, and a display, the accuracy classes of the components of the measuring system can also be used to estimate the instrumental system uncertainty, as illustrated in Fig. 3.11.

3.4.3 Multiple Measurement Uncertainty Components

The method outlined in Figs. 3.9 and 3.10 considers only one single measurement quantity and only the sources covered by only one variable. However, very often uncertainty evaluations have to be related to functional combinations of measured quantities or uncertainty components $y = f(x_1, x_2, x_3, \ldots, x_n)$. In these cases, for uncorrelated (i. e., independent) values, the single uncertainties are combined by applying the law of propagation of uncertainty to give the so-called combined measurement uncertainty

$$u_{combined}(y) = \sqrt{\sum \left(\frac{\partial f}{\partial x_i}\right)^2 u^2(x_i)}\ .$$

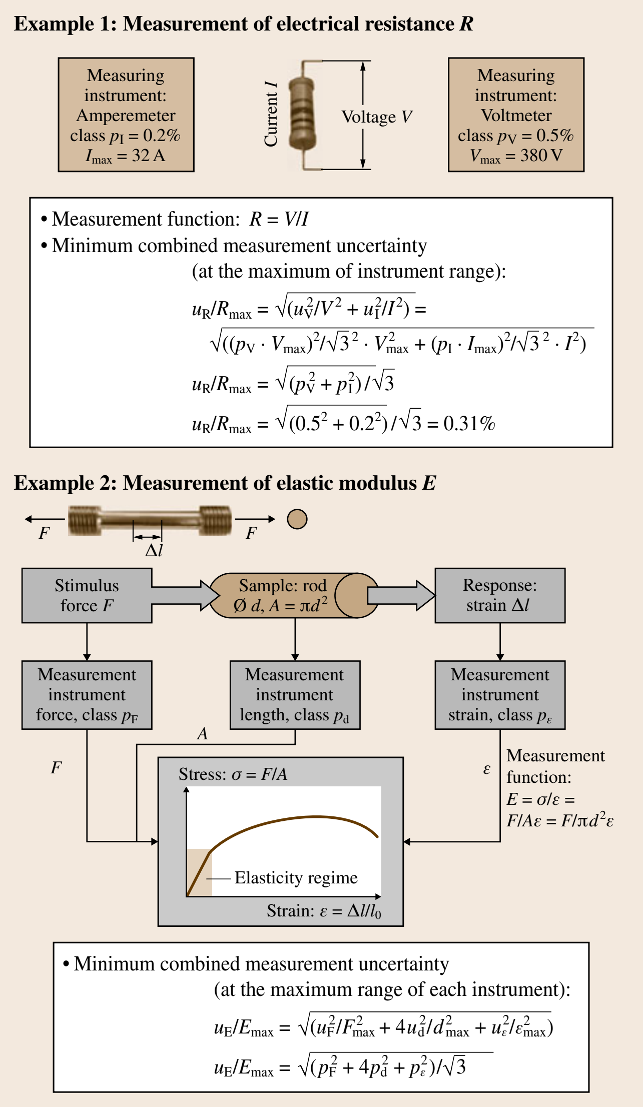

Example 1: Measurement of electrical resistance R
Measuring instrument: Amperemeter class $p_I = 0.2\%$ $I_{max} = 32$ A
Current I
Voltage V
Measuring instrument: Voltmeter class $p_V = 0.5\%$ $V_{max} = 380$ V
• Measurement function: $R = V/I$
• Minimum combined measurement uncertainty (at the maximum of instrument range):
$u_R/R_{max} = \sqrt{(u_V^2/V^2 + u_I^2/I^2)} =$
$\sqrt{((p_V \cdot V_{max})^2/\sqrt{3}^2 \cdot V_{max}^2 + (p_I \cdot I_{max})^2/\sqrt{3}^2 \cdot I^2)}$
$u_R/R_{max} = \sqrt{(p_V^2 + p_I^2)}/\sqrt{3}$
$u_R/R_{max} = \sqrt{(0.5^2 + 0.2^2)}/\sqrt{3} = 0.31\%$
Example 2: Measurement of elastic modulus E
F
Δl
F
Stimulus force F
Sample: rod Ø d, $A = \pi d^2$
Response: strain Δl
Measurement instrument force, class p_F
Measurement instrument length, class p_d
Measurement instrument strain, class p_ε
A
F
ε
Measurement function: $E = \sigma/\varepsilon = F/A\varepsilon = F/\pi d^2\varepsilon$
Stress: $\sigma = F/A$
Elasticity regime
Strain: $\varepsilon = \Delta l/l_0$
• Minimum combined measurement uncertainty (at the maximum range of each instrument):
$u_E/E_{max} = \sqrt{(u_F^2/F_{max}^2 + 4u_d^2/d_{max}^2 + u_\varepsilon^2/\varepsilon_{max}^2)}$
$u_E/E_{max} = \sqrt{(p_F^2 + 4p_d^2 + p_\varepsilon^2)}/\sqrt{3}$

Fig. 3.12 Determination of the combined uncertainty of multiple measurands

From the statistical law of the propagation of uncertainties it follows that there are three basic relations, for which the resulting derivation becomes quite simple

1. for equations of the measurand involving only sums or differences

$$y = x_1 + x_2 + \cdots + x_n \quad \text{it follows}$$

$$u_y = \sqrt{\left(u_1^2 + u_2^2 + \cdots + u_n^2\right)}$$

2. for equations of the measurand involving only products or quotients

$$y = x_1\, x_2 \cdots x_n \quad \text{it follows}$$

$$\frac{u_y}{|y|} = \sqrt{\left(\frac{u_1^2}{x_1^2} + \frac{u_2^2}{x_2^2} + \cdots + \frac{u_n^2}{x_n^2}\right)}$$

3. for equations of the measurand involving exponents

$$y = x_1^a\, x_2^b \cdots x_n^z \quad \text{it follows}$$

$$\frac{u_y}{|y|} = \sqrt{\left(\frac{a^2 u_1^2}{x_1^2} + \frac{b^2 u_2^2}{x_2^2} + \cdots + \frac{z^2 u_n^2}{x^2}\right)}\,.$$

If the parameters are not independent from each other, the mutual dependence has to be taken into account by the covariances; see, e.g., GUM [3.19], but in practise they are often neglected for simplicity.

Also for multiple measurands or measurement instruments, it is possible to use the instrument accuracy class data and other information – if available – for the estimation of the demanded combined measurement uncertainty. The method for the determination of the combined uncertainty is shown in Fig. 3.12, exemplified with simple cases of two and three measurands.

However, for strict application of the measurement uncertainty approach, all uncertainty sources have to be identified and possible additional components not covered have to be considered. This is especially the case in the examples for such uncertainty sources that are not covered by p from the calibration experiment from which p is derived.

3.4.4 Typical Measurement Uncertainty Sources

While in the previous examples only the measurement uncertainty components included in the accuracy class – which is obtained from calibration experiments – were considered, the GUM [3.19] requests to consider all components that contribute to the measurement uncertainty of a measured quantity. The various uncertainty sources and their contributions can be divided into four major groups, as has been proposed by the EUROLAB *Guide to the Evaluation of Measurement Uncertainty for Quantitative Test Results* [3.20]. Measurement uncertainty may depend on

1. the sampling process and sample preparation, e.g.,
   - the sample being not completely representative
   - inhomogeneity effects

- contamination of the sample
- instability/degradation of the sample or other effects during sampling, transport, storage, etc.
- the subsampling process for the measurement (e.g., weighing)
- the sample preparation process for the measurement (dissolving, digestion)

2. the properties of the investigated object, e.g.,
   - instability of the investigated object
   - degradation/ageing
   - inhomogeneity
   - matrix effects and interactions
   - extreme values, e.g., small measured quantity/little concentration
3. the applied measurement and test methods, e.g.,
   - the definition of the measurand (approximations, idealizations)
   - nonlinearities, extrapolation
   - different perception or visualization of measurands (different experimenters)
   - uncertainty of process parameters (e.g., environmental conditions)
   - neglected influence quantities (e.g., vibrations, electromagnetic fields)
   - environment (temperature, humidity, dust, etc.)
   - limits of detection, limited sensitivity
   - instrumental noise and drift
   - instrument limitations (resolution, dead time, etc.)
   - data evaluation, numerical accuracy, etc.
4. the basis of the measurement, e.g.,
   - uncertainties of certified values
   - calibration values
   - drift or degradation of reference values/reference materials
   - uncertainties of interlaboratory comparisons
   - uncertainties from data used from the literature.

All possible sources for uncertainty contributions need to be considered, when the measurement uncertainty is estimated, even if they are not directly expressed in the measurement function. They are not necessarily independent from each other. They are partly of random and partly of systematic character.

3.4.5 Random and Systematic Effects

In the traditional error approach (Sect. 3.4.1) a clear distinction was made between so-called *random errors* and *systematic errors*. Although this distinction is not relevant within the uncertainty approach anymore, as it is not unambiguous, the concept is nevertheless descriptive.

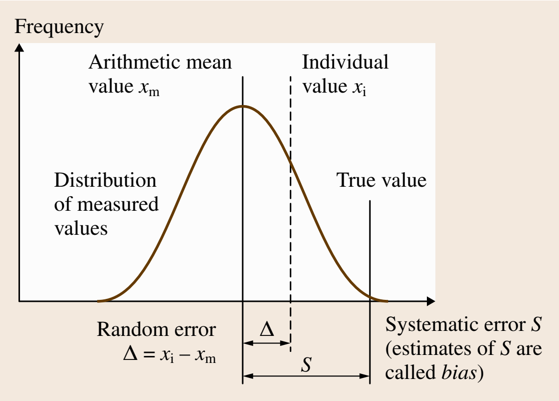

Frequency
Arithmetic mean value x_m
Individual value x_i
Distribution of measured values
True value
Random error $\Delta = x_i - x_m$
Δ
S
Systematic error S (estimates of S are called *bias*)

Fig. 3.13 Illustration of random and systematic errors of measured values

Random effects contribute to the variation of individual results in replicate measurements. Associated uncertainties can be evaluated using statistical methods, e.g., the experimental standard deviation of a mean value (type A evaluation).

Systematic errors result in the center of the distribution being shifted away from the true value even in the case of infinite repetitions (Fig. 3.13).

If systematic effects are known, they should be corrected for in the result, if possible. Remaining systematic effects must be estimated and included in the measurement uncertainty.

The consideration and inclusion of the various sources of measurement errors in the measurement result or the measurement uncertainty is illustrated in Fig. 3.14.

3.4.6 Parameters Relating to Measurement Uncertainty: Accuracy, Trueness, and Precision

The terms *accuracy*, *trueness*, and *precision*, defined in the ISO 3534 international standard characterize a measurement procedure and can be used with respect to the associated uncertainty.

Accuracy as an umbrella term characterizes the closeness of agreement between a measurement result and the true value of the measurand. If several measurement results are available for the same measurand from a series of measurements, accuracy can be split into trueness and precision. Trueness accounts for the closeness of agreement between the mean value and the true

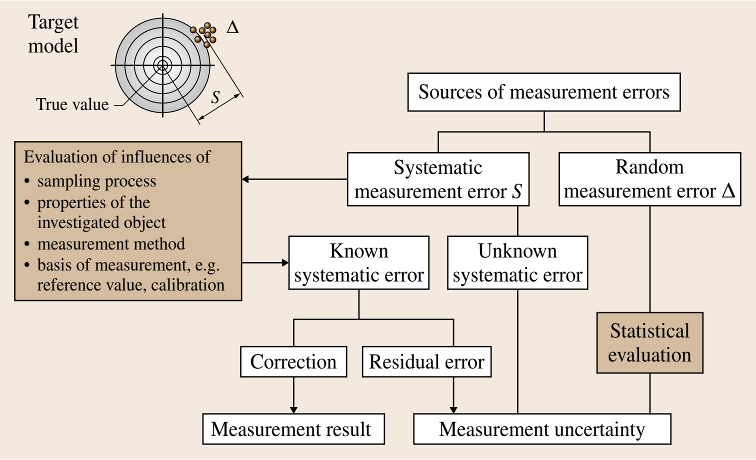

Target
model
Δ
True value
S
Sources of measurement errors
Evaluation of influences of
• sampling process
• properties of the investigated object
• measurement method
• basis of measurement, e.g. reference value, calibration
Systematic measurement error S
Random measurement error Δ
Known systematic error
Unknown systematic error
Statistical evaluation
Correction
Residual error
Measurement result
Measurement uncertainty

Fig. 3.14 Methodology of considering random and systematic errors in measurement

value. Precision describes the closeness of agreement of the individual values themselves.

The target model (Fig. 3.15) visualizes comprehensively the different possible combinations which result from true or wrong and precise or imprecise results.

Estimates of precision are commonly determined for repeated measurements and are valuable information with a view to the measurement uncertainty. They are strongly dependent on the conditions under which precision is investigated: repeatability conditions, reproducibility conditions, and intermediate conditions.

- Repeatability conditions mean that all parameters are kept as constant as possible, e.g.,
  a) the same measurement procedure,
  b) the same laboratory,
  c) the same operator,
  d) the same equipment,
  e) repetition within short intervals of time.
- Reproducibility conditions imply those conditions for a specific measurement that may occur between different testing facilities, e.g.,
  a) the same measurement procedure,
  b) different laboratories,

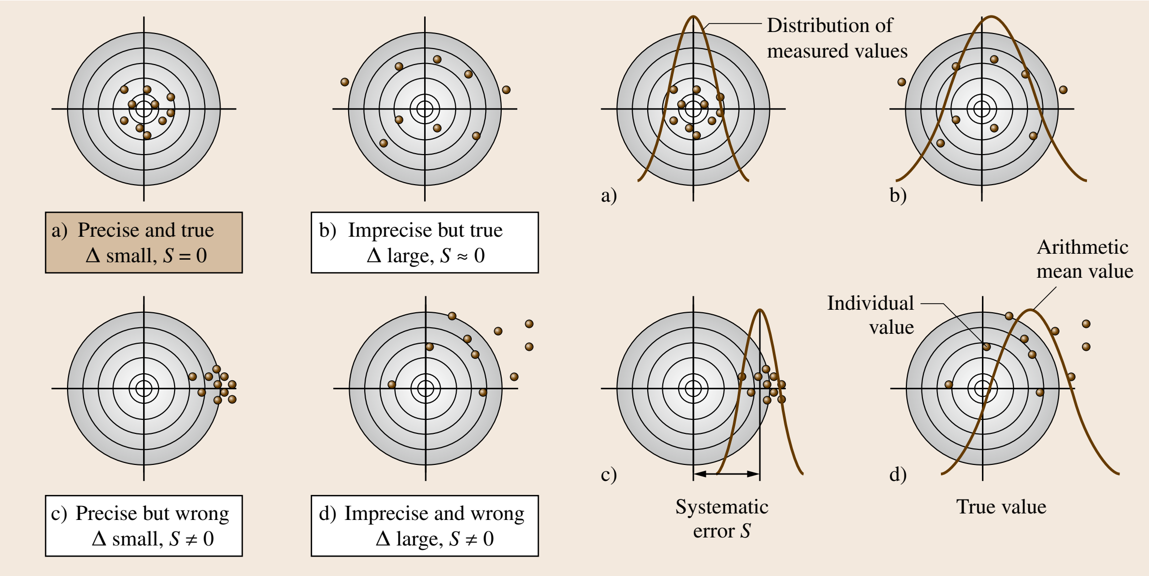


Fig. 3.15 Target model to illustrate trueness and precision. The center of the target symbolizes the (unknown) true value

c) different operators,
d) different equipment.

- Intermediate conditions have to be specified regarding which factors are varied and which are constant. For *within-laboratory reproducibility* the following conditions are used
  a) the same measurement procedure,
  b) the same laboratory,
  c) different operators,
  d) the same equipment (alternatively, different equipment),
  e) repetition within long intervals of time.

3.4.7 Uncertainty Evaluation: Interlaboratory and Intralaboratory Approaches

For the evaluation of measurement uncertainties in practise, often many different approaches are possible. They all begin with the careful definition of the measurand and the identification of all possible components contributing to the measurement uncertainty. This is especially important for the sampling step, as primary sampling effects are often much larger than the uncertainty associated with the measurement of the investigated object.

A convenient classification of uncertainty approaches is shown in Fig. 3.16. The classification is based on distinction between uncertainty evaluation carried out by the laboratory itself (called *intralaboratory approach*) and uncertainty evaluation based on collaborative studies in different laboratories (called *interlaboratory approach*). These approaches are compiled in the EUROLAB Technical Report 1/2007 *Measurement uncertainty revisited: Alternative approaches to uncertainty evaluation* [3.21].

In principle, four different approaches can be applied. The four approaches to uncertainty estimations outlined in Fig. 3.16 are briefly described in the following.

1) The Modeling Approach

This is the main approach to the evaluation of uncertainty and consists of various steps as described in Chap. 8 of the GUM.

For the modeling approach, a mathematical model must be set up, which is an equation defining the quan-

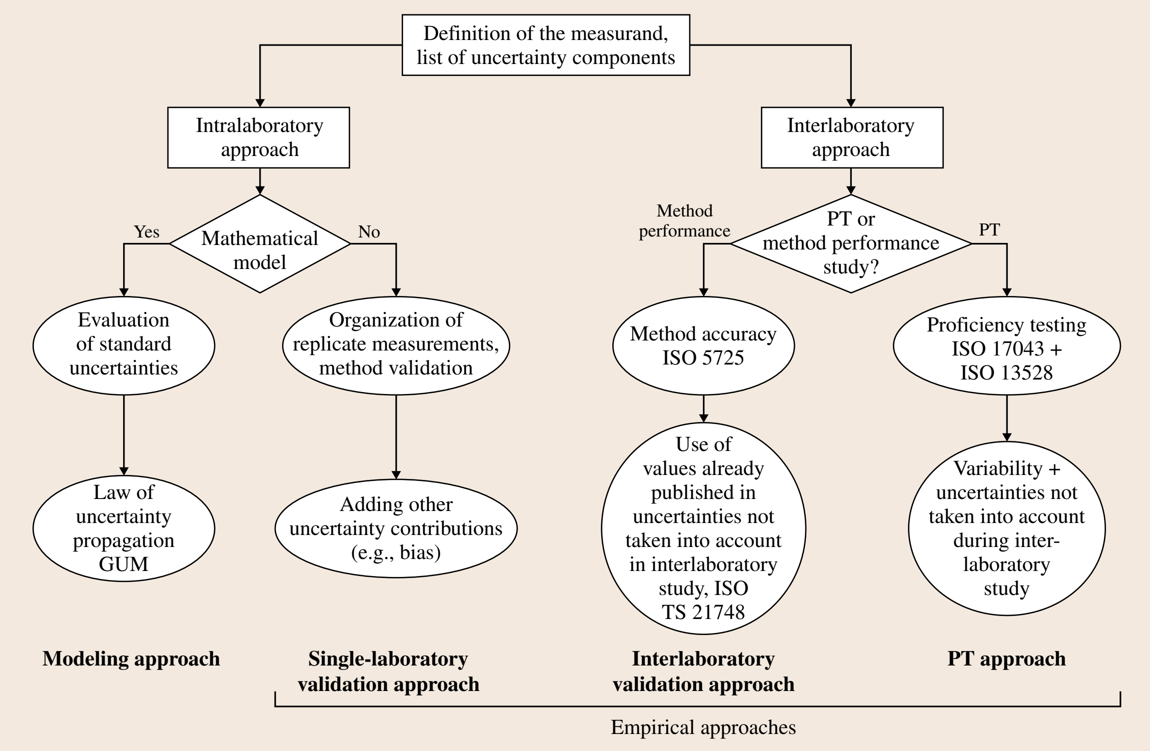

Definition of the measurand,
list of uncertainty components
Intralaboratory approach
Interlaboratory approach
Mathematical model
Yes
No
Method performance
PT or method performance study?
PT
Evaluation of standard uncertainties
Organization of replicate measurements, method validation
Method accuracy ISO 5725
Proficiency testing ISO 17043 + ISO 13528
Law of uncertainty propagation GUM
Adding other uncertainty contributions (e.g., bias)
Use of values already published in uncertainties not taken into account in interlaboratory study, ISO TS 21748
Variability + uncertainties not taken into account during inter-laboratory study
Modeling approach
Single-laboratory validation approach
Interlaboratory validation approach
PT approach
Empirical approaches

Fig. 3.16 A road map for uncertainty estimation approaches according to [3.21]

titative relationship between the quantity measured and all the quantities on which it depends, including all components that contribute to the measurement uncertainty.

Afterwards, the standard uncertainties of all the single uncertainty components are estimated. Standard deviations from repeated measurements are directly the standard uncertainties for the respective components (if normal distribution can be assumed). The combined uncertainty is then calculated by the application of the law of propagation of uncertainty, which depends on the partial derivatives for each input quantity. In strictly following the modeling approach, correlations also need to be incorporated.

Usually the expanded uncertainty U (providing an interval $y-U$ to $y+U$ for the measurand y) is calculated. For normal distribution, the coverage factor $k=2$ is chosen typically. Finally, the measurement result together with its uncertainty should be reported according to the rules of the GUM [3.19]. These last two steps of course also apply to the other approaches (2–4).

Because full mathematical models are often not available or the modeling approach may be infeasible for economic or other reasons, the GUM [3.19] foresees that also alternative approaches may be used. The other approaches presented here are as valid as the modeling approach and sometimes even lead to more realistic evaluation of the uncertainty, because they are largely based on experimental data. These approaches are based on long experience and reflect common practise.

Even though the single-laboratory validation, interlaboratory validation, and PT approaches also use statistical models as the basis for data analysis (which also be described as *mathematical models*) the term *mathematical model* is reserved for the modeling approach, and the term *statistical model* is used for the other approaches. The latter are also called *empirical approaches*.

2) The Single-Laboratory Validation Approach

If the full modeling approach is not feasible, in-house studies for method validation and verification may deliver important information on the major sources of variability. Estimates of bias, repeatability, and within-laboratory reproducibility can be obtained by organizing experimental work inside the laboratory. Quality control data (control charts) are valuable sources for precision data under within laboratory reproducibility conditions, which can be used to serve directly as standard uncertainties. Standard uncertainties of additional (missing) effects can be estimated and combined – see also under point 5). If possible, during the repetition of the experiment, the influence quantities should be varied, and certified reference materials (CRMs) and/or comparison with definitive or reference methods should be used to evaluate the component of uncertainty related to the trueness.

3) The Interlaboratory Validation Approach

Precision data can also be obtained by utilizing method performance data and other published data (other than proficiency testing that the testing laboratory has taken part in itself, as this is considered in the PT approach). The reproducibility data can be used directly as standard uncertainty.

ISO 5725 *Accuracy (trueness and precision) of measurement methods and results* [3.22] provides the rules for assessment of repeatability (repeatability standard deviation s_r), reproducibility (reproducibility standard deviation s_R), and (sometimes) trueness of the method (measured as a bias with respect to a known reference value). Uncertainty estimation based on precision and trueness data in compliance with ISO 5725 [3.22] is extensively described in ISO/TS 21748 *Guidance for the use of repeatability, reproducibility and trueness estimates in measurement uncertainty estimation* [3.23].

4) The PT Approach: Use of Proficiency Testing (EQA) Data

Proficiency tests (external quality assessment, EQA) are intended to check periodically the overall performance of a laboratory. Therefore, the laboratory can compare the results from its participation in proficiency testing with its estimations of measurement uncertainty of the respective method and conditions.

Also, the results of a PT can be used to evaluate the measurement uncertainty. If the same method is used by all the participants in the PT scheme, the standard deviation is equivalent to an estimate of interlaboratory reproducibility, which can serve as standard uncertainty and, if required, be combined with additional uncertainty components to give the combined measurement uncertainty. If the laboratory has participated over several rounds, the deviations of its own results from the assigned value can be used to evaluate its own measurement uncertainty.

Combination of the Different Approaches to Uncertainty Evaluation

It is also possible – and often necessary – to combine the different approaches described above. For example, in the PT approach, sometimes missing components need

Table 3.8 Compilation of relevant documents on measurement uncertainty

Document	Reference	General	Modeling	Single laboratory	Inter-laboratory	PT
ISO (1993/1995), *Guide to the expression of uncertainty in measurement (GUM)*	[3.19]	×	×			
EURACHEM/CITAC (2000), *Quantifying uncertainty in analytical measurement*, 2nd edn.	[3.24]	×	×	×		
EUROLAB technical report no. 1/2002, *Measurement uncertainty in testing*	[3.25]	×				
EUROLAB technical report no. 1/2006, *Guide to the evaluation of measurement uncertainty for quantitative test results*	[3.20]	×	×	×		×
EUROLAB technical report no. 1/2007, *Measurement uncertainty revisited: Alternative approaches to uncertainty evaluation*	[3.21]	×	×	×	×	×
EA 4/16 (2004), *Guidelines on the expression of uncertainty in quantitative testing*	[3.26]	×	×	×	×	×
NORDTEST technical report 537 (2003), *Handbook for calculation of measurement uncertainty in environmental laboratories*	[3.27]			×	×	×
EA-4/02 (1999), *Expression of the uncertainty of measurement in calibration*	[3.28]		×			
ISO 5725 *Accuracy (trueness and precision) of measurement methods and results (six parts)*	[3.22]				×	
ISO 5725-3 *Accuracy (trueness and precision) of measurement methods and results – Part 3: Intermediate measures of the precision of a standard measurement method*	[3.22]			×		
ISO/TS 21748 *Guide to the use of repeatability, reproducibility, and trueness estimates in measurement uncertainty estimation*	[3.23]				×	
AFNOR FD X 07-021, *Fundamental standards – Metrology and statistical applications – Aid in the procedure for estimating and using uncertainty in measurements and test results*	[3.29]		×		×	
Supplement no. 1 to the GUM, *Propagation of distributions using a Monte Carlo method)*	[3.30]		×			
ISO 13528 *Statistical methods for use in proficiency testing by interlaboratory comparison*	[3.31]					×
ISO/TS 21749 *Measurement uncertainty for metrological applications – Repeated measurements and nested experiments*	[3.32]			×		

to be added. This may be the case if the PT sample was a solution and the investigated object is a solid sample that needs to be dissolved first before undergoing the same measurement as the PT sample. Therefore, uncertainty components from the dissolving and possible dilution steps need to be added. These could be estimated by intralaboratory validation data or – especially for the dilution uncertainty – by repeated measurements from the resulting standard deviation.

Concerning the reliability of the methods described, it should be emphasized that there is no hierarchy; i. e., there are no general rules as to which method should be preferred. The laboratory should choose the most fit-for-purpose method of estimating uncertainty for its individual application. Also, the time and effort invested in the uncertainty estimation should be appropriate for the purpose.

Finally there may be cases where none of the approaches described above is possible. For example for fire protection doors repeated measurements are not possible. Also, there may be no PT scheme available. For such cases, experience-based expert estimate (type B evaluation) may be the best option to estimate measurement uncertainty contributions.

A compilation of references (guidelines and standards) for the various approaches is given in Table 3.8 (adopted from the EUROLAB Technical Report 1/2007 [3.21]) together with the reference number and an indication (×) of which uncertainty evaluation approaches are addressed in the respective document.

3.5 Validation

The operation of a testing facility or a testing laboratory requires a variety of different prerequisites and supporting measures in order to produce trustworthy results of measurements. The most central of these operations is the actual execution of the test methods that yield these results. At all times it has therefore been vital to operate these test methods in a skilful and reproducible manner, which requires not only good education and training of the operator in all relevant aspects of performing a test method, but also experimental verification that the specific combination of operator, sample, equipment, and environment yields results of known and fit-for-purpose quality. For this experimental verification at the testing laboratory the term *validation* (also *validation of test methods*) was introduced some 20 years ago. Herein, test method and test procedure are used synonymously.

In the following it is intended to present purpose, rationale, planning, execution, and interpretation of validation exercises in testing. We will, however, not give the mathematical and statistical framework employed in validation, as this is dealt with in other chapters of the handbook.

3.5.1 Definition and Purpose of Validation

Definitions

Although in routine laboratory jargon a good many shades of meaning of *validation* are commonly associated with this word, the factual operation of a validation project encompasses the meaning better than words do. Nevertheless, a formal definition is offered in the standards, and the following is cited from EN-ISO 9000:2000 [3.33].

Validation. Confirmation, through the provision of objective evidence, that the requirements for a specific intended use or application have been fulfilled.

Objective Evidence. Data supporting the existence or verity of something.

Requirement. Need or expectation that is stated, generally implied or obligatory.

In ISO 17025 (*General Requirements for the Competence of Testing and Calibration Laboratories*), validation prominently features in Sect. 5.4 on technical requirements, and the definition is only slightly different: *Validation is the confirmation by examination and the provision of objective evidence that the particular requirements for a specific intended use are fulfilled*.

Although such definitions tend to be given in a language that makes it difficult to see their ramifications in practise, there are a couple of key features that warrant some discussion. Validation is (done for the purpose of) confirmation and bases this confirmation on objective evidence, generally data from measurements. It can be concluded that, in general, only carefully defined, planned, and executed measurements yield data that will permit a judgement on the fulfilment of requirements. The important point here is that the requirements have to be cast in such a way that permits the acquisition of objective evidence (data) for testing the question of whether these requirements are fulfilled.

Verification is frequently used in a manner indistinguishable from validation, so we also want to resort to the official definition in EN-ISO 9000:2000.

Verification. Confirmation, through the provision of objective evidence, that specified requirements have been fulfilled.

The parallels with validation are obvious, as verification is also confirmation, also based on objective evidence, and also tested against specified requirements, but apparently without a specific use in mind, which is part of the definition of validation. In practise, the difference lies in the fact that validation is cited in connection with test methods, while verification is used in connection with confirmation of data.

As the formal definitions are not operationally useful, it may be helpful to keep in mind the essentials offered from ISO 17025 that appear to be summarized in Chap. 5.4.5.3:

> *The range and accuracy of the values obtainable from validated methods (e.g. the uncertainty of the results, detection limit, selectivity of the method, linearity, limit of repeatability and/or reproducibility, robustness against external influences and/or cross-sensitivity against interference from the matrix of the sample/test object) as assessed for the intended use shall be relevant to the clients' needs.*

This statement makes clear that there must be an assessment for the intended use, although the various figures of merit in parenthesis are inserted in a rather artificial manner into the sentence.

The view of Cooperation International Traceability in Analytical Chemistry (CITAC)/EURACHEM on validation is best summarized in Chap. 18 of the *Guide to Quality in Analytical Chemistry* [3.34], where the introductory sentence reads:

> *Checks need to be carried out to ensure that the performance characteristics of a method are understood and to demonstrate that the method is scientifically sound under the conditions in which it is to be applied. These checks are collectively known as validation. Validation of a method establishes, by systematic laboratory studies that the method is fit-for-purpose, i. e. its performance characteristics are capable of producing results in line with the needs of the analytical problem . . .*

At this point we shall leave the normative references and try to develop a general-purpose approach to validation in the following.

Purpose

The major purpose in line with the formal definition is confirmation. Depending on the party concerned with the testing, the emphasis of such a confirmation may be slightly different.

