CODEX ALIMENTARIUS COMMISSION E

Food and Agriculture Organization of the United Nations

World Health Organization

Viale delle Terme di Caracalla, 00153 Rome, Italy - Tel: (+39) 06 57051 - E-mail: codex@fao.org - www.codexalimentarius.org

Agenda Item 6 **CX/MAS 20/41/8**
February 2020

JOINT FAO/WHO FOOD STANDARDS PROGRAMME

CODEX COMMITTEE ON METHODS OF ANALYSIS AND SAMPLING

41st Session
Budapest, Hungary, 11 - 15 May 2020

INFORMATION DOCUMENT: *GUIDELINES ON MEASUREMENT UNCERTAINTY*

Prepared by Germany

Introduction

1. The 39th Session of the Committee on Methods of Analysis and Sampling (CCMAS39) agreed to start new work on the revision of the *Guidelines on Measurement Uncertainty* (CXG 54 – 2004) and it was noted that an information document containing examples would support the revision of CXG 54 – 2004.[1]

2. CCMAS40 took several decisions on the revised CXG 54-2004 including the transfer of examples to the information document.

3. CCMAS40 noted that the information document was intended to give some examples on the procedures for estimating measurement uncertainty and to provide the user with some references on the general topics. The information document was not considered by CCMAS40 as it had to take into account changes made in the revision of CXG54 at the session. The document would be presented to CCMAS41 consideration.[2]

4. The information document was revised taking into account decisions of CCMAS40 and is presented in Appendix I.

Recommendation

5. The Committee is invited to consider the proposed draft information document (Appendix I).

[1] REP18/MAS, paras 60 - 61
[2] REP19/MAS, paras 65 - 66

Appendix I

Draft Information Document on Procedures for the Estimation of Measurement Uncertainty

1 Introduction

Every measurement is subject to error. A measurement result should thus always be accompanied by information regarding its uncertainty. Such information provides an indication of the quality of the measurement result and allows meaningful comparison to other measurement results or reference values. Without a statement of measurement uncertainty, a measurement result is essentially incomplete and cannot be properly interpreted.

This document provides guidance regarding those sources of uncertainty which originate in the laboratory itself, i.e. in connection with the procedures and conditions starting with the laboratory sample and ending with the measurement result. In particular: the question of sampling uncertainty and the extent to which laboratory samples are representative of the content in the container will not be addressed. Such questions are addressed in CXG 50-2004 [12].

Measurement uncertainty is defined as a parameter "…that characterizes the dispersion of the values which could reasonably be attributed to the measurand", see 2.2.3 in GUM [1]. This document aims to clarify what is meant in this definition and to provide the information which is necessary to understand how different approaches for the evaluation of measurement uncertainty relate to one another. This should allow the reader to make informed decisions regarding the best procedure to adopt in any given case.

Accordingly, the present document provides background information and clarifies basic notions which are central to a correct evaluation and interpretation of measurement uncertainty. First, the top-down and bottom-up approaches are described and compared. Then, the basic model for the top-down approach is presented. This constitutes a convenient framework within which to elucidate some of the basic conceptual aspects of measurement uncertainty. In the course of the discussion, it will become increasingly clear how important it is to understand what is involved in specifying the measurand and due clarifications will be given. The relationship between the top-down and bottom-up approaches will be further clarified on the basis of a more general classification of uncertainty sources. The question of the statistical uncertainty in estimating dispersion parameters – such as standard deviation values – will be addressed; and the effect of the number of observations on this statistical uncertainty will be examined. Specific designs for the evaluation of the different components of the top-down approach will then be provided, including designs for the evaluation of subsampling and matrix effects. Finally, examples will illustrate how measurement uncertainty influences sampling plans.

2 Top-down versus bottom-up approaches

The term "bottom-up approach" is used to denote any approach in which the measurement uncertainty is calculated on the basis of an equation expressing the relationship between input variables and the measurement result. In the phrasing from Section 4.1.1 of the *Guide to the expression of uncertainty in measurement* (GUM) [1]: "In most cases, the measurand Y is not measured directly, but is determined from N other quantities $X_1, X_2, \ldots, X_N$ through a functional relationship f:

$$Y = f(X_1, X_2, \ldots, X_N)$$

It must be emphasized that, in this approach, the measurement result Y is *calculated* from the input variables $X_1, X_2, \ldots, X_N$. Analyte concentration is an example of a measurement result; optical density, peak area and signal height are examples of input variables.

An alternative approach – described e.g. in EURACHEM/CITAC Guide CG4 [2] and in ISO 21748 [3] – consists in making use of available *method validation* data. In the words of Section 7.6.1 in the EURACHEM Guide [2]: "A collaborative study carried out to validate a published method […] is a valuable source of data to support an uncertainty estimate." In this approach, there is no "functional relationship" between input variables and the measurement result. Rather, results are obtained under different measurement conditions, and total observed variation is partitioned into individual components. This approach is often referred to as the *top-down* approach.

In order to obtain measures of precision which can subsequently be used to "support an uncertainty estimate" following the top-down approach, two main types of experiments can be conducted: single-lab (in-house) and multi-lab (collaborative) studies. It must be emphasized that precision measures obtained in these two types of studies are not always comparable. Nonetheless, if relevant uncertainty sources have not been taken into account, it is often expedient to complement the information from a multi-lab study by means of subsequent single-lab experiments.

The main distinction between the two approaches is that whereas the bottom-up approach starts from a physico-chemical consideration of the actual measurement mechanism, the top-down approach starts from a data set in which the variation between different measurement results is directly observable. In this sense, it can be said that the bottom-up approach is *theoretical* whereas the top-down approach is *empirical*.

A related distinction is that, in the bottom-up approach, the starting point is the relationship between the measurement result and input variables, whereas, in the top-down approach, the starting point is the relationship between total variation and individual components of variation.

