TECHNICAL SPECIFICATION

ISO/TS 21749

First edition
2005-02-15

Corrected version
2005-07-15

Measurement uncertainty for metrological applications — Repeated measurements and nested experiments

Incertitude de mesure pour les applications en métrologie — Mesures répétées et expériences emboîtées

Reference number
ISO/TS 21749:2005(E)

ISO copyright office
Case postale 56 • CH-1211 Geneva 20
Tel. + 41 22 749 01 11
Fax + 41 22 749 09 47
E-mail copyright@iso.org
Web www.iso.org

Published in Switzerland

Contents

Page

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

International Standards are drafted in accordance with the rules given in the ISO/IEC Directives, Part 2.

The main task of technical committees is to prepare International Standards. Draft International Standards adopted by the technical committees are circulated to the member bodies for voting. Publication as an International Standard requires approval by at least 75 % of the member bodies casting a vote.

In other circumstances, particularly when there is an urgent market requirement for such documents, a technical committee may decide to publish other types of normative document:

— an ISO Publicly Available Specification (ISO/PAS) represents an agreement between technical experts in an ISO working group and is accepted for publication if it is approved by more than 50 % of the members of the parent committee casting a vote;

— an ISO Technical Specification (ISO/TS) represents an agreement between the members of a technical committee and is accepted for publication if it is approved by 2/3 of the members of the committee casting a vote.

An ISO/PAS or ISO/TS is reviewed after three years in order to decide whether it will be confirmed for a further three years, revised to become an International Standard, or withdrawn. If the ISO/PAS or ISO/TS is confirmed, it is reviewed again after a further three years, at which time it must either be transformed into an International Standard or be withdrawn.

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO shall not be held responsible for identifying any or all such patent rights.

ISO/TS 21749 was prepared by Technical Committee ISO/TC 69, *Applications of statistical methods*, Subcommittee SC 6, *Measurement methods and results*.

This corrected version of ISO/TS 21749:2005 incorporates the correction of the title.

Introduction

Test, calibration and other laboratories are frequently required to report the results of measurements and the associated uncertainties. Evaluation of uncertainty is an on-going process that can consume time and resources. In particular, there are many tests and other operations carried out by laboratories where two or three sources of uncertainty are involved. Following the approach in the *Guide to the expression of uncertainty of measurement* (*GUM*) to combining components of uncertainty, this document focuses on using the analysis of variance (ANOVA) for estimating individual components, particularly those based on Type A (statistical) evaluations.

An experiment is designed by the laboratory to enable an adequate number of measurements to be made, the analysis of which will permit the separation of the uncertainty components. The experiment, in terms of design and execution, and the subsequent analysis and uncertainty evaluation, require familiarity with data analysis techniques, particularly statistical analysis. Therefore, it is important for laboratory personnel to be aware of the resources required and to plan the necessary data collection and analysis.

In this Technical Specification, the uncertainty components based on Type A evaluations can be estimated from statistical analysis of repeated measurements, from instruments, test items or check standards.

A purpose of this Technical Specification is to provide guidance on the evaluation of the uncertainties associated with the measurement of test items, for instance as part of ongoing manufacturing inspection. Such uncertainties contain contributions from the measurement process itself and from the variability of the manufacturing process. Both types of contribution include those from operators, environmental conditions and other effects. In order to assist in separating the effects of the measurement process and manufacturing variability, measurements of check standards are used to provide data on the measurement process itself. Such measurements are nominally identical to those made on the test items. In particular, measurements on check standards are used to help identify time-dependent effects, so that such effects can be evaluated and contrasted with a database of check standard measurements. These standards are also useful in helping to control the bias and long-term drift of the process once a baseline for these quantities has been established from historical data.

Clause 4 briefly describes the statistical methods of uncertainty evaluation including the approach recommended in the *GUM*, the use of check standards, the steps in uncertainty evaluation and the examples in this Technical Specification. Clause 5, the main part of this Technical Specification, discusses the Type A evaluations. Nested designs in ANOVA are used in dealing with time-dependent sources of uncertainty. Other sources such as those from the measurement configuration, material inhomogeneity, and the bias due to measurement configurations and related uncertainty analyses are discussed. Type B (non-statistical) evaluations of uncertainty are discussed for completeness in Clause 6. The law of propagation of uncertainty described in the *GUM* has been widely used. Clause 7 provides formulae obtained by applying this law to certain functions of one and two variables. In Clause 8, as an example, a Type A evaluation of uncertainty for a gauge study is discussed, where uncertainty components from various sources are obtained. Annex A lists the statistical symbols used in this Technical Specification.

Measurement uncertainty for metrological applications — Repeated measurements and nested experiments

1 Scope

This Technical Specification follows the approach taken in the *Guide to the expression of the uncertainty of measurement* (*GUM*) and establishes the basic structure for stating and combining components of uncertainty. To this basic structure, it adds a statistical framework using the analysis of variance (ANOVA) for estimating individual components, particularly those classified as Type A evaluations of uncertainty, i.e. based on the use of statistical methods. A short description of Type B evaluations of uncertainty (non-statistical) is included for completeness.

This Technical Specification covers experimental situations where the components of uncertainty can be estimated from statistical analysis of repeated measurements, instruments, test items or check standards.

It provides methods for obtaining uncertainties from single-, two- and three-level nested designs only. More complicated experimental situations where, for example, there is interaction between operator effects and instrument effects or a cross effect, are not covered.

This Technical Specification is not applicable to measurements that cannot be replicated, such as destructive measurements or measurements on dynamically varying systems (such as fluid flow, electronic currents or telecommunications systems). It is not particularly directed to the certification of reference materials (particularly chemical substances) and to calibrations where artefacts are compared using a scheme known as a "weighing design". For certification of reference materials, see ISO Guide 35[14].

When results from interlaboratory studies can be used, techniques are presented in the companion guide ISO/TS 21748[15]. The main difference between ISO/TS 21748 and this Technical Specification is that the ISO/TS 21748 is concerned with reproducibility data (with the inevitable repeatability effects), whereas this Technical Specification concentrates on repeatability data and the use of the analysis of variance for its treatment.

This Technical Specification is applicable to a wide variety of measurements, for example, lengths, angles, voltages, resistances, masses and densities.

2 Normative references

The following referenced documents are indispensable for the application of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO 3534-1:1993, *Statistics — Vocabulary and symbols — Part 1: Probability and general statistical terms*

ISO 3534-3:1999, *Statistics — Vocabulary and symbols — Part 3: Design of experiments*

ISO 5725-1, *Accuracy (trueness and precision) of measurement methods and results — Part 1: General principles and definitions*

ISO 5725-2, *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*

ISO 5725-3, *Accuracy (trueness and precision) of measurement methods and results — Part 3: Intermediate measures of the precision of a standard measurement method*

ISO 5725-4, *Accuracy (trueness and precision) of measurement methods and results — Part 4: Basic methods for the determination of the trueness of a standard measurement method*

ISO 5725-5, *Accuracy (trueness and precision) of measurement methods and results — Part 5: Alternative methods for the determination of the precision of a standard measurement method*

ISO 5725-6, *Accuracy (trueness and precision) of measurement methods and results — Part 6: Use in practice of accuracy values*

Guide to the expression of uncertainty in measurement (GUM), BIPM, IEC, IFCC, ISO, IUPAC, IUPAP, OIML, 1993, corrected and reprinted in 1995

3 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO 3534-1, ISO 3534-3, ISO 5725 (all parts) and the following apply.

3.1
measurand
well-defined physical quantity that is to be measured and can be characterized by an essentially unique value

3.2
uncertainty of measurement
parameter or an estimate of the parameter, associated with the result of a measurement, that characterizes the dispersion of the values that could reasonably be attributed to the quantity being measured

3.3
Type A evaluation
method of evaluation of uncertainty by using statistical methods

3.4
Type B evaluation
method of evaluation of uncertainty by means other than statistical methods

3.5
standard uncertainty
uncertainty expressed as a standard deviation associated with a single component of uncertainty

3.6
combined standard uncertainty
standard deviation associated with the result of a particular measurement or series of measurements that takes into account one or more components of uncertainty

3.7
expanded uncertainty
combined standard uncertainty multiplied by a coverage factor which usually is an appropriate critical value from the t-distribution which depends upon the degrees of freedom in the combined standard uncertainty and the desired level of coverage

3.8
effective degrees of freedom
degrees of freedom associated with a standard deviation composed of two or more components of variance

NOTE The effective degrees of freedom can be computed using the Welch-Satterthwaite approximation (see *GUM*, G.4).

3.9
nested design
experimental design in which each level (i.e. each potential setting, value or assignment of a factor) of a given factor appears in only a single level of any other factor

NOTE 1 Adapted from ISO 3534-3:1999, definition 2.6.

NOTE 2 See ISO 3434-3:1999, 1.6, for the definition of level.

3.10
fixed effects
⟨factors⟩ effects resulting from the preselection of levels of each factor over the range of values of the factors

3.11
random effects
⟨factors⟩ effects resulting from the sampling at each level of each factor from the population of levels of each factor

3.12
balanced nested design
nested design experiment in which the number of levels of the nested factors is constant

[ISO 3534-3:1999, definition 2.6.1]

3.13
mean square for random errors
sum of squared error divided by the corresponding degrees of freedom

NOTE See ISO 3534-1:1993, 2.85 for the definition of the degrees of freedom.