The (future) operator of a test method has the need to acquire enough skill for performing the method and may also care to optimize the routine execution of this method. The laboratory manager needs to know the limits of operation of a test method, as well as the performance characteristics within these limits. The prospective customer, who will generally base decisions on the outcome of the testing operation, must know the limits and performance characteristics as well, in order to make an educated judgement on the reliability of the anticipated decisions. He/she must be the one to judge the fitness for purpose, and this can only be done on the basis of experimental trials and a critical appraisal of the data thereby generated.

In a regulated environment, such as the pharmaceutical or car industry, regulatory agencies are additional stakeholders. These frequently take the position that a very formalized approach to validation assures the required validity of the data produced. In these instances very frequently every experimental step to be taken is prescribed in detail and every figure to be reported is unequivocally defined, thereby assuring uniform execution of the validation procedures.

On a more general basis one can argue that validation primarily serves the following purposes.

1. Derivation of performance characteristics
2. Establishment of short- and long-term stability of the method of measurement, and setting of control limits
3. Fine-tuning of the standard operating procedure (SOP)
4. Exploitation of scope in terms of the nature and diversity of samples and the range of the values of the measurand
5. Identification of influence parameters
6. Proof of competence of the laboratory.

In simple words, validation for a laboratory/operator is about getting to know your procedure.

3.5.2 Validation, Uncertainty of Measurement, Traceability, and Comparability

Relation of Uncertainty, Traceability, and Comparability to Validation

Validation cannot be discussed without due reference to other important topics covered in this handbook. We therefore need to shed light on the terms uncertainty, traceability, and comparability, in order to demonstrate their relationship to method validation.

The existence of a recognized test method is the prerequisite for the mutual recognition of results. This recognition is based on reproducibility and traceability, whereby traceability to a stated and (internationally) accepted reference is an indispensable aid in producing reproducible results. This links a locally performed measurement to the world of internationally accepted standards (references, scales) in such a way that all measurements linked to the same standards give results that can be regarded as fractions and multiples of the same unit. For identical test items measured with the same test method this amounts to identical results within the limits of measurement uncertainty. Measurement uncertainty cannot be estimated without due consideration of the quality of all standards and references involved in the measurement, and this in turn necessitates the clear stating of all references, which has been defined as traceability earlier in this paragraph. In a way a tight connection of a result to a standard is realized by very well-defined fractions and multiples, all carrying small uncertainties. Well-defined fractions and multiples are thus tantamount to small measurement uncertainty.

Formal Connection of the Role of Validation and Uncertainty of Measurement

In a certain way, validation is linked to measurement uncertainty through optimization of the method of measurement: validation provides insight into the important influence quantities of a method, and these influence quantities are those that generally contribute most to measurement uncertainty. As a result of validation, the reduction of measurement uncertainty can be affected in one of two ways: (a) by tighter experimental control of the influence quantity, or (b) by suitable numerical correction of the (raw) result for the exerted influence. By way of example, we consider the influence of temperature on a measurement. If this is significant, one may control the temperature by thermostatting, or alternatively, one can establish the functional dependence of the measurement, note the temperature at the time of measurement, and correct the raw result using the correction function established earlier. Both these actions can be regarded as a refinement of the measurement procedure, constitute an improvement over the earlier version of the method (*optimization*), and necessitate changes in the written SOP.

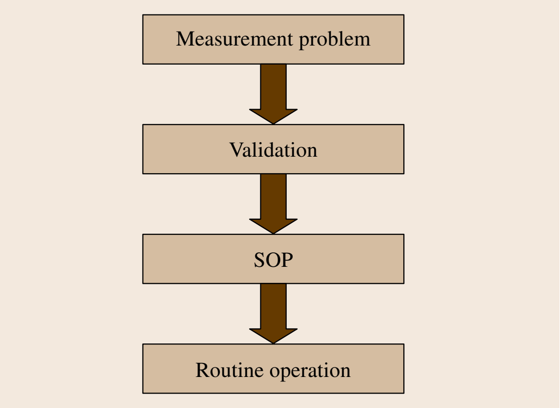


Fig. 3.17 Validation has a central place in the operation of a test method

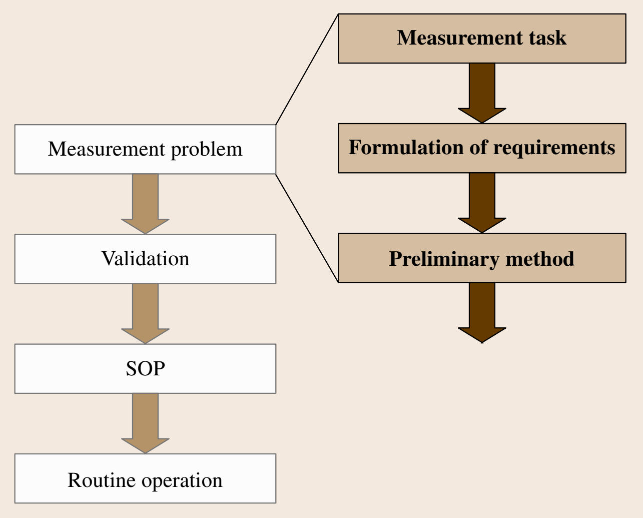


Fig. 3.18 Blown-up view of the measurement problem validation leads from the preliminary method to routine operation

A good measurement provides a close link of the result to the true value, albeit not perfectly so. Prior to validation, the result is

$$x_{ijk} = \mu + \varepsilon_{ijk} \,,$$

$$x_{ijk} \ldots \text{result} \,;\ \mu \ldots \text{true value} \,;\ \varepsilon_{ijk} \ldots \text{deviation} \,.$$

The deviation is large and unknown in size and sign, and will give rise to a large uncertainty of measurement. A major achievement in a successful validation exercise is the identification of influence quantities and their action on the result. If, for instance, three (significant) influence quantities are identified, the result can be viewed as biased by these effects δ

$$x_{ijk} = \mu + \delta_i + \delta_j + \delta_k + \varepsilon_{ijk} \,,$$

and in so doing the residual (and poorly understood) deviation ε_{ijk} is now greatly reduced as the effects of the identified quantities δ are quasi-extracted from the old ε_{ijk}. As the bias is now known in sign and size, δs can be used for correcting the original x_{ijk}, which after validation can be viewed as the uncorrected raw result,

$$x_{ijk} - \delta_i - \delta_j - \delta_k = \mu + \varepsilon_{ijk} \,.$$

Alternatively – as is occasionally done in chemistry with recovery – the corrections can be ignored, and thus the raw results are left uncorrected.

Figure 3.17 highlights the central position of validation in the introduction of a new method of measurement in a testing laboratory.

From a blown-up view of the measurement problem (Fig. 3.18) one can see that, in reality, it can be broken down into three distinct steps: the measurement task as formulated by the customer, the formulation of the requirements in technical terms derived from the communicated measurement task, and the preliminary method devised from experience and/or literature that will serve as basis for the (first round of) validation.

3.5.3 Practice of Validation

Validation Plan

Prior to validation, a plan needs to establish which performance characteristics have to be established. This serves the purpose that in a clear succession of experiments test are applied that ultimately allow the assessment of the performance of the method with respect to the client's needs.

Therefore, a written plan must be laid out for the experiments performed in the course of validation, and the criteria to be met by the method in the course of validation must be established in this plan beforehand. It is then immediately obvious whether the method validated is suitable in the particular instance or not.

Validation is frequently planned by senior supervisory personnel or by staff specialized in validation with a good grasp of the client's needs. In regulated areas, such as pharmacy or food, fixed validation plans may be available as official documents and are not to be altered by laboratory personnel.

In any case it is advisable to have a separate standard operating procedure to cover the generic aspects of validation work.

Method Development and Purpose of Basic Validation

For a specified method as practised in a laboratory under routine conditions, method validation marks the end of preliminary method development. It serves the purpose of establishing the performance characteristics of a suitably adapted and/or optimized method, and also the purpose of laying down the limitation of reliable operation by either environmental or sample-related conditions. These limits are generally chosen such that the influence of changes in these conditions is still negligible relative to the required or expected measurement uncertainty. In chemical terms, this makes transparent which analytes (measurands) can be measured with a specific method in which (range of) matrices in the presence of which range of potential interferents.

If a method is developed for a very specific purpose, method validation serves to provide experimental evidence that the method is suitable for this purpose, i. e., for solving the specific measurement problem. In a sense, method validation is interlinked with method development, and care must be taken to draw a clear line experimentally between those steps. The validation plan forms the required delimitation.

Implicitly it is assumed that experiments for the establishment of performance characteristics are executed with apparatus and equipment that operate within permissible specifications, work correctly, and are calibrated. Such studies must be carried out by competent staff with sufficient knowledge in the particular area in order to interpret the obtained results properly, and base the required decision regarding the suitability of the method on them.

In the literature there are frequent reports regarding results of interlaboratory comparison studies being used for the establishment of some method characteristics. There is, however, also found the situation in which a single laboratory requires a specific method for a very special purpose. The Association of Official Analytical Chemists (AOAC), which is a strong advocate of interlaboratory studies as a basis for method validation, established in 1993 the peer-verified method program [3.35], which serves to validate methods practised by one or a few laboratories only.

For an analytical result to be suitable for the anticipated purpose, it must be sufficiently reliable that every decision based on it will be trustworthy. This is the key issue regarding method validation performance and measurement uncertainty estimation.

Regardless of how good a method is and how skillfully it is applied, an analytical problem can only be solved by analyzing samples that are appropriate for this problem. This implies that sampling can never be disregarded.

Once a specific analytical question is defined by the client, it must be decided whether one of the established (and practised) methods meets the requirements. The method is therefore evaluated for its suitability. If necessary, a new method must be developed/adapted to the point that it is regarded as suitable. This process of evaluating performance (established by criteria such as selectivity, detection limits, decision limits, recovery, accuracy, and robustness) and the confirmation of the suitability of a method are the essence of method validation.

The questions considered during the development of an analytical procedure are multifaceted: Is a qualitative or a quantitative statement expected? What is the specific nature of the analyte (measurand)? What matrix is involved? What is the expected range of concentrations? How large a measurement uncertainty is tolerable? In practise, limitations of time and money may impose the most stringent requirements.

Confronted with altered or new analytical queries, the adaptation of analytical procedures for a new analyte, a new matrix, another concentration range or similar variations is frequently required. General analytical trends may also require modifications or new developments of analytical procedures; a point in case are trends in miniaturization, as experienced in high-performance liquid chromatography (HPLC), flow photometry, capillary electrochromatography, hyphenation, etc.

Many analytical procedures are described in the scientific literature (books, journals, proceedings, etc.). These sources are frequently an appropriate basis for the development of new procedures. In many cases, there are also standards available with detailed documentation of the experimental procedure. If only general documentation is provided, it might be suitable as a starting point for the development of customized laboratory procedures.

Alternatively, the interchange of ideas with friendly laboratories can give impetus to the development of new or modified analytical procedures. Occasionally, a new combination of established analytical steps may lead to a new method.

There is also an increasing need for good cooperation between several disciplines, as it is hardly possible for a single person to independently develop complex methods in instrumental analysis. Also, the great flood of data from multichannel detection systems cannot be captured or evaluated by conventional procedures of data treatment, so additional interfaces are needed, particularly to information technology.

The basic validation exercise cannot cover the complete validation process, but is concerned mainly with those parts of validation that are indispensable in the course of development of an analytical procedure. Most importantly the scope of the method must be established, inter alia with respect to analytes, matrices, and concentration range within which measurements can be made in a meaningful way. In any case, the basic validation comprises the establishment of performance characteristics (also called figures of merit), with a clear emphasis on data supporting the estimation of measurement uncertainty.

Depth and Breadth of Validation

Regarding the depth and breadth of validation, ISO 17025 states that *validation shall be as extensive as is necessary to meet the needs in the given application or field of application*.

However, how does this translate into practical experimental requirements? As already stated earlier, it is clear that every analytical method must be available in the written form of an SOP. Until fitness for the intended use is proven through validation, all methods must be regarded as preliminary. It is not uncommon that the results of validation require revision of the SOP with regard to the matrix and the concentration range. This can be understood as laboratory procedures are based on a combination of SOP and validation as delimited by matrix, analyte, and this particular SOP. Here too, the close connection of SOP and validation is noteworthy.

Besides the type and number of different matrices and the concentration range for the application of the method, the extent of validation also depends markedly on the number of successive operations; for a multistage procedure, the extent and consequently the effort of validation will be much larger than for a single-stage procedure.

For the time sequence of basic validation there are also no fixed rules, but it seems appropriate to adopt some of the principles of method development for validation.

- Test of the entire working range, starting from one concentration
- Reverse inclusion of the separate stages into validation, starting with the study of the final determination
- Testing of all relevant matrices, starting with the testing of standards.

In all phases of validation, it must be ascertained that the method is performing satisfactorily, for instance, by running recovery checks alongside.

The final step must be the proof of trueness and reproducibility, e.g., on the basis of suitable reference materials.

Performance Characteristics

The importance of performance characteristics has been mentioned repeatedly in this text. These parameters generally serve to characterize analytical methods and – in the realm of analytical quality assurance – they serve to test whether a method is suitable to solve a particular analytical problem or not. Furthermore, they are the basis for the establishment of control limits and other critical values that provide evidence for the reliable performance of the method on an everyday basis.

The latter use of performance characteristics is a very significant one, and it is obvious that these performance characteristics are only applicable if established under routine conditions and in real matrices, and not under idealized and unrealistic conditions.

The actual selection of performance characteristics for validation depends on the particular situation and requirements. Table 3.9 gives an overview of the most relevant ones.

Table 3.9 Performance characteristics (after *Kromidas* [3.36])

Parameter	Comment
Trueness, accuracy of the mean, freedom from bias	The older English literature does not distinguish between accuracy and trueness
Precision – repeatability – reproducibility	ISO 5725 series: Accuracy, trueness, and precision
Linearity	
Selectivity	
Recovery	
Limit of detection (LOD)	
Limit of quantification (LOQ) Limit of determination (LOD) Limit of decision (LOC)	
Robustness, ruggedness	
Range	
Sensitivity	
Stability	
Accuracy	See *trueness*
Specificity	Often used synonymously with *selectivity*
Uncertainty of measurement expanded uncertainty	
Method capability	
Method stability/process stability	

Different emphasis is given to many of these parameters in the various standards. The most significant recent shift in importance is seen in ISO 17025, where the previously prominent figures of merit (*accuracy* and *precision*) are replaced by *uncertainty in measurement*.

The following performance characteristics are specifically emphasized in the CITAC/EURACHEM *Guide to Quality in Analytical Chemistry* of 2002.

- Selectivity and specificity (description of the measurand)
- Measurement range
- Calibration and traceability
- Bias/recovery
- Linearity
- Limit of detection/limit of quantitation
- Ruggedness
- Precision.

In ISO 17025 the performance characteristics are listed exemplarily:

> *e.g. the uncertainty of the results, detection limit, selectivity of the method, linearity, limit of repeatability and/or reproducibility, robustness against external influences and/or cross-sensitivity against interference from the matrix of the sample/test object.*

From this wording it can be understood that the actual set of figures must be adapted to the specific problem. Selection criteria for the best set in a given situation will be discussed later.

Some of these performance characteristics are discussed in the following.

Accuracy, Precision, and Trueness. There are different approaches for the proof of *accuracy* of results from a particular method. The most common one is by testing of an appropriate reference material, ideally a certified reference material, with certified values uncontested as known and true. A precondition, however, is obviously that such a material is available. It must also be noted that, when using this approach, most sampling and some of the sample preparation steps are not subjected to the test.

Numerically, the comparison of the results of a test method with a certified value is most frequently carried out using a t-test, or alternatively the Doerffel test for significant deviations can be applied.

Trueness of results can also be backed up by applying a completely different measurement principle. This alternative method must be a well-established and recognized method. In this approach, only those steps that are truly independent of each other are subjected to a serious test for accuracy. For instance, if the same method of decomposition is applied in both procedures, this step cannot be taken as independent in the two procedures and therefore cannot be regarded as having been tested for accuracy by applying the alternative method of measurement.

In practise, the differences between the results from the two procedures are calculated, these differences are averaged, and their standard deviation is computed. Finally, a t-value from these results is obtained and compared with a critical t-value from the appropriate table. If the computed t-value is greater than the tabulated one, it can be assumed with a previously determined probability (e.g., 95%) that the difference between the two methods is indeed significant.

Another method to check the accuracy is the use of recovery studies (particularly useful for checking separation procedures), or balancing the analyte, applying mass balances or plausibility arguments.

In all of these considerations, there must always be due regard to the fact that trueness and precision are hardly independent from each other.

Precision can be regarded as a measure of dispersion between separate results of measurements. The standard deviation under repeatability, reproducibility or intermediate conditions, but also the relative standard deviation and variance, can be used as measures of precision. From the standard deviation, repeatability or reproducibility limits can be obtained according to ISO 5725.

Which of these measures of precision is actually used is up to the analyst. It is, however, recommended to use the repeatability, reproducibility or intermediate standard deviation according to ISO 5725.

The values of precision and trueness established in practise are always estimates that deviate in the operation of successive interlaboratory comparison studies or proficiency testing rounds.

Precision can therefore be regarded as a measure of dispersion (typical statistical measure: standard deviation) and trueness as a measure of location (typical statistical measure: arithmetic average), adding up to a combined measure of accuracy as a measure of dispersion and location: the deviation of a single value from the true one.

To avoid misunderstanding in the practical estimation of trueness and precision, the description of the experimental data underlying the computations must be done most carefully. For instance, it is of significant importance to know whether the data used are results of single determinations, or whether they were obtained from duplicate or triplicate measurements. Equally, the conditions under which these measurements were made must be meticulously documented, either as part of the SOP or independently. Important but neglected parameters might be the temperature constancy of the sample, constant time between single measurements, extraction of raw data, etc.

Calibration and Linearity. Valid calibration can be regarded as a fundamental prerequisite for a meaningful analytical measurement. Consequently, calibration frequently constitutes the first step in quality assurance. In short, the challenge is to find the dependence between signal and amount of substance (or concentration). Preconditions for reliable calibration are

- standards with (almost) negligible uncertainty (independent variable x),
- constant precision of the entire working range,
- useful model (linear or curved),
- random variation of deviations in signals,
- deviations distributed according to the normal distribution.

These criteria are ranked in order of decreasing importance. This means that all analytical work is meaningless unless there is a firm idea about the reliability of standards. Many methods of analysis have poorer precision at higher concentrations (larger absolute standard deviation) than at lower concentrations. In practise, this means that the working range must either be reduced or be subdivided into several sections, each with its own calibration function. Alternatively, the increase of standard deviation with increasing concentration can be established and used for calibration on the basis of weighted regression. In this case a confidence band cannot be given.

In all cases, it is advantageous to position the calibration function so that the majority of expected concentrations fall in the middle part of the curve.

The calibration function is therefore the mathematical model that best describes the connection between signal and concentration, and this function can be straight or curved. The linearity of a measurement method determines the range of concentrations within which a straight line is the best description for the dependence of the signal on concentration.

A large deviation from linearity can be visually detected without problems. Alternatively, the dependence of signal on concentration is modeled by appropriate software in the way that best describes this dependence. A statistical F-test then shows deviations from linearity, or the correlation coefficients of the different models can be compared with each other; the closer the value is to 1, the better the fit of the model.

Recovery. Recovery is the ratio of a measured mean value under repeatability conditions to the true value of the analyte in the sample

$$R = \frac{\overline{x}}{x_\mathrm{t}}100\,,$$

where R is recovery (in %), $\bar{x}$ is the mean value, and x_t is the true value.

Recovery data can be useful for the assessment of the entire method, but in a specific case it is applicable just for the experimental conditions, i.e., the matrix, etc. for which the mean value was determined.

If R is sufficiently well established, it can also be used for the correction of the results.

The following are the most important procedures for determining recoveries.

- *Certified reference materials*: the certified value is used as the true value in the formula above.
- *Spiking*: this procedure is widely practised either on a blank sample or on a sample containing the analyte.
- *Trueness*: if spiking is used at several concentration levels or if several reference materials are available over the concentration range of interest, R can be estimated from a regression line by testing the trueness of a plot of true (spiked) values versus measured values.
- *Mass balance*: tests are conducted on separate fractions of a sample. The sum of the results on each fraction constitute 100%. This tedious method is applied only in special cases.

Robustness. A method is robust (or rugged) if minor variations in the practise of the method do not lead to changes in the data quality. Robustness therefore is the degree of independence of the results from changes in the different influence factors. It is easily seen that robustness is becoming a very major issue in routine operation of analytical methods.

For the determination of robustness, two different approaches are feasible.

Interlaboratory studies: The basic reasoning behind the usefulness of interlaboratory studies for robustness testing is the fact that the operation of a specific method in a sufficiently large number of laboratories (≥ 8) will always lead to random deviations in the experimental parameters.

Experimental design in a single laboratory: In a carefully designed study the relevant experimental parameters are varied within foreseen or potential tolerances and the effects of these perturbations on the results are recorded.

- Experimental parameters (also called factors) that are most likely to have an influence on the result are identified.
- For each experimental parameter the maximum deviation from the nominal value that might be seen in routine work is laid down.
- Experiments are run under these perturbed conditions.
- The results are evaluated to identify the truly influential experimental parameters.
- A strategy is devised to optimize the procedure with respect to the identified influences.

Relationship Between the Objective of a Method and the Depth of Validation

To present the basic considerations considered so far in a concrete form, it is useful to classify analytical methods according to their main purposes.

1. Methods for qualitative analysis
2. Methods for measuring main components, assaying
3. Methods for trace analysis
4. Methods for the establishment of physicochemical properties.

The requirements for validation that follow for the different classes of applications are given in Table 3.10.

These performance characteristics have already been described in an earlier part of the chapter and do not require further discussion. It should be stressed, however, that selectivity must be demonstrated in the course of validation by accurate and reliable measurements on real samples. A test of selectivity is at the same time a test of the influence of interference on the results.

Particular attention should also be drawn to the fact that the working range of a method of analysis is never

Table 3.10 Purpose of a method of measurement and the relevant performance characteristics in validation

	(a) Qualitative	(b) Main component/assay	(c) Trace analysis	(d) Phys.-chem. properties
Trueness		×	×	×
Precision		×	×	×
Linearity/working range		×	×	×
Selectivity	×	×	×	
Limit of detection	×		×	
Limit of determination			×	
Robustness	×	×	×	×

larger than that tested on real samples in the course of validation. Extrapolation to smaller or larger values cannot be tolerated.

In practise, this leads to a definition of the limit of determination by the sample with the smallest content for which data on trueness and precision are available. The lower limit of the working range therefore also defines the limit of determination; the upper limit of the working may sometimes be extended by suitable dilutions.

Frequency of Validation

The situation regarding the frequency of validation is comparable to the situation for the appropriate amount of validation; there are no firm and generally applicable rules, and only recommendations can be offered that help the person responsible for validation with a competent assessment of the particular situation. Some such recommendations can also be found in ISO 17025 Chap. 5.4.5. Besides those cases where a basic validation is in order, e.g., at the beginning of the lifecycle of a method, there is the recommendation to validate

> *standard methods used outside their intended scope, and amplifications and modifications of standard methods to confirm that the methods are fit for the intended use,*

and

> *when some changes are made in the validated nonstandard methods, the influence of such changes carried out should be documented and if appropriate a new validation should be carried out.*

In routine work, regular checks are required to make sure that the fitness for the intended use is not compromised in any way. In practise this is best done by control charts.

It is fair to state that, in essence, the frequency and extent of revalidation depend on the problem and on the magnitude of changes applied to previously validated methods. In a way it is therefore not time but a particular event that triggers the quest for revalidation. For a simple orientation and overview, some typical examples are addressed in Table 3.9.

If a new sample is analyzed, this might constitute the simplest event calling for validation measures. Depending on the method applied, this might be accomplished by adding an internal standard, by the method of standard additions, or by calling for duplicate measurements. If a new batch of samples is to be analyzed, it may be appropriate to take some additional actions, and it is easily seen that the laboratory supervisory personnel must incorporate flexibility in the choice of the appropriate revalidation action.

A special case is the training of new laboratory personnel, as the workload necessary may be significant, for instance, if difficult clean-up operations are involved. It may be advisable to have a backup operator trained in order to have a smooth transition from one operator to another without interruption of the laboratory workflow.

System Suitability Test

A system suitability test (SST) is an integral part of many analytical methods. The idea behind an SST is to view equipment, electronics, analytical operations, and samples as one system that can therefore be evaluated in total. The particular test parameters of an SST therefore critically depend on the type of method to be validated. In general, an SST must give confidence that the test system is operating without problems within specified tolerances. An SST is carried out with real samples, and it therefore cannot pinpoint problems with a particular subsystem.

Details can be found in pharmaceutical science literature, particularly in pharmacopeia. In the literature there are several examples for SST, e.g., for HPLC.

If an SST is applied regularly, it is generally laid down in a separate standard operating procedure.

Table 3.11 Event-driven actions in revalidation; adapted from [3.37]

Event	Action taken for revalidation
A new sample	Internal standard Standard additions Duplicate analysis
Several new samples (a new batch)	Blank(s) Recalibration Measurement of a standard reference material or a control check sample
New operator	Precision Calibration Linearity Limit of detection Limit of determination Control check sample(s)
New instrument	General performance check Precision Calibration Limit of detection Limit of determination Control check samples
New chemicals/standards	Identity check for critical parameters Laboratory standards
New matrix	Interlaboratory comparisons New certified reference material Alternative methods
Small changes in analytical methodology	Proof of identical performance over the concentration range and range of matrices (*method comparison*)

Report of Validation and Conclusions

Around the globe, millions of analytical measurements are performed daily in thousands of laboratories. The reasons for doing these measurements are extremely diverse, but all of them have in a common the characteristic that the cost of measurement is high, but the decisions made on the basis of the results of these measurements involve yet higher cost. In extreme cases, they can lead to fatal consequences; points in case are measurements in the food, toxicological, and forensic fields.

Results of analytical measurements are truly of foremost importance throughout life, demonstrating the underlying responsibility to ensure that they are correct. Validation is an appropriate means to demonstrate that the method applied is truly fit for purpose. For every method applied, a laboratory will have to rely on validation for confidence in the operation of the method.

The elements of validation discussed in this chapter must ascertain that the laboratory produces, in every application of a method, data that are well defined with respect to trueness and precision. The basics of quality management aid in providing this confidence. Therefore, every laboratory should be prepared to demonstrate its competence on the basis of internal data not only for methods it has devised itself, but also for standard methods of analysis. Revalidation will eventually be required for all methods to keep this data up to date.

A laboratory accredited according to ISO 17025 must be able, at any time, to demonstrate the required performance by well-documented validation results.

3.6 Interlaboratory Comparisons and Proficiency Testing

Interlaboratory comparisons (ILCs) are a valuable quality assurance tool for measurement laboratories, since they allow direct monitoring of the comparability of measurement and testing results. Proficiency tests (PTs) are interlaboratory comparisons that are organized on a continuing or ongoing basis. PTs and ILCs are therefore important components in any laboratory quality system. This is increasingly recognized by national accreditation bodies (NABs) in all parts of the world, who are increasingly demanding that laboratories participate in PTs or ILCs where these are available and appropriate.

PTs and ILCs enable laboratories to benchmark the quality of their measurements. Firstly, in many ILCs, a laboratory's measurement results may be compared with reference, or true, values for one or more parameters being tested. Additionally, where applicable, the associated measurement uncertainties may also be compared. These reference values will be the best estimate of the true value, traceable to national or international standards or references. Reference values and uncertainties are determined by expert laboratories; these will often be national measurement institutes (NMIs).

However, not all ILCs and PTs will be used to determine reference values. In most of these cases, a laboratory will only be able to benchmark their results against other laboratories. In these situations, a consensus value for the true value will be provided by the organizer, which will be a statistical value based upon the results of the participating laboratories or a value derived form extended validation.

3.6.1 The Benefit of Participation in PTs

The primary benefit from participating in PTs and ILCs for a laboratory is the ability to learn from the experience. Organizers of PTs and ILCs usually see themselves in the role of teachers rather than policemen. PTs and ILCs are therefore viewed as educational tools, which can help the participating laboratories to learn from their participation, regardless of how successful the participation is.

There are many quality assurance tools available to laboratories, including

- appropriate training for staff,
- validation of methods for testing and calibration,
- use of certified reference materials and artifacts,
- implementation of a formal quality system and third-party accreditation,
- participation in appropriate PTs and ILCs.

It is usually recommended that all these tools be used by measurement laboratories. However, laboratories are now recognizing the particular importance of participation in PTs and ILCs as a quality tool. Of the tools listed above, it is the only one that considers a laboratory's outputs, i. e., the results of its measurements. The other tools are input tools, concerned with quality assurance measures put in place to provide the infrastructure necessary for quality measurements.

As a consequence of this, appropriate use of participation in PTs and ILCs is of great value to laboratories in assessing the validity of the overall quality management system. Appropriate participation in PTs and ILCs can highlight how the quality management system is operating, where any problems may be found that have an effect on the measurement results expected. Regular participation can therefore form a continuous feedback mechanism, enabling the quality management system to be monitored and improved on an ongoing basis. In particular, following poor performance in a PT or ILC, laboratories should institute an investigation, which may result in corrective action being taken. This corrective action may involve changes to the quality management system and its documentation.

3.6.2 Selection of Providers and Sources of Information

There are literally thousands of PTs and ILCs offered during any year, across all measurement sectors, by reputable organizations across the world. Laboratories can gain information about available PTs and ILCs from a number of sources. These include

- the European Proficiency Testing Information System (EPTIS),
- national accreditation bodies (NABs),
- international accreditation bodies [e.g., the Asian Pacific Accreditation Cooperation (APLAC), ILAC, and the European Cooperation for Accreditation (EA)],
- peer laboratories.

National accreditation bodies (NABs) will hold, as part of their normal laboratory surveillance and assessment activities, a great deal of information about PTs and ILCs (or organizations that run ILCs). They will have noted, during laboratory surveillance visits, what these PTs and ILCs cover, how they operate, and how relevant they are to the laboratory's needs. NABs are therefore in a good position to provide information about available and appropriate PTs and ILCs and, in some cases, may advise on the suitability and quality of these. Some NABs also accredit PT providers, usually against ISO guide 43 part 1 (1997) and ILAC guide G13 (2000). These NABs will therefore have more detailed information regarding accredited PTs, which they can pass on to laboratories.

International accreditation bodies such as APLAC, ILAC or EA will also have a significant body of information regarding international or regional PTs and ILCs. Additionally they may organize PTs and ILCs themselves, or associate themselves with specific PTs and ILCs, which they use for their own purposes, such as monitoring the efficacy of multilateral agreements (MLAs) or multiregional agreements. APLAC, for example, associates itself with a number of ILCs, which are usually organized by member accreditation bodies. EA may be involved with independent PT and ILC organizers, such as the Institute of Reference Materials and Measurements (IRMM) in Geel, Belgium, who organize the International Measurement Evaluation Programme (IMEP) series of ILCs.