Finally, another distinction between both approaches is that while the number of components in the top-down approach is usually low[1], the number of input variables in the bottom-up approach can be quite high. For this reason, in the bottom-up approach, it will often be impractical to conduct an experiment in which estimates for the uncertainties associated with all the input variables can be reliably obtained. Indeed, the bottom-up approach explicitly allows the inclusion of *prior information* regarding the size of the errors which can be expected to arise in connection with each source (Type B evaluation).

In the case of the bottom-up approach (and in the case that there are no correlations between the different input variables), the combined (i.e. total) measurement uncertainty – expressed as a standard deviation – is obtained as follows:

$$u_c = \sqrt{\sum_{i=1}^{N} c_i \cdot u_i^2}$$

where u_c denotes the combined uncertainty, u_i denotes the uncertainty associated with input variable i and c_i denotes the corresponding sensitivity coefficient, usually obtained via partial differentiation $\left(c_i = \left(\frac{\partial f}{\partial X_i}\right)^2\right)$, see 5.1.2 and 5.1.3 in GUM [1].

In the case of the top-down approach, the total measurement uncertainty is obtained by summing different variance components, such as between-laboratory variance and repeatability variance. The number of replicate measurements should be taken into consideration. For instance, in the simplest case, the total standard uncertainty is obtained as

$$u = \sqrt{s_L^2 + \frac{s_r^2}{m}}$$

where s_L denotes the between-laboratory standard deviation, s_r denotes the repeatability standard deviation and m denotes the number of replicates whose mean value is taken as the final measurement result. For further information, the reader is referred to ISO 21748 [3].

3 Basic model for the top-down approach

In this section, the basic model for the top-down approach is discussed. The model is premised on the assumption that data from an interlaboratory validation study (also known as a collaborative study) are available. Such a study is conducted in order to characterize the performance of an analytical method. In particular, the characterization of the *precision*[2] of an analytical method can be used "to support an uncertainty estimate". The reader is referred to the ISO 5725 series – in particular to Part 2 [4] – for background information.

The basic model is as follows:

$$Y = \textit{true value} + \textit{method bias (average across labs and matrices)} + \textit{matrix-specific bias} + \textit{laboratory bias} + \textit{repeatability error}$$

For further details, the reader is referred to [5] and [6].

In the following, the individual terms of the basic model are discussed.

[1] The number of components follows directly from the experimental design of the method validation study.

[2] Precision is defined (paraphrasing 2.15 in [7]) as the degree of agreement between independent measurement results obtained under specified conditions. For instance, reproducibility precision characterizes the agreement between results from different laboratories, while repeatability precision characterizes the agreement between results obtained under near-identical conditions in the same laboratory. Precision can be used to derive a measurement uncertainty estimate – but it must not be confused with measurement uncertainty.

True value

In general, the true value is not known. It can be estimated by averaging e.g. across methods, samples and laboratories. However, it is crucial to note that in the GUM [1], measurement uncertainty is defined *without any reference to a true value*; rather, it is defined as a parameter "… that characterizes the dispersion of the values which could reasonably be attributed to the measurand", see 2.2.3 in GUM [1]. This definition has since been adopted in all other relevant standards and guidance documents (EURACHEM [2], VIM [7]). This does not mean that the true value no longer plays a role in the evaluation of measurement uncertainty. However, it is not the (unavailable) difference between true value and measurement result, but *the uncertainty of bias correction* which must be taken into account in the evaluation of measurement uncertainty. In other words, the focus shifts from the (unavailable) true value itself to the uncertainty in the estimation of the bias. Note that if a certified reference value is available along with a reference uncertainty value, the latter can be included in the uncertainty of bias correction.

Method bias (average across labs and matrices)

The method bias across both labs and matrices can be estimated by averaging across laboratories and matrices. As explained in the discussion of the true value, the corresponding contribution to the calculation of measurement uncertainty will consist in the uncertainty in the estimate of this bias.

Matrix-specific bias

In many cases, a method's bias depends on the sample being examined. In other words: bias varies from sample to sample. Such effects occur when the extraction of analyte is affected by the matrix, so that a part of the analyte is not recovered; or when a part of the matrix is extracted along with the analyte and interacts with the measurement's physico-chemical mechanism, resulting in a bias. The corresponding component of total variability is called the matrix standard deviation. It is important to note that all the uncertainty sources listed in Section 7 contribute to this term of the basic model.

Laboratory bias

In many cases, a method's bias depends on the laboratory which is performing the measurement. In other words, the bias varies from laboratory to laboratory. The corresponding component of total variability is called the laboratory standard deviation.

Repeatability error

This term represents variation across replicate measurements (i.e. independent measurements performed under near-identical test conditions).

4 **Specifying the measurand**

The concept "measurand" clearly plays a central role in the definition of measurement uncertainty and will shed further light on the connection between validation data and measurement uncertainty.

Leaving aside the technicalities of the definition of a measurand[3], it is sufficient to note that the specification of a measurand has three separate components:

- specification of a property, e.g. *mean arsenic concentration.* Note that the concept "analyte" corresponds to this part of the specification of the measurand
- specification of a phenomenon, body or substance which the property is associated with, e.g. *a given batch of apple juice*. Note that the concept "matrix", used in the previous section, corresponds to this part of the specification of the measurand
- and specification of a reference framework regarding the manner in which the property is characterized, e.g. [ng/ml]

Loosely phrased, specifying a measurand thus involves stating (1) *what* is to be measured, (2) what is it to be measured *in*, and (3) *how* should the measurement result be expressed in order to ensure comparability to other measurement results or relevant values?

[3] In the VIM [7], measurand is defined (definition 2.3) as "quantity intended to be measured". Quantity, in turn, is defined (definition 1.1) as "property of a phenomenon, body, substance, where the property has a magnitude that can be expressed as a number and a reference". An example given directly under this definition is "amount-of-substance concentration of ethanol in wine sample i". The term "reference" in this definition is explained in NOTE 2 as: "A reference can be a measurement unit, a measurement procedure, a reference material, or a combination of such."

