CCQM Guidance note: Estimation of a consensus KCRV and associated Degrees of Equivalence

Version: 10

Date: 2013-04-12

Status: Released for reference

Preface

This document provides guidance on the choice and calculation of Key Comparison Reference Values (KCRVs) for CCQM Pilot and Key comparisons [1] for which no independent KCRV is available. It implements the principles set out in the CCQM paper "Data Evaluation Principles for CCQM Key Comparisons" [2].

The document is not intended as a prescriptive or mechanical process. Rather, it provides guidance on the selection of appropriate calculation methods for a KCRV. It is based on the idea that working groups should generally discuss study results in the light of both measurement science and statistical information, with both being given due weight and involving appropriate expertise.

The general approach to data analysis involves

- Selection of the set of participants whose results are to be considered in forming the KCRV. This set may include all participants, or a qualified subset selected by the working group. Selection of qualified participants can only be made by the working group. This document does not advise on that decision, although CCQM-09-03 does set out some principles that working groups are expected to consider.
- Review of the reported data, both to facilitate discussion of the results and to identify the main features of the data set. In particular, this review considers whether results can be considered mutually consistent, or whether the KCRV calculation needs to allow for the presence of unexpectedly extreme values or unexpectedly large dispersion of reported values. Statistical checks are recommended to assist the discussion, but are expected to be viewed in the light of all available information pertaining to the study materials, methods used, participant experience and any other relevant information.
- Selection of a KCRV calculation method that is appropriate for the particular data set and, where required, allows the calculation of degrees of equivalence.
- Calculation of the KCRV with associated uncertainty and degrees of equivalence.

These outline steps are intended to fit well into the planning, Draft A and Draft B stages in CCQM comparisons.

This document is written in part to guide the data analyst(s) responsible for advising a working group on statistical issues for a particular study, and some of the details of implementation consequently assume expertise in the analysis of data from inter-laboratory studies, including sufficient knowledge of the statistical methods used.

Contents

Estimation of a consensus KCRV and associated Degrees of Equivalence

1 Introduction

This document provides guidance on the choice, calculation and use of Key Comparison Reference Values based on consensus of reported results in CCQM Pilot and Key comparisons within the scope of the CIPM MRA [1]. It implements the principles set out in the CCQM paper "Data Evaluation Principles for CCQM Key Comparisons" [2].

2 Scope

This guidance is applicable to the calculation of a single KCRV from a set of reported laboratory results relating to measurement of a measurand, namely, a specific property of a material under consideration. The results from each laboratory constitute a measured value* and an associated standard uncertainty or an expanded uncertainty with stated coverage factor. It is assumed throughout that the intent is to obtain the best available estimate of the value of the measurand, taken as the KCRV, and its associated uncertainty.

Note: The KCRV and associated expanded uncertainty define an interval that will not usually encompass all reported values.

The guidance provides information on procedures used to provide the KCRV, evaluate the associated standard uncertainty, and calculate the degrees of equivalence (DoEs). Correlation associated with the KCRV and individual laboratory measured values is taken into account in providing DoE uncertainties.

Note: A degree of equivalence has two components, a value and an associated uncertainty at the 95 % level of confidence.

It is assumed that most (if not all) the deviations from the KCRV can be regarded as outcomes of a Normal distribution, with the remainder having possibly extreme values, or that the data have been appropriately transformed to achieve underlying Normality. It is further assumed that uncertainties are reported in accordance with the Guide to the Expression of Uncertainty in measurement (the GUM) [3].

This guidance does not describe

- detailed methods for review of data on multiple measurands;
- graphical methods of KC data analysis;
- use of information on stability of the materials used.

Note: The coordinator is normally responsible for setting transport and storage conditions that assure the stability of test materials and if necessary confirming that test materials do not change during the course of the study.

3 General approach

This document is based on the following general approach:

1. The working group identifies which laboratories should be considered as candidates for inclusion in the KCRV calculation (the 'candidate set') and provides any additional information to be taken into account in preliminary data analysis. In particular the working group advises whether there is good

* "measured value" (or "measured quantity value") is the term used by the VIM for a single value reported by a laboratory and representing a measurement result, and which may be accompanied by an uncertainty. In this document, the unqualified word "value" refers to a "measured value" as defined by the VIM. To emphasise that the value in question is reported by a study participant, "reported value" is also sometimes used.

reason to expect general consistency of the set of measured values, having regard to the reported uncertainties, for the candidate set and for the particular materials used.

2. Graphical and numerical tests for consistency are performed, and an initial screening carried out in which any apparently anomalous results are identified and checked (for example for transcription or other remediable errors).
3. One or more candidate KCRVs and associated standard uncertainties are calculated using accepted procedures appropriate to the assumptions in force and supported by the data.
4. The working group reviews the results of the initial screening and candidate KCRV(s) and approves the final selection of laboratories to be used in the calculation of the KCRV together with the method of calculation.
5. The KCRV and its associated standard uncertainty are obtained for the selected laboratories, DoEs calculated (including allowance for correlation) and the results incorporated in the Draft B report.