The *European Proficiency Testing Information System* (EPTIS) is the leading international database of PTs and ILCs. EPTIS was originally set up with funding from the European Commission, and is now maintained by the German Federal Institute of Materials Testing (BAM) in Berlin. EPTIS contains over 800 PTs and ILCs across all measurement sectors excluding metrology. Although originally established as a database for the pre-May 2004 countries within the European Union, plus Norway and Switzerland, it has now been extended to include the new European Union (EU) countries, the USA, as well as other countries in South and Central America and Asia. EPTIS now enjoys the support of the International Laboratory Accreditation Conference (ILAC), and has the goal of extending its coverage to include potentially all providers of PTs and ILCs throughout the world. The database, however, is searchable by anyone, anywhere in the world. It is accessed online at www.eptis.bam.de. It can be searched for PTs by country, test sample type, measurement sector, or determinand. The details contained in EPTIS for each PT and ILC are comprehensive. These include

- organizer,
- frequency,
- scope,
- test samples,
- determinands,
- statistical protocol,
- quality system,
- accreditation status,
- fees payable.

Many of the entries also contain a link to the home page of the provider so that more in-depth information can be studied.

EPTIS also provides more general information on the subject of proficiency testing. Any laboratory wishing to find a suitable PT or ILC in which to participate is strongly advised to search EPTIS first. One warning must, however, be given. Although there is no cost to PT providers to have an entry on EPTIS, it is voluntary, and therefore there are a small number of PTs in the countries covered by EPTIS which are not listed.

Peer laboratories are a good source of information about available and appropriate PTs and ILCs. A laboratory working in the same field as your own may be a good source of information, particularly if they already participate in a PT, or have investigated participation in a PT or ILC. Although such laboratories may be commercial competitors, a PT or ILC that is appropriate for them is very likely to be appropriate for all similar laboratories in that measurement sector.

When a laboratory has obtained the information about available ILCs and PTs, there may be a need to make a decision.

- Is there more than one ILC/PT available? If so, which is the most appropriate for my laboratory?
- There is only one ILC/PT that covers my laboratory's needs. Is it appropriate for my laboratory to participate?

There are many issues that are appropriate to both the above questions. In order to make the correct decision, there are a number of aspects of the ILCs/PTs that must be understood. To select the most appropriate ILC or PT, or determine if an ILC or PT is appropriate for a specific laboratory, the following factors need to be considered.

- Test samples, materials or artifacts used.
- Measurands, and the magnitude of these measurands.
- What is the frequency of distribution for a PT scheme?
- Who are the participants?
- What quality system is followed by the organizer?
- In which country is the ILC or PT organized, and what language is used?
- What is the cost of participation?

We will consider these factors individually below.

Test Samples, Materials or Artifacts Used

The laboratory must satisfy itself that the test samples, materials or artifacts used in the PT or ILC are appropriate to their needs. The test materials should be of a type that the laboratory would normally or routinely test. They should be materials that are covered by the scope of the laboratory's test procedures. If the materials available in the PT or ILC are not fully appropriate – they may be quite similar but not ideal – the laboratory must make a judgement as to whether participation would have advantages. The laboratory could also contact the PT or ILC organizer to ask if the type of material appropriate to them could be included.

Measurands and the Levels of These Measurands

If the test materials in the PT or ILC are appropriate for the laboratory, then the question of the measured properties (measurands) needs to be taken into consideration. The measurands available should be the same as the laboratory would routinely measure. Of course, for those materials where many tests could be carried out, the PT or ILC may not routinely provide all of these. Again, the laboratory must make a judgement about whether the list of tests available is appropriate and fits sufficiently well with the laboratory's routine work to make participation worthwhile.

The origin of the samples is also important to many laboratories. The laboratory needs to know where and how they were prepared, or from which source they were obtained. For example, it is important to know whether they have been tested for homogeneity and/or stability. If so, where there is more than one measurand required for that material, the laboratory needs to know for which measurands. A good-quality PT or ILC will prepare sufficient units that surplus samples are available for participants later, particularly those who need them following poor performance.

What Is the Frequency of Distribution for a PT Scheme?

For PT schemes, rather than ILCs, the frequency of distributions, or rounds, is important. The frequency of PTs does vary from scheme to scheme and from sector to sector. Most PTs are distributed between two and six times a year, and a frequency of three or four rounds per year is quite common. The frequency is important for laboratories, in case of unsatisfactory performance in a PT, when the efficacy of corrective actions must be studied to ensure any problem has been properly corrected.

Who Are the Participants?

For any PT or ILC, it is important that a laboratory can compare its results with peer laboratories. Peer laboratories may not always be those who carry out similar tests.

Laboratories in different countries may have different routine test methods – these may be specified by regulation. In some cases, these test methods will be broadly equivalent technically, but in other cases their performance may be significantly different. In fact, in this case, this situation may not be recognized by laboratories or expert sectoral bodies. Comparison with results generated using such methods will be misleading.

Even within any individual country, there may be differences in the test methods used by laboratories. The PT or ILC organizer should be able to offer advice on which test methods may be used by participants, how these vary in performance, and what steps the organizer will follow to take these into account when evaluating the results.

The type of laboratories participating in a PT or ILC is also important. For a small nonaccredited laboratory, comparison with large, accredited laboratories or national measurement institutes (NMIs) may not be appropriate. The measurement capabilities of these different types of laboratories, and the magnitude of their estimated measurement uncertainties will probably be significantly different. The actual end use of results supplied by different types of laboratories to their customers will usually determine the level of accuracy and uncertainty to which these laboratories will work.

What Quality System Is Followed by the Organizer?

For laboratories who may rely significantly on participation in PTs or ILCs, or if they are accredited and are required to participate by their national accreditation body (NAB), as a major part of their quality system, it is important that the schemes they use are of appropriate quality. This gives laboratories a higher degree of confidence in the PT or ILC, and hence the actions they may need to take as a result of participating.

In recent years the concept of quality for PTs has gained more importance. ISO/IEC guide 43 parts 1 and 2 were reissued in 1997, and many PT and ILC organizers claim to follow this. In practise, this guide is very generic, but compliance with it does confer a higher level of quality. The development of the ILAC guide G13:2000 has, however, enabled many accreditation bodies throughout the world (including in countries such as The Netherlands, Australia, the UK, Spain, Sweden, and Denmark) to offer accreditation of PT scheme providers as a service. Most accreditation bodies who offer this service accredit providers against a combination of ISO/IEC guide 43 and ILAC G13. Guide G13 is a considerably more detailed document and is generally used as an audit protocol. Not all NABs accredit PT and ILC organizers using these documents; some NABs in Europe prefer the approach of using ISO/IEC 17020, considering the PT or ILC organizers to be inspection bodies. In Europe, the policy of the EA is that it is not mandatory for NABs to provide this service, but that, if they do, they should accredit using a combination of ISO guide 43 part 1 (1997) and the ILAC guide G13:2000, which is also the preferred approach within APLAC.

Information on quality is listed on EPTIS, and now information on accreditation status is also included, at the request of ILAC.

Laboratories need to make a judgement on whether an accredited scheme is better than a nonaccredited scheme where a choice is available.

The quality of a PT or ILC is important, as the operation of such an intercomparison must fit well with the requirements of participating laboratories. All PTs and ILCs should have a detailed protocol, available to all existing and potential participating laboratories. The protocol clearly illustrates the modus operandi of the PT or ILC, including timescales, contacts, and the statistical protocol. The statistical protocol is the heart of any intercomparison, and should comprehensively show how data should be reported (e.g., number of replicates and reporting of measurement uncertainty), how the data is statistically evaluated, and how the results of the evaluation are reported to participating laboratories. Laboratories need to understand the principles of the statistical protocol of any PT or ILC in which they participate. This is necessary in order to understand how their results are evaluated, which criteria are used in this evaluation, and how these fit with the laboratory's own criteria for the quality and fitness for purpose of results. It is therefore important to find a PT or ILC that asks for data in an appropriate format for the laboratory and evaluates the data in a way that is broadly compatible with the laboratory's own procedures.

In Which Country Is the ILC or PT Organized, and What Language Is Used?

Where a laboratory has a specific need which cannot be met by a PT or ILC in their own country, or where a choice between PTs or ILCs exists where one or more of these are organized in countries outside their own, the country of origin may be important.

The modus operandi of many PTs and ILCs may vary significantly between countries, particularly with regard to the statistical evaluation protocol followed. This may be important where a laboratory wants to take part in a PT or ILC that fits well with their own internal quality procedures.

More important for many laboratories is the language in which the PT or ILC documentation is written. A number of PTs or ILCs may be aimed mainly at laboratories in their own country and will use only their native language. Laboratories wishing to participate in such a PT or ILC will need to ensure that they have members of staff who can use this language effectively. Other PTs and ILCs are more international in nature, and may use more than one language. In particular, many of these will issue documents in English as a second language.

What Is the Cost of Participation?

If a laboratory has researched the available PTs and ILCs and has found more than one of these that could be appropriate, the final decision may often be made on the basis of cost.

Some laboratories see participation in PTs and ILCs as another cost that should be minimized. Some accredited laboratories see participation as an extra cost on top of what they already pay for accreditation.

Therefore, cost is an important factor for some laboratories. However, it should be noted that a less expensive scheme may not always provide the quality or service that is required for all the many benefits of participation in PTs and ILCs to be realized.

Some laboratories successfully negotiate with the organizers where cost is a real issue for them (e.g., very small laboratories, university laboratories, laboratories in developing economies, etc.). Laboratories should note that the cost of participation is not just the subscription that is paid to the organizer. The cost in time and materials of testing PT and ILC test materials or samples also needs to be taken into account.

What if There is no Appropriate PT or ILC for a Laboratory's Needs?

When the right PT or ILC does not exist, a laboratory can participate in one which is the best fit, or decide not to participate at all. In this case, reliance on other quality measures will be greater. A laboratory can approach a recognized organizer of PTs and ILCs to ask if an appropriate intercomparison can be organized. Also, a laboratory may collaborate with a group of laboratories with similar needs (these groups will nearly always be quite small, otherwise a PT or ILC will probably already have been organized), to organize small intercomparisons between themselves.

3.6.3 Evaluation of the Results

It is important for laboratories, when they have participated in any PT or ILC, to gain the maximum benefit from this. A major aspect of this is in the interpretation of the results from a PT or ILC, and how to use these results to improve the quality of measurements in the laboratory.

There are a number of performance evaluation procedures used in PT schemes. Two of the most widely used of these are outlined here.

1. Z-scores
2. E_n numbers.

Z-scores are commonly used in many PT schemes across the world, in many sectors. This performance evaluation technique is probably the most widely used on an international basis.

E_n numbers incorporate measurement uncertainty and are used in calibration studies and by many ILCs where the measurement uncertainty is an important aspect of the measurement process. E_n numbers are therefore used more commonly in physical measurement ILCs and PTs, where the measurement uncertainty concept is much better understood.

More examples of performance evaluation techniques can be found in the ISO standard for statistics used in proficiency testing, ISO 13528 (2005).

Z-Scores

Z-scores are calculated according to the following equation:

$$Z = (x_\text{I} - X)/s\ ,$$

where x_I is the individual result, X is the assigned or true value, and s is a measure of acceptability. For example, s can be a percentage of X: if X is 10.5, then if results should be within 20% of this to be awarded a satisfactory Z-score, the s will be 10% of 10.5, i. e., 1.05. It could also be a value considered by the organizer to be appropriate from previously generated precision data for the measurement. s may also be a statistically calculated value such as the standard deviation, or a robust measure of the standard deviation.

The assigned value can be either a *reference* value or a *consensus* value. Reference values are traceable and can be obtained, for example, from

- formulation (the test sample is prepared in a quantitative manner so that its properties and/or composition are known),
- reference measurement (the test sample has been characterized using a primary method, or traceable to a measurement of a certified reference material of a similar type).

Consensus values are obtained from the data submitted by participants in a PT or ILC.

Most schemes will classify Z-scores as

- satisfactory ($|Z| \leq 2$),
- questionable ($2 > |Z| > 3$),
- unsatisfactory ($|Z| \geq 3$).

These are broadly equivalent to internal quality control charts, which give warning limits (equivalent to a *questionable* result) and action limits (equivalent to an *unsatisfactory* result).

E_n Numbers

The equation for the calculation of E_n numbers is

$$E_n = \frac{x - X}{\sqrt{U_\text{lab}^2 + U_\text{ref}^2}}\ ,$$

where the assigned value X is determined in a reference laboratory, U_ref is the expanded uncertainty of X, and U_lab is the expanded uncertainty of a participant's result x.

E_n numbers are interpreted as follows.

- Satisfactory ($E_n \leq 1$)
- Unsatisfactory ($E_n > 1$).

Laboratories are encouraged to learn from their performance in PTs and ILCs. This includes both positive and negative aspects.

Action should be considered

- when an unsatisfactory performance evaluation has been obtained (this is mandatory for laboratories accredited to ISO/IEC 17025), or
- when two consecutive questionable results have been obtained for the same measurement, or
- when nine consecutive results with the same bias against the assigned value, for the same measurement, have been obtained. This would indicate that, although the measurements may have been very precise, there is a clear bias. Deviations from this

situation could easily take the measurements out of control.

The above guidelines should enable laboratories to use PT and ILC results as a way of monitoring measurement quality and deciding when action is necessary.

When interpreting performance in any PT or ILC, there are a number of factors that need to be considered to enable the performance to be placed into a wider context. These include

- the overall results in the intercomparison from all participating laboratories,
- the performance of different testing methods,
- any special characteristics or problems concerning the test sample(s) used in the intercomparison,
- bimodal distribution of results,
- other factors concerning the PT or ILC organization.

It is always advisable to look at any unsatisfactory performance in the context of all results for that measurement in the intercomparison. For example, if the majority of the results have been evaluated as satisfactory, but one single result has not, then this is very serious. However, if many participating laboratories have also been evaluated as unsatisfactory, then for each laboratory with an unsatisfactory performance, there is still a problem but it is less likely to be specific to each of those laboratories.

It is also a good idea to look at how many results have been submitted for a specific measurement. When there are only a few results, and the intercomparison has used a consensus value as the assigned value for the measurement, the confidence in this consensus value is greatly reduced. The organizer should provide some help in interpreting results in such a situation and, in particular, should indicate the minimum number of results needed.

3.6.4 Influence of Test Methods Used

In some cases, an unsatisfactory performance may be due, at least in part, to the test method used by the laboratory being inappropriate, or having lower performance characteristics than other methods used by other laboratories in the intercomparison.

If the PT or ILC organizer has evaluated performance using the characteristics of a standard method, which may have superior performance characteristics, then results obtained using test methods with inferior performance characteristics will be more likely to be evaluated as unsatisfactory. It is always suggested that in such situations participating laboratories should compare their results against other laboratories using the same test method.

Some PTs and ILC will clearly differentiate between the various test methods used in the report, so the performance of each test method can be compared in order to see if there is a difference in precision of these test methods, and any bias between test methods can also be evaluated. The performance of all participating laboratories using the same test method can be studied, which should give laboratories information about both the absolute and relative performance of that test method in that intercomparison.

As has been previously stated, the test samples used in PTs and ILCs should be similar to those routinely measured by participating laboratories. A PT scheme may cover the range of materials appropriate to that scheme, so some may be unusual or extreme in their composition or nature for some of the participating laboratories. Such samples or materials should ideally be of a type seen from time to time by these laboratories. These laboratories should be able to make appropriate measurements on these test samples satisfactorily, if only to differentiate them from the test samples they would normally see. These unusual samples can, however, present measurement problems for laboratories when used in a PT or ILC, and results need to be interpreted accordingly.

In some cases, the value of the key measurands may be much higher or lower than what is considered to be a normal value. This can cause problems for laboratories, and results need to be interpreted appropriately, and lessons should be learned from this. If the values are in fact outside the scope of a laboratory's test methods, then any unsatisfactory performance may not be surprising, and investigation or corrective actions do not always need to be carried out.

One consequence of divergence of performance of different test methods, which may not necessarily be related to the test samples, is that a bimodal distribution of results is obtained. This is often caused by two test methods which should be, or are considered by experts in the appropriate technical sector to be, technically equivalent showing a significant bias. This could also arise from two different interpretations of a specific test method, or the way the results are calculated and/or reported. Problems that are typically encountered with reporting include the units or number of significant figures. When the assigned value for this measurement is a consensus value, this will have a more significant effect on result evaluation. Automatically, any smaller group of laboratories will be evaluated as unsatisfactory, regardless. In extreme cases, the two distributions will contain the same number of results, and then the consensus value will lie between them, and probably most, if not all, results will be evaluated as unsatisfactory.

In these cases, the organizer of the PT or ILC should take action to ensure that the effect of this is removed or minimized, or no evaluation of performance is carried out in order that laboratories do not misinterpret the evaluations and carry out any unnecessary investigations or corrective actions.

Although organizers of PTs and ILCs should have a quality system in place, occasionally some problems will arise that affect the quality of the evaluation of performance that they carry out. These can include, for example,

- transcription errors during data entry,
- mistakes in the report,
- software problems,
- use of inappropriate criteria for evaluation of performance.

In these cases, the evaluation of the performance of participating laboratories may be wrong, and the evaluation must either be interpreted with caution or, in extreme situations, ignored. The organizer of the PT or ILC should take any necessary corrective action once the problem has been identified.

3.6.5 Setting Criteria

In setting the criteria for satisfactory performance in a PT or ILC, the organizer, with the help of any technical steering group, may need to make some compromises in order to set the most appropriate criteria that will be of value to all participating laboratories. These criteria should be acceptable and relevant to most laboratories, but for a small minority these may be inappropriate. From a survey carried out by the author in 1997, some laboratories stated that they chose to use their own criteria for performance evaluation, rather than those used by the PT or ILC organizer. For most of these laboratories, the criteria they chose were tighter than those used in the PT or ILC.

Laboratories are normally free to use their own criteria for assessing their PT results if those used by the scheme provider are not appropriate, since the PT provider can obviously not take any responsibility for participating laboratories' results. These criteria should be fit for purpose for the individual laboratory's situation, and should be applied consistently. Interpretation of performance using these criteria should be carried out in the same manner as when using the criteria set by the PT or ILC organizer.

3.6.6 Trends

It is very useful to look at trends in performance in a PT that is carried our regularly. This is particularly useful when a laboratory participates at a relatively high frequency (e.g., once every 3 months).

Performance over time is the major example of this. The example in Fig. 3.19 shows how this may be illustrated graphically. This approach is recommended by experts rather than using statistical procedures, which may produce misleading information or hide specific problems.

The chart shows an example from a Laboratory of the Government Chemist (LGC) PT scheme of a graph showing performance over time. *Z*-scores for one measurement are plotted against the round number. In this case, the laboratory has reported results using three different test methods. This graph can be used to assess trends and to ascertain whether problems are individual in nature or have a more serious underlying cause.

Where more than one test method has been used, these can also be used to see if there is a problem with any individual method, or whether there is a calibration problem, which could be seen if more than one test method shows a similar trend.

In many PTs and ILCs there may be measurements that are requested to be measured using the same method, or are linked to each other technically in some way. Where all results for such linked measurements are unsatisfactory, the problem is likely to be generic,

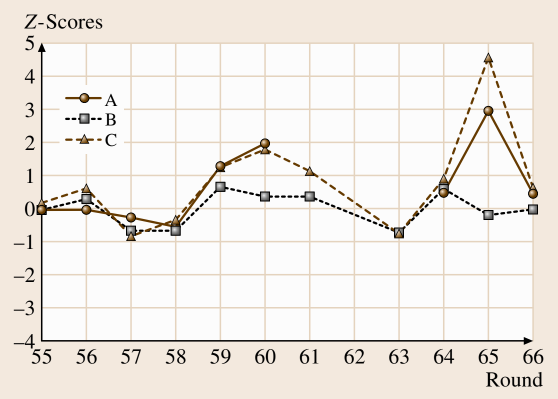


Fig. 3.19 Example graphical presentation of performance over time

and only one investigation and corrective action will be necessary.

Laboratory managers can gain information about the performance of individual staff on PT or ILC test samples. Information on each member of staff can be collated from PT and ILC reports and interpreted together with the information they should hold about which member of staff carried out the measurements.

Alternatively, where the test sample is of an appropriate nature, the laboratory manager can give the PT/ILC test sample(s) to more than one member of staff. Only one set of results needs to be reported to the organizer, but the results of appropriate members of staff can then be compared when the report is published.

Samples provided by the organizer should be tested in the same way as routine samples in order to get optimum feedback on performance. If this is not done, the educational benefit will be limited.

3.6.7 What Can Cause Unsatisfactory Performance in a PT or ILC?

There are many potential causes of unsatisfactory performance in any PT or ILC. These fall into two distinct categories.

- Analytical problems with the measurement itself
- Nonanalytical problems that usually occur after the measurement has been made.

Analytical errors include

- problems with calibration (e.g., the standard materials prepared to calibrate a measurement, or the accuracy/traceability of the calibration material),
- instrument problems (e.g., out of specification),
- test sample preparation procedures not being carried out properly,
- poor test method performance. This may be due to problems with the member of staff carrying out the measurement, or the appropriateness of the test method itself.

Nonanalytical errors include

- calculation errors,
- transcription errors,
- use of the wrong units or format for the reported result.

Any result giving rise to an unsatisfactory performance in a PT or ILC indicates that there is a problem in the laboratory, or a possible breakdown of the laboratory's quality system. It does not matter if the cause of this unsatisfactory result was analytical or nonanalytical as the result has been reported. At this point, it must be remembered that the PT or ILC organizer is acting in the role of the laboratory's customer and is providing a service to examine the laboratory's quality system thoroughly by means of an intercomparison.

The author's own experience of the organization of PTs over 10 years has shown that 35–40% of unsatisfactory results are due to nonanalytical errors.

3.6.8 Investigation of Unsatisfactory Performance

Participation in appropriate PTs and ILCs is strongly recommended by most national accreditation bodies for accredited laboratories and those seeking accreditation. Some NABs will stipulate that participation is mandatory in certain circumstances. Additionally, some regulatory authorities and, increasingly, customers of laboratories, will also mandate participation in certain PTs and ILCs in order to assist in the monitoring of the quality of appropriate laboratories.

It is mandatory under accreditation to ISO/IEC 17025 that an investigation be conducted for all instances of unsatisfactory performance in any PT or ILC, and to implement corrective actions where these are considered appropriate. All investigations into unsatisfactory performance in an intercomparison, and what, if any, corrective actions are implemented must be fully documented.

Some measurement scientists believe that unsatisfactory performance in any PT or ILC is in itself a noncompliance under ISO/IEC 17025. This is not true, although there are a few exceptions in regulatory PTs where participation is mandatory and specified performance requirements are stated. However, failure to investigate an unsatisfactory result is certainly serious noncompliance for laboratories accredited to ISO/IEC 17025.

It is generally recommended to follow the policy for the investigation of unsatisfactory performance in PTs and ILCs given by most national accreditation bodies, and the subsequent approach to taking corrective actions. All investigations should be documented, along with a record of any corrective actions considered necessary and the outcome of the corrective action(s).

There are a number steps that it is recommended should be taken when investigating unsatisfactory performance in any intercomparison. This should be done in a logical manner, working backwards.

Firstly, it should be checked that the PT or ILC organizer is not at fault. This should be done by ensuring that the report is accurate, that they have not entered any of the laboratory's data incorrectly, and that they have carried out all performance evaluations appropriately.

If the organizer has not made any errors, then the next check is to see that the result was properly reported. Was this done accurately, clearly, and in the correct units or format required by the PT or ILC?

If the result had been reported correctly and accurately, the next check is on any calculations that were carried out in producing the result.

If the calculations are correct, the next aspect to check is the status of the member of staff who carried out the measurement. In particular, was he or she appropriately trained and/or qualified for this work, and were the results produced checked by their supervisor or manager?

This should identify most sources of nonanalytical error. If no nonanalytical errors can be found, then analytical errors must be considered. When it appears that an unsatisfactory result has arisen due to analytical problems, there are a number of potential causes that should be investigated, where appropriate.

Poor calibration can lead to inaccurate results, so the validity of any calibration standards or materials must be checked to ensure that these are appropriate and within their period of use, and that the calibration values have been correctly recorded and used.

If the measurement has been made using an instrument – which covers many measurements – the status of that instrument should be checked (i. e., is it within its calibration period, and when was it last checked?). It is also recommended to check that the result was within the calibration range of the instrument.

Any CRM, RM or other QC material measured at the same time as the PT test sample should be checked with the result. If the result for such a material is acceptable, then a calibration or other generic measurement problem is unlikely to be the cause of the unsatisfactory performance.

Finally, the similarity of the test sample to routine test samples or, where appropriate, other samples tested in the same batch, should be noted.

This is not an exhaustive list, but covers the main causes.

When an investigation into unsatisfactory performance has indicated a potential cause, one or more corrective actions may need to be implemented. These include

- modifying a test method – which may then need revalidating,
- recalibration or servicing of an instrument,
- obtaining new calibration materials,
- changing the procedure for checking and reporting test results,
- considering whether any members of staff need further training, or retraining in particular test methods or techniques.

3.6.9 Corrective Actions

Corrective actions are not always necessary. Investigation of the situation may in fact conclude that

- no problem can be readily identified, and that the unsatisfactory result is just a single aberration – this needs monitoring, however, to ensure that this is not the beginning of a trend,
- there is a problem external to the laboratory – for example with the organize of the PT or ILC,
- the test sample from the PT or ILC is very unusual for the laboratory compared with the test samples they normally receive so that any corrective action will be of little or no value.

In some cases, it can prove very difficult for a laboratory to find the causes of unsatisfactory performance. Many PT and ILC organizers provide help to laboratories in such situations. It is always recommended to contact the organizer to ask for confidential help to solve such a problem. Many organizers have the expertise to give valuable advice, or can obtain the advice in strictest confidence from third parties.

Whatever is – or is not – done should be documented fully.

When corrective actions have been implemented, the laboratory needs to know that the actions have been successful. The corrective actions therefore need to be validated. The easiest way is to reanalyze the PT or ILC test sample. (If there is none remaining, some organizers will be able to provide another sample.) This will not, of course, be appropriate for checking nonanalytical errors. If the result from retesting agrees with the assigned value in the report, the corrective action can be considered to be successful. Alternatively (this is particularly true for more frequent PTs), it may be more appropriate to wait for the next round to be distributed and carry out the testing of the sample, so the efficacy of the corrective action can be assessed when the report is received. Doing both is the ideal situation, where appropriate, and will give greater confidence that the corrective action has been effective.

In some cases, the nature of the problem is such that there must be significant doubt about the quality of results made for the test under investigation, and that this problem may have existed for some weeks or months. In fact, the problem will certainly have occurred since the last PT or ILC where satisfactory performance for the test had been obtained.

The investigation in such a situation therefore needs to be deeper in order to ascertain which results within this timeframe have a high degree of confidence, and which may be open to questions as to their validity.

There are other, secondary, benefits from participation in appropriate PTs or ILCs. These include

- help with method validation,
- demonstration of competence to internal and external customers, accreditation bodies, and regulatory bodies,
- evaluation of technical competence of staff, which can be used in conjunction with a staff training programme.

3.6.10 Conclusions

Participation in PTs and ILCs is a very good way for a laboratory to demonstrate its competence at carrying out measurements. This may be for internal use (giving good news and confidence to senior management, for example) or giving positive feedback to the staff who carried out the measurements. Alternatively it may be used externally. Accreditation bodies, of course, will ask for evidence of competence from the results of PTs and ILCs. Regulatory authorities may ask for a level of PT or ILC performance from laboratories carrying out measurements in specific regulated areas. Customers of laboratories may require evidence of PT or ILC performance as part of their contractual arrangements. The laboratory can also be proactive in providing data to existing and potential customers to show their competence.

PT can also be used effectively in the laboratory as a tool for monitoring the performance of staff. This is particularly valuable for staff undergoing training, or who have been recently trained. The results obtained in an intercomparison can be used for this purpose, and appropriate feedback can be given. Where performance has been good, these results can be used as a specific example in a training record, and positive feedback should be given to the individual. Where performance has been less than satisfactory, it should be used constructively to help the individual improve, as part of any corrective action.

To conclude, PTs and ILCs are very important quality tools for laboratories. They can be used very effectively in contributing to the assessment of all aspects of a laboratory's quality system. The most valuable use of PTs and ILC participation is in the educational nature of proficiency testing.

3.7 Reference Materials

3.7.1 Introduction and Definitions

Role of Reference Materials in Quality Assurance, Quality Control, and Measurement

Reference materials (RMs) are widely used for the calibration of measuring systems and the validation of measurement procedures, e.g., in chemical analysis or materials testing. They may be characterized for nominal properties (e.g., chemical structure, fiber type, microbiological species, etc.) and for quantitative values (e.g., hardness, chemical composition, etc.). Nominal property values are used for identification of testing objects, and assigned quality values can be used for calibration or measurement trueness control. The measurand needs to be clearly defined, and the quantity values need to be, where possible, traceable to the SI units of measurement, or to other internationally agreed references such as the values carried by certified reference material [3.38].

The key characteristics of RMs, and therefore the characteristics whose quality needs to be assured, include the following: definition of the measurand, metrological traceability of the assigned property values, measurement uncertainty, stability, and homogeneity.

Users of reference materials require reliable information concerning the RM property values, preferably in the form of a certificate. The user and accreditation bodies will also require that the RM has been produced by a competent body [3.39, 40].

The producers of reference materials must be aware that the values they supply are invariably an indispensable link in the traceability chain. They must implement all procedures necessary to provide evidence internally and externally (e.g., by peer review, laboratory intercomparison studies, etc.) that they have met the conditions required for obtaining traceable results at all times.

There are a number of authoritative and detailed texts on various aspects of reference materials, and these are listed in Sect. 7.3.4. Reference materials are an important tool in realizing a number of aspects of measurement quality and are used for method validation, calibration, estimation of measurement uncertainty, training, and for internal quality control (QC) and external quality assurance (QA) (proficiency testing) purposes.

Different types of reference materials are required for different functions. For example, a certified reference material would be desirable for method validation, but a working-level reference material would be adequate for QC [3.39].

Definition of RM and CRM [3.41]

Reference material (RM) is a material, sufficiently homogeneous and stable with reference to specified properties, which has been established to be fit for its intended use in measurement or in examination of nominal properties.

Certified reference material (CRM) is a reference material, accompanied by documentation issued by an authoritative body and providing one or more specified property values with associated uncertainties and traceabilities, using valid procedures.

Related Terms.

- *Quantity*: property of a phenomenon, body or substance, where the property has a magnitude that can be expressed as a number and a reference.
- *Quantity value*: number and reference together expressing the magnitude of a quantity.
- *Nominal property*: property of a phenomenon, body or substance, where the property has no magnitude.
- *Measurand*: quantity intended to be measured.
- *Metrological traceability*: property of a measurement result whereby the result can be related to a reference through a documented unbroken chain of calibrations, each contributing to the measurement uncertainty.
- *Measurement standard (etalon)*: realization of the definition of a given quantity, with stated quantity value and associated measurement uncertainty, used as a reference.
- *Reference material producer*: technically competent body (organization or firm, public or private) that is fully responsible for assigning the certified or property values of the reference materials it produces and supplies which have been produced in accordance with ISO guides 31 and 35 [3.42].
- *European reference material* (ERM): new standard in certified reference materials issued by three European reference materials producers (IRMM, BAM, LGC).
- *In-house reference material*: material whose composition has been established by the user laboratory by several means, by a reference method or in collaboration with other laboratories [3.43].
- *Primary method* [3.44]: method having the highest metrological qualities, whose operation can be completely described, and understood, and for which a complete uncertainty statement can be written in terms of SI units. A primary direct method measures the value of an unknown without reference to a standard of the same quantity. A primary ratio method measures the ratio of an unknown to a standard of the same quantity; its operation must be completely described by a measurement equation. The methods identified as having the potential to be primary methods are: isotope dilution mass spectrometry, gravimetry (covering gravimetric mixtures and *gravimetric analysis*), titrimetry, coulometry, determination of freezing point depression, differential scanning calorimetry, and nuclear magnetic resonance spectroscopy. Other methods such as chromatography, which has extensive applications in organic chemical analysis, have also been proposed.
- *Standard reference materials* (SRMs): are certified reference materials issued by the National Institute of Standards and Technology (NIST) of the USA. SRM is a trademark.
- *Validation*: confirmation, through the provision of objective evidence, that the requirements for a specific intended use or application have been fulfilled [3.33].