In particular, the specification of the measurand should include information as to whether analyte concentration is to be measured in a laboratory sample or in a “larger sample” or a batch of products in a container. Only in the latter case is *sampling* uncertainty relevant (see Section 7 for an overview of the different sources of uncertainty). Similarly, if measurement results from several laboratory samples are used to assess the conformity of bulk material from a container, it is the measurement uncertainty of the mean value across the results corresponding to the individual laboratory samples which is relevant.

More generally, while measurement uncertainty is always determined on the basis of the laboratory sample, it is nevertheless important to include all available information about the laboratory sample in the evaluation of measurement uncertainty, e.g.

- Where does the material come from (e.g. container)?
- Have other samples from the same origin been tested?
- What is the intended use of the measurement result (e.g. conformity assessment for the individual laboratory sample or for the container)?

For example, determining the contribution to uncertainty which arises from the material’s heterogeneity (e.g. fundamental variability, see Section 9.4) may require a considerable amount of work, depending on the analyte, concentration and grain/particle size. If the origin of the material is known, it may be possible to use previously obtained results regarding the heterogeneity contribution to uncertainty instead of obtaining a new estimate from scratch.

The specification of the measurand should also make it possible to determine whether bias/recovery correction is required, and what form this correction should take. For example, if the measurand is specified in terms of the amount of analyte recovered, then recovery correction may not be appropriate. On the other hand, if the measurand is specified in terms of the total amount of analyte present in a test sample, then recovery correction may be necessary.

Finally, it may be impractical or impossible to provide an exhaustive specification of the measurand. For this reason, it may be necessary to include an extra component of measurement uncertainty, called “definitional uncertainty” (see definition 2.27 in VIM [7]), in order to account for any ambiguity (“finite amount of detail”) in the specification of the measurand. However, in most cases, the definitional uncertainty can be considered negligible.

5 Relation between measurand and validation data

If the results of a validation study are to be used to determine measurement uncertainty, it must be ensured that the study refers to the same measurand.

Example 1: Measurement uncertainty is being evaluated in a given laboratory for a measurand specified in terms of analyte concentration in test samples. The analytical method used has been validated for the same analyte, but on the basis of extracts rather than test samples. In other words, the measurand for the validation study is analyte concentration in extracts. It follows that the measurand for which measurement uncertainty must be evaluated is different from the measurand from the validation study. Accordingly, the measurement uncertainty cannot be evaluated on the basis of the characterization of the dispersion of measurement results from the validation study.

Example 2: Measurement uncertainty is being evaluated in a given laboratory for a measurand which is specified in terms of a range of matrices. The analytical method used has been validated for the same analyte, but for only one of the matrices. It follows that the measurand for which measurement uncertainty must be evaluated is different from the measurand from the validation study. Accordingly, the measurement uncertainty cannot be evaluated on the basis of the characterization of the dispersion of measurement results from the validation study (the matrix bias term is missing).

The conditions under which validation data can be used to support a measurement uncertainty estimate can be stated as follows:

If…

the measurement result is obtained using a validated method

and the *measurand* is included in the scope of the validation

and precision within the laboratory which is evaluating measurement uncertainty is comparable to the method's precision as characterized in the validation study

then…

→ the precision estimates from the validation study can be used in the calculation of measurement uncertainty.

The reader is referred to Section 7 in EURACHEM [2] for further guidance regarding using validation data in the evaluation of measurement uncertainty.

6 Empirical versus rational methods

In the definition of the measurand, the specification of the property must include sufficient information to allow an appropriate reference (see 1.1 in the VIM [7]) to be selected. In particular, it is important to distinguish between

- Empirical method (type I methods in the CODEX system)
- Rational method (type II-IV methods in the CODEX system)

In Section 5.4 of EURACHEM [2], the following explanation is provided: *"In analytical measurement, it is particularly important to distinguish between measurements intended to produce results which are independent of the method used, and those which are not so intended. The latter are often referred to as empirical methods or operationally defined methods."*

In Section 5.5 of the same document, it is explained that non-empirical methods are sometimes called rational methods. This distinction is closely related to that between *operationally defined* and *non-operationally defined* measurands found in Section 9.2.3 of ISO Guide 35 [8]. The reader is also referred to Section 3.1 in the EURACHEM Guide to Metrological Traceability in Chemical Measurement [20].

As far as the evaluation of measurement uncertainty is concerned, this distinction has the following important implication: for *empirical* methods (*operationally defined* measurands), there is no method bias term in the basic model for the top-down approach described in Section 3. (Please note that the bottom-up approach does not allow the distinction *method* versus *other* bias components).

7 Uncertainty sources in the top-down and bottom-up approaches

In the *top-down* approach, total variation observed in a data set is partitioned into different components. In the *bottom-up* approach, the total uncertainty is obtained from uncertainty values associated with individual input variables. The following question arises: what is the *relationship* between the components from a top-down model and the uncertainty sources included in a bottom-up model?

In order to answer this question, an overview of different types of uncertainty sources – *independently of the approach* – is now provided. The intention is to distinguish broad categories of uncertainty sources. Apart from shedding further light on the relationship between the top-down and bottom-up approaches, this overview may prove useful for determining which sources may be relevant in any given case, and whether all relevant sources have been included in the evaluation of measurement uncertainty.