4 Statistical methods of uncertainty evaluation

4.1 Approach of the *Guide to the expression of uncertainty of measurement*

The *Guide to the expression of uncertainty of measurement* (*GUM*) recommends that the result of measurement be corrected for all recognized significant systematic effects, that the result accordingly be the best (or at least unbiased) estimate of the measurand and that a complete model of the measurement system exists. The model provides a functional relationship between a set of input quantities (upon which the measurand depends) and the measurand (output quantities). The objective of uncertainty evaluation is to determine an interval that can be expected to encompass a large fraction of the distribution of values that could reasonably be attributed to the measurand. Since a bias cannot be quantified exactly, when a result of measurement is corrected for bias, the correction has an associated uncertainty.

The general approach, beginning from the modelling process, is the following.

NOTE The approach here relates to input quantities that are mutually independent. It is capable of a further generalization to mutually dependent input quantities (see the *GUM*, 5.2).

a) Develop a mathematical model (functional relationship) of the measurement process or measurement system that relates the model input quantities (including influence quantities) to the model output quantity (measurand). In many cases, this model is the formula (or formulae) used to calculate the measurement result, augmented if necessary by random, environmental and other effects such as bias correction that may affect the measurement result.

b) Assign best estimates and the associated standard uncertainties (uncertainties expressed as standard deviations) to the model input quantities.

c) Evaluate the contribution to the standard uncertainty associated with the measurement result that is attributable to each input quantity. These contributions shall take into account uncertainties associated with both random and systematic effects relating to the input quantities, and may themselves involve more detailed uncertainty evaluations.

d) Aggregate these standard uncertainties to obtain the (combined) standard uncertainty associated with the measurement result. This evaluation of uncertainty is carried out, according to GUM, using the law of propagation of uncertainty, or by more general analytical or numerical methods when the conditions for the law of propagation of uncertainty do not apply or it is not known whether they apply.

e) Where appropriate, multiply the standard uncertainty associated with the measurement result by a coverage factor to obtain an expanded uncertainty and hence a coverage interval for the measurand at a prescribed level of confidence. The *GUM* provides an approach that can be used to calculate the coverage factor. If the degrees of freedom for the standard uncertainties of all the input quantities are infinite, the coverage factor is determined from the normal distribution. Otherwise, the (effective) degrees of freedom for the combined standard uncertainty is estimated from the degrees of freedom for the standard uncertainties associated with the best estimates of the input quantities using the Welch-Satterthwaite formula.

The *GUM* permits the evaluation of standard uncertainties by any appropriate means. It distinguishes the evaluation by the statistical treatment of repeated observations as a Type A evaluation of uncertainty, and the evaluation by any other means as a Type B evaluation of uncertainty. In evaluating the combined standard uncertainty, both types of evaluation are to be characterized by variances (squared standard uncertainties) and treated in the same way.

Full details of this procedure and the additional assumptions on which it is based are given in the *GUM*.

The purpose of this Technical Specification is to provide additional detail on the evaluation of uncertainty by statistical means, concentrating on b) above, whether obtained by repeated measurement of the input quantities or of the entire measurement.

In this Technical Specification the term “artefact” is often used in the context of measurement. This usage is to be given a general interpretation in that the measurement may also relate to a bulk or chemical item, etc.

4.2 Check standards

A check standard is a standard required to have the following properties.

a) It shall be capable of being measured periodically.

b) It shall be close in material content and geometry to the production items.

c) It shall be a stable artefact.

d) It shall be available to the measurement process at all times.

Subject to its having these properties, an ideal check standard is an artefact selected at random from the production items, if appropriate, and reserved for this purpose.

Examples of the use of check standards include

— measurements on a stable artefact, and

— differences between values of two reference standards as estimated from a calibration experiment.

Methods for analysing check standard measurements are treated in 5.2.3.

In this Technical Specification, the term “check standard” is to be given a general interpretation. For instance, a bulk or chemical item may be used.

4.3 Steps in uncertainty evaluation

4.3.1 The first step in the uncertainty evaluation is the definition of the measurand for which a measurement result is to be reported for the test item. Special care should be taken to provide an unambiguous definition of the measurand, because the resulting uncertainty will depend on this definition. Possibilities include

— quantity at an instant in time at a point in space,

— quantity at an instant in time averaged over a specified spatial region,

— quantity at a point in space averaged over a time period.

For instance, the measurands corresponding to the hardness of a specimen of a ceramic material are (very) different

a) at a specified point in the specimen, or

b) averaged over the specimen.

4.3.2 If the value of the measurand can be measured directly, the evaluation of the standard uncertainty depends on the number of repeated measurements and the environmental and operational conditions over which the repetitions are made. It also depends on other sources of uncertainty that cannot be observed under the conditions selected to repeat the measurements, such as calibration uncertainties for reference standards. On the other hand, if the value of the measurand cannot be measured directly, but is to be calculated from measurements of secondary quantities, the model (or functional relationship) for combining the various quantities must be defined. The standard uncertainties associated with best estimates of the secondary quantities are then needed to evaluate the standard uncertainty associated with the value of the measurand.

The steps to be followed in an uncertainty evaluation are outlined as follows.

a) Type A evaluations:

   1) If the output quantity is represented by Y, and measurements of Y can be replicated, use an ANOVA model to provide estimates of the variance components, associated with Y, for random effects from

      — replicated results for the test item,

      — measurements on a check standard,

      — measurements made according to a designed experiment.

   2) If measurements of Y cannot be replicated directly, and the model

      $$Y = f(X_1, X_2, ..., X_n)$$

      is known, and the input quantities X_i can be replicated, evaluate the uncertainties associated with the best estimates x_i of X_i; then the law of propagation of uncertainty can be used.

   3) If measurements of Y or X_i cannot be replicated, refer to Type B evaluations.

b) Type B evaluations: evaluate a standard uncertainty associated with the best estimate of each input quantity.

c) Aggregate the standard uncertainties from the Type A and Type B evaluations to provide a standard uncertainty associated with the measurement result.

d) Compute an expanded uncertainty.

4.4 Examples in this Technical Specification

The purpose of the examples in various clauses of this Technical Specification and the more detailed case study in Clause 8 is to demonstrate the evaluation of uncertainty associated with measurement processes having several sources of uncertainty. The reader should be able to generalize the principles illustrated in these sections to particular applications. The examples treat the effect of both random effects and systematic effects in the form of bias on the measurement result. There is an emphasis on quantifying uncertainties observed over time, such as those for time intervals defined as short-term (repeatability) and for intermediate measures of precision such as day-to-day or run-to-run, as well as for reproducibility. For the reader's purpose, the time intervals should be defined in a way that makes sense for the measurement process in question.

To illustrate strategies for dealing with several sources of uncertainty, data from the Electronics and Electrical Engineering Laboratory of the National Institute of Standards and Technology (NIST), USA, are featured. The measurements in question are volume resistivities (Ω·cm) of silicon wafers. These data were chosen for illustrative purposes because of the inherent difficulties in measuring resistivity by probing the surface of the wafer and because the measurand is defined by an ASTM test method and cannot be defined independently of the method.

The intent of the experiment is to evaluate the uncertainty associated with the resistivity measurements of silicon wafers at various levels of resistivity (Ω·cm), which were certified using a four-point probe wired in a specific configuration. The test method is ASTM Method F84. The reported resistivity for each wafer is the average of six short-term repetitions made at the centre of the wafer.

5 Type A evaluation of uncertainty

5.1 General

5.1.1 Generally speaking, any observation that can be repeated (see *GUM*, 3.1.4 to 3.1.6) can provide data suitable for a Type A evaluation. Type A evaluations can be based on (for example) the following:

— repeated measurements on the item under test, in the course of, or in addition to, the measurement necessary to provide the result;

— measurements carried out on a suitable test material during the course of method validation, prior to any measurements being carried out;

— measurements on check standards, that is, test items measured repeatedly over a period of time to monitor the stability of the measurement process, where appropriate;

— measurements on certified reference materials or standards;

— repeated observations or determination of influence quantities (for example, regular or random monitoring of environmental conditions in the laboratory, or repeated measurements of a quantity used to calculate the measurement result).

5.1.2 Type A evaluations can apply both to random and systematic effects (*GUM*, 3.2). The only requirement is that the evaluation of the uncertainty component is based on a statistical analysis of series of observations. The distinction with regard to random and systematic effects is that

— random effects vary between observations and are not to be corrected,

— systematic effects can be regarded as essentially constant over observations *in the short term* and can, theoretically at least, be corrected or eliminated from the result.

Sometimes it is difficult to distinguish a systematic effect from random effects and it becomes a question of interpretation and the use of related statistical models. In general, it is not possible to separate random and systematic effects.

The *GUM* recommends that generally all systematic effects are corrected and that consequently the only uncertainty from such sources are those of the corrections. The role of time in the evaluation of Type A uncertainty using nested designs is discussed in 5.2. The uncertainties associated with measurement configuration and material inhomogeneity, respectively, are discussed in 5.3 and 5.4. Guidance on how to assess and correct for bias due to measurement configurations and to evaluate the associated uncertainty is given in 5.5. The manner in which the source of uncertainty affects the reported value and the context for the uncertainty determine whether an analysis of a random or systematic effect is appropriate.

Consider a laboratory with several instruments of a certain type, regarded as representative of the set of all instruments of that type. Then the differences among the instruments in this set can be considered to be a random effect if the uncertainty statement is intended to apply to the result of any particular instrument, selected at random, from the set.

Conversely, if the uncertainty statement is intended to apply to one (or several) specific instrument, the systematic effect of this instrument relative to the set is the component of interest.

5.2 Role of time in Type A evaluation of uncertainty

5.2.1 Time-dependent sources of uncertainty and choice of time intervals

Many random effects are time-dependent, often due to environmental changes. Three levels of time-dependent fluctuations are discussed and can be characterized as

a) short-term fluctuations (repeatability or instrument precision),

b) intermediate fluctuations (day-to-day or operator-to-operator or equipment-to-equipment, known as intermediate precision),

c) long-term fluctuations [run-to-run or stability (which may not be a concern for all processes) or intermediate precision].