3.7.2 Classification

Principles of Categorization

Physical, chemical character:

- Gases, liquids, solutions
- Metals, organics
- Inorganics

Preparation:

- Pure compounds, code of reference materials
- Natural or synthetic mixtures

- Artifacts and simulates
- Enriched and unenriched real-life samples

Function:

- Calibration of apparatus and measurement systems
- Assessment of analytical methods
- Testing of measurement devices
- Definition of measuring scales
- Interlaboratory comparisons
- Identification and qualitative analysis
- Education and training

Application field (this principle is mainly used in the catalogs of RM producers):

- Food and agriculture (meat, fish, vegetable, etc.)
- Environment (matter, soil, sediment, etc.)
- Biological and clinical (blood, urine, etc.)
- Metals (ferrous, nonferrous, etc.)
- Chemicals (gas, solvents, paints, etc.)
- Pure materials (chromatography, isotopes, etc.)
- Industrial raw materials and products (fuels, glass, cement, etc.)
- Materials for determination of physical properties (optical, electrical properties, etc.)

Metrological qualification CMC:

- Primary, secondary, and tertiary standards
- Reference, transfer, and working standards
- Amount of substance standards
- Chemical composition standards
- Gases, electrochemistry, inorganic chemistry, organic chemistry

Reliability:

1. Certified reference materials of independent institutions (NIST, IRMM, BAM, LGC)
2. CRM traceable to 1. of reliable producers (Merck, Fluka, Messer-Grießheim)
3. Reference materials derived from 1. or 2. (in-house RM, dilution, RM preparations)

3.7.3 Sources of Information

CRM Databases

Information about reference materials is available from a number of sources. The international database for certified reference materials Code d'Indexation des Materiaux de Reference (COMAR) contains information on about 10 500 CMC from about 250 producers in 25 countries. It can be accessed via the Internet [3.45]. Advisory services assist users identify the type of material required for their task and identify a supplier. A number of suppliers provide a comprehensive range of materials including materials produced by other organizations and aim to provide a one-stop shop for users. An additional Internet database of natural matrix reference materials is published by the International Atomic Energy Agency (IAEA) [3.46].

Calibration and Measurement Capabilities (CMC) of the BIPM [3.47]

In 1999 the member states of the Metre Convention signed the mutual recognition arrangement (MRA) on measurement standards and on calibration and measurement certificates issued by national metrology institutes. Appendix C of the CIPM MRA is a growing collection of the calibration and measurements capabilities (CMC) of the national metrology institutes. The CMC database is available for everyone on the website of the Bureau International des Poids et Mesures (BIPM) and includes reference materials as well as references methods.

The methods used are proved by key comparisons between the national metrology institutes. For chemical measurements the Comité Consultative pour la Quantité de Matière (CCQM) has been established. The CMC database provides a reliable service for customers all over the world to establish traceability.

Conferences and Exhibitions (Election)

PITTCON:	annual; largest RM conference and exhibition in the USA
ANALYTICA:	biannual; Munich
BERM:	biannual; biological and environmental RM

Guides (Selection)

- ISO guide 30:1992/Amd 1:2008 – Terms and definitions used in connection with reference materials [3.38]
- ISO guide 31:2000 – Contents of certificates of reference materials [3.48]
- ISO guide 32:1997 – Calibration of chemical analysis and use of certified reference materials
- ISO guide 33:2000 – Uses of certified reference materials
- ISO guide 34:2009 – General requirements for the competence of reference material producers [3.42]
- ISO guide 35:2006 – Certification of reference materials – General and statistical principles
- ISO/AWI guide 79 – Reference materials for qualitative analysis – Testing of nominal properties

- ISO/CD guide 80 – Minimum requirements for in-house production of in-house-used reference materials for quality control
- ISO/NP guide 82 – Reference materials – Establishing and expressing metrological traceability of quantity values assigned to reference materials
- ISO/TR 10989:2009 – Reference materials – Guidance on, and keywords used for, RM categorization
- ISO/WD TR 11773 – Reference materials transportation
- ILAC-G9:2005 – Guidelines for the selection and use of reference materials
- ISO/REMCO (ISO-Committee on Reference Materials) document N 330 – List of producers of certified reference materials, information by task group 3 (*promotion*)
- 4E-RM guide (B. King) – Selection and use of reference materials [3.39]
- European Commission document BCR/48/93 (Dec. 1994) – Guidelines for the production and certification of Bureau Communautaire de Référence (BCR) reference materials
- ISO/REMCO – List of producers of certified reference materials
- RM report (RMR) (http://www.rmreport.com/)
- European Commission document BCR/48/93 (Dec. 1994) – Guidelines for the production and certification of BCR reference materials
- NIST publication 260-100 (1993) – Standard reference materials – handbook for SRM users
- IUPAC orange book – Recommended reference materials for the realization of physicochemical properties (ed. K. N. Marsh, Blackwell Scientific, 1987)
- World Health Organization (WHO) – Guidelines for the preparation and characterization and establishment of international and other standards and reference reagents for biological substances, technical report series no. 800 (1990)

3.7.4 Production and Distribution

Requirements on RM Producers [3.42]

All or some of the following activities can be crucial in RM production, and their quality assessment can be crucial to the quality of the final RM.

- Assessment of needs and specification of requirements
- Financial planning and cost–benefit analysis
- Subcontracting and selection of collaborators
- Sourcing of materials including synthesis
- Processing of materials including purification, grinding, particle size separation, etc.
- Packaging, storage, and design of dispatch processes
- Homogeneity and stability testing
- Development and validation of measurement methods, including consideration of the traceability and measurement uncertainty of measurement results
- Measurement of property values, including evaluation of measurement uncertainty
- Certification and sign-off of the RM
- Establishment of shelf-life
- Promotion, marketing, and sales of RM
- Postcertification stability monitoring
- Postcertification corrective action
- Other after-sales services
- QC and QA of quality systems and technical aspects of the work.

Certification Strategies [3.49]

Interlaboratory Cooperation Approach. The producer organizes interlaboratory comparisons of selected experienced laboratories, contributing independent measurements. Systematic uncertainties can be identified and minimized.

Elite Group Method Approach. Only a few qualified laboratories contribute to the certification by validated, independent measurement methods.

Primary Method Approach. Only primary methods (CIPM definition [3.44]) are used for certification. A blunder check is recommended.

Most BCR, BAM, and EURONORM reference materials are certified by the interlaboratory cooperation approach. NIST prefers, however, the latter methods.

Homogeneity and Stability [3.48]

The homogeneity of an RM has to be estimated and noted on the certificate. It describes the smallest amount (of a divisible material) or the smallest area (of a reference object) for which the certified values are accurate in the given uncertainty range.

The stability or a RM has to be stated in the certificate and has to be tested by control measurements (e.g., control charts). Time-dependent changes of the certified values within the uncertainty range are tolerated.

List of Suppliers (Examples)

Institutes. NIST (USA), LGC (UK), National Physical Laboratory (NPL, UK), Laboratoire d'Essais (LNE, France), BAM (Germany), PTB (Germany), NMU (Japan), Netherlands Measurement Institute (NMi, The Netherlands), National Research Center for Certified Reference Materials (NRC-CRM, China), UNIM (Russia), Canadian Centre for Mineral and Energy Technology (CANMET, Canada), South African Bureau of Standards (SABS, South Africa), Orzajos Meresugyi Hivatal (OMH, Hungary), Slovenski Metrologicky Ustav (SMU, Slovak), Swedish National Testing and Research Institute (SP, Sweden), Glowny Urzad Miar (GUM, Poland), IRMM (Europe).

Associations. Pharmacopeia, the European Network of Forensic Science (ENFS), Bureau Communantaire de Référence (BCR), European Committee for Iron and Steel Standardization (ECISS), Codex Alimentarius Committee (food standard program), Environmental Protection Agency (EPA, environment), UBA (Bundesumweltamt, environment), GDMB, Verein Deutscher Eisenhüttenleute (VDEh).

Companies (Branches). Sigma-Aldrich, LGC-Promochem, Merck, Fluka, Polymer Standard Service GmbH, Ehrenstorfer, Brammer Standard Company, Messer-Grießheim (gas), Linde (gas).

3.7.5 Selection and Use

Requirements on RM

Generally, the demand for reference materials exceeds supply in terms of the range of materials and availability. It is rare to have a choice of alternative RMs, and the user must choose the most suitable material available. It is important, therefore, that users and accreditation bodies understand any limitations of reference materials employed.

There are, however, several hundred organizations producing tens of thousands of reference materials worldwide. Producers include internationally renowned institutions such as NIST, collaborative government-sponsored programs such as the EU BCR program, semicommercial sectoral or trade associations such as the American Oil Chemicals Association, and an increasing number of commercial organizations. The distinction between government institutes and commercial businesses is disappearing with the privatization of a number of national laboratories.

Not all materials that are used as reference materials are described as such. Commercially available chemicals of varying purity, commercial matrix materials, and products from research programs are often used as standards or reference materials. In the absence of certification data provided by the supplier, it is the responsibility of the user to assess the information available and undertake further characterization as appropriate. Guidance on the preparation of reference materials is given in ISO guides 31, 34, and 35, and guides on the preparation of working-level reference materials are also available.

The suitability of a reference material depends on the details of the analytical specification. Matrix effects and other factors such as concentration range can be more important than the uncertainty of the certified value as detailed. The factors to consider include

- measurand, including analyte,
- measurement range (concentration),
- matrix match and potential interferences,
- sample size,
- homogeneity and stability,
- measurement uncertainty,
- value assignment procedures (measurement and statistical),
- the validity of the certification and uncertainty data,
- track record of both,
- availability of certificate.

The validity of the certification and uncertainty data, including conformance to key procedures of ISO guide 35.

Track record of both the producer and the material. For example, when an RM in use has been subjected to an interlaboratory comparison, cross-checked by the use of different methods, or there is experience of use in a number of laboratories over a period of years. Availability of a certificate and report conforming to ISO guide 31 is needed.

All or some of the requirements may be specified in the customer and analytical specification, but often it will be necessary for the analyst to use professional judgement. Finally, quality does not necessarily equate to small uncertainty, and fitness-for-purpose criteria need to be used [3.39].

Certificates and Supporting Reports. Ideally, a certificate complying with ISO guide 31 and a report covering the characterization, certification, and statistical analysis procedures, complying with ISO guide 35, will be available. However, many RM, particularly older materials and materials not specifically produced as RM, may not fully comply with ISO guides 31 and 35. Alternative, equivalent information in whatever form available and that provides credible evidence of compliance can be considered acceptable. Examples include the following: technical reports, trade specifications, papers in journals or reports of scientific meetings, and correspondence with suppliers.

Assessment of the Suitability of Reference Materials. Laboratories must be able to explain and justify the basis of selection of all RMs and of course any decision not to use an RM. In the absence of specific information it is not possible to assess the quality of an RM. The rigor with which an assessment needs to be conducted depends on the criticality of the measurement, the level of the technical requirement, and the expected influence of the particular RM on the validity of the measurement. Only where the choice of RM can be expected to affect measurement results significantly is a formal suitability assessment required.

Requirements of ISO/IEC 17025 on Laboratories

Measurement Traceability (§ 5.6 of ISO/IEC 17025) General (§ 5.6.1). (The symbol § refers to parts of ISO 17025.) All equipment used for tests and/or calibrations, including equipment for subsidiary measurements (e.g., for environmental conditions) having a significant effect on the accuracy or validity of the result of the test, calibration or sampling, shall be calibrated before being put into service. The laboratory shall have an established program and procedure for the calibration of its equipment.

Note that such a program should include a system for selecting, using, calibrating, checking, controlling, and maintaining measurement standards, reference materials used as measurement standards, and measuring and testing equipment used to perform tests and calibrations.

Specific Requirements (§ 5.6.2) Calibration (§ 5.6.2.1). § 5.6.2.1.1. For calibration laboratories, the program for calibration of equipment shall be designed and operated so as to ensure that calibrations and measurements made by the laboratory are traceable to the International System of Units [Système International d'Unités (SI)].

§ 5.6.2.1.2. There are certain calibrations that currently cannot be strictly made in SI units. In these cases calibration shall provide confidence in measurements by establishing traceability to appropriate measurement standards such as: the use of certified reference materials provided by a competent supplier to give a reliable physical or chemical characterization of a material; the use of specified methods and/or consensus standards that are clearly described and agreed by all parties concerned. Participation in a suitable programme of interlaboratory comparisons is required where possible.

Testing (§ 5.6.2.2). § 5.6.2.2.1. For testing laboratories, the requirements given in § 5.6.2.1 apply for measuring and test equipment with measuring functions used, unless it has been established that the associated contribution from the calibration contributes little to the total uncertainty of the test result. When this situation arises, the laboratory shall ensure that the equipment used can provide the uncertainty of measurement needed. Note that the extent to which the requirements in § 5.6.2.1 should be followed depends on the relative contribution of the calibration uncertainty to the total uncertainty. If calibration is the dominant factor, the requirements should be strictly followed.

§ 5.6.2.2.2. Where traceability of measurements to SI units is not possible and/or not relevant, the same requirements for traceability to, for example, certified reference materials, agreed methods, and/or consensus standards, are required as for calibration laboratories (§ 5.6.2.1.2). (e.g., breath alcohol, pH value, ozone of air).

Reference Standards and Reference Materials (§ 5.6.3). Reference standards (§ 5.6.3.1). The laboratory shall have a programme and procedure for the calibration of its reference standards. Reference standards shall be calibrated by a body that can provide traceability as described in § 5.6.2.1. Such reference standards of measurement held by the laboratory shall be used for calibration only and for no other purpose, unless it can be shown that their performance as reference standards would not be invalidated. Reference standards shall be calibrated before and after any adjustment.

Reference materials (§ 5.6.3.2). Reference materials shall, where possible, be traceable to SI units of measurement, or to certified reference materials. Internal reference materials shall be checked as far as is technically and economically practicable.

Assuring the Quality of Test and Calibration Results (§ 5.9 of ISO/IEC 17025). The laboratory shall have quality control procedures for monitoring the validity of tests and calibrations undertaken. The resulting data shall be recorded in such a way that trends are detectable, and where practicable, statistical techniques shall be applied to the reviewing of the results. This monitoring shall be planned and reviewed and may include, but not be limited to, the following.

1. Regular use of certified reference materials and/or internal quality control using secondary reference materials
2. Participation in interlaboratory comparison or proficiency testing programmes
3. Replicate tests or calibrations using the same or different methods
4. Retesting or recalibration of retained items; correlation of results for different characteristics of an item.

Note that the selected methods should be appropriate for the type and volume of the work undertaken.

Application Modes

Method Validation and Measurement Uncertainty. Estimation of bias (the difference between the measured value and the true value) is one of the most difficult elements of method validation, but appropriate RMs can provide valuable information, within the limits of the uncertainty of the RM certified value(s) and the uncertainty of the method being validated. Although traceable certified values are highly desirable, the estimation of bias differences between two or more methods can be established by use of less rigorously certified RM. Clearly the RM must be within the scope of the method in terms of matrix type, analyte concentration, etc., and ideally a number of RM covering the full range of the method should be tested. Where minor modifications to a well-established method are being evaluated, less-rigorous bias studies can be employed.

Replicate measurements of the RM, covering the full range of variables permitted by the method being validated, can be used to estimate the uncertainty associated with any bias, which should normally be corrected for.

The uncertainty associated with an RM should be no greater than one-third of that of the sample measurement [3.38, 50].

Verification of the Correct Use of a Method. Successful application of a valid method depends on its correct use, with regard to both operator skill and the suitability of equipment, reagents, and standards. RM can be used for training, for checking infrequently used methods, and for trouble-shooting when unexpected results are obtained.

Calibration. Normally, a pure substance RM is used for calibration of the measurement stage of a method. Other components of the test method, such as sample digestion, separation, and derivatization, are, of course, not covered, and loss of analyte, contamination, and interferences and their associated uncertainties must be addressed as part of the validation of the method. The uncertainty associated with RM purity will contribute to the total uncertainty of the measurement. For example, an RM certified as 99.9% pure, with an expanded uncertainty $U(k=2)$ of 0.1%, will contribute an uncertainty component of 0.1% to the overall measurement uncertainty budget. In the case of trace analysis, this level of uncertainty will rarely be important, but for assay work, it can be expected to be significant.

Some other methods, such as x-ray-fluorescence (XRF) analysis, use matrix RM for calibration of the complete analytical process. In addition to a close matrix match, the analyte form must be the same in the samples and RM, and the analytical concentrations of the RM must span that of the samples.

ISO guide 32 provides additional useful information.

Quality Control and Quality Assurance (QC and QA). RM should be characterized with respect to homogeneity, stability, and the certified property value(s). For in-house QC, however, the latter requirement can be relaxed, but adequate homogeneity and stability are essential. Similar requirements apply to samples used to establish how well or badly measurements made in different laboratories agree. In the case of proficiency testing, homogeneity is essential and sample stability within the time scale of the exercise must be assessed and controlled. Although desirable, the cost of certifying the property values of proficiency testing samples often prohibits this being done, and consensus mean values are often used instead. As a consequence, there often remains some doubt concerning the reliability of assigned values used in proficiency testing schemes. This is because, although the consensus mean of a set of data has value, the *majority* is not necessarily correct and as a consequence the values carry some undisclosed element of uncertainty. The interpretation of proficiency testing data thus needs to be carried out with caution.

Errors and Problems of RM Use

Election of RM.

- Certificate not known
- Certificate not complete
- Required uncertainty unknown

- Contribution of calibration to total uncertainty of measurement unknown
- Wrong matrix simulation
- Precision of measurement higher than precision of certification of RM
- No need for a certified RM

Handling of RM.

- Amount of RM too small
- Stability date exceeded
- Wrong preparation of in-house RM
- Wrong preparation of sample
- Matrix of sample and RM differ too much

Assessment of Values.

- Wrong correction of matrix effect
- Use of incorrect quantities (e.g., molality for unspecified analyte)
- Uncertainty budget wrong

3.7.6 Activities of International Organizations

Standardization Bodies

ISO. The International Organization for Standardization (ISO) is a worldwide federation of national standards bodies from some 130 countries. The scope of the ISO covers standardization in all fields except electrical and electronic standards, which are the responsibility of the IEC (see below).

IEC. The International Electrotechnical Commission (IEC), together with the ISO, forms a specialized system for worldwide standardization – the world's largest nongovernmental system for voluntary industrial and technical collaboration at the international level.

ISO REMCO. REMCO is ISO's committee on reference materials, responsible to the ISO technical management board [3.51]. The objectives of REMCO are

- to establish definitions, categories, levels, and classification of reference materials for use by ISO,
- to determine the structure of related forms of reference materials,
- to formulate criteria for choosing sources for mention in ISO documents (including legal aspects),
- to prepare guidelines for technical committees for making reference to reference materials in ISO documents,
- to propose, as far as necessary, action to be taken on reference materials required for ISO work,
- to deal with matters within the competence of the committee, in relation with other international organizations, and to advice the technical management board on action to be taken.

ASTM. The American Society for Testing and Materials (ASTM) is the US standardization body with international activities. The committees of the ASTM are also involved in determining reference materials, providing cross-media standards, and working in other associated fields.

Accreditation Bodies

ILAC. International Laboratory Accreditation Cooperation (ILAC) and the International Accreditation Forum (IAF) are international associations of national and regional accreditation bodies. ILAC develops guides for production, selection, and use of reference materials.

EA. The European Cooperation for Accreditation (EA) is the regional organization for Europe. EA is directly contributing to the international advisory group on reference materials.

Metrology Organizations (Chap. 2)

BIPM. In 1875, a diplomatic conference on the metre took place in Paris, where 17 governments signed a treaty (the Metre Convention). The signatories decided to create and finance a scientific and permanent institute, the Bureau International des Poids et Mesures (BIPM).

CIPM. The Comité Internationale des Poids et Mesures (CIPM) supervises the BIPM and supplies chairmen for the consultative committees.

CCQM. The consultative committee for amount of substance (CCQM) is a subcommittee of the CIPM. It is responsible for international standards in chemical measurements, including reference materials.

OIML. The International Organization of Legal Metrology (OIML) was established in 1955 on the basis of a convention in order to promote global harmonization of legal metrology procedures. OIML collaborates with the Metre Convention and BIPM on international harmonization of legal metrology.

User Organizations (Users of RM)

EUROLAB. The European Federation of National Associations of Measurement, Testing, and Analytical Laboratories (EUROLAB) promotes cost-effective services, for which the accuracy and quality assurance requirements should be adjusted to actual needs. EUROLAB contributes to the international advisory group on reference materials.

EURACHEM. The European Federation of National Associations of Analytical Laboratories (EURACHEM) promotes quality assurance and traceability in chemical analysis. EURACHEM also contributes to the international advisory group on reference materials.

CITAC. The Cooperation for International Traceability in Analytical Chemistry (CITAC), a federation of international organizations, coordinates activities of international comparability of analytical results, including reference materials.

IAGRM. The International Advisory Group on Reference Materials (IAGRM) is the successor of the 4E/RM group (selection and use of reference materials). It coordinates activities of users, producers, and accreditation bodies in the field of reference materials. IAGRM published guides and policy papers. Presently, accreditation of reference materials producers according to ISO guide 34 is being discussed.

AOAC International. The Association of Official Analytical Chemists (AOAC) International also has a reference materials committee to develop RM for analytical chemistry.

IFCC. The International Federation of Clinical Chemistry and Laboratory Medicine (IFCC) develops concepts for reference procedures and reference materials for standardization and traceability in laboratory medicine.

Pharmacopeia. Pharmacopeias [European and US Pharmacopeia (USP)] provide analysts and researchers from the pharmaceutical industry and institutes with written standards and certified reference materials.

Codex Alimentarius Commission. This commission of the Food and Agriculture Organization (FAO) of the United Nations and the World Health Organization (WHO) deals with safety and quality in food analysis, including reference materials.

ENFSI. The Network of Forensic Science Institutes (ENFSI) recommends standards and reference materials for forensic analysis.

BCR. The Bureau Communautaire de Référence (BCR) of the European Commission has, since 1973, set up programs for the development of reference materials needed for European directives. The Institute of Reference Materials and Measurement (IRMM) in Geel is responsible for distribution.

3.7.7 The Development of RM Activities and Application Examples

Activities for the development of reference materials started as early as 1906 at the US National Bureau of Standards (NBS). In 1912, the first iron and steel reference materials were certified for carbon content in Germany by the Royal Prussian Materials Testing Institute MPA, predecessor of BAM, the Federal Institute for Materials Research and Testing.

As in other parts of the world, the production of RM in Europe was primarily organized nationally, but as early as 1958 three institutes and enterprises of France (F) and Germany (D) combined their efforts in issuing exclusively iron and steel RM under the common label EURONORM. In 1973, a supplier from the UK, and in 1998 a company from Sweden (S), joined this group (Fig. 3.20).

To overcome national differences, to avoid duplicate work, and to improve mutual acceptance, a new class of European reference materials (ERM) has been created. In October 2003, this initiative was launched by three major reference material producers in Europe: the Institute for Reference Materials and Measurements (IRMM), BAM, Germany, and the Laboratory of the Government Chemist (LGC), UK. ERM are certified reference materials that undergo uncompromising peer evaluation by the ERM Technical Board to ensure the highest quality and reliability according to the state of the art.

A similar initiative to commonly produce CRM in a harmonized way is currently taking place in the Asian Pacific region.

To illustrate reference materials and their impact for technology, industry, economy, and society some examples from sectors such as

1. currency,
2. industry,
3. food,
4. environment

are briefly presented.

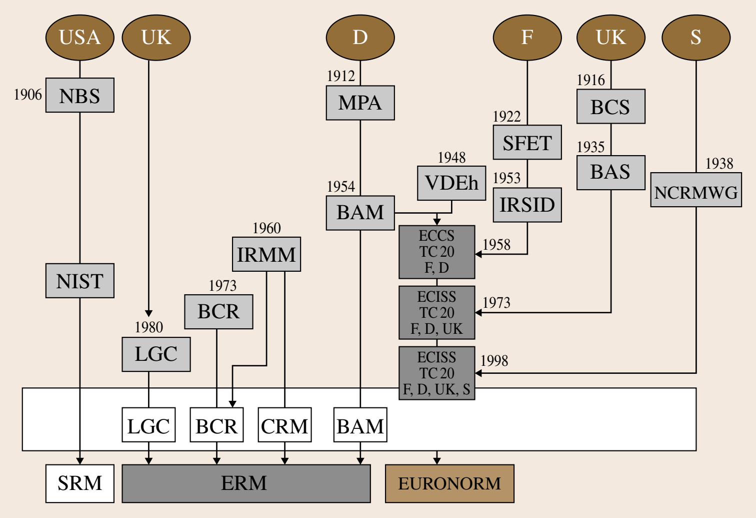


Fig. 3.20 Historical development of reference material activities in the USA and Western Europe (excerpt)

Fig. 3.21 Variety of reference materials highlighted by six fields of application

Currency

Since 2002, Europe has a new common currency: the Euro (€). To control and assure the alloy quality of the coins, several ERM have been issued (Fig. 3.22).

Industry

The automobile sector is an important industrial factor in all economies. There is a demand for automobiles to be exported also to countries with deviating standards for exhaust emission. Comparable, correct measurements are not only a national goal but a challenge with international implications. To support the detection of sulfur in gasoline, certified reference materials have been developed which cover the present legal limits in the European Union and in the USA (Fig. 3.23). These certified reference materials have two unique features: They are the first CRM made from commercial gasoline, and they offer lower uncertainties than presently available materials.

In addition to CRM, also interlaboratory comparisons are needed to assess reliably the determination of harmful substances such as sulfur in diesel fuel (Fig. 3.24). While the International Measurement Evaluation Programme (IMEP) is open to any laboratory, in the key comparison studies of the Consultative Committee for the Amount of Substance (CCQM-K) only national metrology institutes are accepted as participants.

Food

Toxic components in food affect health and endanger quality of life. Foodstuffs and a large number of other goods cross national borders. Legislation sets out limit values to protect consumers. RM such as ERM-BD475 *Ochratoxin A in roasted coffee* enable control (Fig. 3.25).

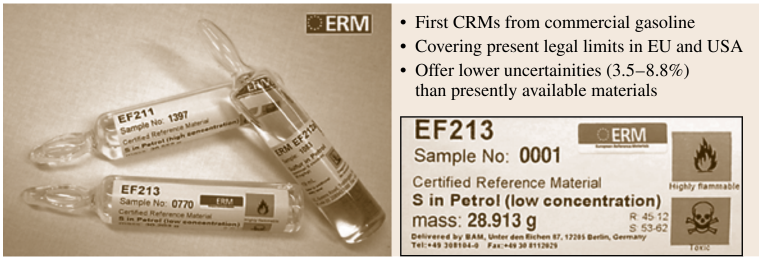
Environment

Harmful substances in industrial products may detrimentally influence technical functionality and may harm both man and the environment (Fig. 3.26). Consequently, CRM are needed to assess toxicity or show that industrial products are environmentally benign for the benefit of society and the economy.

3.7.8 Reference Materials for Mechanical Testing, General Aspects

In the area of mechanical testing, certified reference materials (CRM) are important tools to establish confidence and traceability of test results, as has been explained in Chap. 1 (Fig. 1.4). Usually, testing methods are defined in international ISO standards. In these standards, special focus is laid on direct calibration of all parts of the testing equipment as well as the related traceability of all measured values to national and/or international standards. Annual direct calibration is used to demonstrate this update of the measurement capabilities. Within the calibration interval only a few

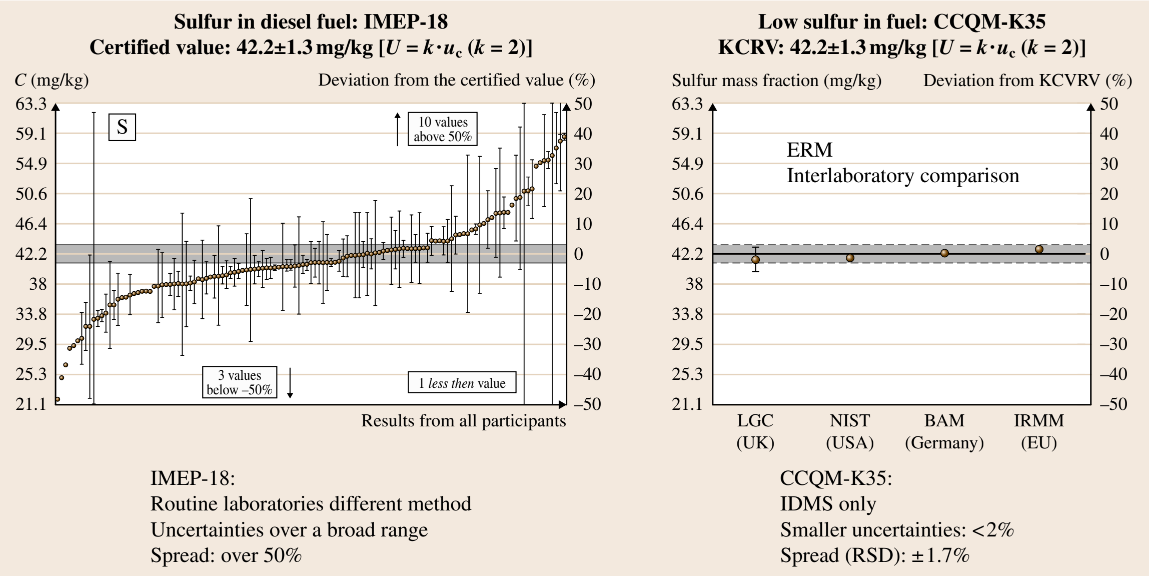
Fig. 3.22 Certified reference materials representing Euro coin alloys


ERM
EF211
Sample No: 1397
Certified Reference Material
S in Petrol (high concentration)
EF213
Sample No: 0770
Certified Reference Material
S in Petrol (low concentration)
ERM EF212

- First CRMs from commercial gasoline
- Covering present legal limits in EU and USA
- Offer lower uncertainities (3.5–8.8%) than presently available materials

EF213
Sample No: 0001
ERM
Certified Reference Material
S in Petrol (low concentration)
mass: 28.913 g
R: 45-12
S: 53-62
Delivered by BAM, Unter den Eichen 87, 12205 Berlin, Germany
Tel:+49 308104-0 Fax:+49 30 8112029
Highly flammable
Toxic

Fig. 3.23 Certified reference material for sulfur content in gasoline



IMEP-18:
Routine laboratories different method
Uncertainties over a broad range
Spread: over 50%

CCQM-K35:
IDMS only
Smaller uncertainties: <2%
Spread (RSD): ±1.7%

Fig. 3.24 International comparison results of sulfur measurements in fuel

Ochratoxin A in roasted coffee
ERM-BD475
Legal limit in EU: 5 μg/kg
Certified value: (6.1±0.6) μg/kg
Material produced by suspension spiking
(by 17 international collaborators)
Storage at –20 °C

Fig. 3.25 Certified reference material meets legal limit: ochratoxin A in coffee

laboratories use the in-house specimen or rarely available certified reference material. Increasing demands from quality management systems and customers, and lower acceptable tolerances, will require effective use of CRM in the field of mechanical testing in the future as well. As a first result, new CRMs have been developed in the past few years, and their use is required or at least recommended in the updated ISO test standards.