Sources of uncertainty are conveniently classified under six main headings:

- Sampling (The question of sampling uncertainty is not addressed in the present document. The reader is referred to CXG 50-2004 [12])
- Storage/transportation
- Subsampling
- Measurement conditions
- Measurement procedure
- Computational effects

Source of uncertainty	Role in measurement uncertainty
Sampling	If the measurand is defined in terms of e.g. analyte concentration in a container or in a batch of products, then sampling is required, and its contribution to measurement uncertainty must be assessed, see Section 7.6 in ISO 17025 [9]. If the measurand is defined in terms of a single test material (laboratory sample), then there is no contribution to uncertainty due to sampling. There may be a contribution from subsampling, however (i.e. obtaining test portions from the laboratory sample). *Fundamental variability* is one of the "subcomponents" of sampling uncertainty, see the discussion in Section 9.4.
Storage/transportation	If different storage or shipping conditions have an effect on measurement results, then the corresponding contribution to the total uncertainty must be taken into account.
Subsampling	This term denotes taking test portions from the laboratory sample. If the latter is not homogeneous (finely ground in case of solid matter, mixed or agitated in case of liquids and semi-solids), then it cannot be ensured that the subsampling uncertainty is negligible. Accordingly, appropriate homogenisation is required before subsampling in order to reduce this uncertainty source. *Fundamental variability* is one of the "subcomponents" of subsampling uncertainty, see the discussion in Section 9.4.
Measurement conditions	It must be emphasized that the term measurement as used here includes any sample preparation and clean-up procedures. If different measurement conditions (e.g. different time of year, different technician, different reagents, different equipment) contribute to measurement uncertainty, this source must be taken into consideration.
Measurement procedure	This term denotes the intrinsic or irreducible uncertainty component associated with the physical/chemical/biochemical mechanisms involved in the measurement procedure (including sample preparation and clean-up procedures), e.g. extraction efficiency. The input variables in the bottom-up approach can be considered to belong under this heading.
Computational effects	Inaccurate calibration model and calculation methods, peak integration procedures and rounding will also contribute to measurement uncertainty.

8 Requirements regarding data size

If a standard deviation is calculated on the basis of a series of measurement results, how well does it characterize the actual dispersion of the values? Indeed, if several measurement series are performed and a separate standard deviation value is calculated for each, these standard deviation values will differ. In other words, a given standard deviation, obtained on the basis of empirical data, only represents an *estimate* of the "true" standard deviation. Just as in the case of the measurement uncertainty of a measurement result, the uncertainty of a given standard deviation value can be characterized in terms of a confidence interval. Table 3 in CXG 59 [10] provides confidence intervals for standard deviation values calculated from empirical data for different values of N (number of observations). For instance, with $N = 5$ values, the confidence interval for the standard deviation is $[0.35 \cdot s, 1.67 \cdot s]$, where s denotes the standard deviation calculated on the basis of the available data. With $N = 7$ values, the confidence interval for the standard deviation is $[0.45 \cdot s, 1.55 \cdot s]$, which is still very large.

Accordingly, it is recommended that standard deviations be computed on the basis of a minimum of $N = 12$ values (corresponding to 11 degrees of freedom for the estimation of the standard deviation), in which case the confidence interval for the standard deviation is $[0.59 \cdot s, 1.41 \cdot s]$.

As far as the simultaneous estimation of e.g. between-laboratory (or between-matrix) standard deviation and repeatability standard deviation is concerned, this recommendation means that measurement results from at least 12 laboratories (or matrices) should be available, each with at least two replicates per laboratory (or matrix).

It is required that data from at least 8 laboratories must be available (see Section 6.3.4 in ISO 5725-1 [17] where 8-15 laboratories is proposed as a "common" figure).

In the case that different uncertainty sources are *simultaneously* taken into consideration, say in the bottom-up approach, the requirement regarding data size can be applied via the Satterthwaite formula. More specifically: take the case that 2 different uncertainty sources are included in the calculation of the combined uncertainty, u_1 and u_2. Say that each was obtained by applying the formula for the sample standard deviation on the basis of n_1 and n_2 measurement results, respectively. The number of degrees of freedom for the combined uncertainty can then be computed as

$$\text{Degrees of freedom for combined uncertainty} = \frac{(u_1^2/n_1 + u_2^2/n_2)^2}{\frac{(u_1^2/n_1)^2}{n_1 - 1} + \frac{(u_2^2/n_2)^2}{n_2 - 1}}$$

The recommendation is to ensure a minimum of 11 degrees of freedom for the combined uncertainty.

In the case that prior information is used for an individual u_i value (Type B variable) and that no information regarding data size is available, it is suggested to use $n_i = 7$; the approximate ± 50 % uncertainty which corresponds to this data size is intended to reflect the fact that, in the case of Type B variables, distributional assumptions are often based on "educated guesses".

Example of the application of the Satterthwaite formula

Take the case that measurement uncertainty must be evaluated on the basis of the following functional relationship, where the measurement result Y is expressed as a function of 4 input variables:

$$Y = f(X_1, X_2, X_3, X_4) = X_1 + X_2 + X_3 + X_4$$

Table 1: Data size and uncertainty values for the input variables

Input variable	Type	n	u^2
X_1	A	3	4
X_2	B	30	15

X_3	B	30	15
X_4	B	Not available Take $n_4 = 7$	5

The Satterthwaite formula can now be applied.
Degrees of freedom for combined uncertainty

$$= \frac{(u_1^2/n_1 + u_2^2/n_2 + u_3^2/n_3 + u_4^2/n_4)^2}{\frac{(u_1^2/n_1)^2}{n_1 - 1} + \frac{(u_2^2/n_2)^2}{n_2 - 1} + \frac{(u_3^2/n_3)^2}{n_3 - 1} + \frac{(u_4^2/n_4)^2}{n_4 - 1}}$$

$$= 9.4$$

9 Simple procedures for evaluating uncertainty components

If validation data are incomplete (i.e. some of the relevant sources of uncertainty have not been characterized), further experiments must be conducted before the top-down approach can be applied.

For instance, in a collaborative study, each participating laboratory should ideally receive samples representing different matrices and different analyte concentrations. However, due to restrictions in material availability, collaborative studies are often conducted on the basis of a single sample per participant. In such a case, almost no conclusions can be drawn regarding the impact of matrix effects. Accordingly, the characterization of the matrix-specific bias term from the basic model must often be performed in a separate experiment.

In the following, simple procedures are described for characterizing different components of variation – such as the matrix-specific bias.

More sophisticated procedures for simultaneously estimating several components of variation are provided in [11]. The reader is also referred to CD ISO 5725-3 [18] and DTS 23471 [19].