This characterization is only a guideline. It is necessary for the user to define the time increments that are of importance in the measurement process of concern, whether they are minutes, hours or days.

One reason for this approach is that much modern instrumentation is exceedingly precise (repeatable measurements) in the short term, but changes over time, often caused by environmental effects, can be the dominant source of uncertainty in the measurement process. An uncertainty statement may be inappropriate if it relates to a measurement result that cannot be reproduced over time. A customer is entitled to know the uncertainties associated with the measurement result, regardless of the day or time of year when the measurement was made.

Two levels of time-dependent components are sufficient for describing many measurement processes. Three levels may be needed for new measurement processes or processes whose characteristics are not well understood. A three-level design is considered, with a two-level design as a special case.

Nested designs having more than three levels are not considered in this Technical Specification, but the approaches discussed can be extended to them. See ISO 5725-3.

5.2.2 Experiment using a three-level design

5.2.2.1 A three-level nested design is generally recommended for studying the effect of sources of variability that manifest themselves over time. Data collection and analysis are straightforward, and there is usually no need to estimate interaction terms when dealing with time-dependent effects. Nested designs can

be operated at several levels. Three levels are recommended for measurement systems where sources of uncertainty are not well understood and have not previously been studied.

The following levels are based on the characteristics of many measurement systems and should be adapted to a specific measurement situation as required:

a) Level 1: measurements taken over a short-time to capture the repeatability of the measurement;

b) Level 2: measurements taken over days (or other appropriate time increment);

c) Level 3: measurements taken over runs separated by months.

Symbols relating to these levels are defined thus:

— Level 1: J (J > 1) repetitions;

— Level 2: K (K > 1) days;

— Level 3: L (L > 1) runs.

The following balanced three-level nested design is recommended for collecting data on this basis. It describes the long-term fluctuations in the measurement process:

$$Y_{lkj} = \mu + \gamma_l + \delta_{lk} + \varepsilon_{lkj}$$

Here the measurements are represented by Y_{lkj}(l = 1,...,L; k = 1,...,K; j = 1,...,J) for the jth repetition on the kth day, which are repeated for the lth run. The subscripted terms in the model represent random effects in the measurement process that fluctuate with runs, days and short-term time intervals. The purpose of the experiment is to estimate the variance components that quantify these sources of variability. Let the variance components of the day and run effects for δ and γ be σ_D^2 and σ_R^2, respectively, and the variance of the measurement error ε be σ^2. These variance components form the basis for providing the standard uncertainties.

Table 1 — ANOVA table for a three-level nested design

Source	Degrees of freedom ν	Sum of squares SS	Mean square MS	Expected mean square
Run	$L - 1$	SS_R	MS_R	$\sigma^2 + J\sigma_D^2 + JK\sigma_R^2$
Day (run)	$L(K - 1)$	$SS_{D(R)}$	$MS_{D(R)}$	$\sigma^2 + J\sigma_D^2$
Error	$LK(J - 1)$	SS_E	MS_E	σ^2
The sources of variation, the sum of squares (SS), and the corresponding degrees of freedom (ν), are listed in the first, third and the second columns, respectively. The mean squares (MS), which are obtained from the sums of squares divided by the corresponding degrees of freedom, are listed in the fourth column. The last column lists the expected mean squares.				

Figure 1 depicts a design with J = 4, K = 3 and L = 2.

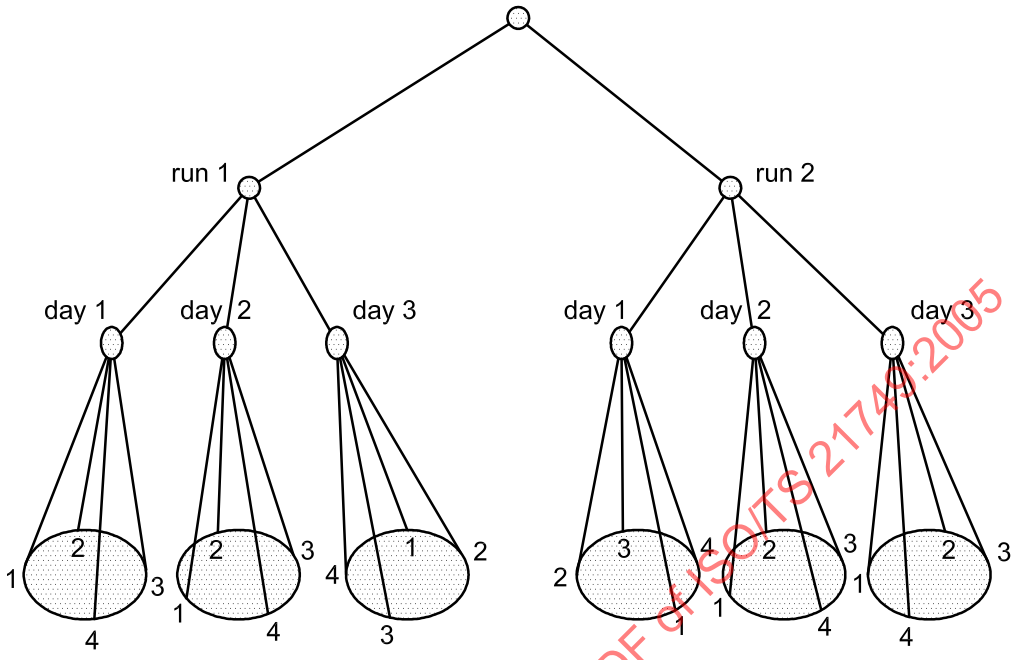


Figure 1 — Three-level nested design

5.2.2.2 The design can be repeated for Q (Q > 1) check standards (for check standards, see 5.2.3) and for I (I > 1) gauges (measuring instruments) if the intent is to characterize several similar gauges. Such a design has advantages in ease of use and computation. In particular, the number of repetitions at each level need not be large because information is being gathered on several check standards.

The measurements should be made with a *single* operator. The operator is not usually a consideration with automated systems. However, systems that require decisions regarding line, edge or other feature delineations may be operator-dependent. If there is reason to think that results might differ significantly between operators, “operators” can be substituted for “runs” in the design. Choose L (L > 1) operators at random from the pool of operators who are capable of making measurements at the same level of precision. (Conduct a small experiment with operators making repeatability measurements, if necessary, to verify comparability of precision among operators.) Then complete the data collection and analysis as outlined. In this case, the Level 3 standard deviation estimates operator effect.

Randomize with respect to gauges for each check standard, i.e. choose the first check standard and randomize the gauges, choose the second check standard and randomize the gauges, and so forth.

Record the average and standard deviation from each group of J repetitions by check standard and gauge.

The results should be recorded together with pertinent environmental readings and identifications for significant factors. A recommended way to record this information is in one computer file with one line or row of information in fixed fields for each check standard measurement. A spreadsheet is useful for this purpose. A list of typical entries follows:

a) month;

b) day;

c) year;

d) operator identification;

e) check standard identification;

f) gauge identification;

g) average of J repetitions;

h) short-term standard deviation from J repetitions;

i) degrees of freedom;

j) environmental readings (if pertinent).

From the model above, the standard deviation of the error with $LK(J-1)$ degrees of freedom is estimated using the mean square for random errors, MS_E, which is calculated as follows:

$$\hat{\sigma} = S = \sqrt{MS_E} = \sqrt{\frac{\sum_{l=1}^{L}\sum_{k=1}^{K}\sum_{j=1}^{J}(Y_{lkj} - \bar{Y}_{lk\bullet})^2}{LK(J-1)}}$$

where

$\bar{Y}_{lk\bullet} = \frac{1}{J}\sum_{j=1}^{J} Y_{lkj}$ is the average from each group of J repetitions.

The mean square for the day effect, $MS_{D(R)}$, with $L(K-1)$ degrees of freedom, is calculated as follows:

$$MS_{D(R)} = J\frac{\sum_{l=1}^{L}\sum_{k=1}^{K}(\bar{Y}_{lk\bullet} - \bar{Y}_{l\bullet\bullet})^2}{L(K-1)}$$

where

$$\bar{Y}_{l\bullet\bullet} = \frac{1}{K}\sum_{k=1}^{K}\bar{Y}_{lk\bullet}$$

The mean square for the run effect, MS_R, with $L-1$ degrees of freedom is calculated as follows:

$$MS_R = JK\frac{\sum_{l=1}^{L}(\bar{Y}_{l\bullet\bullet} - \bar{Y}_{\bullet\bullet\bullet})^2}{L-1}$$

where

$$\bar{Y}_{\bullet\bullet\bullet} = \frac{1}{L}\sum_{l=1}^{L}\bar{Y}_{l\bullet\bullet}$$

From the ANOVA Table 1, the estimator of the standard deviation for days is

$$\hat{\sigma}_D = S_D = \sqrt{\frac{MS_{D(R)} - MS_E}{J}}$$

and the estimator of the standard deviation for runs is

$$\hat{\sigma}_{\mathrm{R}} = S_{\mathrm{R}} = \sqrt{\frac{\mathrm{MS}_{\mathrm{R}} - \mathrm{MS}_{\mathrm{D(R)}}}{JK}}$$

if the differences under the square root sign are positive. Otherwise, $\hat{\sigma}_{\mathrm{D}}$ or $\hat{\sigma}_{\mathrm{R}}$ or both is (are) taken as zero, as appropriate.