The increasing demand for CRM in the field of mechanical testing is driven by growing requirements from quality management systems. The reliability of test results is no longer a question of yearly direct calibration and demonstrated traceability. Regulatory demands regarding product safety place higher requirements on the producer regarding the reliability of test results. The major question today is the documented, daily assurance that a test system is working properly in the defined range. Three main streams are driving the development of CRM in this field.

- Comparability of test results within company laboratories, producers, and customers must be reliably demonstrated. In this framework the validation of the capabilities of test methods is necessary.
- Customers and the market demand reduced product tolerances. This is only possible when the test method itself allows a judgement on the level of reduced values for trueness and precision.
- To establish measurement uncertainty budgets. Mathematical models are usually not practical in the field of mechanical testing because of the complexity of the parameters affecting the results.

Modern test systems, for example, for tensile testing of metals, are a combination of hardware, the measurement sensors themselves, additional measurement equipment, and computer hardware and software. Direct calibration reflects only one aspect of the overall functionality of the complete and complex test system. Additional measures and proofs are necessary to demonstrate that the system is working properly. The following independent aspects can be verified using CRM.

- The ability of the test system to produce true values. The calculated bias between the certified reference value and the mean value from a defined number of repeated tests using the CRM is calculated. The acceptable range for the bias is defined in the test standard itself (hardness test, Charpy impact test) or by the user (tensile test).
- The ability of the test system to produce precise results can be demonstrated. Usually the standard deviation of repeated tests using the CRM is calculated as a measure for the precision of the test system. Limitations of this value are defined in the test standard itself or by the user.
- The use of CRMs to establish the measurement uncertainty of a test system is an accepted procedure. The known uncertainty of the CRM in combination with the uncertainty calculated from the use

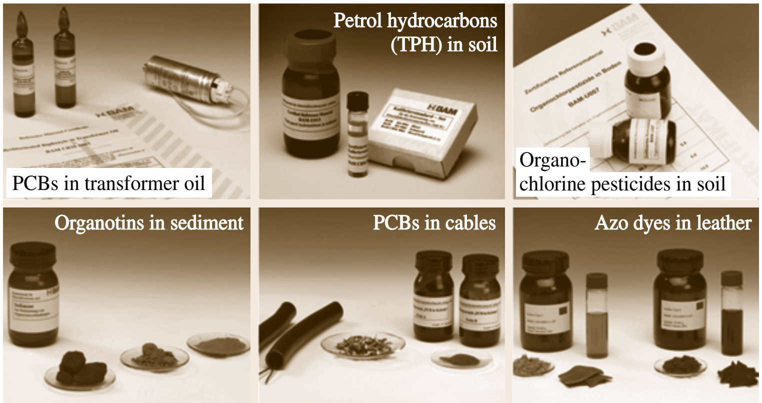


Fig. 3.26 Matrices of environmental or industrial origin certified for contents of organic toxins

of this CRM in the test systems is defined in the corresponding ISO test standard. With this uncertainty budget, smallest measurement tolerances can be established. The stability of an up-to-date test system must be documented. Quality control charts are rarely used in mechanical testing laboratories.

Accredited Producers of Reference Materials for Mechanical Testing

Certified Reference Material for Charpy Impact Test.

- EU Joint Research Centre Institute for Reference Materials and Measurements
  Retieseweg 111
  2440 Geel, Belgium

Certified Reference Material for Hardness Testing.

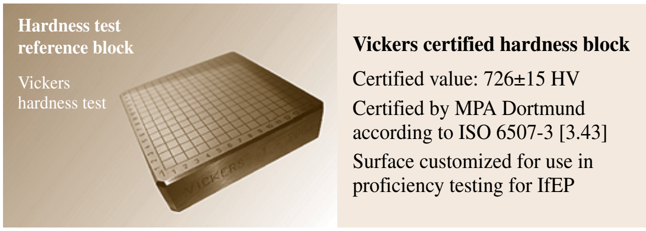
- MPA NRW Materialprüfungsamt Nordrhein-Westfalen
  Marsbruchstraße 186
  44287 Dortmund, Germany
- MPA-Hannover
  Materialprüfanstalt für Werkstoffe und Produktionstechnik
  An der Universität 2
  30823 Garbsen, Germany

Certified Reference Material for Charpy Impact Test and Tensile Test.

- IfEP GmbH
  Institut für Eignungsprüfung GmbH
  Daimlerstraße 8
  45770 Marl, Germany

3.7.9 Reference Materials for Hardness Testing

In hardness testing of metals (Sect. 7.3) indirect verification with certified hardness reference blocks is mandatory. The related standards ISO 6506 Brinell [3.52], ISO 6507 Vickers [3.53], and ISO 6508 Rockwell [3.54] define the relevant and acceptable criteria for a test system when using a CRM. After direct calibration, a final check of the whole system is done by using a material of defined hardness. The parameters assessed are the precision and repeatability of the measurements. In the related standards of the ISO 650X series individual requirements are defined for every test method. Prior to a test series, the certified reference block (Fig. 3.27) should be used to verify the trueness and precision of the measurement capability of the testing machine under the specified test conditions. If the result shows an error or the repeatability exceeds the limits defined in the test standard, tests shall not be performed.



Vickers certified hardness block

Certified value: 726±15 HV

Certified by MPA Dortmund according to ISO 6507-3 [3.43]

Surface customized for use in proficiency testing for IfEP

Fig. 3.27 Example of a hardness reference block

Example: Vickers Hardness Test According to ISO 6507-1

The evaluation criteria are based on ISO 6507-2 [3.55], Table 4 (permissible repeatability of the testing machine r and r_{rel}) and Table 5 (error of the testing machine E_{rel}). The error of the testing machine E_{rel} is calculated according to (3.1)

$$E_{\text{rel}} = \frac{\bar{H} - H_{\text{C}}}{H_{\text{C}}} \cdot 100\% \,. \tag{3.1}$$

Examples of permissible error of the testing machine (3.2) stated in ISO 6507-2, Table 5 are

$$\begin{aligned} &\text{HV}10: -3\% \le E_{\text{rel}} \le 3\% \\ &\text{HV}30: -2\% \le E_{\text{rel}} \le 2\% \,. \end{aligned} \tag{3.2}$$

$\bar{H}$ is the (arithmetic) mean value of the measurements on a given hardness block of certified reference value H_{C}.

For the determination of the repeatability (r and r_{rel}) *both* values of (3.3) must be calculated

$$\begin{aligned} r_{\text{rel}} &= \frac{d_{\max} - d_{\min}}{\bar{d}} \cdot 100\% \,, \\ r &= H_{\max} - H_{\min} \,. \end{aligned} \tag{3.3}$$

$d_{\max / \min}$ are the maximum/minimum measured diagonal, and $H_{\max / \min}$ are the maximum/minimum measured hardness in HV10/HV30.

According to ISO 6507-2, Table 4 the permissible repeatability is given by

$$r_{\text{rel}} < 2\% \,, \tag{3.4a}$$

$$r < 30\text{HV}10/\text{HV}30 \,. \tag{3.4b}$$

Both requirements must be fulfilled to guarantee an acceptable status of the testing machine prior to the test series.

Determination of Measurement Uncertainty. The results of testing the CRM are also used to establish the measurement uncertainty budget of the test procedure. The determination of the measurement uncertainty according to ISO 6507-1 is based on the UNCERT Code of Practice Nr. 14 [3.55] and GUM [3.18]. Additional to the CRM, this requires the measurement of hardness on a standard material. The results of these measurements are the mean values and the standard deviation. The expanded measurement uncertainty for the measurement done by one laboratory on the standard material is calculated according to (3.5) and (3.6).

$$U = 2\sqrt{u_{\mathrm{E}}^2 + u_{\mathrm{CRM}}^2 + u_{\bar{H}}^2 + u_{\bar{x}}^2 + u_{\mathrm{ms}}^2}\,, \tag{3.5}$$

$$\tilde{U} = \frac{U}{\bar{X}_{\mathrm{CRM}}} 100\% \tag{3.6}$$

with

U	Expanded measurement uncertainty
$\tilde{U}$	Relative expanded measurement uncertainty
u_{E}	Standard uncertainty according to the maximum permissible error
u_{CRM}	Standard measurement uncertainty of the certified reference block
$u_{\bar{H}}$	Standard measurement uncertainty of the laboratory testing machine measuring the hardness of the certified reference block
$u_{\bar{x}}$	Standard measurement uncertainty from testing the material
u_{ms}	Standard measurement uncertainty according to the resolution of the testing machine
$\bar{X}_{\mathrm{CRM}}$	Certified reference value of the certified reference block.

The minimum level of relative expanded measurement uncertainty $\tilde{U}$ is given by the combination of the fixed factors u_{E}, u_{CRM}, and u_{ms}.

This approach is used in the same manner for other hardness methods.

3.7.10 Reference Materials for Impact Testing

The Charpy impact test, also known as the Charpy V-notch test, is a standardized high-strain-rate test that determines the amount of energy absorbed by a material during fracture (Sect. 7.4.2). This absorbed energy is a measure of a given material's toughness and acts as a tool to study temperature-dependent brittle–ductile transition. It is widely applied in industry, since it is easy to prepare and conduct, and results can be obtained quickly and cheaply. However, a major disadvantage is that all results are only comparative. This may be commercially important when values obtained by these machines are so different that one set of results does meet a defined specification while another, tested on a second machine, does not meet the requirements. To avoid disagreements, in the future, all machines have to be verified by testing certified reference test pieces. A testing machine is in compliance with the ISO 148-1:2008 international standard [3.56] when it has been verified using direct and indirect methods. Methods of verification (ISO 148-2:2008) [3.57] are

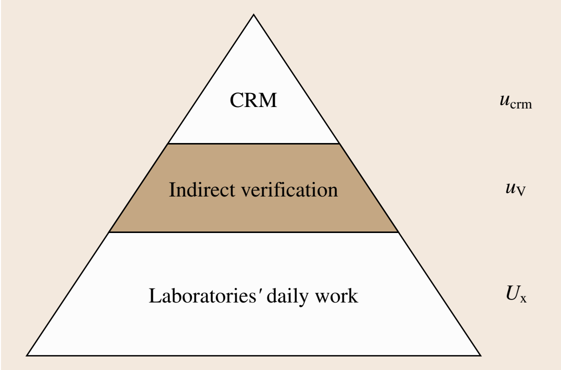


Fig. 3.28 Traceability chain of ISO 148

- The first method uses instruments for direct verification that are traceable to national standards. All specific parameters are calibrated individually. Direct methods are used yearly, when a machine is installed or repaired, or if the indirect method gives a nonconforming result.
- The second method is indirect verification, using certified reference test pieces to verify points on the measuring scale.

Additionally, the results of the indirect verification are used to establish the measurement uncertainty budget of the test system (Fig. 3.28).

Requirements for Reference Material and Reference Test Pieces. The preparation and characterization of Charpy test pieces for indirect verification of pendulum impact testing machines are defined in ISO 148-3:2008 [3.58].

The specimen shall be as homogeneous as possible. The ranges of absorbed energy that should be used in indirect verification are specified in ISO 148-3:2008 [3.58] and displayed in Table 3.12.

Table 3.12 Requirements for certified reference material in Charpy testing, according to ISO 148-3

Energy level	Range of absorbed energy
Low	$< 30\,\mathrm{J}$
Medium	$\geq 30-110\,\mathrm{J}$
High	$\geq 110-200\,\mathrm{J}$
Ultra high	$\geq 200\,\mathrm{J}$

Table 3.13 Permissible standard deviation in homogeneity testing

Energy KV_R	Standard deviation
$< 40\,\mathrm{J}$	$\geq 2.0\,\mathrm{J}$
$\geq 40\,\mathrm{J}$	$\geq 5\%$ of KV_R

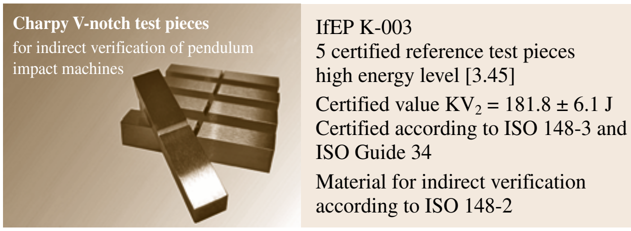
One set of reference pieces (Fig. 3.29) contains five specimens. This set is accompanied by a certificate, which gives information on the production procedure, the certified reference value, and the uncertainty value.

Certified Energy of Charpy Reference Materials. Charpy RM specimens are produced in a batch of up to 2000 pieces. From this bach a representative number of samples are tested. The samples are destroyed to measure the absorbed energy. The average of all test results is defined as the certified value KV_R.

Qualification Procedure. The certified value can be determined using any method which is defined in ISO guides 34 and 35 [3.59].

Reference Machine. Sets of at least 25 test pieces are randomly selected from the batch. These sets are tested on one or more reference machines. The grand average of the results obtained from the individual machines is taken as the reference energy. The standard deviation in homogeneity testing is calculated according to ISO 148-3 [3.58] and must meet the requirements of Table 3.13, where KV_R is the certified KV value of the Charpy reference material.

Intercomparison Among Several Charpy Impact Machines. To reduce the effect of machines on the certified reference value, it is possible to perform tests on different impact testing machines; ISO guide 35 recommends at least six laboratories. The larger the number of testing machines used to assess the average of a batch of samples, the more likely it is that the average of the values obtained is true and unbiased. It is necessary that individual participating pendulums are high-quality instruments and that the laboratory meets minimum quality requirements, e.g., accreditation according to ISO/IEC 17025 [3.59].


Charpy V-notch test pieces
for indirect verification of pendulum impact machines

IfEP K-003
5 certified reference test pieces
high energy level [3.45]
Certified value $KV_2 = 181.8 \pm 6.1\,\mathrm{J}$
Certified according to ISO 148-3 and ISO Guide 34
Material for indirect verification according to ISO 148-2

Fig. 3.29 Charpy reference test pieces according to ISO 148-3, for 2 mm striker (after [3.58])

Uncertainty of the Certified Energy Value of Charpy Reference Material. The uncertainty budget of the reference material is calculated using the basic model from ISO guide 35, which is in compliance with GUM. The uncertainty of the certified value of the Charpy reference material can be expressed as (3.7)

$$U_\mathrm{RM} = \sqrt{u_\mathrm{char}^2 + u_\mathrm{hom}^2 + u_\mathrm{lts}^2 + u_\mathrm{sts}^2}\,. \tag{3.7}$$

Here, u_lts means uncertainty due to long-term stability. Although steel properties are supposed to be stable, some producers limit their material to 5 years, within which u_lts is negligible.

u_sts means short-term stability. As stability is given for at least 5 years, this is negligible, too.

u_hom is given by (3.8)

$$u_\mathrm{hom} = \frac{s_\mathrm{RM}}{\sqrt{n_\mathrm{V}}}\,, \tag{3.8}$$

with

s_RM Standard deviation of the homogeneity study
n_V Number of specimens in one set of CRM (here five)

u_char is calculated according to (3.9), usually based on an interlaboratory comparison

$$u_\mathrm{char} = \frac{s_\mathrm{p}}{\sqrt{p}}\,, \tag{3.9}$$

with

s_p Standard deviation of the interlaboratory comparison
p Number of participants

The better the within-instrument repeatability and between-instrument reproducibility, the smaller u_char will be.

The standard deviation of the interlaboratory comparison is calculated using (3.10)

$$s_p = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2}\,, \tag{3.10}$$

with

X_i Laboratories' mean value
$\bar{X}$ Grand mean
n Number of participants

The coverage factor k is calculated using the Welch–Satterthwaite equation (3.11). The confidence level is usually set at 95%.

$$\begin{aligned} k &= t_{95}\left(\nu_{\overline{KV}}\right) \quad \text{with} \\ \nu_{\overline{KV}} &= \frac{u_{\text{RM}}^4}{u_{\text{char}}^4/\nu_{\text{char}} + u_{\text{hom}}^4/\nu_{\text{hom}}}\,. \end{aligned} \tag{3.11}$$

Indirect Verification of the Impact Pendulum Method by Use of Reference Test Pieces

Indirect verification of an industrial machine is done using five specimens in random order and including all results in the average. The indirect verification shall be performed at least every 12 months. Substitution or replacement of individual test pieces by test pieces of another reference set is not permitted. These reference test pieces are used:

- For comparison between test results obtained with the machine and reference values obtained from the procedure described in ISO 148-3:2008 [3.58].
- To monitor the performance of a testing machine over a period of time, without reference to any other machine. This is done by laboratories to assure the internal quality of testing.

The indirect verification shall be performed at a minimum of two energy levels within the range of the testing machine. The absorbed energy level of the reference samples used shall be as close as possible to the lower and upper levels of the range of use in the laboratory. When more than two absorbed energy levels are used, other levels should be uniformly distributed between the lower and upper limits, subject to the availability of reference test pieces. The indirect verification shall be performed at the time of installation, or after moving the machine, or when parts have been replaced.

Evaluation of the Result. $KV_1, KV_2, \ldots, KV_{n_\text{V}}$ are the absorbed energies at rupture of the n_V reference test pieces of a set, numbered in order of increasing value. The repeatability of the machine performance under the particular controlled conditions is characterized by (3.12)

$$b = KV_{n_\text{V}} - KV_1\,, \quad \text{i. e.} \quad KV_{\max} - KV_{\min}\,. \tag{3.12}$$

The maximum allowed repeatability values are given in Table 3.14.

Bias. The bias of the machine performance under the particular controlled conditions is characterized by (3.13)

$$B_\text{V} = \overline{KV}_\text{V} - KV_\text{R}\,, \tag{3.13}$$

with

$$\overline{KV}_\text{V} = \frac{\sum KV_i + \ldots + KV_{n_\text{V}}}{n_\text{V}} \tag{3.14}$$

and KV_R = certified reference value. The maximum allowed bias values are given in Table 3.14.

Measurement Uncertainty of the Results of Indirect Verification. The primary result of an indirect verification is the estimate of the instrument bias B_V (3.13). The standard uncertainty of the bias value $u(B_\text{V})$ is equal to the combined standard uncertainties of the two terms in (3.15)

$$u(B_\text{V}) = \sqrt{\left(\frac{s_\text{V}}{\sqrt{n_\text{V}}}\right) + u_{\text{RM}}^2}\,. \tag{3.15}$$

As a general rule, bias should be corrected for. However, due to wear of the anvil and hammer parts, it is difficult to obtain a perfectly stable bias value throughout the period between two indirect verifications. This is why the measured bias value is considered an uncertainty contribution, to be combined with its own uncertainty to obtain the uncertainty of the indirect verification result u_V (3.16)

$$u_\text{V} = \sqrt{u^2(B_\text{V}) + B_\text{V}^2}\,. \tag{3.16}$$

Table 3.14 Permissible limits in indirect verification according to ISO 148-2 [3.57]

Absorbed energy level	Repeatability b	Bias $\lvert B_\text{V}\rvert$
$< 40\,\text{J}$	$\leq 6\,\text{J}$	$\leq 4\,\text{J}$
$\geq 40\,\text{J}$	$\leq 15\%$ of KV_R	$\leq 10\%$ of KV_R

To correct for the absorbed energy values measured with a pendulum impact testing machine, a term equal to $-B_\mathrm{V}$ can be added. This requires that the bias value be firmly established and stable. Such a level of knowledge on the performance of a particular pendulum impact testing machine can only be achieved after a series of indirect verification and control chart tests, which should provide the required evidence regarding the stability of the instrument bias. Therefore, this practise is likely to be limited to the use of reference pendulum impact testing machines.

The coverage factor k is calculated using the Welch–Satterthwaite equation (3.17). The confidence level is usually set at 95%.

$$k = t_{95}(\nu_\mathrm{V}) \quad \text{with}$$
$$\nu_\mathrm{V} = \frac{u_\mathrm{V}^4}{u^4(\overline{KV}_\mathrm{V})/\nu_\mathrm{B} + u^4_\mathrm{RM}/\nu_\mathrm{RM} + B^4_\mathrm{V}/\nu_\mathrm{B}} \,. \tag{3.17}$$

The value of ν_B is $n_\mathrm{V}-1$; the value of ν_RM is taken from the reference material certificate. The number of verification test samples is most often five, and the heterogeneity of the samples is not insignificant. This is why the number of effective degrees of freedom is most often not large enough to use a coverage factor of k equal to 2.

Determination of the Uncertainty of a Related Test Result. This approach requires the results of the indirect verification process. This is the normative method of assessing the performance of the test machine with certified reference materials. The following principle factors contribute to the uncertainty of the test result.

- Instrument bias, identified by the indirect verification
- Homogeneity of the tested material
- Instrument repeatability
- Test temperature

Instrument Bias. Measured values are allowed to be corrected for if the bias is stable and well known. This is the case only when an acceptable number of repeated verifications have been performed. More often, a reliable bias is not known. In this case the bias is not corrected for, but it contributes to the uncertainty budget (3.16).

Homogeneity of the Test Material and Instrument Repeatability. The uncertainty of the test result $u(\bar{X})$ is calculated using equation (3.18)

$$u(\bar{X}) = \frac{s_X}{\sqrt{n}} \,, \tag{3.18}$$

where s_X is the standard deviation of the values obtained on the n test samples.

In this factor the sample-to-sample heterogeneity of the material and the repeatability of the test method are cofounded. They cannot be identified individually. The value s_x is a conservative measure for the variation due to the material tested.

Temperature Bias. The effect of temperature bias on the measured absorbed energy is extremely material dependent. A general model cannot be formulated to solve the problem in terms of the uncertainty budget. It is recommended to report the test temperature and the related uncertainty in the test report. During the testing phase the temperature shall be kept as constant as possible.

Machine Resolution. Usually, the influence of the machine resolution r is negligible compared with the other factors. Only when the resolution is large and the measured values are low can the corresponding uncertainty be calculated using (3.19)

$$u(r) = \frac{r}{\sqrt{3}} \,, \tag{3.19}$$

where r is the machine resolution. The corresponding number of degrees of freedom is ∞.

Combined and Expanded Uncertainty. To calculate the overall uncertainty the individual parts shall be combined according to (3.20)

$$u(\overline{KV}) = \sqrt{u^2(\bar{x}) + u_\mathrm{V}^2 + u^2(r)} \,. \tag{3.20}$$

The number of tested samples in the Charpy impact test is usually low. In addition, the heterogeneity of the material leads to high values for $u(x)$. For this reason, the coverage factor shall not be selected as $k=2$. To calculate the expanded uncertainty, the combined uncertainty is multiplied by a k-factor which depends on the degrees of freedom, calculated using (3.21)

$$k = t_{95}(\nu_\mathrm{V}) \quad \text{with } \nu_\mathrm{V} = \frac{u^4 u(\overline{KV})}{u^4(\bar{X})/\nu_{\bar{X}} + u_\mathrm{V}^4/\nu_\mathrm{V}} \,. \tag{3.21}$$

With this number, the coverage factor k can be determined using tables published in GUM. Examples are shown in Table 3.15.

Table 3.15 Typical values of k with given ν

Degrees of freedom ν	Corresponding coverage factor k
8	2.31
9	2.26
10	2.23
11	2.20
12	2.18
13	2.16
14	2.14
15	2.13
16	2.12
17	2.11

3.7.11 Reference Materials for Tensile Testing

Tensile testing of metals according to ISO 6892-1:2009 [3.60] is one of most important methods to characterize materials and products. The methodology of tensile testing is described in Sect. 7.4.1 and illustrated in Fig. 3.30. The direct calibration of load and displacement is done on a regular base. However, the complexity of modern test systems requires additional measures to guarantee acceptable test results, including knowledge about measurement uncertainty. The international standard ISO 6892-1:2009 recommends use of reference materials to demonstrate the functionality of the whole test system. This is part of the concept to prove the capability of the whole measuring process and ensures the reliability of the tensile test system. In the majority of tensile tests, the proof strength $R_{\mathrm{p0,2}}$, the ultimate strength R_{m}, and the elongation A are the resulting parameters.

Concept to Prove the Capability of a Tensile Test System

Level 1 – Requirements of the Test Standard. ISO 6892-1:2009 defines criteria regarding the basic status of acceptable test equipment. Specific requirements are formulated for the load and length measurements as well as for the elongation measuring device. All measurements must be in class 1, with maximum deviation of 1% (class 1) over the whole measurement range. The corresponding values are determined in the direct calibration process on a regular base, usually once a year. The weakness of the calibration process is that it is not possible to demonstrate the full functionality of the system. Many influencing factors such as the dimensions of the specimen used, the test speed, and the software settings are not evaluated in this process.

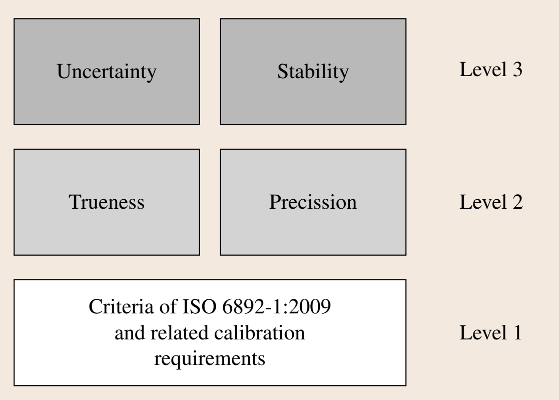

Uncertainty
Stability
Level 3
Trueness
Precission
Level 2
Criteria of ISO 6892-1:2009 and related calibration requirements
Level 1

Fig. 3.30 Schematic presentation of the IfEP accuracy concept

Level 2 – Trueness. The trueness of the measured values and the calculated results for strength and elongation can only be checked using reference material (Fig. 3.31). After the direct calibration, 25 specimens (round or flat) are tested under realistic laboratory conditions. The reference material used should have similar characteristics to material tested regularly.

The results are used to calculate the systematic deviation, the bias b for all characteristics (R_{p}, R_{m}, A, Z) as a measure of trueness using (3.22)

$$b = \bar{y} - \mu\,, \tag{3.22}$$

using $\bar{y}$, the mean value of 25 tests, and μ, the certified reference value.

According to ISO 5725-6 [3.61], Chap. 7.2.3.1.3 a judgement on the systematic deviation can be defined based on (3.23)

$$|b| < 2\sqrt{\sigma_{\mathrm{R}}^2 - \sigma_{\mathrm{r}}^2\frac{(n-1)}{n}}\,, \tag{3.23}$$

using

σ_{R} Reproducibility standard deviation
σ_{r} Repeatability standard deviation
n Number of repetitions

σ_{R} and σ_{r} are defined in the certification process. Table 3.16 shows an example of the allowed bias for 25 repetitions.

The use of 25 specimens allows reliable determination of the bias. This bias can be corrected for. If it is not corrected, it is included in the measurement uncertainty budget of the test system.

Table 3.16 Example for the allowed bias b, testing 25 certified specimens; calculation according to ISO 5725-6

Parameter	Maximum $\|b\|$
$R_{p0,2}$	7.5 MPa
R_m	8.0 MPa
A	3.3%

Table 3.17 Example for maximum limits of repeatability standard deviation using 25 reference specimens, calculated according to ISO 5725-6

Parameter	S_{max}
$R_{p0,2}$	4.2 MPa
R_m	2.6 MPa
A	0.8%

Precision. The precision of the test system can be evaluated using ISO 5725-6:2002, Chap. A7.2.3 *Measurement method for which reference material exists*. In this approach the standard deviation of laboratories' results for the tested reference material, S_r, is divided by the repeatability standard deviation, σ_r, of the certification process. The result is compared with the tabulated c^2 distribution according to (3.24)

$$\frac{S_r^2}{\sigma_r^2} < \frac{\chi^2_{(1-\alpha)}(\nu)}{\nu} \,, \tag{3.24}$$

using

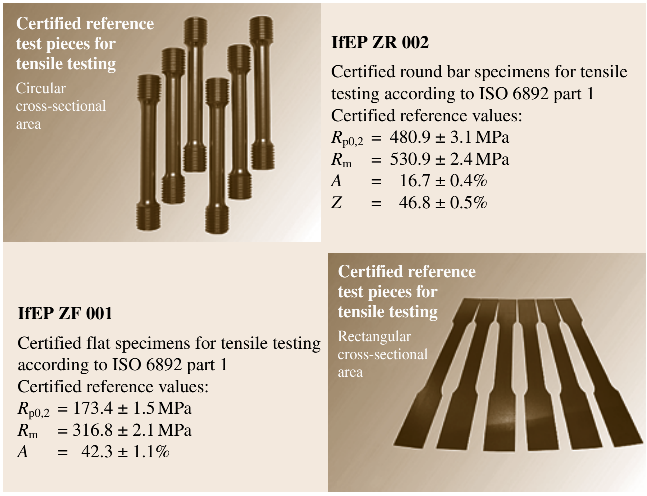
S_r	Standard deviation when testing the CRM
σ_r	Repeatability standard deviation of the certification process
$\chi^2_{(1-\alpha)}(\nu)$	$(1-\alpha)$ quartile of the χ^2 distribution
α	Significance level, here 0.01 or 1%
ν	$n-1$ degrees of freedom

σ_r is defined in the certification process.

To define an acceptable maximum standard deviation for the test system (3.24) is transformed to define S_{max} (3.25)

$$S_{max} < \sqrt{\frac{\chi^2_{(1-\alpha)}(\nu)}{\nu}\sigma_r^2} \,. \tag{3.25}$$

For a significance level of 1% and 25 specimens tested, Table 3.17 shows example maximum standard deviations for various parameters of a tensile test system. The defined limits guarantee state-of-the-art test results.


Certified reference test pieces for tensile testing
Circular cross-sectional area
Certified reference test pieces for tensile testing
Rectangular cross-sectional area

IfEP ZR 002

Certified round bar specimens for tensile testing according to ISO 6892 part 1
Certified reference values:
$R_{p0,2}$ = 480.9 ± 3.1 MPa
R_m = 530.9 ± 2.4 MPa
A = 16.7 ± 0.4%
Z = 46.8 ± 0.5%

IfEP ZF 001

Certified flat specimens for tensile testing according to ISO 6892 part 1
Certified reference values:
$R_{p0,2}$ = 173.4 ± 1.5 MPa
R_m = 316.8 ± 2.1 MPa
A = 42.3 ± 1.1%

Fig. 3.31 Certified reference test pieces for the tensile test (after [3.58])

Level 3 – Measurement Uncertainty. The concepts of the guide to the expression of uncertainty in measurement, ISO/IEC guide 98-3:2008 [3.62], are used to establish the uncertainty budget.