9.1 Procedure for characterizing in-house variation

If the analytical method is an in-house method, then an in-house (single-lab) validation study is conducted. If validation data are incomplete or unavailable, in-house components of variation can be characterized on the basis of a further experiment (or QC data, as long as such data are available and have an appropriate structure).

Total in-house variation is called intermediate precision and should reflect all relevant uncertainty sources except matrix bias[4] – in particular, variation arising from different measurement conditions (i.e. operator, reagent batch, etc.) within the laboratory, along with repeatability.

The structure of the experimental or QC data must allow the distinction between in-house repeatability conditions and intermediate conditions (different day, different technician, different reagent batch, etc.). The uncertainty can then be calculated as follows:

$$u = \sqrt{s_I^2 - s_{r,inhouse}^2 + \frac{s_{r,inhouse}^2}{k}}$$

where s_I denotes the intermediate standard deviation, $s_{r,inhouse}$ denotes the repeatability estimate and k denotes the number of replicates whose mean value is taken as the final measurement result.

As explained in Section 8, it is recommended that, at a minimum, $N = 12$ different in-house measurement conditions (e.g. different days) be represented in the data set.

[4] By definition, intermediate precision does not include matrix bias, see 2.22 in VIM [7]. If matrix bias is included, then the term in-house reproducibility is used.

In the following example, we take the case that QC data are available for 20 different days. (If appropriate QC data are not available and a further experiment is required, $N = 12$ days are sufficient).

Table 2: In-house QC data for the calculation of intermediate (in-house) and repeatability standard deviation values

	Result 1	Result 2
Day 1	10.72	12.29
Day 2	4.56	0.90
Day 3	8.79	9.75
Day 4	10.08	6.51
Day 5	12.29	11.32
Day 6	7.95	6.79
Day 7	13.06	14.54
Day 8	11.23	12.09
Day 9	7.31	9.51
Day 10	5.85	5.08
Day 11	7.48	9.12
Day 12	12.59	10.65
Day 13	7.55	6.59
Day 14	12.05	11.15
Day 15	4.86	6.48
Day 16	6.99	7.10
Day 17	7.40	6.75
Day 18	8.85	11.15
Day 19	11.93	10.17
Day 20	8.50	8.29

The between-day and repeatability standard deviation values are calculated as follows.
First we introduce the following notation: the days are indexed $i = 1, \ldots, m$ (in this example, $m = 20$); the replicates within each day are indexed $j = 1, n$ (in this example, $n = 2$); and the individual measurement results are denoted x_{ij}.
First, compute the overall mean value $\bar{x}$, and the day-specific mean values $\bar{x}_i$. Then compute the between-day sum of squares:

$$SSB = n \cdot \sum_{i=1}^{m} (\bar{x}_i - \bar{x})^2$$

and the within-day sum of squares:

$$SSW = \sum_{i=1}^{m} \sum_{j=1}^{n} (x_{ij} - \bar{x}_i)^2$$

The in-house repeatability standard deviation $s_{r,inhouse}$ is then obtained as

$$s_{r,inhouse} = \sqrt{\frac{SSW}{m \cdot (n-1)}}$$

and the between-day standard deviation s_D is obtained as

$$s_D = \sqrt{\frac{1}{n}\left(\frac{SSB}{m-1} - s_{r,inhouse}^2\right)}.$$

(If the value under the square root sign is negative, then $s_D = 0$.)
Finally, the intermediate (in-house) standard deviation is calculated as:

$$s_I = \sqrt{s_D^2 + s_{r,inhouse}^2}.$$

For the data from Table 2, the calculation results are as follows:

Table 3: Calculation of SSB and SSW on the basis of in-house QC data

Overall mean value $\bar{\bar{x}}$	Day-specific mean values $\bar{x}_i$	Differences $\bar{x}_i - \bar{\bar{x}}$	SSB	Differences $x_{ij} - \bar{x}_i$	Differences $x_{ij} - \bar{x}_i$	SSW
8.91	11.51	2.60	283.05	-0.79	0.79	29.95
	2.73	-6.18		1.83	-1.83	
	9.27	0.36		-0.48	0.48	
	8.29	-0.61		1.79	-1.79	
	11.80	2.90		0.49	-0.49	
	7.37	-1.54		0.58	-0.58	
	13.80	4.90		-0.74	0.74	
	11.66	2.75		-0.43	0.43	
	8.41	-0.50		-1.10	1.10	
	5.46	-3.44		0.39	-0.39	
	8.30	-0.61		-0.82	0.82	
	11.62	2.72		0.97	-0.97	
	7.07	-1.83		0.48	-0.48	
	11.60	2.69		0.45	-0.45	
	5.67	-3.24		-0.81	0.81	
	7.05	-1.86		-0.06	0.06	
	7.08	-1.83		0.32	-0.32	
	10.00	1.09		-1.15	1.15	
	11.05	2.14		0.88	-0.88	
	8.40	-0.51		0.10	-0.10	

The following precision estimates are obtained:

Table 4: Precision estimates obtained from in-house QC data

$s_{r,inhouse}$	s_D	s_I
1.22	2.59	2.86

9.2 Procedures for characterizing variation across matrices

In this section it is assumed that heterogeneity between laboratory samples is negligible, and that the measurand is specified in terms of a number of matrices, from which N matrices are selected[5]. Selection should be based on the method's intended use/scope. As explained in Section 8, it is recommended that, at a minimum, $N = 12$ matrices be included.

A simple approach for characterizing variation across matrices consists in spiking the N matrices and obtaining duplicate measurement results in a single laboratory for each matrix. In this manner, variation between the matrices (matrix-specific bias) can be distinguished from variation within each matrix (repeatability error). In this procedure, the matrix is modelled as a random effect, and the result is a standard deviation characterizing variation across all the matrices included in the specification of the measurand.

[5] For instance, a number of different apple types, or a number of different cattle breeds.