Sometimes, a two-level nested design is suggested for collecting data on short-term and day-to-day fluctuations in the measurement process. The data that is collected in this experiment is similar to that collected on the check standard in the next section. If more than one check standard is used, the factor of "check standard" may be treated as a random factor, since the factor of "run" in the three-level case and the model and analysis are the same.

5.2.3 Check standard for assessing two levels of variability

5.2.3.1 Check standard procedure

Measurements on a single check standard are recommended for studying the effect of sources of variability that manifest themselves over time. Data collection and analysis are straightforward, and there is usually no need to estimate interaction terms when dealing with time-dependent errors. The measurements are made at two levels, which should be sufficient for characterizing many measurement systems. The following levels are based on the characteristics of many measurement systems and should be adapted to a specific measurement situation as required:

— Level 1 measurements, taken over a short term to estimate gauge precision;

— Level 2 measurements, taken over days to estimate longer-term variability.

A schedule for making check standard measurements over time (once a day, twice a week, or whatever is appropriate for sampling all conditions of measurement) should be established and followed. The check standard measurements should be structured in the same way as values reported on the test items. For example, if the reported values are the averages of two repetitions made within 5 min of each other, the check standard values should be averages of the two measurements made in the same manner. One exception to this rule is that there should be at least $J = 2$ repetitions per day, etc. Without this redundancy, there is no way to check the short-term precision of the measurement system.

5.2.3.2 Model

The statistical model that explains the sources of uncertainty being studied is a balanced two-level nested design:

$$Y_{kj} = \mu + \delta_k + \varepsilon_{kj}$$

Measurements on the test items are denoted by $Y_{kj} (k = 1,...,K; j = 1,...,J)$ with the first index identifying day and the second index the repetition number. The subscripted terms in the model represent random effects in the measurement process that fluctuate with days and short-term time intervals. The purpose of the experiment is to estimate the variance components that quantify these sources of variability.

5.2.3.3 Time intervals

The two levels discussed in this subclause are based on the characteristics of many measurement systems and can be adapted to a specific measurement situation as required. A typical design is shown in Figure 2, where there are J = 4 repetitions per day with the following levels:

— Level 1 J (J > 1) short-term repetitions to capture gauge precision;

— Level 2 K (K > 1) days (or other appropriate time increment).

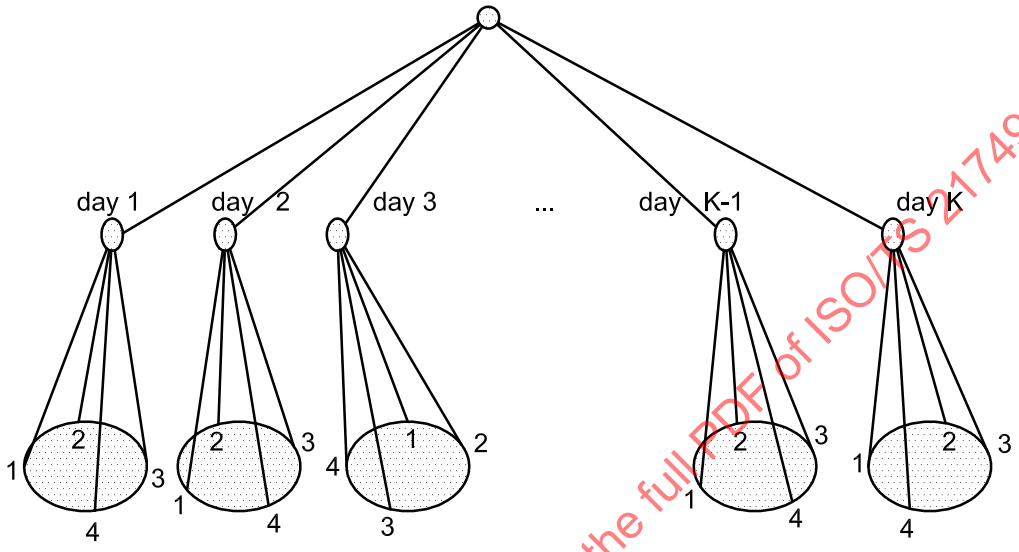


Figure 2 — Two-level nested design

5.2.3.4 Data collection

It is important that the design be truly nested as shown in Figure 2, so that repetitions are nested within days. It is sufficient to record the average and standard deviation for each group of J repetitions, with the following information:

a) month;

b) day;

c) year;

d) operator identification;

e) check standard identification;

f) gauge identification;

g) average of J repetitions;

h) repeatability standard deviation from J repetitions;

i) degrees of freedom;

j) environmental readings (if pertinent).

For this two-level nested design, the ANOVA table, Table 2, can be obtained from the three-level case.

Table 2 — ANOVA table for a two-level nested design

Source	Degrees of freedom ν	Sum of squares SS	Mean square MS	Expected mean square
Day	$K-1$	SS_D	MS_D	$\sigma^2 + J\sigma_D^2$
Error	$K(J-1)$	SS_E	MS_E	σ^2

The standard deviation of error with $K\,(J-1)$ degrees of freedom is calculated from

$$\hat{\sigma} = S = \sqrt{MS_E} = \sqrt{\frac{1}{K(J-1)}\sum_{k=1}^{K}\sum_{j=1}^{J}(Y_{kj} - \bar{Y}_{k\bullet})^2}$$

where

$$\bar{Y}_{k\bullet} = \frac{1}{J}\sum_{j=1}^{J} Y_{kj}$$

The mean square for the day effect, MS_D, with $K-1$ degrees of freedom is

$$MS_D = J\frac{\sum_{k=1}^{K}(\bar{Y}_{k\bullet} - \bar{Y}_{\bullet\bullet})^2}{K-1}$$

where

$$\bar{Y}_{\bullet\bullet} = \frac{\sum_{k=1}^{K}\bar{Y}_{k\bullet}}{K}$$

The standard deviation that explains day-to-day variability is

$$\hat{\sigma}_D = S_D = \sqrt{\frac{MS_D - MS_E}{J}}$$

if the difference under the square root sign is positive. Otherwise, $\hat{\sigma}_D$ is taken as zero.

A consequence of the use of the classical estimator covered here is that it can give rise to variance estimates that are negative. Other estimates may not have this property and may be used if appropriate.

5.3 Measurement configuration

5.3.1 Other sources of uncertainty

Measurements on test items are usually made on a single day, with a single operator, on a single instrument, etc. If the uncertainty is to be used to characterize all measurements made in the laboratory, it should account for any differences due to

— instruments,

— operators,

— geometries,

— other.

The effect of uncontrollable environmental conditions in the laboratory can often be estimated from check standard data taken over a period of time. Methods for calculating the associated components of uncertainty are discussed elsewhere in this Technical Specification. Uncertainties resulting from controllable factors, such as operators or instruments chosen for a specific measurement, are discussed in this subclause.

Note that operators should be studied only once, either under time-dependent types of experiments or under measurement configuration. Examples of causes for differences within a well-maintained laboratory are as follows:

— differences among instruments for measurements of derived units, such as sheet resistance of silicon, where the instruments cannot be directly calibrated to a reference standard;

— differences among operators for optical measurements that are not automated and that depend strongly on operator sightings;

— differences among geometrical or electrical configurations of the instrumentation.

Calibrated instruments do not normally fall in this class because uncertainties associated with the calibration are often reported by Type B evaluations, and the instruments in the laboratory should agree within the calibration uncertainties. Instruments whose responses are not directly calibrated to the defined unit are candidates for Type A evaluations. This covers situations where the measurement is defined by a test procedure or standard practice using a specific instrument type.

However, it should be noted that some systematic effects cannot be eliminated by calibration, for example, matrix effects in analytical chemistry.

5.3.2 Importance of context for the uncertainty

The differences mentioned at the beginning of this 5.3.1 are treated either as random differences or as bias differences. The approach depends primarily on the context for the uncertainty statement. For example, if instrument effect is the concern, one approach is to regard, say, the instruments in the laboratory as a random sample of instruments of the same type and to evaluate an uncertainty that applies to all results regardless of the particular instrument with which the measurements are made. In this case, the two-level nested design in 5.2.3 can be applied, where the second level is for one of the sources of influence such as the source of instruments. The other approach is to evaluate an uncertainty that applies to results using a specific instrument, which is treated as an analysis of systematic effect or bias in 5.5.

Below is a simple approach using two-level random effect nested design to evaluate the uncertainty for one source of influence.

5.3.3 Data collection and calculation of variance component

To evaluate the uncertainty of a measurement process due to instruments, select a random sample of $I (I > 1)$ instruments from those available. Make measurements on $Q (Q > 1)$ artefacts with each instrument. Given the $I \times Q$ measurements, the standard deviation that describes the differences among instruments is computed as follows, from the average for each instrument, and has $I - 1$ degrees of freedom:

$$S_{\text{inst}} = \sqrt{\frac{\sum_{i=1}^{I} (\bar{Y}_{\bullet i} - \bar{Y}_{\bullet\bullet})^2}{I - 1}}$$

where, for the ith instrument:

$$\bar{Y}_{i\bullet} = \frac{1}{Q} \sum_{q=1}^{Q} Y_{iq} , \qquad \bar{Y}_{\bullet\bullet} = \frac{1}{I} \sum_{i=1}^{I} \bar{Y}_{i\bullet}$$

5.3.4 Example of analysis of random differences

A two-way table of resistivity measurements (Ω·cm) with five probes (numbered 1, 281, 283, 2 062, 2 362) on $Q = 5$ wafers (numbered 138, 139, 140, 141, 142) is shown in Table 3. The same data are analyzed for the bias of Probe No. 2362 in 5.5. The average for each probe across artefacts is shown. The standard deviation (of the averages of resistivity measurements) for probes is 0,021 9 Ω·cm with four degrees of freedom. Thus, $S_{\text{inst}} = 0{,}021\ 9$.