The test standard ISO 6892-1:2009 recommends in appendix J.4 the use of a reference material to establish the uncertainty budget of the test system, to incorporate all elements of the test process itself. The concept to establish the uncertainty budget uses the approach defined in ISO 148-1:2009 for Charpy impact testing, which uses certified reference materials as well.

Uncertainty of the Systematic Deviation. The uncertainty budget is calculated using the uncertainty of the reference material itself combined with the standard deviation of the tested 25 reference specimens. The laboratory has to define their acceptable uncertainty level to judge the result for each individual parameter.

First, the uncertainty of the calculated bias from level 2 must be evaluated under the condition that the bias is within the defined limits of acceptable trueness (3.22). If this requirement is fulfilled, the uncertainty can be calculated using (3.26)

$$U_b = k\sqrt{\left(\frac{s_V}{\sqrt{n_V}}\right)^2 + u_{RM}^2} \,. \tag{3.26}$$

s_V is the standard deviation of the results from the 25 (n_V) reference specimens tested. The uncertainty of the

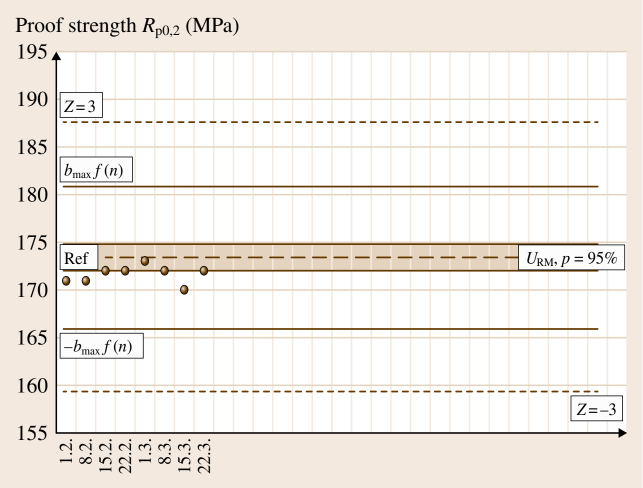


Fig. 3.32 Example quality control chart (proof strength, $R_{p0,2}$)

reference material is u_{RM}, defined in the certification process and stated in the certificate, divided by the coverage factor k_{RM}, also stated in the certificate.

Determination of the Coverage Factor k for 95% Confidence Level. The coverage factor k is calculated using (3.27)

$$k = t(\nu_{eff}) , \tag{3.27}$$

where t is the value of the t-distribution for the effective number of degrees of freedom ν_{eff}. The common confidence level is 95%.

The number of effective degrees of freedom ν_{eff} is calculated using (3.28)

$$\nu_{eff} = \frac{u_b^4}{\left(s_V/\sqrt{n_V}\right)^4/n_V - 1 + u_{RM}^4/\nu_{RM}} , \tag{3.28}$$

where ν_{RM} is the number of degrees of freedom from the certification process, stated in the certificate of the reference material. The corresponding t-factor is tabulated in ISO/IEC guide 98-3:2008, Table G2.

Evaluation of the Influence of the Bias b on the Uncertainty Budget. Three different cases have to be analyzed.

- A laboratory is correcting their test results using the calculated bias b of level 2. This bias must be constantly checked afterwards using a quality control chart. In this case the bias is not an element of the uncertainty budget.
- A laboratory is not correcting their test results with their bias. The bias b is within the range of the established uncertainty budget. In this case the bias b is not included in the uncertainty budget.
- A laboratory is not correcting their test results with their bias. The bias b is outside of the range of the established uncertainty budget. In this case the bias must be included in the uncertainty budget using (3.29). This approach is similar to the procedure used in ISO 6507-1:2006.

$$U_{B_{ukorr}} = \left(U_{B_V} + |B|\right) . \tag{3.29}$$

The uncertainty is considerably higher in this case, and it might be questionable whether this machine is working according to the state of the art.

Stability. The stability of the test system must be checked regularly. The established tool for this is a quality control chart in which the results of repeated tests are documented and evaluated. The basic condition is the use of the same reference material that was used to establish the trueness and precision of the test system. The number of repetitions (daily, weekly, monthly) depends on the level of confidence a laboratory wants to demonstrate. An example for a quality control chart is shown in Fig. 3.32. It must be noted that the limit values depend on the number of repeated tests in one run. The number of test specimens in one run (for example, 1 to 6) must be defined to establish statistically correct limits in the control chart. Basically, the maximum allowed bias b_{max} is calculated using (3.23). In case of only one specimen tested per run, the proof of trueness reduces to (3.30)

$$|b| < 2\sigma_R . \tag{3.30}$$

3.8 Reference Procedures

3.8.1 Framework: Traceability and Reference Values

Establishing *traceability* of measurement results (Sect. 3.2) is a shared activity. A laboratory performing measurements of materials properties has to

- identify all the laboratory references that are relevant for a given result, e.g., *specify the working* *standard used for calibrating the measuring system*,

- specify the relationship between the result and any laboratory reference used, e.g., *specify how the value of the working standard is used in a correction*,
- ensure that all those laboratory references are fit for purpose, e.g., *check the calibration status and the uncertainty of the working standard*.

As a complementary activity, the provider of a laboratory reference (e.g., *a calibration laboratory, performing calibration of working standards against reference standards*) has to

- ensure that the specification of the laboratory reference is valid, e.g., *operate a quality control program for the calibration of working standards*,
- establish traceability of the laboratory reference to a primary reference, e.g., *ensure traceability of the reference standard used for calibrating working standards against a primary standard*.

Laboratory references are by no means restricted to measurement standards used for calibration, reference materials used for bias investigation, and the like. Rather than relating measurement results to devices or materials, traceability relates measurement results to *reference values*, which may be associated with devices or materials but may also be of other origin. They include any data except for those generated in the measurement process and subsidiary measurements (e.g., for control of environmental conditions) which are utilized explicitly or implicitly in any stage of the measurement process, from preparation of measuring objects to data evaluation, and whose values are taken for granted.

Traceability relates measurement results to reference values.

Having raised the issue of a reference value, here are two current definitions.

- Definition 1: Quantity value used as a basis for comparison with values of quantities of the same kind. (VIM, 3rd edition, 2008 [3.63])
- Definition 2: Property value of a specified material or product that has been determined with an accuracy fit for use as a source of traceability for test results obtained on comparable materials or products. (Eurolab Position Paper, 2007 [3.64])

While originating from different fields – metrology and testing – these definitions are in fact very close, with complete agreement concerning the basic requirement: the uncertainty of reference values must be known and fit for the intended use. Given this, the relevance of reference values with regard to the accuracy of specified measurements may be assessed as follows: a reference value is relevant to a measurement result, if the uncertainty associated with the reference value contributes significantly to the overall uncertainty of the measurement result.

Reference values need specified uncertainties.

Depending on the type of material and the property under consideration, there are three basic sources of reference values for materials properties: *reference data compilations*, *reference materials* (Sect. 3.7), and *reference procedures*.

- For
  - well-defined and commonly available materials (e.g., pure copper), and
  - well-investigated quantities (e.g., thermophysical)

  reference values may be taken from a recognized reference data compilation.
- For
  - certified reference materials (e.g., a copper alloy CRM CuZn37), and
  - certified properties (e.g., the mass fraction of nickel)

  reference values may be taken from the certificate of the reference material.
- For
  - real-life materials (e.g., a sample from a batch of raw copper), and
  - well-defined quantities (for the purpose at hand)

  reference values may be measured using a reference procedure, if available.

As a remark in passing, uncertainty statements in reference data compilations are generally poor. Strategies for improving this situation are

- *New measurements* – Example: in the framework of a Japanese national project [3.65], new measurements were made on thermophysical property data of key industrial materials to generate reference data with state-of-the-art uncertainty.
- *Reevaluation of original measurements* – Example: for an international standardization project [3.66], publications on measurements of virial coefficients of pure gases were reevaluated to estimate the un-

certainty of virial coefficient data used in a previous standard.

3.8.2 Terminology: Concepts and Definitions

Unlike for reference materials (Sect. 3.7), there are currently no internationally harmonized terminology, fundamental concepts, and basic requirements for reference measurements. Before reviewing current concepts and definitions for the notion of a reference procedure, there is an obvious question to address first: why use the term *procedure* instead of the more common term *method*?

Method versus procedure: According to current perception, a measurement method provides a generic description of measurements, specifying the kind of equipment to be used, the kind of measuring objects, and a sequence of operations, while a measurement procedure provides a detailed description of how to actually carry out measurements according to a given method using specified equipment on specified measuring objects.

Although this distinction is also made in the VIM [3.63], it is not yet part of standard terminology and is often disregarded even in current publications.

Method versus procedure: a subtle but important difference

In the scientific/technical literature, various different definitions of the term *reference procedure* are in use, falling into three main categories as follows.

1. Measurement procedure with established high quality of results, which can be used for the assessment of other measurement procedures; *. . . results obtained using the procedure (proper implementation and validation assumed) serve as a benchmark for alternative procedures.*
2. Measurement procedures defining a quantity subject to measurement; *. . . the property under consideration is in fact defined by the procedure.*
3. Measurement procedure prescribed by legal regulations for specific measurements; *. . . to obtain valid results for regulatory purposes, the procedure must be used.*

In the last decade, approach (1) has clearly become the dominant concept, and in this section the term reference procedure is used in this sense. Below are three definitions with international scope [3.38, 67, 68]. Irrespective of their widely different origins – metrology, laboratory medicine, and reference materials – these definitions are very close and agree completely with respect to scope and requirements.

The term *reference procedure* has a wide range of connotations.

Obviously the term reference procedure applies equally to measurement, testing, and chemical analysis, i. e., all procedures for determining quantitative properties of materials, but the concept may also be extended to other fields.

Qualitative testing, i. e., determination of qualitative properties: qualitative properties are expressed by

Table 3.18 Definitions related to the concept of reference procedures

Reference procedure	Reference measurement procedure	Reference method
Procedure of testing, measurement or analysis, thoroughly characterized and proven to be under control, intended for quality assessment of other procedures for comparable tasks, or characterization of reference materials including reference objects, or determination of reference values. The uncertainty of the results of a reference procedure must be adequately estimated and appropriate for the intended use. (Euramet Guide [3.67])	Thoroughly investigated measurement procedure shown to have an uncertainty of measurement commensurate with the intended use, especially in assessing the trueness of other measurement procedures for the same quantity and in characterising reference materials. (ISO 15195 [3.68])	Thoroughly investigated method, clearly and exactly describing the necessary conditions and procedures, for the measurement of one or more property values that has been shown to have accuracy and precision commensurate with its intended use and that can therefore be used to assess the accuracy of other methods for the same measurement, particular in permitting the characterisation of a reference material. (ISO Guide 30 [3.38])

membership to one of several specified classes, and the uncertainty of a qualitative test result is an estimate of the probability that the result is incorrect.

Preparation of reference materials or reference objects: for example, the gravimetric preparation of mixtures of pure substances as reference materials of chemical composition, or the preparation of reference defects by spark erosion on metal specimens to be used as reference objects for nondestructive testing of pressure vessels.

Procedures for data evaluation: for example, uncertainty propagation, i. e., the calculation of combined standard uncertainty according to the GUM, is based on approximations that may not always be valid in practise. In case of doubt, Monte Carlo simulation may be used as a reference procedure to investigate whether uncertainty propagation is valid [3.69].

According to the definitions above, reference procedures for materials measurements can be used to

- validate (or calibrate) other measurement procedures that are used for a similar task and to determine their uncertainty,
- determine reference values for material properties that are embodied by a reference material or a reference object,
- determine reference data of materials properties for reference data compilations.

Another application of reference procedures could be measurements as the basis for important decisions, e.g., for authoritative evidence.

It should be noted that there are a number of alternative terms with a similar meaning. The term *primary method of measurement* has been developed for use in the Consultative Committee for Amount of Substance (CCQM), i. e., the international committee of the Metre Convention for metrology in chemistry. The term *definitive method* originates from the International Union of Pure and Applied Chemistry (IUPAC).

Primary Method of Measurement

A method having the highest metrological qualities, whose operation can be completely described and understood, and for which a complete uncertainty statement can be written down in terms of SI units (*Milton* and *Quinn* [3.44]).

Primary Reference Measurement Procedure

A reference measurement procedure used to obtain a measurement result without relation to a measurement standard for a quantity of the same kind [3.63]. From the conceptional point of view, this is identical to the term primary method of measurement.

Definitive Method of Measurement

A method of exceptional scientific status that is sufficiently accurate to stand alone in the determination of a given property for the certification of a reference material. Such a method must have a firm theoretical foundation so that systematic error is negligible relative to the intended use. The property values must be measured directly in terms of the base units of measurement, or indirectly related through sound theoretical equations (adapted from IUPAC [3.70]).

3.8.3 Requirements: Measurement Uncertainty, Traceability, and Acceptance

The definition of a reference procedure presumes the existence of several procedures for a specified task. Given this, a reference procedure is qualified by the uncertainty of results that are proven to be fit for the purpose of providing reference values of the property under consideration.

Before addressing requirements for measurement quality, the basic condition that the procedure be one among several procedures for the same task requires a second thought. For measurement of *rationally defined properties*, i. e., properties that are defined independently of the method of measurement, the concept is clear. There is a variety of measurement techniques available, and a reference procedure implements the best available technique, i. e., that with the capability of providing the most reliable results with the best available uncertainty.

Reference procedures for operationally defined properties?

However, the concept is also applicable to operationally defined properties, i. e., properties that are (to a significant degree) defined by a particular measurement method, often specified in a standard. In this case a reference procedure is a *reference implementation* of the defining measurement method, used to provide reference values for bench-marking *routine implementations*.

The scope of reference procedures implies a number of key requirements, concerning *measurement uncertainty*, *measurement traceability*, and *acceptance*, as follows. In addition to these, measurement quality control must be designed and executed at an appropriate level that is commensurate with provision of reliable reference values.

1. *The uncertainty of measurement results must be known.* The uncertainty of measurement is evaluated and specified in accordance with the GUM [3.71]. This entails that the uncertainty attributed to a measurement result accounts for all significant sources of uncertainty, and that every effort is taken to ensure that there is no significant residual bias that could compromise measurement results. The uncertainty of measurement may be evaluated using any of the following approaches or combinations thereof
   - Propagation of input uncertainties, based on a detailed mathematical model of the measurement process, as specified in the GUM [3.71];
   - Utilizing data from interlaboratory method-validation studies according to ISO 5725 [3.22];
   - Utilizing data from within-laboratory method-validation studies [3.26].
2. *The validity of the uncertainty claim must be demonstrated by interlaboratory comparisons.* Comparison with results obtained by other proficient laboratories, preferably using reference procedures based on independent measurement techniques, is generally accepted as the best way to challenge the uncertainty claimed for the results of a reference procedure. Agreement is taken to support the claim. Disagreement requires further action to identify the reason, e.g., underestimation of the uncertainty or significant bias not detected so far. As a recent example, the so-called key comparisons (KC) provide the basis for the acceptance of calibration and measurement capability (CMC) claims of national metrology institutes in the various peer reviews prior to publication (Sect. 3.8.6).
3. *The uncertainty of measurements results must be commensurate with their intended use as reference values.* The traditional policy concerning the use of reference values, e.g., in calibration, has been: the uncertainty of reference values (e.g., the uncertainty of a reference standard used to calibrate working standards) should be negligible compared with the target uncertainty of the result under consideration (e.g., the uncertainty of a working standard) and must not exceed one-third of that target uncertainty. The rationale for this requirement is unknown to the authors, but as a conjecture, the driver behind this requirement may have been the lack of techniques for the statistical evaluation of data with specified uncertainty, e.g., least-squares regression with uncertainties in the independent variable. When using techniques designed for data without uncertainty on data with an associated uncertainty, one has to make sure that they still give valid results. To this end, the uncertainty associated with the data has to be negligible against the relevant statistical variability. Similarly the 1 : 3 rule mentioned above may be justified using the root-sum-of-squares addition of standard uncertainties

$$\sqrt{u^2+\left(\frac{u}{3}\right)^2}=\sqrt{1.11u^2}=1.05u\approx u\ .$$

Fitness for purpose concerning requirements on measurement uncertainty

Given appropriate techniques for uncertainty-based data evaluation, there is no need to require that the uncertainty of reference values be negligible. Instead, their contribution to the uncertainty of the final result may be determined, and limits on the uncertainty of reference values may be derived from target uncertainties.

4. *Measurement results must be traceable to recognized references.* Reference values constitute the endpoint of within-laboratory traceability chains. Unless a reference value happens to be a primary reference that is recognized by the relevant community, its traceability to an appropriate higher-level reference value has to be established. Reference measurements carried out by national metrology institutes according to measurement capabilities specified in the BIPM database (Sect. 3.8.6) are internationally recognized as providing traceability to SI units. Traceability of reference values embodied by certified reference materials (CRM) is currently under debate in various fora, concerning technical issues such as the traceability of consensus values from interlaboratory certification studies and quality management issues such as third-party assessments of CRM quality and accreditation of CRM producers. The BIPM database contains a number of entries for CRMs provided by national metrology institutes, which may therefore be utilized as internationally recognized sources of traceability.
5. *Reference procedures must be accepted as such by the relevant target groups.* Reference procedures have the potential to stand alone in the provision of reference values for the property in

question. Reference procedures are qualified as such in a validation study, where the performance of the procedure is investigated in appropriate detail, and the values of the relevant performance characteristics – in particular measurement uncertainty – are determined. However, as an equally important requirement, the relevant target groups have to be convinced of the merits of a candidate reference procedure before it can be used as such. These target groups will normally include customers of reference measurements, laboratories, research institutes, and technical associations in the respective field, but may also involve accreditation bodies and regulators. Results of appropriate interlaboratory comparisons will normally be the most convincing arguments in favor of a newly proposed reference procedure.

3.8.4 Applications for Reference and Routine Laboratories

Developing and maintaining a reference procedure often requires significant effort, money, and manpower, and the costs for operating a reference procedure are often significantly higher than for other procedures for the same task. Therefore, reference procedures are most often operated by reference laboratories (Sect. 3.8.7), e.g., for the provision of reference measurement services, the determination of reference data, and the characterization of reference materials or reference objects. However, it may also be profitable for laboratories engaged in routine measurement and testing to operate a reference procedure for dedicated tasks such as validation or calibration of routine procedures.

The key qualification of reference procedures is their potential to stand alone in the determination of reference values for the property in question. This is due to

- the high capability of the particular measurement technology, with respect to trueness (absence of bias), precision, robustness to interfering effects (measuring conditions and measuring objects), and
- the high level of performance evaluation, in particular concerning measurement uncertainty, and ongoing internal and external quality control for the reference procedure developed using this technology.

The main fields of application for reference procedures were already mentioned in the various definitions (Sect. 3.8.2).

Euramet Guide [3.67].

- Quality assessment of other procedures for comparable tasks
- Characterization of materials, including reference objects
- Determination of reference values

ISO 15195 [3.68].

- Assessing the trueness of other measurements procedures for the same quantity
- Characterizing reference materials

ISO Guide 30 [3.38].

- Assessing the accuracy of other methods for the same measurement
- Characterization of a reference material

VIM [3.63].

- Assessing the measurement trueness of measured quantity values obtained from other measurement procedures for quantities of the same kind
- Calibration
- Characterization of a reference material

Reference Data

Determination of reference data for materials property databases is a typical reference laboratory activity, requiring reference procedures with established uncertainty and traceability of measurement, and an appropriate level of measurement quality control.

Reference Measurements

An increasingly important application of reference procedures is the provision of reference values for proficiency testing of samples or specimens, as an activity for a single reference laboratory or a small group of reference laboratories. Utilizing traceable reference values with specified uncertainty as the target values in proficiency tests (PT) has the advantage of providing information to the participants about their level of measurement bias, while PTs using a consensus value of the participants' results only provide information about their comparability.

Reference values for proficiency testing: a key application field

The International Measurement Evaluation Programme (IMEP) of the IRMM, Geel [3.72] is a prominent example of a proficiency test utilizing reference values. Recently, provision of reference values for PT samples or specimens has become a major activity for national metrology institutes for the dissemination of their measurement capabilities (Sect. 3.8.6).

Reference Materials

Reference materials are often characterized in interlaboratory studies that include a major number of expert laboratories utilizing a broad range of different measurement techniques. Alternative approaches include a small group of reference laboratories, or even a single reference laboratory, utilizing appropriate reference procedures as follows.

- Stand-alone determination of the property value under consideration in a single laboratory, utilizing an appropriate reference procedure, with check measurements using an independent method as a safeguard against blunders. In the chemistry field, this approach is often advocated by national metrology institutes, using primary methods of measurement such as isotope-dilution mass spectrometry.
- Collaborative determination, in a single laboratory using several independent reference procedures or several laboratories, using a reference procedure each. Given agreement within uncertainty limits, the results are combined into a mean value and an associated uncertainty.

Further information on strategies and procedures for characterization of reference materials is given in Sect. 3.7.

Validation of Routine Procedures

As another major field of application, reference procedures may be used to validate routine procedures for the same task. Similarly, routine procedures may be calibrated against references procedures for the same task. To this end, measurements using the routine procedure and the reference procedure are carried out in parallel on appropriate samples or specimens, and the results are compared. The great benefit of this approach is that real-life samples may be used, thus avoiding the problem of mismatch between calibrants and measuring objects, which is often encountered using reference materials.

Validation and calibration of routine procedures: an alternative to the use of reference materials

When validating a routine procedure against a reference procedure, the task is to investigate whether the results of the routine procedure (y) agree with those of the reference procedure (x). For this purpose, measurements using both procedures are carried out on a series of samples or specimens $S_1, S_2, \ldots, S_N$ with varying values of the property under consideration, yielding paired results $(y_1, x_1), (y_2, x_2), \ldots, (y_N, x_N)$. These data are examined to see whether they are compatible with a straight line $y = x$. To this end, a straight line $y_i = ax_i + b$ is fitted to the calibration data (y_i, x_i) using, e.g., the method of least squares, and the joint confidence region for the slope a and the intercept b is determined. If this confidence region contains the point $(a = 1, b = 0)$, the test is positive, i. e., the results obtained using the routine procedure and the reference procedure are compatible with a straight line $y = x$. This means that the routine procedure is not significantly biased against the reference procedure, thus both procedures measure the same quantity, although most often with different uncertainties.

Calibration of Routine Procedures

Calibration of a routine procedure against a reference procedure makes use of the same experimental design, with a slightly different task: the measurement results are utilized to derive a correction for the results of the routine procedure, and to evaluate the uncertainty for the corrected results.

Often calibration is performed for a narrow measuring range, using a single sample or specimen S. With x_S denoting the result of the reference procedure and y_S that of the routine procedure, the difference $\Delta_S = y_S - x_S$ may be utilized to correct the results obtained using the routine procedure on similar samples or specimens according to $y_{\text{corr}} = y_{\text{meas}} - \Delta_S$. The uncertainty of the corrected result can be calculated by combining the precision of the routine procedure and the uncertainty associated with the correction. The latter uncertainty includes the measurement uncertainty of the reference procedure.

For extended measuring ranges, calibration often requires parallel measurement of several samples or specimens $S_1, S_2, \ldots, S_N$. The results are used to derive a correction curve, e.g., by way of fitting a straight line $y_i = ax_i + b$ to the calibration data (y_i, x_i) and subsequent inversion according to $y_{\text{corr}} = (y_{\text{meas}} - b)/a$. The uncertainty of the corrected results may be calculated from the precision of the routine procedure and the uncertainty associated with the parameters of the calibration line. The latter uncertainty includes contributions from the measurement uncertainty of the reference procedure.

The approach outlined above is also applicable to another type of calibration, where, instead of a correction, the input/output behavior of a measuring system is determined. Here the task is to establish the relationship between the target quantity and the response (output) of the measuring system, e.g., temperature and current in case of a thermoelement. To this end, measurements using the procedure subject to calibration and the reference procedure are carried out in parallel on appropriate samples or specimens. A calibration curve is fitted to the calibration data, and this is used to convert measured responses into values of the target quantity. Again the advantage of using a reference procedure instead of reference materials or reference objects is that calibration can be carried out on real-life samples or specimens.

3.8.5 Presentation: Template for Reference Procedures

When dealing with a range of measurement or test procedures, it is useful to have the most important characteristics presented in a standard format. Most often these standards are sector specific, as e.g., ISO 78-2 [3.73], which specifies the presentation of procedures for chemical analysis. A generic template for the presentation of reference procedures was developed at the Federal Institute for Materials Research and Testing, Germany and is utilized in their catalog [3.74].

BAM catalog of reference procedures

According to this template, reference measurement procedures are presented as follows.

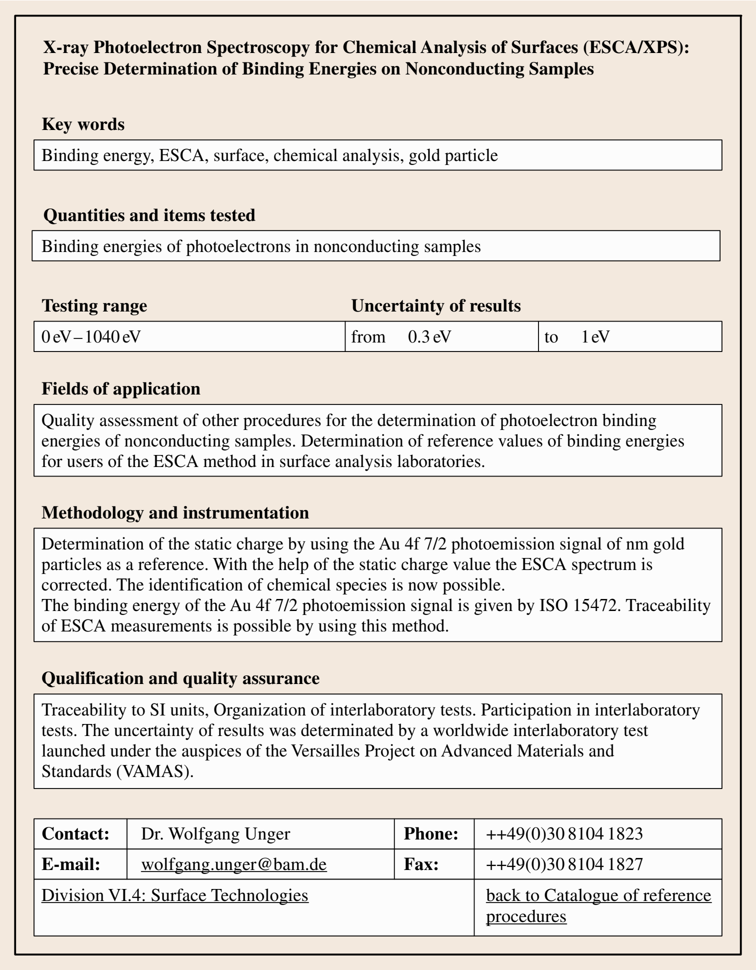
- *Title* – the title of the procedure.
- *Key words* – key words for the procedure.
- *Quantities and items measured* – the quantities measured, the kind of objects that can be investigated, and important measuring conditions.
- *Measuring range and uncertainty of measurement* – the range of values, i. e., the working range as a reference procedure, for each quantity measured, and an associated uncertainty range. Uncertainty of measurement is an expanded uncertainty according to the GUM [3.71] for an approximate confidence level of 0.95 (coverage factor $k = 2$), expressed as a relative uncertainty in % or as an absolute uncertainty in units of the quantity measured. For expressing an uncertainty range, different conventions may be used
  - referring to the lower and upper limit of the measuring range, respectively, or
  - specifying best-case and worst-case uncertainty.
- *Fields of application* – the kind of tasks for which the procedure is currently in use, or may be utilized as a reference: quality assessment of other procedures, characterization of reference materials and objects, determination of reference values, and other high-level tasks.
- *Methodology and instrumentation* – the method of measurement and essentials of the measuring system.
- *Qualification and quality assurance* – supporting evidence for the critical figures of merit, in particular measurement uncertainty (e.g., successful participation in interlaboratory comparisons), and efforts taken for measurement quality control.
- *Further information* – free-style presentation of additional information, in particular including pictures, diagrams, and references.

This template is also applicable to reference procedures for testing, even in cases of qualitative characteristics. Figure 3.33 shows an example taken from the BAM catalog [3.74].

3.8.6 International Networks: CIPM and VAMAS

One of the main services of national metrology institutes is to calibrate reference standards against national standards. Typically, these calibrations are performed by reference measurements, i. e., measurements using reference procedures of measurement.

On 14 October 1999 a multilateral agreement (the so-called CIPM MRA) was signed in Paris between 38 member states of the Metre Convention as well as two international organizations (the IAEA and IRMM). Currently the agreement has been signed by 48 member states, 32 associates of the General Conference on Weights and Measures (CGPM), and 3 international organizations (WMO, in addition to the above mentioned). This agreement had been drawn up by the International Committee of Weights and Measures (CIPM), its subject being the mutual recognition of national measurement standards and of the calibration and measurement certificates issued by national metrology institutes. The mutual recognition of national traceability systems aims to create a secure technical foundation for wider agreements in international trade, commerce, and regulatory affairs.



X-ray Photoelectron Spectroscopy for Chemical Analysis of Surfaces (ESCA/XPS): Precise Determination of Binding Energies on Nonconducting Samples

Key words

Binding energy, ESCA, surface, chemical analysis, gold particle

Quantities and items tested

Binding energies of photoelectrons in nonconducting samples

Testing range	Uncertainty of results	
0 eV – 1040 eV	from 0.3 eV	to 1 eV

Fields of application

Quality assessment of other procedures for the determination of photoelectron binding energies of nonconducting samples. Determination of reference values of binding energies for users of the ESCA method in surface analysis laboratories.

Methodology and instrumentation

Determination of the static charge by using the Au 4f 7/2 photoemission signal of nm gold particles as a reference. With the help of the static charge value the ESCA spectrum is corrected. The identification of chemical species is now possible.
The binding energy of the Au 4f 7/2 photoemission signal is given by ISO 15472. Traceability of ESCA measurements is possible by using this method.

Qualification and quality assurance

Traceability to SI units, Organization of interlaboratory tests. Participation in interlaboratory tests. The uncertainty of results was determinated by a worldwide interlaboratory test launched under the auspices of the Versailles Project on Advanced Materials and Standards (VAMAS).

Contact:	Dr. Wolfgang Unger	**Phone:**	++49(0)30 8104 1823
E-mail:	wolfgang.unger@bam.de	**Fax:**	++49(0)30 8104 1827
Division VI.4: Surface Technologies			back to Catalogue of reference procedures

Fig. 3.33 Example presentation of a reference procedure

This mutual recognition is far more than a formal act. The national measurement institutes concerned have to prove their competence and the reliability of their results in international intercomparisons – so-called key comparisons (KC) – and must operate an appropriate quality management system. For maximum transparency the results of these intercomparisons are published with free access on the Internet [3.75] by the International Office of Weights and Measures (BIPM) in Paris. In addition, the BIPM publishes another (freely accessible) Internet database [3.76] specifying the reference measurement services of the national metrology institutes – the so-called calibration and measurement capabilities (CMC).

BIPM databases for intercomparisons and reference measurements

These activities take place in close cooperation with the national metrology institutes and related designated institutes in the respective committees of the CIPM, the so-called consultative committees (CCs), one associated with each base quantity of the SI. Further information on the Metre Convention and the BIPM is given in Chap. 2.