Example

Table 5: Data from an experiment for the calculation of the matrix bias

	MV1	MV2
Matrix 1	114.51	112.24
Matrix 2	120.25	111.59
Matrix 3	88.46	86.62
Matrix 4	118.93	102.35
Matrix 5	74.06	80.91
Matrix 6	117.50	102.69
Matrix 7	120.96	109.35
Matrix 8	96.05	92.92
Matrix 9	98.43	87.09
Matrix 10	107.99	117.42
Matrix 11	117.34	126.87
Matrix 12	76.56	109.79

Applying the same calculation procedure as in Section 9.1, the following precision estimates are obtained:

Table 6: Precision estimates for the calculation of matrix bias

s_r	s_{matrix}
9.53	12.24

*9.3 **Procedures for characterizing between-laboratory variation***

Procedure 1: Conduct an interlaboratory validation study with a minimum of $N = 12$ laboratories and with duplicate measurement results within each laboratory. It is necessary to ensure that heterogeneity between laboratory samples is negligible. In this manner, variation between the laboratories (lab bias) can be distinguished from variation within the laboratories (repeatability error). In this procedure, the laboratory is modelled as a random effect, and the result is a standard deviation characterizing variation across laboratories.

Example

Table 7: Data from an experiment for the calculation of the lab bias

	MV1	MV2
Lab 1	0.981	1.238
Lab 2	0.182	0.601
Lab 3	1.107	0.994
Lab 4	1.471	1.532
Lab 5	1.169	0.674
Lab 6	0.491	1.271
Lab 7	1.717	0.970
Lab 8	0.931	1.171
Lab 9	1.017	1.248
Lab 10	0.909	0.723
Lab 11	0.812	1.312
Lab 12	1.375	1.719

Applying the same calculation procedure as in Section 9.1, the following precision estimates are obtained:

Table 8: Precision estimates for the calculation of lab bias

s_r	s_{lab}
0.30	0.23

Procedure 2: If PT data are available, and a sufficient number of participants (ideally, at least 12) have used the same method – then these data can be used to characterize variation across laboratories. In order to ensure neutral data evaluation and avoid conflicts of interest, the data should come from PT schemes run by competent authorities.

9.4 Procedures for characterizing fundamental variability

Fundamental variability is a subcomponent of the repeatability error term from the basic model in Section 3 and denotes the irreducible variation between samples which remains even under the highest achievable degree of homogeneity. Fundamental variability reflects heterogeneity at the level of the sample's constituent particles; it has an influence on the uncertainty of measurement results when the target analyte is located on sparsely distributed carrier particles. Fundamental variability appears twice: first, during sampling, and second, during subsampling in the laboratory, i.e. extraction of a test portion after homogenization of the laboratory sample. In practice, nonnegligible fundamental variability can be reduced by modifying the testing procedure in two respects: first, by finer grinding or comminuting or mixing of the test material, and second, by increasing the test portion size.

It should be noted that, while a correct partitioning of observed variability between sampling, subsampling and other uncertainty components is achievable in theory, doing so is difficult in practice *when the fundamental variability is significant*. Take the case that several laboratory samples arecollected from the container and assume that the number of carrier particles in the laboratory samples varies randomly between 0 and 10. The fundamental variability between subsamples (test portions) will thus depend on which laboratory sample they were collected from. In such a situation, a correct characterization of fundamental variability would be quite involved. It would be much more efficient to ensure variation regarding carrier particle numbers between laboratory samples were negligible – in other words, to ensure that every single laboratory sample were representative of the container or batch of products, thus eliminating the sampling fundamental variability from the equation. Often, this may be achieved by increasing laboratory sample size; but a more general point is that a correct evaluation of fundamental variability requires an appropriate inclusion of the sampling step, i.e. a consideration of the different steps from sampling to analysis as one single process[6].

The question thus arises: how can we decide whether fundamental variability is significant? Fundamental variability cannot be characterized by means of classical homogeneity studies such as the standard designs described in ISO 13528 [21] and Guide 35 [8]. Indeed, in these designs, it is not possible to distinguish fundamental variability from sample heterogeneity *per se*, so that the former may be mistaken for the latter.

The following procedure, originally proposed in Uhlig (2020) **Error! Reference source not found.**, allows a characterization of fundamental variability.

[6] Consider the following hypothetical example: a 5 t container contains one single carrier particle with a content of 5 mg, translating to 1 µg/kg analyte average concentration in the container. A 5 kg laboratory sample is collected from the container. Thus, with 99.9 % probability, the laboratory sample will contain no carrier particle, and there will be no fundamental variability in the subsampling step. However, with 0.1 % probability, the laboratory sample will contain the single carrier particle. In such a case, if a 500 g test portion is taken from the laboratory sample, then the analyte concentration in the test portion will be either 0 mg/kg (nine times out of ten) or 10 mg/kg (one time out of ten). This corresponds to a fundamental standard deviation of 3mg/kg for the subsampling step – whereas the actual fundamental standard deviation for the complete sampling + subsampling step is 0.1 mg/kg only. This results from the fact that the analyte concentration in the test portion is either 0 mg/kg (with 99,99 % probability) or 10 mg/kg (with 0,01 % probability).

This example shows how restricting the calculation of fundamental variability to the subsampling step can lead to gross misestimation.

Step 1

Check whether one of the following criteria are met:

Criterion 1: The in-house repeatability standard deviation is larger than 3 times the expected value.

Criterion 2: The in-house repeatability standard deviation is larger than the Horwitz SD value.

Criterion 3: Conspicuous “upper” outliers are present in QC data. For instance, in the QC data provided in Table 2 (Section 9.1), the Day 7 value of 14.54 could be considered such an “upper” outlier. The presence of such outliers constitutes a further indication that the unexpectedly large observed variability may be due to fundamental variability.

If at least one of these criteria is met, proceed to Step 2.