Table 3 — Measurements of the resistivity of five wafers using five probes

Values in Ω·cm

Probe No.	Wafer identification number					Average
	138	139	140	141	142	
1	95,154 8	99,311 8	96,101 8	101,124 8	94,259 3	97,190 5
281	95,140 8	99,354 8	96,080 5	101,074 7	94,290 7	97,188 3
283	95,149 3	99,321 1	96,041 7	101,110 0	94,248 7	97,174 2
2 062	95,112 5	99,283 1	96,049 2	101,057 4	94,252 0	97,150 8
2 362	95,092 8	99,306 0	96,035 7	101,060 2	94,214 8	97,141 9

For a graphical analysis, differences between the measured values and the average for each probe can be plotted against the wafer, for each probe, with probes being individually identified by a particular plotting symbol. The plot is examined to determine whether some instruments always read high or low relative to the other instruments and if this behaviour is consistent across probes. The graph given in Figure 3, established from data taken from Table 4 (see 5.5.2.2), shows that there are differences among the probes, with Probe No. 2062 and No. 2362, for example, consistently reading low relative to the other probes.

5.4 Material inhomogeneity

5.4.1 Problems generated by inhomogeneities

Artefacts, electrical devices, chemical substances, etc. can be inhomogeneous relative to the quantity that is being characterized. Inhomogeneity can be a factor in the uncertainty evaluation where

a) an artefact is characterized by a single value, but is inhomogeneous over its surface, etc., and

b) a lot of items is assigned a single value from a few samples from the lot and the lot is inhomogeneous from sample to sample.

An unfortunate aspect of this situation is that inhomogeneity may be the dominant source of uncertainty. Even if the measurement process itself is very precise and in statistical control, the combined uncertainty may still be unacceptable for practical purposes because of material inhomogeneities. Detailed discussions on the homogeneity study for reference materials are given in ISO Guide 35[14].

5.4.2 Strategy for random inhomogeneities

Random inhomogeneities are assessed using statistical methods for quantifying random effects. An example of inhomogeneity is a chemical reference material that cannot be sufficiently homogenized with respect to isotopes of interest. Isotopic ratio must be determined from measurements on a few bottles drawn randomly from the lot.

5.4.3 Data collection and calculation of component for inhomogeneity

A simple scheme for identifying and quantifying the effect of inhomogeneity on a measurement result is a balanced two-level nested design. K (K > 1) test items are drawn at random from a lot and J(J > 1) measurements are made per test item. The measurements are denoted by

$$Y_{k1}, Y_{k2}, \ldots, Y_{kJ}, \ldots, Y_{K1}, Y_{K2}, \ldots, Y_{KJ},$$

with index k =1, …, K relating to test items for Level 2 and j = 1, …, J to repetitions within a test item for Level 1.

The inhomogeneity (between test items) variance, defined as the variance component due to the inhomogeneity of the test items, is calculated as in 5.2.3 using an ANOVA technique, where

$$s_{\text{inh}}^2 = \frac{\text{MS}_{\text{item}} - \text{MS}_{\text{E}}}{J}$$

$$= \frac{1}{K-1}\sum_{k=1}^{K}\left(\bar{Y}_{k\bullet} - \bar{Y}_{\bullet\bullet}\right)^2 - \frac{1}{KJ(J-1)}\sum_{k=1}^{K}\sum_{j=1}^{J}(Y_{kj} - \bar{Y}_{k\bullet})^2$$

where

$$\bar{Y}_{k\bullet} = \frac{1}{J}\sum_{j=1}^{J} Y_{kj};$$

$$\bar{Y}_{\bullet\bullet} = \frac{1}{K}\sum_{k=1}^{K} \bar{Y}_{k\bullet};$$

MS_{item} is the mean square due to the test items;

s_{inh}^2 is an estimator of the variance component due to material inhomogeneity or test items.

If S^2_{inh} is negative, the effect of inhomogeneity is statistically regarded as being equal to zero and there is no contribution to uncertainty. That is, the uncertainty associated with inhomogeneity is reported as

$$u_{\text{inh}} = \max(S_{\text{inh}}, 0)$$

5.4.4 Evaluation of uncertainty associated with inhomogeneity

The uncertainty evaluation depends on the use of the measurement result. Typically, inhomogeneity is important when the mean for a number of test items from a larger batch is obtained, and that mean value is assigned to each test item in the batch. For a measurement result calculated as the mean of results from K different test items, the standard uncertainty u_{inh} arising from inhomogeneity and associated with the mean result is calculated from S_{inh} according to

$$u_{\text{inh}} = \max\left(\frac{S_{\text{inh}}}{\sqrt{K}}, 0\right)$$

However, for a measurement result calculated as the mean of results from K different test items and applied to *each* of the items in the remainder of the batch, the standard uncertainty u_{inh} arising from inhomogeneity and associated with the prediction interval for each item of the remainder of the batch is based on the prediction interval (see reference [5]) and given by

$$u_{\text{inh}} = \max\left(\sqrt{1+\frac{1}{K}} \times S_{\text{inh}}, 0\right)$$

5.4.5 Strategy for systematic inhomogeneities

Systematic inhomogeneities require a somewhat different approach. For example, roughness can vary systematically over the surface of a 50 mm square metal piece prepared to have a specific roughness profile. The certification laboratory can measure the piece at several sites, but unless it is possible to characterize roughness as a function of position on the piece, it is necessary to assess inhomogeneity as a source of uncertainty.

In this situation, one strategy is to compute the reported value as the average of measurements made over the surface of the piece and assess an uncertainty for departures from the average. The component of uncertainty can be assessed by one of several methods for Type B evaluation of uncertainty given in the *GUM*.

5.5 Bias due to measurement configurations

5.5.1 General

In statistics, for a parameter θ to be estimated, the bias of an estimator $\hat{\theta}$ is defined as the difference between the expectation of $\hat{\theta}$ and the true value θ. Namely, $b = E[\hat{\theta}] - \theta$. Since the true value θ is unknown, b is unknown. When an estimate of b is available, it is termed a correction and denoted by $\hat{b}$. A correction can be with respect to a reference value or to some kind of an average value. Given a set of corrections, $\hat{b}_1,\ldots,\hat{b}_n$, the bias of the estimator can be estimated by the average of the corrections, which is

$$\hat{b} = \frac{\sum_{i=1}^{n} \hat{b}_i}{n}$$

If the corrections are treated as random, a probability distribution such as a normal distribution can be assumed for the correction.

If the corrections are clustered about zero, it is often assumed that the probability distribution for the corrections has a mean of zero and the case is often called “zero” correction. In this case, $\hat{\theta}$ is an unbiased estimator of θ. If the corrections are normally distributed or there are a large number of corrections available, the uncertainty of bias can be estimated by the standard deviation of the sample mean of the corrections. If there is not much information concerning the distribution, it can be assumed that the corrections $\{\hat{b}_i, i = 1,...,n\}$ are uniformly distributed between $-a$ and a. The bias is thus estimated by zero. The quantity a can be estimated by

$$\hat{a} = \frac{n+1}{n-1}\left(\frac{\max\{\hat{b}_i\} - \min\{\hat{b}_i\}}{2}\right)$$

The standard deviation of the estimator of the bias, $\hat{b}$, is estimated by

$$S_{\hat{b}} = \frac{n+1}{(n-1)\sqrt{3n}}\,\frac{\max\{\hat{b}_i\} - \min\{\hat{b}_i\}}{2}$$

Sources of bias discussed in this Technical Specification in the metrology context cover specific measurement configurations. Measurements on test items are usually made on a single day, with a single operator, with a single instrument, etc. Even if the uncertainty is to be used to characterize only those measurements made in one specific configuration, it is necessary to account for any significant differences due to

a) instruments,

b) operators,

c) geometries,

d) other.

Calibrated instruments do not normally fall in this class because uncertainties associated with the calibration are often reported by Type B evaluations, and the instruments in the laboratory should agree within the calibration uncertainties. Instruments whose responses are not directly calibrated to the defined unit are candidates for Type A evaluations. This covers situations where the measurement is defined by a test procedure or standard practice using a specific instrument type.

If measurements for only one configuration are of interest, such as measurements made with a specific instrument, or if a smaller uncertainty is required, the differences among, say, instruments are treated as biases. One strategy in this situation is to correct all measurements made with a specific instrument to the average for the instruments in the laboratory and evaluate a Type A uncertainty for the correction. This strategy, of course, relies on the assumption that the instruments in the laboratory represent a random sample of all instruments of a specific type.

However, suppose that it is only possible to make comparisons, for example between two instruments, and neither is known to be “unbiased”. This scenario requires a different strategy because the average will not necessarily give an unbiased result. The recommended strategy, if there is a significant difference between the instruments (and this should be tested), is to apply a “zero” correction and evaluate a Type A uncertainty associated with the correction.

The discussion above is intended to point out that there are many possible scenarios for biases and that they should be treated on a case-by-case basis. A plan is needed for:

— gathering data;

— testing for biases (graphically or statistically);

— estimating biases;

— evaluating uncertainties associated with significant biases.

Without loss of generality, in this Technical Specification the instruments are treated as the only source of bias. Consider first the situation with a measurement model for one instrument. Suppose $Y_1,...,Y_n$ are independent measurements of the true value θ of a measurand based on a single instrument. The average of a large number of independent measurements using this instrument is denoted by μ. Therefore

$$Y_i = \mu + e_i$$

where $e_1,...,e_n$ are assumed to be independently distributed random errors with mean zero and variance σ^2.