Currently, the CMC database contains some 24 000 entries, specifying reference measurement procedures of the national metrology institutes and designated institutes for various metrology areas.

- *Acoustics, ultrasound, and vibration*
- *Electricity and magnetism*, including direct-current (DC) and alternating-current (AC) measurements, impedance, electric and magnetic fields, radio frequencies, and measurements on materials

- *Length*, including laser frequencies and dimensional metrology
- *Mass and related quantities*, including mass standards, force, pressure, density, hardness, torque, gravity, viscosity, and fluid flow
- *Photometry and radiometry*, including fiber optics and properties of detectors, sources, and materials
- *Amount of substance*, including high-purity chemicals, inorganic solutions, organic solutions, gases, water, metals and metal alloys, advanced materials, biological materials, food, fuels, sediments, soils, ores and particulates, other materials, pH, electrolytic conductivity, surfaces, films, and engineered nanomaterials
- *Ionizing radiation*, including dosimetry, radioactivity, and neutron measurements
- *Thermometry*, including temperature, humidity, and thermophysical quantities
- *Time and frequency*

In several of these CMC directories, entries for reference measurements of materials properties can be found. However, the reference procedures for materials measurements are a long way from covering the needs of materials metrology. Therefore, an initiative was recently taken by some of the leading national materials research institutes towards the International Committee of Weights and Measures, proposing to consider dedicated activities in the field of materials metrology.

A well-known international network for materials research and technology is the Versailles project on advanced materials and standards(VAMAS). VAMAS was founded in 1982, as a follow-up action from a G7 economic summit in Versailles, to provide a framework for international collaboration on prestandardization materials research, with scope to provide a technical basis for agreement on methodologies prior to the formal development of standards.

The Versailles project on advanced materials and standards

The main objective of VAMAS is to support trade in high-technology products through international collaborative projects aimed at providing the technical basis for drafting codes of practise and specifications for advanced materials. The scope of the collaboration embraces all agreed aspects of science and technology concerned with advanced materials including materials technology, test methods, design methods, and materials databases that are required as a precursor to the drafting of standards.

Current member countries are Australia, Brazil, Canada, Taiwan, France, Germany, India, Italy, Japan, Mexico, South Africa, South Korea, UK, USA, and also the EC, but VAMAS is open to organizations from other countries to participate in research activities.

Technical work is done in various working groups, the so-called technical work areas (TWA), of which there are 16 currently

- Surface chemical analysis
- Polymer composites
- Superconducting materials
- Mechanical property measurement of thin films and coatings
- Performance-related properties of electroceramics
- Full-field optical stress and strain measurement
- Spectrometry of synthetic polymers
- Nanomechanics applied to scanning probe microscopy
- Tissue engineering
- Creep, crack, and fatigue growth in weldments
- Modulus measurements
- Polymer nanocomposites
- Nanoparticle populations
- Materials databases interoperability
- Organic electronics
- Quantitative microstructural analysis

Further information about VAMAS activities may be obtained from the VAMAS website [3.77].

Considering the need to involve the metrology community, VAMAS approached the CIPM. The CIPM established an ad hoc Working Group on Materials Metrology (WGMM) in 2005. The WGMM published its final report in 2008 [3.78], leading to the signature of a memorandum of understanding (MoU) between VAMAS and the BIPM. This MoU provides the framework for the activities on materials metrology of all CIPM Consultative Committees (CC), including the establishment of special working groups in selected CC, identification of key issues and priority areas, and the development of collaborative studies for the validation of reference test procedures and of CIPM appropriate pilot studies for addressing comparability. The current status of material metrology, including the activities of the CIPM and of VAMAS and covering a broad scope of application fields, is comprehensively outlined in a special edition of *Metrologia* [3.79].

3.8.7 Related Terms and Definitions

In Sect. 3.8.4 the term *reference laboratory* was introduced and utilized without any further explanation. Similarly to the term *reference procedure*, this term has recently been used on an increasing scale, but a widely accepted definition is not available yet. Table 3.19 gives two definitions, originating from rather different fields: testing and laboratory medicine [3.64, 68].

The concept of reference laboratories, as defined above, has a lot of similarity with that of calibration laboratories and may be seen as a generalization. Unfortunately, the authors are not aware of any officially designated reference laboratories in the field of materials measurements. Examples from other fields include

- EU community and national reference laboratories for regulatory residue analysis of veterinary drugs and for air pollution control,
- reference laboratories for clinical measurements.

The concept of reference laboratories could become a key issue in the future development of ISO/IEC 17025, but this will require comprehensive discussion in various fora. In the authors' opinion, the concept should designate functions rather than a formal status, with functions centered on the provision of reference values. This could include provision of

- reference measurements/tests, preferably using reference procedures,
- reference materials and reference objects,
- proficiency testing schemes.

Another basic term that is intimately related to many of the issues of this section is that of a *measurement standard.* Here are two definitions from the International Vocabulary of Basic and General Terms in Metrology (VIM).

Measurement Standard. Realization of the definition of a given quantity, with stated quantity value and associated measurement uncertainty, used as a reference (VIM, 2008 [3.63]).

Reference Measurement Standard. Measurement standard designated for the calibration of other measurement standards for quantities of the same kind in a given organization or at a given location (VIM, 2008 [3.63]).

The concept of a measurement standard embraces both reference materials and reference procedures (of measurement). Currently the notion of national measurement standards maintained and disseminated by national metrology institutes is moving away from tangible objects to measurement service delivery capabilities.

Table 3.19 Definitions of the term reference laboratory

Reference (testing) laboratory	Reference measurement laboratory
Testing laboratory which – in arrangement with a specified laboratory community or through appointment by a competent organisation provides reference values in a specified technical field, i. e. property values of materials or products to which test results can be related or traced back and whose quality is fit for this purpose. (Eurolab Position Paper, 2007 [3.64])	Laboratory that performs a reference measurement procedure and provides results with stated uncertainties. (ISO 15195 [3.68])

3.9 Laboratory Accreditation and Peer Assessment

3.9.1 Accreditation of Conformity Assessment Bodies

Accreditation is the formal recognition that laboratories and other conformity assessment bodies such as inspection bodies and bodies certifying products, quality management systems, environmental management systems or personnel are competent to perform specified tasks such as tests or measurements.

Modern technologies in processing goods, just-in-time manufacturing of goods, subcontracting or outsourcing of activities to specialized organizations are demanding an established system to demonstrate competence and reliability of testing laboratories.

The word *accreditation* comes from the Latin *credere* (to believe, to be confident in) in which the word *dare*, meaning to give or to offer, is found. Accreditation therefore means *to give confidence*. People must be confident that accredited bodies provide their services with competence and reliability.

Accreditation should be understood as a system to maintain and continuously develop a laboratory's competence through competent assessments provided by accreditation bodies. It is designed to be a transparent process in which all the interested parties should have the possibility to become involved effectively. In this context, processes should also be understood as a host of small innovative steps with the goal of causing something new to arise. (For more about processes, see Sect. 3.9.)

Accreditation is defined in standards that are developed by ISO/CASCO. CASCO is the ISO committee on conformity assessment. ISO is the International Standards Organization, and CASCO is its conformity assessment cooperation. These standards specify the accreditation body's competence as well as the laboratory's competence. While accreditation bodies work according to the ISO/IEC 17011 standard, laboratories operate according to the ISO/IEC 17025 standard. In some special fields, such as in the field of the medical laboratories sector, specific standards have been developed as well. In other cases, sector-specific criteria are given in annexes to ISO/IEC 17025.

The goal of all of these standards is to ensure that the respective organizations are technically competent to perform well-defined services and that they have an adequate management system in place to ensure daily quality as defined by their clients.

Competence according to the spirit of all of these standards can be assured in the long term only when the accredited companies – and the accreditation bodies themselves – including all the persons involved, go through a permanent and conscious process.

It is obvious that conformity assessment bodies that continuously develop their competence in a systematic way have a large advantage in an increasingly competitive market. Accreditation recognizes the constant development of their competence and service quality. So far, accreditation should be seen as a structure to implement processes leading to competence, quality, and efficiency.

The task of an accreditation service is to assess whether a conformity assessment body (CAB) has the necessary technical competence, infrastructure, and organization to provide reliable services. The CAB's procedures should be clearly structured and controlled in a defined way. CABs need specialized knowhow and to develop an adapted structure to allow systematic decisions and learning processes.

Having a look at how accreditation works makes the significance of this evident. Let us take cable transportation plants and operators as an example.

- Accredited calibration laboratories calibrate measuring equipment for material testing
- Accredited testing laboratories test mechanical, hydraulic, and electronic system components and assess their quality
- Accredited inspection bodies inspect the critical technical installations for final approval of cable transportations
- Accredited certification bodies for products certify security components of cable transportation. In delivering credible conformity certificates they recognize compliance of such components with security requirements
- Accredited certification bodies evaluate quality and environmental management systems of cable transportation manufacturers and operators
- Accredited certification bodies for personnel examine training and technical competence of specialists.

3.9.2 Measurement Competence: Assessment and Confirmation

Measurement and testing are important elements of conformity assessment, the formal recognition that a material or a product conforms with specified requirements. The institutions performing conformity assessment tasks are called conformity assessment bodies (CABs). CABs are organizations providing the following conformity assessment services: calibration, testing, inspection, management system certification, personnel certification, and product certification. Accreditation is a third-party attestation related to a conformity assessment body conveying formal demonstration of its competence to carry out specific conformity assessment tasks. An authoritative body that performs accreditation is called an accreditation body (AB, ISO/IEC 17011). The goals of accreditation are

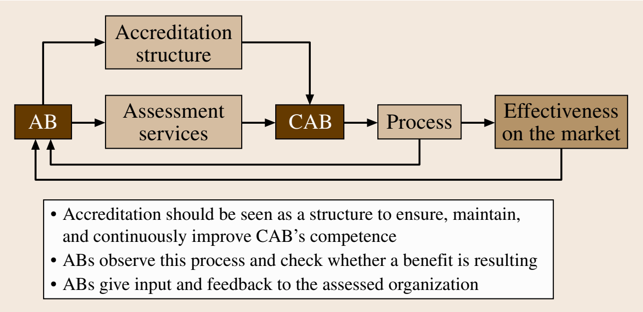
- to assess and confirm the competence and service quality of conformity assessment bodies,
- to ensure worldwide acceptance of their reports and certificates through a highly reliable confidence-building process,

- to support the competitiveness of accredited bodies on the national and global market.

In the entire chain of production and services one must have confidence that accredited testing and conformity assessment bodies provide their services in a competent and reliable manner.

The following examples illustrate the tasks fulfilled by accredited bodies.

- Measuring instruments of manufacturers and service providers have to be calibrated in order to allow correct measurements. *Accredited calibration laboratories* execute this task with the required precision and traceability to the international system of units (ISO 31).
- In their daily life, consumers depend on reliable analyses and tests, e.g., in the fields of food products, electrotechnical appliances, and medical diagnostic analyses. *Accredited testing laboratories* carry out these analyses and tests in a reliable manner.
- Inspections have to be carried out in order to detect safety oversights in food products, medicine, and technical installations at a given moment. *Accredited inspection bodies* fulfill these tasks reliably.



- Accreditation should be seen as a structure to ensure, maintain, and continuously improve CAB's competence
- ABs observe this process and check whether a benefit is resulting
- ABs give input and feedback to the assessed organization

Fig. 3.34 Accreditation should be seen as a structure to implement processes leading to competence quality and efficiency

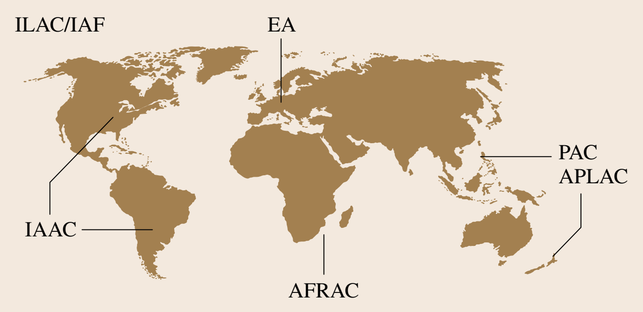


Fig. 3.35 Organizations of accreditation bodies in regions and worldwide

- Consumers and industry depend on the compliance of products with defined requirements. *Accredited certification bodies for products* attest to this conformity, after evaluation, in a credible way.
- *Accredited certification bodies for quality systems* deliver conformity certificates attesting to the conformity of quality systems to the standard requirements. This allows clients to be confident that these firms have structures and working procedures at their disposal that ensure a service that respects deadlines and corresponds to the agreed quality.
- *Accredited certification bodies for environmental management systems* attest through their conformity certificates that firms present the necessary conditions in order to continuously improve their environmental performance.
- A large number of firms depend on recognized specialists whose competence has to comply with defined criteria. Amongst these specialists we find welders, auditors, business consultants, and project managers. *Accredited certification bodies for personnel* attest to their technical competence so the economy can rely on these specialists.

In the light of the considerable costs of assessments and the long-term acceptance of accreditation as a tool to build confidence in reports and certificates, accreditation bodies are fully aware of the fact that the necessary added value can only be achieved if

- assessments are aimed at real objectives and values and not at formalism of standards,
- assessors are able to provide useful feedback to the ongoing processes in a competent and independent way.

Today, trade has become global, and export trade is vital to the development of any country's economy. The increasing development of globalization, the reduction of technical barriers to trade (TBTs), and the recognition of conformity certificates worldwide is leading to increasing competition not only between companies, but also between marketplaces and economic regions. Governments, industry, and the whole economy is challenged to face this situation. Specialist knowledge, use of modern technologies, and experience and competence in management and in the realization of modern scientific technical solutions are the way to face this development. Today, a competent and well-recognized infrastructure of conformity assessment bodies is a fundamental key factor for the economical success of a marketplace.

The technical barriers to trade (TBT) agreement encourages members of the World Trade Organization (WTO) to enter into mutual recognition agreements (MRAs) to enable assessments carried out in one country to be recognized in the other, based on reliable conformity assessments (TBT, Article 6):

> *Members shall ensure, whenever possible, that results of conformity assessment procedures in other Member's states are accepted, even when those procedures differ from their own, provided they are satisfied that those procedures offer an assurance of conformity with applicable technical regulations or standards equivalent to their own procedures. It is recognized that prior consultations may be necessary in order to arrive at a mutually satisfactory understanding.*

The TBT agreements were one of the reasons for the establishment of accreditation bodies in the major trade groupings of the world (Fig. 3.33).

ILAC: International Laboratory Accreditation Cooperation (covers the field of laboratory accreditation and, together with the IAF, the field of accreditation of inspection bodies).
IAF: International Accreditation Forum (covers the field of the accreditation of certification bodies and, together with ILAC, the field of accreditation of inspection bodies).
EA: European Cooperation for Accreditation.
IAAC: Inter American Cooperation for Accreditation.
PAC: Pacific Accreditation Cooperation (active in the field of accreditation of certification bodies).
APLAC: Asian Pacific Accreditation Cooperation (active in the field of laboratory accreditation).
AFRAC: African Accreditation Cooperation.

These agreements help to eliminate major technical barriers to trade and, at the same time, reduce costs by removing the need for duplicate testing of products by both exporters and importers. Similar provisions, consistent with the TBT agreement, are being encouraged by regional trade groupings such as, e.g., the European Union (EU), the European Free Trade Association (EFTA), the Asia-Pacific Economic Cooperation (APEC) or the North America Free Trade Association (NAFTA) and the Association of South-East-Asian Nations (ASEAN) Free Trade Association (AFTA).

MRAs may be *bilateral* or *multilateral*. Over the past 20 years, a network of bilateral and multilateral MRAs have been established between laboratory accreditation bodies throughout the major trade regions.

An example illustrates this: within the framework of the *new and global approach concept* of the European Union, national standards are replaced by harmonized European directives, which define the fundamental requirements, especially with regard to the safety of a product. Products in compliance with these requirements are marked with the Communauté Européenne (CE) logo. They can be put into free circulation in the European Economic Area. In order to implement and maintain this concept globally, all states that participate need an infrastructure of competent and reliable calibration, testing, inspection, and certification bodies.

Multilateral Agreements. Agreements within accreditation bodies within regional trade groupings such as Europe (Fig. 3.36) ensure worldwide harmonization of accreditation. They create the basis for international recognition of testing reports and conformity certificates, especially on the technical level. Therefore, an important task of each country's accreditation bodies is to maintain the interests of the respective country in view of international organizations such as the European Cooperation for Accreditation (EA) in Europe, the International Laboratory Accreditation Cooperation (ILAC), and the International Accreditation Forum (IAF). Together with all their stakeholders, these organizations continuously establish recommendations and guidance for harmonization of accreditation criteria and evaluation testing procedures to ensure worldwide reli-

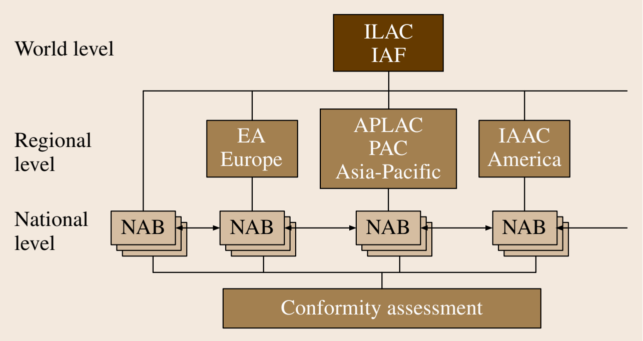


Fig. 3.36 The structure of the MLAs: The competence of the regional organizations is evaluated by ILAC and IAF, the regional organizations evaluate the national accreditation bodies, and these accredit the conformity assessment bodies

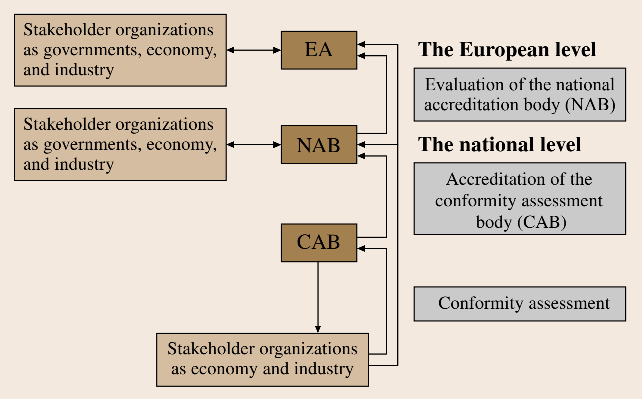

Stakeholder organizations as governments, economy, and industry
EA
The European level
Evaluation of the national accreditation body (NAB)
Stakeholder organizations as governments, economy, and industry
NAB
The national level
Accreditation of the conformity assessment body (CAB)
CAB
Conformity assessment
Stakeholder organizations as economy and industry

Fig. 3.37 Accreditation, an infrastructure offered to all interested parties

ability and recognition of testing reports and conformity certificates.

As a result, accreditation benefits from professional input and provides a structure to all the interested parties in order to

- establish confidence in testing reports and conformity certificates,
- develop competence in a continuous way,
- allow mutual recognition of testing reports and conformity certificates.

3.9.3 Peer Assessment Schemes

Peer assessment schemes follow basically the same goal as the accreditation process. Peer assessments schemes are private agreements with the aim of mutually assessing each other's competence. They are normally designed and performed by professional organizations in order to build confidence in their members' services.

By definition, such peer assessment schemes are second-party recognition schemes, while accreditation is clearly designed as a third-party recognition scheme. However they can provide a big help for its members.

3.9.4 Certification or Registration of Laboratories

The ISO 9001 standard is the well-known standard in the field of quality management systems (QMS) certification. This standard deals with quality management as generally applied to any types of companies. The ISO/IEC 17025 standard and laboratory accreditation are concerned with competence and quality and are specific for laboratories. An official communique (IAF/ILAC JWG/129) characterizes the differences between accreditation and certification as follows.

Certification

- means compliance with a standard or specification
- uses management system auditors who are certified by an independent body that meets internationally agreed criteria
- considers the total activities of an organization, and the scope of recognition is general.

Accreditation

- is the recognition of specific competence, and its scope is normally highly specific
- evaluates people skills and knowledge
- uses assessors who are recognized specialists in their fields
- evaluates the supporting management system for a specific activity
- involves practical tests as appropriate (proficiency testing and measurement audits).

3.10 International Standards and Global Trade

Standards are important tools for metrology and testing. A *standard* (French: *norme*, German: *Norm*) is defined as follows.

> *Standard: document, established by consensus and approved by a recognized body, that provides, for common and repeated use, rules, guidelines, or characteristics for activities and their result, aimed at the achievement of the optimum degree of order in a given context (ISO guide 2).*

The world principal forum for standardization is ISO, the International Organization for Standardization, an international-standard-setting body composed of representatives from various national standards organizations (http://www.iso.org). Founded in 1947 and with head-

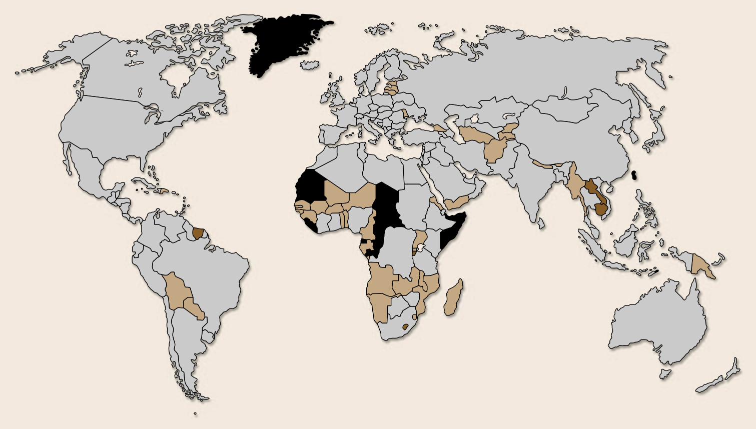


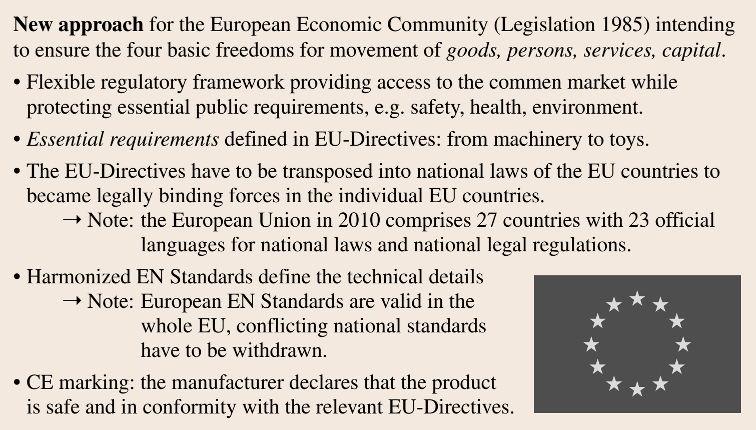
Fig. 3.38 A map of standards bodies who are ISO members (*grey*: members, *light brown*: correspondent members, *dark brown*: subscriber members, *black*: nonmembers)

quarters in Geneva, Switzerland, the organization publicizes worldwide proprietary industrial and commercial standards. While the ISO defines itself as a nongovernmental organization, it acts as a consortium with strong links to governments, setting standards that often become law, either through treaties or national standards. ISO has 163 national members, out of the 203 total countries in the world (Fig. 3.38).

International standards are standards developed by international standards organizations. They are available for consideration and use worldwide, and may be used either by direct application or by a process of modifying an international standard to suit local conditions. International standards are one way of overcoming technical barriers in international trade caused by differences among technical regulations and standards developed independently and separately by each nation, national standards organization, or companies.

3.10.1 International Standards and International Trade: The Example of Europe

Consider, for example, the evolution of the European standardization system. A coherent collection of European standards has been developed by a large number of stakeholders, including industry, public authorities, research organizations, and special interest groups such as consumers, representing a wide range of sectors. Technical barriers to trade have thus been removed, ensuring the free movement of goods – an essential prerequisite for the economic functioning of the Common European Market (Fig. 3.39).



New approach for the European Economic Community (Legislation 1985) intending to ensure the four basic freedoms for movement of *goods, persons, services, capital*.

- Flexible regulatory framework providing access to the commen market while protecting essential public requirements, e.g. safety, health, environment.
- *Essential requirements* defined in EU-Directives: from machinery to toys.
- The EU-Directives have to be transposed into national laws of the EU countries to became legally binding forces in the individual EU countries.
  → Note: the European Union in 2010 comprises 27 countries with 23 official languages for national laws and national legal regulations.
- Harmonized EN Standards define the technical details
  → Note: European EN Standards are valid in the whole EU, conflicting national standards have to be withdrawn.
- CE marking: the manufacturer declares that the product is safe and in conformity with the relevant EU-Directives.

Fig. 3.39 The role of harmonized standards for trade, the example of Europe

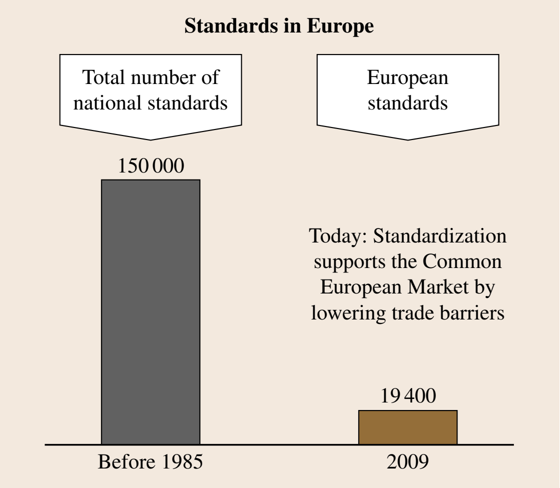


Fig. 3.40 Number of standards before and after formation of the European Union and the evolution of the Common European Market

There are now only three standardization organizations that coordinate, monitor, and support the relevant activities in Europe.

- CENELEC in the field of electrotechnical standardization (http://www.cenelec.eu)
- ETSI in the telecommunications sector (http://www.etsi.eu)
- CEN in all other fields of standardization (http://www.cen.eu).

The development of standardization in Europe is illustrated in Fig. 3.40. It shows that, before 1985, there were about 150 000 national standards in the European countries. Due to harmonization efforts of standards in the evolution process of the European Union (EU), there are now fewer than 20 000 harmonized standards in the EU. The business domains relating to the European standards in 2010, as compiled by the bureaus of the CEN/CENELEC Management Centre, are shown in Fig. 3.41.

As a prerequisite for global trade, international cooperation in standardization is necessary. The modalities of technical cooperation and exchange of information between the CEN and the ISO are laid down in the *Vienna Agreement*, while the *Dresden Agreement* regulates the cooperation between CENELEC and IEC (Fig. 3.42).

The purpose of these agreements is

- to make best use of available resources and thus avoid duplication of work, and
- to ensure that international and European standards are not only compatible, but identical.

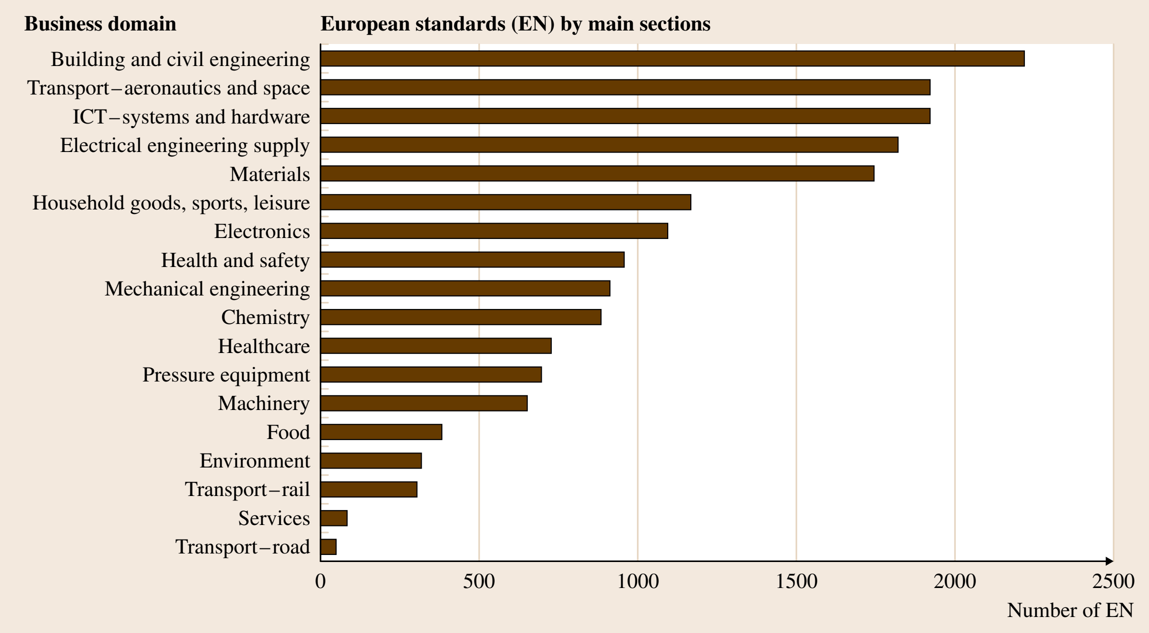


Fig. 3.41 Business domains and number of European standards

Vienna and Dresden agreements:

Where possible, carry out specialist work at one level of standardization, and use parallel voting procedures to achieve simultaneous adoption as ISO/IEC and EN standards.

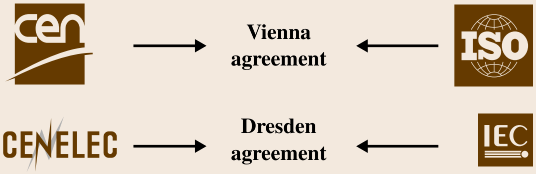


Fig. 3.42 International agreements in standardization

3.10.2 Conformity Assessment

The regulation of international trade is done through the World Trade Organization (WTO) at the global level, and through several other regional arrangements such as MERCOSUR in South America, the North American Free Trade Agreement (NAFTA) between the USA, Canada, and Mexico, and the European Union (EU). International standards have to follow the principles of *the World Trade Organization (WTO) Committee on Technical Barriers to Trade*, i. e., transparency, openness, impartiality and consensus, effectiveness, and relevance.