Step 2

Conduct the following experiment:

1. Obtain 20 test results under repeatability conditions. Calculate the corresponding variance s_1^2.
2. Increase test portion size by a factor k (e.g. triple test portion size, $k = 3$). If it is not possible or practical to increase test portion size, grinding and homogenizing a volume corresponding to a k-fold increase in test portion size prior to taking a test portion with the original size is another option.
3. Obtain 20 test results under repeatability conditions on the basis of the finely ground test material / increased test portion size. Calculate the corresponding variance s_2^2.
4. If the ratio $\frac{s_1^2}{s_2^2}$ is greater than 2.17, then calculate the SD characterizing fundamental variability as follows:

$$s_F = \sqrt{\frac{k}{(k-1)} \cdot (s_1^2 - s_2^2)}$$

Example

Table 9: Data from an experiment for the calculation of fundamental variability

	Experiment 1: Original test portion size	Experiment 2: Test portion size is tripled
Sample 1	14.0	15.1
Sample 2	11.9	13.8
Sample 3	10.5	11.8
Sample 4	14.9	14.0
Sample 5	13.1	11.4
Sample 6	9.5	15.7
Sample 7	15.6	12.4
Sample 8	18.3	11.5
Sample 9	12.5	12.1
Sample 10	16.4	13.7
Sample 11	18.0	15.8
Sample 12	14.0	12.5
Sample	13.0	12.8

13		
Sample 14	20.8	15.1
Sample 15	10.2	11.8
Sample 16	21.5	10.6
Sample 17	13.9	11.1
Sample 18	17.8	12.9
Sample 19	7.7	11.4
Sample 20	12.2	16.3

Note that, in Experiment 1, several conspicuously large values are obtained – an indication that fundamental variability is non-negligible.

The following variances and corresponding ratio are obtained:

Table 10: Variances and their ratio

s_1^2	s_2^2	s_1^2/s_2^2
13.54	3.05	4.44

As can be seen, the ratio s_1^2/s_2^2 is greater than the value 2.17. Accordingly, the fundamental variability is calculated as

$$s_F = \sqrt{\frac{3}{2} \cdot (s_1^2 - s_2^2)} = 3.97.$$

10 Influence of measurement uncertainty on sampling plans: examples

In the *General guidelines on sampling* [12], it is stated that "Codex Methods of Sampling are designed to ensure that fair and valid sampling procedures are used when food is being tested for compliance with a particular Codex commodity standard". Sample size and acceptance number / acceptability constant for inspection by attributes / variables are determined on the basis of procedures and sampling plans described in ISO standards and/or CODEX guidelines. While measurement uncertainty may be considered irrelevant for inspection by attributes, its impact on inspection by variables must be accounted for.

In the introduction to ISO 3951-1:2013, it is stated that "[i]t is assumed in the body of this part of ISO 3951 that measurement error is negligible […]". Nonetheless, procedures for increasing the sample size are provided in Annex B of ISO 3951-1 [13] and Annex P of ISO 3951-2 [14] for the case that measurement uncertainty is non-negligible. It is important to note that these procedures are only applicable if "the measurement method is unbiased, i.e. the expected value of the measurement error is zero" (see Annex P.1 in ISO 3951-2:2013 [14]). In such a case, total variability is expressed as

$$\sigma_{total} = \sqrt{\sigma^2 + \sigma_m^2}$$

where σ denotes the process standard deviation and σ_m denotes the measurement standard deviation.

If σ_m is non-negligible (i.e. greater than one tenth of the sampling standard deviation s or process standard deviation σ), the sample size n must be increased to either $n^* = n \cdot (1 + \gamma^2)$ where $\gamma = \sigma_m/\sigma$ (the process standard deviation σ is known) or $n^* = n \cdot (1 + \tilde{\gamma}^2)$ where $\tilde{\gamma}$ is an estimated upper bound of $\gamma = \sigma_m/\sigma$ (the process standard deviation σ is unknown). The acceptability

constant k remains unchanged. For further details, see Annex P in ISO 3951-2:2013 [14].

Example

A lot of 500 items of pre-packaged mineral water is assessed for sodium content. If the measurement uncertainty is not taken into consideration, for an agreed AQL of 2.5 % (maximum concentration 200 mg/L), general inspection level II (default level) a sample of 30 items should be collected for assessment, (ISO 3951-2 [14], Annex A, Table A1 and Annex B, Table B1). The production is well under control and the control charts give a process standard deviation σ of 2 mg/L. The measurement uncertainty standard deviation σ_m is 1 mg/L and is thus non-negligible. With $\gamma = \sigma_m/\sigma = 0.5$ and $1 + \gamma^2 = 1.25$ the sample size must be increased to 38.

If there is a bias, the above procedure must be modified. One possibility would be to proceed as follows[7]. The standard deviation of $\bar{x}$, the mean across the n measurement results, is expressed as

$$\sigma_{\bar{x}} = \sqrt{\frac{\sigma^2 + \sigma_0^2}{n} + \sigma_b^2}$$

where σ denotes the process standard deviation, σ_0 denotes the repeatability component of measurement uncertainty (calculated on the basis of the n items sampled from the lot), and σ_b represents available information (e.g. the between-lab standard deviation from a method validation study) used to estimate the bias term.

The modified procedure is as follows:

1. Increase the sample size under the assumption that there is no measurement error
2. Calculate $d = \frac{1}{n} - \frac{\sigma_b^2}{\sigma^2}$
3. If $d \leq 0$, inflated variability due to a bias cannot be compensated for via an increase in sample size.
4. If $d \leq \frac{1}{2n}$, bias compensation via an increase in sample size may not be appropriate due to the large number of samples required. It is then suggested to reduce bias or to use another measurement method.
5. If $d > \frac{1}{2n}$, calculate the new sample size as $n^* = \frac{1+\frac{\sigma_0^2}{\sigma^2}}{d} = \frac{\sigma^2+\sigma_0^2}{\frac{\sigma^2}{n}-\sigma_b^2}$

Example (continued from previous example)

It is now assumed that there is a method bias and that a σ_b estimate of 0.2 mg/L is available. Accordingly, on the basis of the previously calculated value of $n = 38$, d is calculated as $d = 0.016$. Since $d > \frac{1}{2n} = 0.013$, the new sample size is calculated as $n^* = 77$ (with $\sigma_0 = \sigma_m = 1$ mg/L).