The sample mean $\overline{Y} = \dfrac{\sum_{i=1}^{n} Y_i}{n}$, is used to estimate θ and as the result to be reported for the instrument under consideration. The correction or error due to using $\overline{Y}$ as the result of measuring θ may be decomposed as

$$\overline{Y} - \theta = (\mu - \theta) + \overline{e} = b + \overline{e}$$

where

$$\overline{e} = \frac{\sum_{i=1}^{n} e_i}{n}$$

and the bias

$$b = \mu - \theta$$

Equivalently,

$$\overline{Y} = \theta + b + \overline{e}$$

The term $\overline{e}$ is the random component and b is the systematic component or instrument bias component. The uncertainty associated with the random component $\overline{e}$ is usually estimated by $u(\overline{e}) = S/\sqrt{n}$, assuming a normal distribution, where S is the sample standard deviation of $Y_1,...,Y_n$. The uncertainty associated with $\hat{b}$, an estimator of b, is evaluated based on scientific judgment (Type B evaluation of uncertainty) or on statistical methods (Type A evaluation of uncertainty). For the one-instrument case, it is often convenient to quantify the uncertainty in $\hat{b}$ by associating with it a distribution whose mean is zero. If the mean of $\hat{b}$ is thought to be a known quantity, b, then each measured value Y_i may be corrected by an amount b, i.e., Y_i is replaced by $Y_i - b$, so the assumption that $\hat{b}$ has mean equal to zero is not a restriction. The form of the distribution of $\hat{b}$ may be taken to be normal or uniform or some other appropriate distribution. The combined uncertainty in $\overline{Y}$ as an estimate of θ is then calculated using

$$u(\overline{Y}) = \sqrt{u_b^2 + \frac{S^2}{n}}$$

where u_b is the uncertainty associated with b based on Type A and/or Type B evaluations. The corresponding degrees of freedom are calculated using the Welch-Satterthwaite formula.

Consider now measurements made by K instruments. $Y_{ki}(k=1,...,K;i=1,...,n)$ is the ith independent measurement made by the kth instrument. The corresponding statistical model is

$$Y_{ki} = \theta + b_k + e_{ki}$$

where b_k is the bias corresponding to the kth ($k=1,...,K$) instrument and e_{ki} are the random errors. The objective is to estimate b_k and the associated uncertainty. The solution to the problem depends on the assumptions made on b_k and is discussed in 5.5.2 and 5.5.3. In 5.5.4, bias with sparse data is discussed briefly.

5.5.2 Consistent bias

5.5.2.1 General

Bias can be treated as consistent or inconsistent. When a bias is significant and persists consistently over time and has the same magnitude for a specific instrument it is called consistent bias, and should be corrected if it can be reliably estimated from repeated measurements. This assumes the level or magnitude of the bias due to the instruments is essentially the same for all materials of interest. Given the measurements

$$Y_{ki}(k=1,\cdots,K;i=1,\cdots,n)$$

on n artefacts with K instruments, the statistical model given in 5.5.1 is

$$Y_{ki} = \theta + b_k + e_{ki}$$

where θ is the value of the measurand Y, b_k is the bias of the kth instrument and e_{ki} the random error. Take the bias b_k as non-random or fixed. From the model above and the assumption that $\sum_{k=1}^{K} b_k = 0$

$$\bar{Y}_{\bullet i} = \theta + \bar{e}_{\bullet i}$$

where

$$\bar{Y}_{\bullet i} = \frac{\sum_{k=1}^{K} Y_{ki}}{K}$$

For the ith artefact, θ can be estimated by $\bar{Y}_{\bullet i}$, and a correction of the ith artefact by the kth instrument is $\hat{b}_{ki} = Y_{ki} - \bar{Y}_{\bullet i}$. Here, the measurement of the ith artefact by the kth instrument is corrected with respect to the average for all K instruments. From the model above, b_k, the bias for the kth instrument, is estimated by the average of the corrections:

$$\hat{b}_k = \frac{1}{n}\sum_{i=1}^{n}\hat{b}_{ki} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_{ki} - \bar{Y}_{\bullet i}\right) = \frac{1}{n}\sum_{i=1}^{n}\left(Y_{ki} - \bar{Y}_{\bullet\bullet}\right)$$

where

$$\bar{Y}_{\bullet\bullet} = \frac{1}{n}\sum_{i=1}^{n}\bar{Y}_{\bullet i}$$

The correction that should be applied to measurements made with kth instrument is

$$Y_{\text{corrected}} = Y_{\text{measured}} - \hat{b}_k$$

The uncertainty of the bias (or of the average of corrections) for the kth instrument is

$$S_{\hat{b}_k}(k) = \frac{1}{\sqrt{n}}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(\hat{b}_{ki} - \hat{b}_k)^2} = \frac{1}{\sqrt{n}}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(Y_{ki} - \bar{Y}_{\bullet i} - \hat{b}_k\right)^2}$$

Depending on the application, a statistical test can be performed to test whether the bias is significant.

5.5.2.2 Example of consistent bias

This example considers the case where measurements are made with one instrument, and the reported values will be corrected for bias due to this instrument. The case where any one of the probes could be used to make measurements is treated as analysis of random effects.

In Table 4, the average for each wafer was subtracted from each value measured. The resistivity measurements (Ω·cm) were performed using five probes on each of five silicon wafers. The correction, as shown, represents the differences for each probe with respect to the other probes, i.e. $\hat{b}_{ki} = Y_{ki} - \bar{Y}_{\bullet i}$, for the kth probe and the ith wafer. The quantities $\hat{b}_{5i}$ $(i = 1,...,5)$ for Probe No. 2362 are persistent and are negative for the five wafers.

Table 4 — Corrections $(\hat{b}_{ki})$ for probes and silicon wafers

Values in Ω·cm

Probe No.	Probe index No.	Wafer identification number				
		138	139	140	141	142
1	1	0,024 76	–0,003 56	0,040 02	0,039 38	0,006 20
181	2	0,010 76	0,039 44	0,018 71	–0,010 72	0,037 61
182	3	0,019 26	0,005 74	–0,020 08	0,024 58	–0,004 39
2 062	4	–0,017 54	–0,032 26	–0,012 58	–0,028 02	–0,001 10
2 362	5	–0,037 25	–0,009 36	–0,026 08	–0,025 22	–0,038 30

For Probe No. 2362:

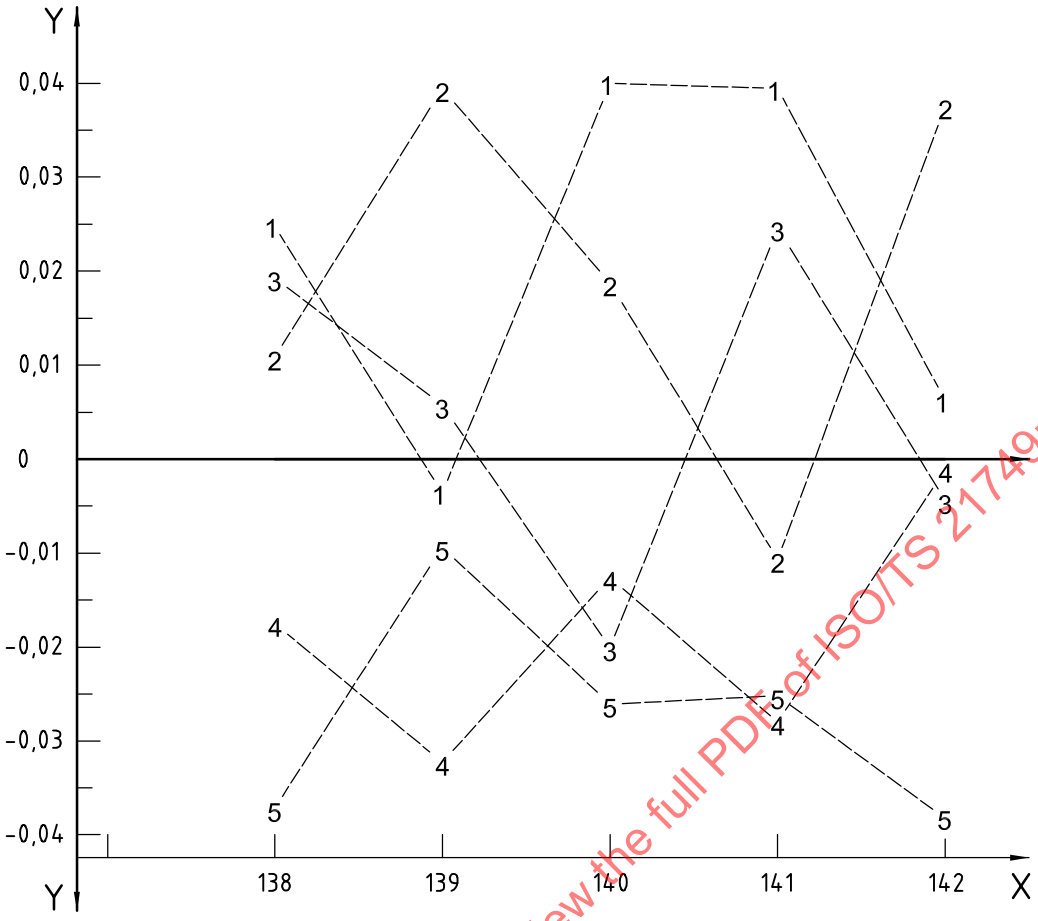
— the bias is equal to $\hat{b}_5 = \dfrac{\sum_{i=1}^{5}\hat{b}_{5i}}{5} = -0{,}027\ 24\ \Omega\cdot\text{cm}$,

— the standard deviation of corrections is equal to $S_{\hat{b}_{5i}}$ = 0,011 71 for any i, and

— the standard deviation of bias $\hat{b}_5$ (or the mean corrections) is equal to $S_{\hat{b}_5} = \dfrac{0{,}011\,71}{\sqrt{5}} = 0{,}005\ 23$.