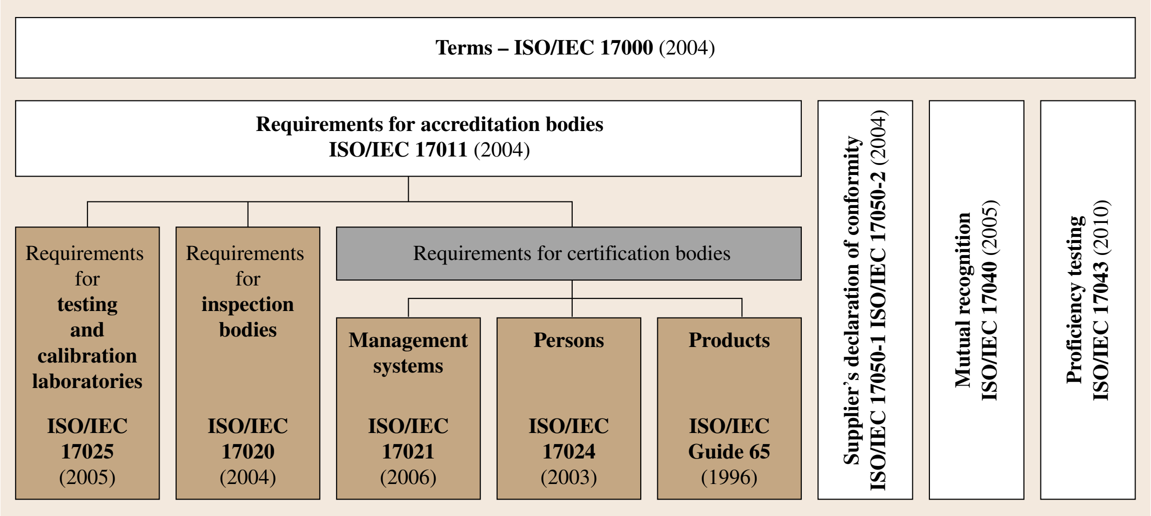
Conformity assessment – performed by *conformity assessment bodies* (CABs) – is any activity to determine, directly or indirectly, that a process, product or service meets relevant standards and fulfills relevant requirements. The types of conformity assessment bodies are

- laboratories
  - testing laboratories
  - calibration laboratories
- inspection bodies
- certification bodies for
  - management systems
  - products
  - persons.

The results of conformity assessment activities are, e.g., certificates, (testing) reports, declarations, and marks of conformity. The documents must be correct, valuable, clear, and easy to understand for all participants in the market. The results must be in such a form that participants in the market and authorities are able to use them in their decisions. The requirements written in standards of conformity assessment activities and the requirements for the CABs must give confidence in the results of the conformity assessment bodies. General requirements in all conformity assessment standards are

- organization and structure,
- impartiality, independence,
- competence of personnel and access to competence,
- equipment, procedures, and methods,
- decision on conformity assessment and reporting,
- management system of the CAB.

An overview on the standards of conformity assessment is given in Fig. 3.43 and Table 3.20.



Terms – ISO/IEC 17000 (2004)

Requirements for accreditation bodies ISO/IEC 17011 (2004)

Requirements for testing and calibration laboratories ISO/IEC 17025 (2005)

Requirements for inspection bodies ISO/IEC 17020 (2004)

Requirements for certification bodies

Management systems ISO/IEC 17021 (2006)

Persons ISO/IEC 17024 (2003)

Products ISO/IEC Guide 65 (1996)

Supplier's declaration of conformity ISO/IEC 17050-1 ISO/IEC 17050-2 (2004)

Mutual recognition ISO/IEC 17040 (2005)

Proficiency testing ISO/IEC 17043 (2010)

Fig. 3.43 Overview of conformity assessment standards

Table 3.20 Conformity assessment standards

ISO/IEC 17000:2004	Terms, *reference document*
ISO/IEC 17011:2004	Accreditation bodies
ISO/IEC 17020:2004	Inspection bodies
ISO/IEC 17021:2006	Certification bodies for management systems
ISO/IEC 17024:2003	Certification bodies for persons
ISO/IEC 17025:2005	Laboratories
EN 45011:1998 (EN ISO/IEC 17065)	Certification bodies for products (possibly to be replaced in the future)
ISO/IEC 17040:2005	Peer assessment
ISO/IEC 17050-1:2000 ISO/IEC 17050-2:2004	Supplier's declaration of conformity

3.11 Human Aspects in a Laboratory

3.11.1 Processes to Enhance Competence – Understanding Processes

Continuous development of the technical competence of all staff necessitates processes that lead to creative and innovative solutions. Such processes must be understood in such a way that continuous progress at all levels and for all functions in an organization is self-evident.

Clients expect from laboratories services and solutions that take into account the latest developments in a respective technical field. Only competent laboratories will be able to perform such services in an economic and reliable way. Materials testing is very seldom based on routine testing. Often, methods have do be developed or adapted to

- new materials,
- new questions to be answered,
- special problems to be solved.

Validation plays a big role, as testing methods must be adapted to given tasks and be proved suitable for these tasks. To prove the fitness for purpose of testing methods in order to achieve highly reliable results, which can in special cases be very complex, requires the continuous development of specialized knowledge. Such knowledge consists not only of the details of the testing methods that one plans to use but also the knowledge about what customers are planning to do with the results and a general knowledge about the materials tested.

In very rare cases the required competence can be organized from outside. Competence must be developed by all people involved in a laboratory, who are faced with the daily questions and problems. This requires staff who are highly motivated and deeply engaged in their daily tasks.

The long-term safeguarding of competence requires the use of processes at all levels of an organization. This necessitates an organizational structure that enables the development of such processes. Processes involving people consist of a host of small steps with the goal of creating something new. Competence, e.g., according to the spirit of the standards relevant for accreditation, can be assured in the long term only if their implementation is considered as an ongoing learning process. Learning processes are highly dependent on management structures, and on the provisions to make sure that the whole staff can feel full responsibility for their work.

We all undergo processes, whether we want to or not, be it consciously or unconsciously, simply because of all the information that reaches us and daily interactions. It is important for the head of an organization to ask the following questions.

- What kinds of processes are going on in the entire company, in parts of it, and within individual people?
- Which of them are important for me, and how can I recognize them?
- What can I do to support processes leading to engaged staff with good motivation and loyalty?
- What is the driving force for enthusiasm in my organization?

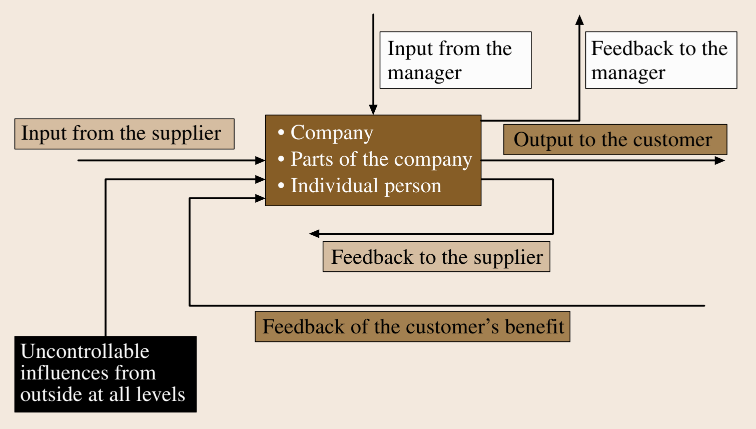


Fig. 3.44 The principle of systematically controlled input and output

To find answers it is important that

- communication principles are discussed regularly,
- conflicts are treated and solved systematically,
- internal audits are carried out in such a way that aspects of the management system and procedures that could hinder learning processes or lead to internal conflicts are discovered.

Being process-orientated means that each staff member may bring in his/her competence and his/her creativity for

- planning procedures
- putting procedures into effect

and therefore participates in the ongoing innovative process.

This is also a reply to those who perceive documented management systems as limiting, restrictive, and conservative, and that, at best, they are able to optimize something already existing, but in no case to create something new and innovative.

Management systems should therefore always, and particularly in the case where the aim is long-term safeguarding of competence and quality, be implemented in such a way that creativity plays an important role. Creativity also depends on how responsibilities are allocated and on the functional behavior of all individuals in an institution; this means that each staff member must have the possibility to take on his or her responsibility fully and be assured of the protection required to provoke unconventional ideas.

3.11.2 The Principle of Controlled Input and Output

Functional behavior therefore necessitates a clear definition of responsibilities. This means that everyone assumes responsibility and respects the responsibility of people who come above or below them in the hierarchy and of colleagues at the same hierarchical level. If this does not happen, good processes will be blocked and creativity cannot arise. Clear structures in a laboratory's management system are therefore an important prerequisite. This structure should ensure that client requests will lead to respective reactions at all levels: the individual staff level, the subunit level, and at the whole organization level. Of course it is important that these reactions are fully controllable and controlled, without taking responsibility away from the individual staff.

The First Control Circle: Input from and Feedback to the Manager (Superior Person)

Input should be given regularly, e.g., by defining targets to be achieved. This can be done on daily, weekly or monthly bases and requires regular feedback from the collaborator. Input and feedback should be systematically planned and given in a structured way. Respected responsibility means that each collaborator has the freedom to organize and to arrange their own activities within his defined frame of responsibility.

The Second Control Circle: Input from and Feedback to the Supplier

All internal and external institutions (parts of a company), and individual people performing the activities prior to and in support of one's own activities, should be considered as suppliers and treated as such. Input and feedback should be defined, planned, and given in a structured way.

The Third Control Circle: Input from and Feedback to the Customer

All internal and external institutions (parts of a company) and individual persons benefiting form one own's activities should be considered as customers and treated as such. Output and feedback should be well defined, planned, and given in a structured way.

The principle of systematically controlled input and output should be valid for the whole organization, for parts of it, and for the individual staff. Each element, the whole organization, parts of it or individual staff members have to deal with information from clients, customers, and superiors. If such a system is applied, it will greatly assist the development of staff entrepreneurship. The system will distinguish between those who are prepared to take responsibility as an entrepreneur and those who are not. It is obvious that staff feeling like entrepreneurs will undergo learning processes to develop the technical and personal skills needed to take the responsibility linked to given hierarchical competencies.

The principle of systematically controlled input and output will automatically lead to intensive interaction of the individuals involved, the team, the subject, and the involved environment and therefore provide good bases for implementation of management systems such as that defined in the ISO/IEC 17025 standard.

3.11.3 The Five Major Elements for Consideration in a Laboratory

Quality management starts with the individual. If the principle of systematically controlled input and output is applied strictly, each entrepreneur in the organization will feel responsible for the quality of his or her product or performed service. The following five elements will help to assure reliable performance at all hierarchical levels, the laboratory, the subunits, and the individual staff.

1. Systematic estimation of opportunities and the corresponding risks
2. Leadership and ability to lead at all levels
3. Traceability of all decisions; decisions should be comprehensive
4. Periodic evaluation of the effectiveness of one's own organization, procedures, etc.
5. A general system to limit damage. Corresponding procedures should be laid down in the overall quality management system.

Regardless of the quality system applied, internal audits should preferably be used to monitor how these five elements are implemented at all the hierarchical levels mentioned above.

3.11.4 Internal Audits

Internal audits are a very helpful tool to observe whether good processes are being carried out. However, they should not be used to establish hidden supervision of staff and their activities. To audit means to observe if the structure of the management system is developed in such a way that conscious learning processes at all levels in a company are possible and provide feedback. This means that auditors have to take into account

- the people,
- their processes,
- the structure of the organization and the management system.

Auditors have to respect the functions and the corresponding hierarchical competencies in an organization. The staff should therefore always be included in the audit teams and be responsible for those audits that are performed within their own area of responsibility. Audit teams should include the following team members.

- The individual persons responsible for a given activity or task
- The superior who wants to gain trust in the work of his collaborator
- A person having the necessary communication skills and who is familiar with
  - the quality system,
  - its structure and procedures, and
  - the audit techniques.

To audit an organization, parts of it, or an individual person means to observe and to learn how the management system is implemented, whether it enables conscious processes at all levels in a company, and to give feedback.

The most important point is that auditors are trained to observe learning processes, and to this end they need to know more about human beings and behavior. If international standards, such as the ISO/IEC 17000ff series of standards, are implemented and audits are used to verify if these standards are implemented effectively, auditors should not only look at how these standards are translated into action but if this translation is of benefit to the company, parts of it, and the individual staff. This means that audits must concentrate on the effects of the defined measures, and after an audit the audited organization or persons should be able to explain what they discovered during the audit, e.g., in the audit report, which is – in the best case – provided by the audited people or the people who are responsible for the audited part of the organization.

3.11.5 Conflicts

Conflicts are normal and will occur in all situations. Conflicts result from disagreements or different opinions. Conflicts can occur if engaged people are convinced about their own solutions or ideas. To be in conflict with somebody is not bad behavior; bad behavior is not to solve these conflicts. Unsolved conflicts lead to situations where learning processes are no longer possible and later where cooperation between staff members becomes nearly impossible.

It is therefore good practise to establish a system to solve conflicts, even the smallest one, in *statu nascendi* and in a systematic way. This can, e.g., be done using the following principles.

1. All the staff (including the general manager) are required to declare a conflict (even a possible once) as soon as it is recognized, to the person(s) with whom a potential conflict exists. The person who recognizes the conflict can be considered as the client, and the other conflicting person as the supplier.
2. The client should discuss this potential conflict as soon as possible with the supplier, while this is possible and with the least emotion possible.
3. If the client and the supplier are not able to solve the conflict, it is important that they involve a third person to act as an arbitrator. Arbitration in an organization should not be considered as a tribunal, but as a good way to translate to one another different opinions that obviously cannot be understood. It is important that the arbitrator does not rate the opinions of the conflicting persons. On the other hand, he is obliged to explain what is already fixed, e.g., in the management system or in technical procedures.
4. Both conflicting people (client and supplier) should afterwards express their wish for future solutions in a comprehensive way such that it is clear what the other person is expecting. The arbitrator ensures that this discussion remains factual.
5. Having learned from each other, both conflicting partners explain to the other person the next steps to realizing good solutions in the future.
6. All three partners explain what they have learned in the process. If one of the partners does not have trust in the solution by the other partner, he will announce this and clarify remaining points immediately.
7. The supplier will minute the discussions, especially the agreed conclusions about the next steps.

It is recommended to lay down such a procedure in the corresponding management system of an organization. This will ensure that conflicts do not remain untouched and that the necessary learning steps are taken; this is an important aspect to maintain the necessary competence of all staff.

3.11.6 Conclusions

The increasing development of globalization, the European and worldwide reduction of technical barriers to trade, and the recognition of conformity certificates worldwide is leading to increasing competition not only between companies, conformity assessment bodies, and laboratories, but also between marketplaces and economic regions. Specialized knowledge, use of modern technologies, and experience and competence in the management and realization of innovative scientific technical solutions are the ways to face this development.

Consideration of human aspects is of utmost important, especially in laboratories where there is a clear need to continuously develop technical competence. Technical competence can only be continuously developed and improved, if the management can base this on engaged staff. There are a few basic principles that must be respected to support the development of engagement. It is highly recommended to invest in these principles. Such investments should be treated in the same way as investments in technical equipment or scientific developments. They are at least of similar importance.

3.12 Further Reading: Books and Guides

The following is a selection of sources and is neither comprehensive nor intended as a recommendation.

Statistics

- NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/ (last updated June 23, 2010, last accessed June 8, 2011)
- G. W. Snedecor, W. G. Cochran: *Statistical Methods* (Iowa State Univ. Press, Ames 1993)
- R. Caulcott: *Statistics in Research and Development* (Chapman Hall, New York 1983)
- V. Barnett, T. Lewis: *Outliers in Statistical Data*, 3rd edn. (Wiley, New York 1994)
- P. J. Huber: *Robust Statistics* (Wiley, New York 1981)

Experimental Design

- G. M. Clarke: *Statistics and Experimental Design* (Edward Arnold, London 1984)
- S. N. Deming, S. L. Morgan: *Experimental Design: A Chemometric Approach* (Elsevier, Amsterdam 1987)

Quality Control

- J. W. Oakland: *Statistical Process Control* (Heinemann, Oxford 1989)
- ISO 8258:1991: *Shewhart Control Charts* (ISO, Geneva 1991)

Uncertainty Estimation

- S. G. Rabinovich: *Measurement Errors and Uncertainties: Theory and Practice* (Springer, New York 2000)
- ISO: *Guide to the Expression of Uncertainty in Measurement*, 2nd edn. (ISO, Geneva 1995)
- ISO TS 21748:2005: *Guide to the Use of Repeatability, Reproducibility and Trueness Estimates in Measurement Uncertainty Estimation* (ISO, Geneva 2005)

Free and Open-Source Statistical Software

- R: *A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna 2005), http://www.r-project.org/ (last accessed June 8, 2011)
- NIST Dataplot: http://www.itl.nist.gov/div898/software/dataplot/homepage.htm (date created June 5, 2001, last updated May 15, 2010, last accessed June 8, 2011)
- AMC Software Excel add-ins: http://www.rsc.org/Membership/Networking/InterestGroups/Analytical/AMC/Software/index.asp (Roy. Soc. Chem., London 2011)

General Statistics

- GenStat (VSN International Ltd.): Video Streaming Network (2011) http://www.vsn-intl.com/?s=genstat (last accessed June 3, 2011)
- Minitab (Minitab Inc.): http://www.minitab.com/en-DE/default.aspx (Minitab, State College 2011)
- SAS/STAT (SAS Institute Inc.): http://www.sas.com/ (SAS Institute Inc., Cary 2011)
- TIBCO Spotfire S+: http://spotfire.tibco.com/products/s-plus/statistical-analysis-software.aspx (TIBCO, Somerville) (last accessed June 8, 2011)
- SPSS (IBM SPSS Inc.): http://www.spss.com/ (last accessed June 8, 2011)
- Statistica (Statsoft Inc.): http://www.statsoft.com/ (last accessed June 8, 2011)

Experimental Design

- Design-Ease (StatEase Inc.): http://www.statease.com/ (Stat-Ease, Inc., Minneapolis 2011)
- MODDE (Umetrics): http://www.umetrics.com/ (Umetrics, Andover 2011)

References

3.1 P.M. Gy: *Sampling of Particulate Materials, Theory and Practice* (Elsevier, Amsterdam 1979) p. 431

3.2 M.H. Ramsey, S.L.R. Ellison (Eds.): *Eurachem/EUROLAB/CITAC/Nordtest/AMC Guide: Measurement Uncertainty Arising from Sampling: A Guide to Methods and Approaches* (Eurachem, Prague 2007)

3.3 ISO: *Guide to the Expression of Uncertainty in Measurement*, 2nd edn. (ISO, Geneva 1995)

3.4 U. Kurfürst, A. Desaules, A. Rehnert, H. Muntau: Estimation of measurement uncertainty by the budget approach for heavy metal content in soil under different land use, Accredit. Qual. Assur. **9**, 64–75 (2004)

3.5 Analytical Methods Committee (AMC): Uncertainty of Measurement: Implications of its Use in Analytical Science, Analyst **120**, 2303–2308 (1995)

3.6 M.H. Ramsey, A. Argyraki: Estimation of measurement uncertainty from field sampling: Implications for the classification of contaminated land, Sci. Total Environ. **198**, 243–257 (1997)

3.7 J.A. Lyn, M.H. Ramsey, S. Coad, A.P. Damant, R. Wood, K.A. Boon: The duplicate method of uncertainty estimation: Are eight targets enough?, Analyst **132**, 1147–1152 (2007)

3.8 M.H. Ramsey: Sampling as a source of measurement uncertainty: Techniques for quantification and comparison with analytical sources, J. Anal. At. Spectrom. **13**, 97–104 (1998)

3.9 S.L.R. Ellison, M. Roesslein, A. Williams (Eds.): *Eurachem/CITAC Guide: Quantifying Uncertainty in Analytical Measurement* (Eurachem, London 2000), Available from the Eurachem secretariat, or from LGC Limited

3.10 M.H. Ramsey, S. Squire, M.J. Gardner: Synthetic reference sampling target for the estimation of measurement uncertainty, Analyst **124**(11), 1701–1706 (1999)

3.11 S. Squire, M.H. Ramsey, M.J. Gardner, D. Lister: Sampling proficiency test for the estimation of uncertainty in the spatial delineation of contamination, Analyst **125**(11), 2026–2031 (2000)

3.12 M. Thompson, T. Fearn: What exactly is fitness for purpose in analytical measurement?, Analyst **121**, 275–278 (1996)

3.13 M.H. Ramsey, P.D. Taylor, J.C. Lee: Optimized contaminated land investigation at minimum overall cost to achieve fitness-for-purpose, J. Environ. Monit. **4**(5), 809–814 (2002)

3.14 M.H. Ramsey, J.A. Lyn, R. Wood: Optimised uncertainty at minimum overall cost to achieve fitness-for-purpose in food analysis, Analyst **126**(10), 1777–1783 (2001)

3.15 B. Geelhoed, H.J. Glass: Comparison of theories for the variance caused by the sampling of random mixtures of non-identical particles, Geostand. Geoanal. Res. **28**(2), 263–276 (2004)

3.16 C.C. Ferguson: The statistical basis for the spatial sampling of contaminated land, Ground Eng. **25**, 34 (1992)

3.17 International vocabulary of metrology — Basic and general concepts and associated terms (VIM), 3rd edn., JCGM 200:2008 (including Corrigendum of Mai 2010)

3.18 BIPM, IEC, IFCC, ISO, IUPAC, IUPAP, OIML: *Guide to the Expression of Uncertainty in Measurement (GUM)*, 1st edn. (BIPM and IEC and IFCC and ISO and IUPAC and IUPAP and OIML, Paris 1993), corrected and reprinted in 1995

3.19 ISO: *Guide to the Expression of Uncertainty in Measurement*, 1st edn. (ISO, Geneva 1995)

3.20 EUROLAB: *Guide to the Evaluation of Measurement – Uncertainty for Quantitative Test Results*, Tech. Rep. 1/2006 (EUROLAB, Paris 2006), http://www.eurolab.org/

3.21 EUROLAB: *Measurement Uncertainty Revisited: Alternative Approaches to Uncertainty Evaluation*, Tech. Rep. 1/2007 (EUROLAB, Paris 2007), http://www.eurolab.org/

3.22 ISO: ISO 5725 (series of standards in 6 parts), Accuracy (trueness and precision) of measurement methods and results (ISO, Geneva 1998–2005)

3.23 ISO/TS 21748, Guide to the use of repeatability, reproducibility and trueness estimates in measurement uncertainty estimation (ISO, Geneva 2004)

3.24 EURACHEM/CITAC: *Quantifying Uncertainty in Analytical Measurement*, 2nd edn. (EURACHEM/CITAC, Prague 2000)

3.25 EUROLAB: *Measurement Uncertainty in Testing*, Tech. Rep. 1/2002 (EUROLAB, Paris 2002)

3.26 EA: *Expression of Uncertainty in Quantitative Testing*, Guideline EA-4/16 (EA, Rijswijk 2003), www.european-accreditation.org/content/publications/pub.htm (last accessed 14 April 2011)

3.27 B. Magnusson, T. Näykki, H. Hovind, M. Krysell: *Handbook for Calculation of Measurement Uncertainty in Environmental Laboratories*, Tech. Rep. 537 (Nordtest, Espoo 2003)

3.28 EA: *Expression of the Uncertainty of Measurement in Calibration*, Guideline EA-4/02 (European Cooperation for Accreditation, Paris 1999)

3.29 AFNOR FD X 07-021: Fundamental standards – Metrology and statistical applications – Aid in the procedure for estimating and using uncertainty in measurements and test results (European Diagnostic Manufacturers Association, Brussels 2006)

3.30 Supplement No. 1 to the GUM: Propagation of distributions using a Monte Carlo method (ISO/IEC Guide 98-3:2008/Suppl 1:2008) (BIPM, Paris 2008)

3.31 ISO 13528: Statistical methods for use in proficiency testing by interlaboratory comparison (ISO, Geneva 2005)

3.32 ISO/TS 21749: Measurement uncertainty for metrological applications (ISO, Geneva 2005)

3.33 ISO: *ISO 9000, International Consensus on Good Quality Management Practices* (ISO, Geneva 2000)

3.34 CITAC/EURACHEM: *Guide to Quality in Analytical Chemistry – An Aid to Accreditation* (CITAC/EURACHEM, Prague 2002)

3.35 AOAC International: *AOAC Peerverified Methods Policies and Procedures* (AOAC International, Gaithersburg 1993)

3.36 S. Kromidas: *Validierung in der Analytik* (Wiley-VCH, Weinheim 1999), (in German)

3.37 W. Wegscheider: Validation of analytical methods. In: *Accreditation of Chemical Laboratories*, ed. by H. Günzler (Springer, Berlin, Heidelberg 1996)

3.38 ISO: *ISO Guide 30, Terms and Definitions Used in Connection with Reference Materials* (ISO, Geneva 1992)

3.39 B. King: 4E-RM-Guide, Selection and use of reference materials – A basic guide for laboratories and accreditation bodies (EURACHEM, Prague 2002) http://www.eurachem.ut.pt/

3.40 IAGRM: *Final Third Party Quality Assessment of Producers of Reference Materials* (IAGRM, Bristol 2004)

3.41 ISO: *International Vocabulary of Basic and General Terms in Metrology* (ISO, Geneva 1999)

3.42 ISO: *ISO Guide 34, General Requirements for the Competence of Reference Materials Producers* (ISO, Geneva 2009)

3.43 United Nation: *Glossary of Terms for Quality Assurance and Good Laboratory Practices* (United Nation, New York 1995)

3.44 M.J.T. Milton, T.J. Quinn: Primary methods for the measurement of amount of substance, Metrologia **38**, 289–296 (2001), http://www.bipm.org/en/metrologia/ (last accessed 14 April 2011)

3.45 BAM: COMAR database (Federal Institute for Materials Research and Testing, Berlin 2011), http://www.comar.bam.de

3.46 International Atomic Energy Agency (IAEA): Nuclear Sciences and Applications (IAEA, Vienna 2011), http://www-naweb.iaea.org

3.47 Bureau International des Poids et Mesures (BIPM): http://www.bipm.org/ (BIPM, Paris 2011)

3.48 ISO: *ISO Guide 31, Contents of Certificates of Reference Materials* (ISO, Geneva 2000)

3.49 A. Zschunke (Ed.): *Reference Materials in Analytical Chemistry* (Springer, Berlin, Heidelberg 2000)

3.50 J. Pauwels, A. Lamberty: CRMs for the 21st century new demands and challenges, Fresenius J. Anal. Chem. **370**, 111–114 (2001)

3.51 M. Parkany, H. Klich, S.D. Rasberry: REMCO, the ISO Council Committee on Reference Materials – its 1st 25 years, Accredit. Qual. Assur. **6**, 226–235 (2001)

3.52 ISO: *ISO 6506-1, Metallic Materials – Brinell Hardness Test - Part 1: Test Method* (ISO, Geneva 2005)

3.53 ISO: *ISO 6507-1, Metallic Materials – Vickers Hardness Test - Part 1: Test Method* (ISO, Geneva 2005)

3.54 ISO: *ISO 6508-1, Metallic Materials – Rockwell Hardness Test - Part 1: Test Method* (ISO, Geneva 2005)

3.55 ISO: *ISO 6507-2, Metallic Materials – Vickers Hardness Test - Part 2: Verification and Calibration of Testing Machines* (ISO, Geneva 2005)

3.56 ISO: *ISO 148-1, Metallic Materials – Charpy Pendulum Impact Test – Part 1: Test Method* (ISO, Geneva 2009)

3.57 ISO: *ISO 148-2, Metallic Materials – Charpy Pendulum Impact Test – Part 2: Verification of Testing Machines* (ISO, Geneva 2008)

3.58 ISO: *ISO 148-3, Metallic Materials – Charpy Pendulum Impact Test – Part 3: Preparation and Characterization of Charpy V-Notch Test Pieces for Indirect Verification of Pendulum Impact Machines* (ISO, Geneva 2008)

3.59 ISO: *ISO Guide 35, Reference Materials – General and Statistical Principles for Certification* (ISO, Geneva 2006)

3.60 ISO: *ISO 6892-1, Metallic Materials – Tensile Testing – Part 1: Method of Test at Room Temperature* (ISO, Geneva 2009)

3.61 ISO: *ISO 5725-6, Accuracy (Trueness and Precision) of Measurement Methods and Results – Part 6: Use in Practice of Accuracy Values (ISO 5725-6:1994 Including Technical Corrigendum 1:2001)* (ISO, Geneva 2002)

3.62 ISO: *ISO/IEC Guide 98-3, Uncertainty of Measurement - Part 3: Guide to the Expression of Uncertainty in Measurement (GUM:1995)* (ISO, Geneva 2008)

3.63 JCGM 200: *International Vocabulary of Metrology – Basic and General Concepts and Associated Terms (VIM)* (JCGM, 2008), http://www.bipm.org/en/publications/guides/vim.html (identical to ISO Guide 99:2007, VIM, 3rd edn.) (last accessed 14 April 2011)

3.64 EUROLAB Position Paper No. 1/2007: Reference laboratories in the field of testing (March 2007) http://www.eurolab.org/pub/i_pub.html (last accessed 14 April 2011)

3.65 A. Ono, T. Baba, K. Fujii: Traceable measurements and data of thermophysical properties for solid materials: A review, Meas. Sci. Technol. **12**, 2023–2030 (2001), (last accessed 14 April 2011)

3.66 ISO: *ISO 14192, Gas Analysis – Conversion of Gas Mixture Composition Data* (ISO, Geneva 1998), (last accessed 14 April 2011)

3.67 EURAMET: *EURAMET Guide, Metrology in Short*, 3rd edn. (EURAMET, 2008), http://www.euramet.org/index.php?id=mis (last accessed 14 April 2011)

3.68 ISO: *ISO 15195, Clinical Laboratory Medicine – Requirements for Reference Measurement Laboratories* (ISO, Geneva 2003)

3.69 JCGM 101: Evaluation of measurement data – Supplement 1 to the "Guide to the expression of uncertainty in measurement", Propagation of distributions using a Monte Carlo method (JCGM, 2008) http://www.bipm.org/en/publications/guides/gum.html (last accessed 14 April 2011)

3.70 IUPAC: *IUPAC Compendium of Chemical Terminology ('Gold Book')*, 2nd edn. (IUPAC, 1997), http://www.iupac.org/indexes/books (last accessed 14 April 2011)

3.71 JCGM 100: Evaluation of measurement data – Guide to the expression of uncertainty in measurement (JCGM, 2008) (identical to ISO/IEC Guide 98-3:2008 Uncertainty of measurement – Part 3: Guide to the expression of uncertainty in measurement (GUM:1995)) http://www.bipm.org/en/publications/guides/gum.html (last accessed 14 April 2011)

3.72 International Measurement Evaluation Programme (IMEP) http://irmm.jrc.ec.europa.eu/html/interlaboratory_comparisons/imep/index.htm (last accessed 14 April 2011)

3.73 ISO: *ISO 78-2 Chemistry – Layout for Standards – Methods of Chemical Analysis* (ISO, Geneva 1999)

3.74 BAM Federal Institute for Materials Research and Testing (Berlin): Catalogue of Reference Procedures http://www.bam.de/en/fachthemen/referenzverfahren/index.htm (contact person: M. Hedrich, BAM) (last accessed 14 April 2011)

3.75 BIPM: Key comparison database – Appendix B (Key and supplementary comparisons) http://kcdb.bipm.org/ (contact person: C. Thomas, BIPM) (last accessed 14 April 2011)

3.76 BIPM: Key comparison database – Appendix C (Calibration and measurement capabilities) http://kcdb.bipm.org/ (contact person: C. Thomas, BIPM) (last accessed 14 April 2011)

3.77 Versailles Project on Advanced Materials and Standards (VAMAS): http://www.vamas.org/(chair person: G. Sims, NPL)

3.78 CIPM: 97th Meeting 2008, Report of the CIPM ad hoc Working Group on Materials Metrology (WGMM), CIPM/08-WGMM "Evolving Need for Metrology in Material Property Measurements" http://www.bipm.org/en/committees/cipm/

3.79 S. Bennett, J. Valdés: Materials metrology, Metrologia **47**(2), S1–S193 (2010), Special edition on Materials metrology