This general approach is intended to inform a decision by the working group on the KCRV and its associated uncertainty. Like the process of uncertainty evaluation, this should neither be seen as a mechanical process driven by simple statistical testing, nor as a decision based entirely on chemical or biological measurement knowledge in which statistical inference plays no part. Rather, the decision should arise from an informed debate that involves all necessary expertise, including expertise in measurement science, in statistics, and in chemistry and biology.

4 Terminology

Terms and definitions used in this guidance document generally follow those in the VIM or in appropriate statistical standards and texts. Terms unique to this guidance document are listed below. Note that the descriptive text is explanatory and should not be taken as a formal definition for each term.

Data set	Set of all measured values and uncertainties reported in a given study
Qualified participant	Participant considered, prior to evaluation of the results, as a candidate for inclusion in the calculation of the KCRV
Candidate set	The set of reported results for a particular measurand (including uncertainties) from qualified participants. Note: The 'Candidate set' may include all the reported results.
Candidate KCRV	A value that could reasonably be considered as a possible KCRV given the assumptions applicable to the data.
Outlier	Any reported value which appears unexpectedly distant from the majority of values or from a candidate KCRV, taking the associated uncertainties into account. Note 1: This usage is intentionally broad, and covers values that appear extreme on visual inspection as well as any identified by particular statistical tests. Note 2: In this document, describing a reported value as an 'outlier' does not of itself imply any judgement about the merit of the particular value. Inevitably, however, outlying values will merit careful attention because they are unexpected. This is discussed further in section 5.
Anomalous value	A value that appears unusual for any reason, whether or not it is an outlier. **Note**: 'anomalous' does not mean 'wrong'. Seriously erroneous values would of course usually appear as anomalous, but correct values in a population of poorer measurements might also appear anomalous on first inspection. The data set as a whole might also exhibit anomalies such as a general lack of agreement, unexpected asymmetry in the distribution of the values, evidence of a laboratory bias across several measurands etc. The emphasis should be on identifying,

discussing, understanding and as far as possible resolving such anomalies.

A list of symbols used in this text can be found in Appendix 3.

5 Procedure

5.1 Identifying Qualified participants

The working group should agree which laboratories should be considered as candidates for inclusion in the calculation of the KCRV for each material and measurand studied. This set may include all participants in the study, or a subset of the participants. The chosen laboratories are the 'qualified participants', and the results from this set of laboratories are the 'candidate set'.

Selection (if any) should be on the basis of demonstrable track record in the relevant field of measurement.

Note 1: For Key Comparisons, the qualified participants must (following CIPM guidance [4]) all be eligible for inclusion in Key Comparisons.

Note 2: It is not guaranteed that all named candidates will eventually be used, as subsequent investigation may show inconsistencies within the candidate set that indicate a need for reduced weighting or even exclusion. However, laboratories excluded at this stage will *not* be used in any part of KCRV calculation. They will, however, be included in data summaries and plots and DoEs will be calculated for them after calculation of the KCRV.

Note 3: If there is only one qualified participant, the measured value and the associated uncertainty from that laboratory are taken as the KCRV and the KCRV uncertainty and the remainder of this guidance note does not apply. (This provision is, of course, only useful if there are additional participants whose results are not in the 'candidate set').

Note 4: Identification of the qualified participants can in principle take place prior to the circulation of test materials but the qualified set should nonetheless be reviewed and confirmed following receipt of results.