The differences between the measurement and the average [i.e. correction $(\hat{b}_{ki})$] are plotted against the wafer identification number for each probe individually identified by its index in Figure 3. The graphs confirm that Probe No. 2362 (Index No. 5 on the graph), which is the instrument of interest for this measurement process, consistently reads low relative to the other probes. This behaviour is consistent over two runs, which are separated by a period of two months.

Because there is significant and consistent bias for Probe No. 2362, measurements made with that instrument should be corrected for average bias relative to the other instruments.



Key

X	index number of wafer
Y	$\hat{b}_{ki}$, Ω·cm
1, ..., 5	index numbers of probes (see Table 4)

Figure 3 — Corrections ($\hat{b}_{ki}$) plotted against the silicon wafer identification number — Gauge study for five probes

5.5.3 Inconsistent bias

5.5.3.1 General

If a bias is significant with a random nature for a specific instrument, operator or configuration, it is treated as inconsistent. Without loss of generality, the mean of the bias can be taken as zero. Otherwise it can be corrected by subtracting an estimate of the bias from the measurements. In this case, the bias changes direction over time. Then a “zero” correction can be assumed. The uncertainty of the bias can be determined depending on the knowledge of the distribution of the corrections such as a normal or uniform distribution as discussed at the beginning of 5.5.1. In 5.5.3.2 is an example of “zero” correction. Another kind of inconsistent bias can be found in 5.5.4.

5.5.3.2 Example of inconsistent bias

The results of resistivity measurements made with five probes on five silicon wafers are obtained. Table 5 gives the correction or bias of Probe No. 283, which is the probe of interest at this level, where the artefacts are 1 Ω·cm wafers calculated based on all probes as shown in 5.5.2. The average correction is negative for Run 1 and positive for Run 2, with the runs separated by a two-month time period.

Table 5 — Biases for Probe No. 283

Values in Ω·cm

Wafer identification number	Run 1	Run 2
11	0,000 034 0	–0,000 184 1
26	–0,000 100 0	0,000 086 1
42	0,000 018 1	0,000 078 1
131	–0,000 070 1	0,000 158 0
208	–0,000 024 0	0,000 187 9
Average	–0,000 028 4	0,000 065 2

Assuming the corrections $\{\hat{b}_{ki}, i = 1,2,3,4,5\}$ are normally distributed, the pair-wise t-test does not reject the hypothesis that Run 1 and Run 2 have the same mean. Combining the corrections for both runs, the computed t-statistic = 0,501 6 with nine degrees of freedom and thus the zero-mean hypothesis is not rejected at 5 % level. The estimate of the bias for Probe No. 283 is 0,000 018 4 Ω·cm and the standard deviation of the bias for Probe No. 283 is $S_{\hat{b}_{283}}$ = 0,000 031 Ω·cm. Alternatively, a conservative assumption is that the corrections could fall somewhere within the limits $\pm a$, where an estimate of a, $\hat{a} = 0{,}000\ 227\ 3$, is obtained from the formula at the beginning of 5.5.1. In this case, the estimate of the bias for Probe No. 283 is zero and the standard deviation of the bias estimate is

$$S_{\hat{b}_{283}} = \frac{11}{9\sqrt{3\times 10}} \frac{\max\{\hat{b}_{283i}\} - \min\{\hat{b}_{283i}\}}{2}$$

$$= \frac{11}{9\sqrt{3\times 10}} \frac{[0{,}000\ 187\ 9 - (-0{,}000\ 184\ 1)]}{2} = 0{,}000\ 042\ \Omega\cdot\text{cm}$$

5.5.4 Bias with sparse data

5.5.4.1 General

This subclause outlines a method for dealing with biases that may be real, but that cannot be estimated reliably because of the scarcity of the data. For example, a test between two, of many possible, configurations of the measurement process cannot produce a sufficiently reliable estimate of bias to permit a correction, but it can reveal problems with the measurement process. If the bias is significant, the strategy depends on whether this is the case of consistent or inconsistent bias.

5.5.4.2 Example of bias from sparse data

An example is given of a study of wiring settings for a single gauge. The gauge, a 4-point probe for measuring resistivity of silicon wafers, can be wired in several ways. Because it was not possible to test all wiring configurations during the gauge study, measurements were made in only two configurations as a way of identifying possible problems.

Measurements were made on five wafers over six days (except for day 2 on Wafer 39), with Probe No. 2062 wired in two configurations. Differences between measurements in the two configurations on the same day are treated as being corrections and are shown in Table 6.

Table 6 — Differences between wiring configurations for Probe No. 2062

Wafer	Wafer identification number	Day	Difference
17	1	1	–0,010 8
		2	–0,011 1
		3	–0,006 2
		4	0,002 0
		5	0,001 8
		6	0,000 2
39	2	1	–0,008 9
		3	–0,004 0
		4	–0,002 2
		5	–0,001 2
		6	–0,003 4
63	3	1	–0,001 6
		2	–0,011 1
		3	–0,005 9
		4	–0,007 8
		5	–0,000 7
		6	0,000 6
103	4	1	–0,005 0
		2	–0,014 0
		3	–0,004 8
		4	0,001 8
		5	0,001 6
		6	0,004 4
125	5	1	–0,005 6
		2	–0,015 5
		3	–0,001 0
		4	–0,001 4
		5	0,000 3
		6	–0,001 7

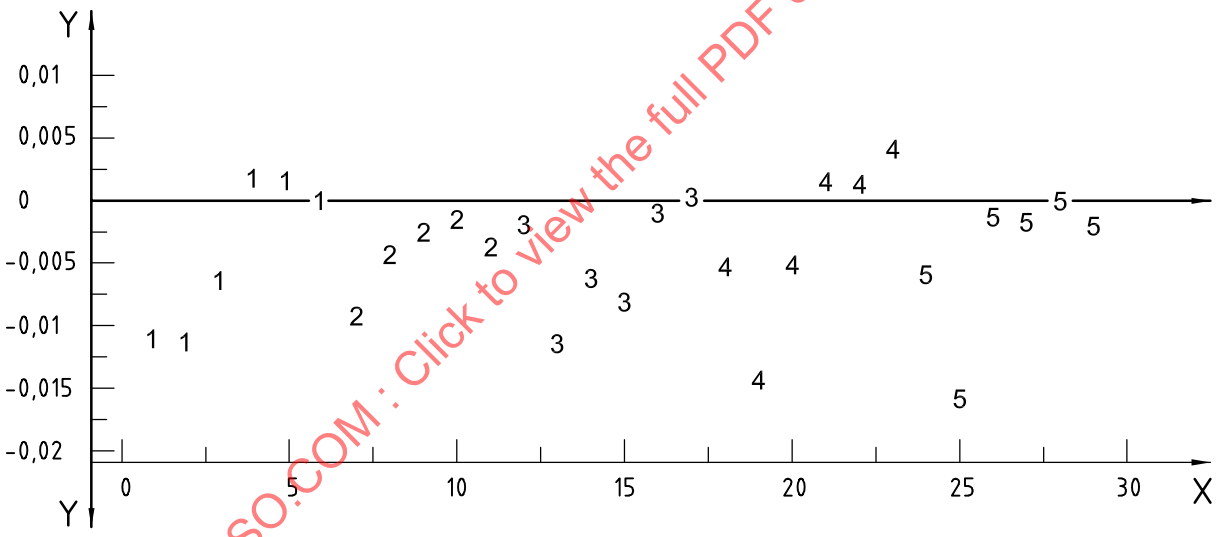
A plot of the differences for the two configurations is shown in Figure 4. This figure shows that the differences are mostly negative. The maximum and minimum of the differences are 0,004 4 and –0,015 5. The bias due to configurations is estimated by the average of the differences or corrections, i.e.

$$\hat{b} = \frac{\sum_{i=1}^{29} \hat{b}_i}{29} = -0{,}003\ 83$$

Since the total number of differences is 29, the uncertainty of the wiring bias based on the sample standard deviation is

$$S_{\hat{b}} = \sqrt{\frac{\sum_{i=1}^{29} (\hat{b}_i - \hat{b})^2}{29 \times 28}} = 0{,}000\ 96\ \Omega\cdot\text{cm}$$

For the 29 corrections ($\hat{b}_i$), the computed t-statistic = –4,013 3. The hypothesis that the probability distribution of the corrections has a zero mean is rejected.



Key

X time, d

Y resistivity differences between two wiring configurations, Ω·cm

1, ..., 5 identification numbers of wafers (see Table 6)

Figure 4 — Differences between two wiring configurations — Run for measurements made with Probe No. 2062 on five wafers for six days

6 Type B evaluation of uncertainty

6.1 Type B evaluations of uncertainty can be applied to both random and systematic effects. The distinguishing feature is that the calculation of the uncertainty component is not based on a statistical analysis of data.

Some examples of sources of uncertainty that lead to Type B evaluations are

— reference standards calibrated by another laboratory,

— physical constants used in the calculation of the reported value,

— environmental effects that cannot be sampled,

— possible configuration/geometry misalignment in the instrument, and

— lack of resolution of the instrument.

6.2 Documented sources of uncertainty, such as calibration reports for reference standards or published reports of uncertainties for physical constants, pose no difficulty in the analysis. The uncertainty will usually be reported as an expanded uncertainty, U, which is converted to the standard uncertainty using the formula:

$$u = \frac{U}{k}$$

If the factor k is not known nor documented, it is probably conservative to assume that k = 2. Sources of uncertainty that are local to the measurement process, but that cannot be adequately sampled to allow a statistical analysis, require Type B evaluations. One technique, which is widely used, is to estimate the worst-case effect from:

— experience;

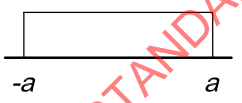
— scientific judgment;

— scant data.