Procedures for bulk sampling are provided in ISO 10725:2000 [16]. As in the case of sampling from packages, these procedures are only valid under the assumption that there is no method bias. Modified procedures for the case that there is a method bias are currently being developed. For now, the discussion is limited to the case that there is no bias.

A *dominant* measurement uncertainty has an effect on the number of test samples per composite sample n_T as well as the number of measurements per test sample n_M. The measurement uncertainty is dominant when both the standard deviation of the sampling increment σ_I and the standard deviation between test samples σ_P are far less (one tenth or less) than the measurement standard deviation σ_M (i.e. the measurement uncertainty), which must be known and stable, see Annex B in ISO 10725 [16]. The number of sample increments per composite sample n_I remains unchanged, no matter whether the measurement uncertainty is dominant or not. The mass of the increments should be sufficiently large to offset the fundamental variability.

[7] This modified procedure is taken from current stage of development of Annex B of ISO/WD ISO 3951-6 [15].

Example

A lot of wheat bulk material is to be assessed for cadmium content (maximum concentration e.g. 0.1 mg/kg). In this example, it is assumed that cadmium concentrations in the lot are homogeneous, resulting in very low standard deviations σ_I and σ_P, estimated as 0.0015 mg/kg and 0.002 mg/kg, respectively. Since the concentrations are very low, a relatively high measurement uncertainty $\sigma_M = 0.025$ mg/kg is obtained. The discrimination interval D (difference between agreed risk-based acceptance and rejection levels) is 0.02 mg/kg. The measurement standard deviation $\sigma_M = 0.025$ mg/kg is thus dominant (d_I is calculated as 0.075). The number of increments per composite sample is $n_I = 6$, the number of test samples per composite sample is $n_T = 2$ and the number of measurements per test sample is $n_M = 2$ (yielding a product $n_T \cdot n_M = 4$, which can be interpreted as a measure of the analytical workload). The combined overall standard deviation σ_0 is calculated as $\sqrt{\frac{n_T \cdot n_M}{n_I}{\sigma_I}^2 + n_M{\sigma_P}^2 + {\sigma_M}^2} \approx 0.03$ mg/kg and divided by the discrimination interval D in order to obtain the relative standard deviation $d_0 = \sigma_0 / D \approx 1.26$. By means of Table B1 in Annex B of ISO 10725 [16], this relative standard deviation d_0 is used to determine the adjusted number of test samples per composite sample $n_T = 2$ (i.e. n_T remains the same) as well as the adjusted number of measurements per test sample $n_M = 3$, yielding a product $n_T \cdot n_M = 6$.

References

[1] Evaluation of measurement data — Guide to the expression of uncertainty in measurement, JCGM 100:2008.

[2] S L R Ellison and A Williams (eds.) EURACHEM/CITAC Guide CG4: Quantifying Uncertainty in Analytical Measurement, Third Edition, QUAM:2012.P1.

[3] ISO 21748:2010, Guidance for the use of repeatability, reproducibility and trueness estimates in measurement uncertainty estimation.

[4] ISO 5725-2:1994, Accuracy (trueness and precision) of measurement methods and results — Part 2: Basic method for the determination of repeatability and reproducibility of a standard measurement method.

[5] B Jülicher, Petra Gowik and Steffen Uhlig (1998) Assessment of detection methods in trace analysis by means of a statistically based in-house validation concept, The Analyst.

[6] B Jülicher, Petra Gowik and Steffen Uhlig (1998) a top-down in-house validation based approach for the investigation of the measurement uncertainty using fractional factorial experiments, The Analyst.

[7] International vocabulary of metrology — Basic and general concepts and associated terms (VIM), JCGM 200:2012.

[8] ISO Guide 35, Fourth edition (2017), Reference materials — Guidance for characterization and assessment of homogeneity and stability.

[9] ISO/IEC 17025:2017, General requirements for the competence of testing and calibration laboratories.

[10] CXG 59-2006, Guidelines on estimation of uncertainty of results.

[11] S Uhlig and P Gowik (2018) Efficient estimation of interlaboratory and in-house reproducibility standard deviation in factorial validation studies, Journal of Consumer Protection and Food Safety.

[12] CXG 50-2004, General guidelines on sampling.

[13] ISO 3951-1:2016, Sampling procedures for inspection by variables — Part 1: Specification of single sampling plans indexed by acceptance quality limit (AQL) for lot-by-lot inspection for a single quality characteristic and a single AQL.

[14] ISO 3951-2:2013, Sampling procedures for inspection by variables — Part 2: General specification for single sampling plans indexed by acceptance quality limit (AQL) for lot-by-lot inspection of independent quality characteristics.

[15] ISO/WD 3951-6:2019, Sampling procedures for inspection by variables — Part 6: Specification for single sampling plans indexed by limiting quality (LQ).
[16] ISO 10725:2000, Acceptance sampling plans and procedures for the inspection of bulk materials.
[17] ISO 5725-1:1994, Accuracy (trueness and precision) of measurement methods and results — Part 1: General principles and definitions.
[18] ISO 5725-3:1994, Accuracy (trueness and precision) of measurement methods and results — Part 3: Intermediate measures of the precision of a standard measurement method.
[19] DTS 23471, Experimental designs for the evaluation of uncertainty – Use of factorial designs for determining uncertainty functions.
[20] S L R Ellison and A Williams (eds.) EURACHEM/CITAC Guide: Metrological Traceability in Chemical Measurement (Second Edition 2019).
[21] ISO 13528:2015, Statistical methods for use in proficiency testing by interlaboratory comparison.
[22] S Uhlig, B Colson and P Gowik (2020) A procedure for estimating fundamental variability, submitted for publication.