6.3 For the situation at hand, an estimated bias or a correction can be regarded as a random draw from an assigned statistical distribution. Then the standard uncertainty is taken as the standard deviation of this distribution. Among the statistical distributions possible, only two distributions are considered.

a) Uniform distribution

Given its end-points, $\pm a$, all values between $-a$ and $+a$ are equally likely:

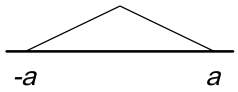


$$S_{\text{source}} = \frac{1}{\sqrt{3}} a$$

The corresponding degrees of freedom may be taken as infinite if a is well quantified. Otherwise, the degrees of freedom should be chosen to reflect the accuracy to which a is known. See *GUM*, G.4.2.

b) Triangular distribution

The triangular distribution gives a smaller standard uncertainty than that provided by the uniform distribution having the same end-points:



$$S_{\text{source}} = \frac{1}{\sqrt{6}} a$$

The degrees of freedom are typically taken as infinite.

7 Propagation of uncertainty

7.1 General

7.1.1 The approach to uncertainty evaluation that has been followed so far has been what is called a top-down approach. Uncertainty components are estimated from direct repetitions of the measurement. To contrast this with the use of the law of propagation of uncertainty, consider the simple example where the area of a rectangular is estimated from replicate *pair* measurements of length, L, and width, W. The area, A,

$$A = L \times W$$

can be computed from each replicate. The standard deviation of the reported area is estimated directly from the replicates of area.

7.1.2 This approach has the following advantages:

— proper treatment of covariances between measurements of length and width;

— proper treatment of unsuspected sources of uncertainty that would emerge if measurements covered a range of operating conditions and a sufficiently long time period.

7.1.3 Sometimes the measurement cannot be replicated directly in a way that reflects all the effects that influence it. Consideration can be given to the use of the law of propagation of uncertainty (*GUM*). The approach in this instance is to compute:

a) the measurement result as the product of the mean of the length measurements and the mean of the width measurements;

b) the standard uncertainty associated with “length” L;

c) the standard uncertainty associated with “width” W;

and combine the two standard uncertainties into a standard uncertainty associated with the measurement result using the approximation for the product of two variables. The formula below is appropriate if there is no covariance between length and width measurements.

$$S_A \approx \sqrt{W^2 \times S_L^2 + L^2 \times S_W^2}$$

7.1.4 In the ideal case, this value will not differ much from that obtained directly from the area measurements. However, in some circumstances they may differ appreciably because of

— unsuspected covariances;

— disturbances that affect the reported value of the measurand;

— approximation error.

In general, the law of propagation of uncertainty applied to the model

$$Y = f\,(X, Z, \ldots)$$

which is a function of one or more variables with measurements $X, Z, \ldots$, gives the following value for the standard deviation associated with Y:

$$S_Y^2 \approx \left(\frac{\partial f}{\partial X}\right)^2 S_X^2 + \left(\frac{\partial f}{\partial Z}\right)^2 S_Z^2 + \cdots + \left(\frac{\partial f}{\partial X}\right)\left(\frac{\partial f}{\partial Z}\right) S_{XZ}^2 + \cdots$$

where

S_X is the standard uncertainty associated with X;

S_Z is the standard uncertainty associated with Z;

S_{XZ} is the covariance associated with X and Z;

$\partial f / \partial X$ is the partial derivative of the function f with respect to X, evaluated at x, z, …, which are the best estimates of X, Z,...,etc.

7.1.5 Covariance terms can be difficult to estimate if measurements are not made in pairs. Sometimes, these terms are omitted from the formula. Guidance on when this is acceptable practice is given below.

a) If the measurements of X, Z are independent, the associated covariance term is zero.

b) Practically speaking, covariance terms should be included in the computation only if they have been estimated from sufficient data or if other information is available to assist in their determination.

Generally, reported values of test items from calibration designs have non-zero covariances, which should be taken into account if Y is a summation such as the mass of two weights, or the length of two gauge blocks end-to-end, etc.

7.2 Formulae for functions of a single variable

Standard deviations of reported values that are functions of a single variable are reproduced in Table 7 from Reference [6]. The reported value Y is a function of the average of N measurements of a single variable.

7.3 Formulae for functions of two variables

Standard deviations of reported values that are functions of measurements of two variables are reproduced in Table 8 from Reference [6]. The reported value Y is a function of averages of N measurements of two variables. The multipliers of the standard deviations are referred to as “sensitivity coefficients”.

Table 7 — Standard deviations for functions of a single variable

Function Y of $\bar{X}$	First order approximation of standard deviation of Y	Notes
$\bar{X}$ is the average of N independent measurements	S_X = standard deviation of X.	
$Y = \bar{X}$	$\frac{1}{\sqrt{N}} S_X$	
$Y = \frac{\bar{X}}{1+\bar{X}}$	$\frac{S_X}{\sqrt{N\left(1+\bar{X}\right)^2}}$	
$Y = \left(\bar{X}\right)^2$	$\frac{2\bar{X}}{\sqrt{N}} S_X$	
$Y = \sqrt{\bar{X}}$	$\frac{S_X}{2\sqrt{N\bar{X}}}$	
$Y = \ln(\bar{X})$	$\frac{S_X}{\sqrt{N}\bar{X}}$	
$Y = e^{\bar{X}}$	$\frac{e^{\bar{X}}}{\sqrt{N}} S_X$	Approximation could be poor if N is small.
$Y = \frac{100 S_X}{\bar{X}}$	$\frac{Y}{\sqrt{2N}}$	Assume that X is normally distributed. See Reference [7].

Table 8 — Standard deviations for functions of two variables

Function Y of $\bar{X}$, $\bar{Z}$	Standard deviation of Y
$\bar{X}$ and $\bar{Z}$ are averages of N measurements	S_X = standard deviation of X S_Z = standard deviation of Z S_{XZ}^2 = covariance [a] of X, Z
$Y = A\bar{X} + B\bar{Z}$	$\frac{1}{\sqrt{N}}\sqrt{A^2 S_X^2 + B^2 S_Z^2 + 2AB \times S_{XZ}^2}$
$Y = \frac{\bar{X}}{\bar{Z}}$	$\frac{\bar{X}}{\sqrt{N}\bar{Z}}\sqrt{\frac{S_X^2}{\bar{X}^2} + \frac{S_Z^2}{\bar{Z}^2} - 2\frac{S_{XZ}^2}{\bar{X}\bar{Z}}}$
$Y = \frac{\bar{X}}{\bar{X}+\bar{Z}}$	$\frac{Y^2}{\sqrt{N}\bar{X}^2}\sqrt{\bar{Z}^2 S_X^2 + \bar{X}^2 S_Z^2 - 2\bar{X}\bar{Z} S_{XZ}^2}$
$Y = \bar{X} \times \bar{Z}$	$\frac{Y}{\sqrt{N}}\sqrt{\frac{S_X^2}{\bar{X}^2} + \frac{S_Z^2}{\bar{Z}^2} + 2\frac{S_{XZ}^2}{\bar{X}\bar{Z}}}$
$Y = \left(\bar{X}\right)^a \left(\bar{Z}\right)^b$	$\frac{Y}{\sqrt{N}}\sqrt{a^2 \frac{S_X^2}{\bar{X}^2} + b^2 \frac{S_Z^2}{\bar{Z}^2} + 2ab\frac{S_{XZ}^2}{\bar{X}\bar{Z}}}$

[a] Covariance term is to be included only if there is a reliable estimate.

8 Example — Type A evaluation of uncertainty from a gauge study

8.1 Purpose and background

The purpose of this case study is to demonstrate the evaluation of uncertainty for a measurement process with several sources of uncertainty. The measurements in question are resistivities (Ω·cm) of silicon wafers. The intent is to calculate an uncertainty associated with the resistivity measurements of approximately 100 silicon wafers that were certified with a 4-point probe wired in a specific configuration, called configuration A, according to ASTM Method F84, which is the defined reference for this measurement. The reported value for each wafer is the average of six short-term repetitions made at the wafer centre. The measurements were made at the National Institute of Standards and Technology (NIST) with Probe No. 2362, which is one of five NIST probes capable of the measurements.

The uncertainty evaluation takes into account the following time-dependent sources of variability:

a) short-term effects from measurements at the centre of the wafer;

b) day-to-day effects;

c) run-to-run effects;

and the following possible sources of bias:

— bias due to Probe No. 2362;

— bias due to wiring configuration A.

8.2 Data collection and check standards

8.2.1 The certification measurements themselves are not the primary source for estimating time-dependent uncertainty components because they do not yield information on day-to-day and long-term effects. The three time-dependent sources of uncertainty are estimated from a 3-level nested design:

— J = 6 measurements at the wafer centre;

— K = 6 days;

— L = 2 runs.

The model for the 3-level nested design is

$$Y_{lkj} = \mu + \gamma_l + \delta_{lk} + \varepsilon_{lkj},$$

where

$l = 1,2$, $k = 1,\ldots,6$, and $j = 1,\ldots,6$.

8.2.2 The experiment is replicated on each of M = 5 wafers chosen at random, for this purpose, from the lot of wafers. These check standards are identified as wafers 138, 139, 140, 141 and 142 in the analysis. The experiment is also replicated over Q = 5 probes, which are identified as Probe No. 1, No. 281, No. 283, No. 2062, No. 2362 in the analysis. The data include

— run number,

— wafer identification,

— month and day of measurement,