5.2 Screening the data for consistency and anomalous values

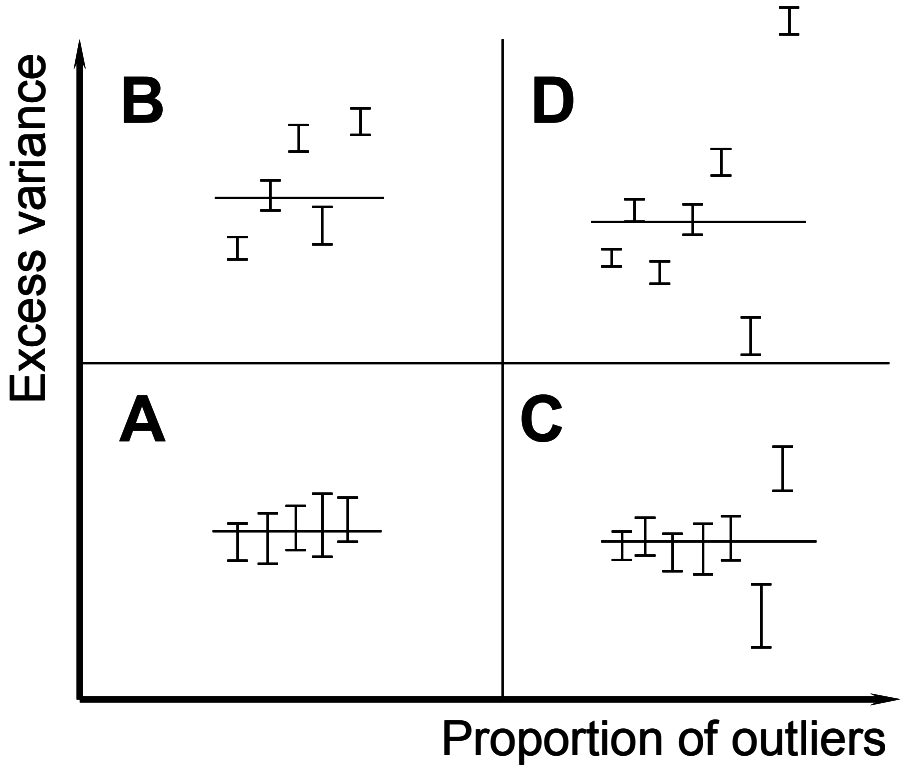
5.2.1 Preliminary inspection

5.2.1.1 The candidate set of values x_i and associated standard uncertainties $u(x_i)$ should be inspected using appropriate graphical methods. The combination of graphical methods used should be capable of identifying:

- Individual values or subsets of the complete set of values whose location appears inconsistent with the majority;
- Reported uncertainties that are unusually large or small;
- Values that deviate substantially relative to their reported standard uncertainties. For example, a plot of $[x_i - \mathrm{med}(x)]/u(x_i)$ quickly identifies points that are far from the median, med(x), relative to their reported uncertainties.

Figure 1 shows some of the main features of metrology comparison data. In addition to identification and checking of individual anomalies, one of the aims of the preliminary inspection is to establish which of the four regions depicted in Figure 1 best represents the data for a particular measurand. Establishing such a region will be important in selecting the method for calculating the KCRV.

Figure 1: Features of metrology comparison data



The figure schematically illustrates two of the main features of importance in metrology comparison data. Vertical bars denote approximate 95 % coverage intervals or expanded uncertainty intervals. A: mutually consistent data; B: Evidence of general over-dispersion (or understatement of uncertainty) affecting most or all participants; C: Generally consistent results with a small number of outlying values: D: General over-dispersion combined with some particularly extreme values.

5.2.1.2 Graphical inspection may be supported by outlier tests, for example, Grubbs' tests, Dixon's test, or tests based on non-parametric or robust statistics. A simple test based on robust statistics when reported uncertainties are essentially identical is to calculate a robust estimate of location $\hat{\mu}$ and dispersion $\hat{\sigma}$, and to consider values as extreme when outside $\hat{\mu} \pm 2\hat{\sigma}$ (corresponding to approximately 95 % confidence). Robust methods are also available for the determination of a robust weighted mean and associated scale parameters, and can be applied where reported uncertainties differ appreciably. A common non-parametric indicator of an outlier, used by default in most box plots [5], is a result outside the inter-quartile interval $[Q_1, Q_3]$ by more than $1.5(Q_3 - Q_1)$[*]. At this stage, any outlier tests should be carried out at approximately the 95 % confidence level, the aim being primarily to check that closer inspection of a visually identified outlier is justified.

5.2.1.3 Where the data set under consideration forms part of a wider study in which several measurands are involved, or several test materials are studied, graphical or other methods for detecting consistent bias across a set of measurands or different materials should be used and taken into account when identifying anomalous values.

[*] This non-parametric outlier indication corresponds to approximately 99 % confidence for large normally distributed data sets, but identifies a higher proportion of identified outliers for data sets of size 5 to 20.

5.2.1.4 It is important to take due account of all available information when screening data, including, for example, knowledge of the measurement methods used in each laboratory and relevant chemical information.

5.2.1.5 Paragraphs 5.2.1.1 to 5.2.1.4 refer only to the candidate set, since outlier tests and some graphical methods (for example, box plots) may be adversely affected by including data other than those from qualified participants. However, it is recognised that it is important to inspect the whole data set carefully, including any data that are not part of the candidate set, in order to refer any additional anomalies to the laboratories or to inform interim reports. This process is usually carried out as part of the preliminary screening.

5.2.1.6 Following inspection, anomalies should be followed up to the extent permitted by current rules for the conduct of key comparisons, and any identifiably erroneous results either corrected or removed from the candidate set.

Note 1: The resulting data set will be a possibly reduced candidate set that may still display unresolved anomalies, including outliers, over-dispersion, or both.

Note 2: The fact that a value appears anomalous is not of itself sufficient to justify removal from the candidate set at this stage, whether or not it is identified as anomalous using a statistical outlier test. 'Identifiably erroneous' indicates that a substantiated error in procedure or reporting has been found or that the participant concerned has chosen to withdraw the value and associated uncertainty from consideration for the KCRV.

5.2.2 Consistency check

5.2.2.1 A check of mutual consistency should be performed to assist selection of the KCRV calculation method. An example of a common consistency check is provided as Appendix 1.

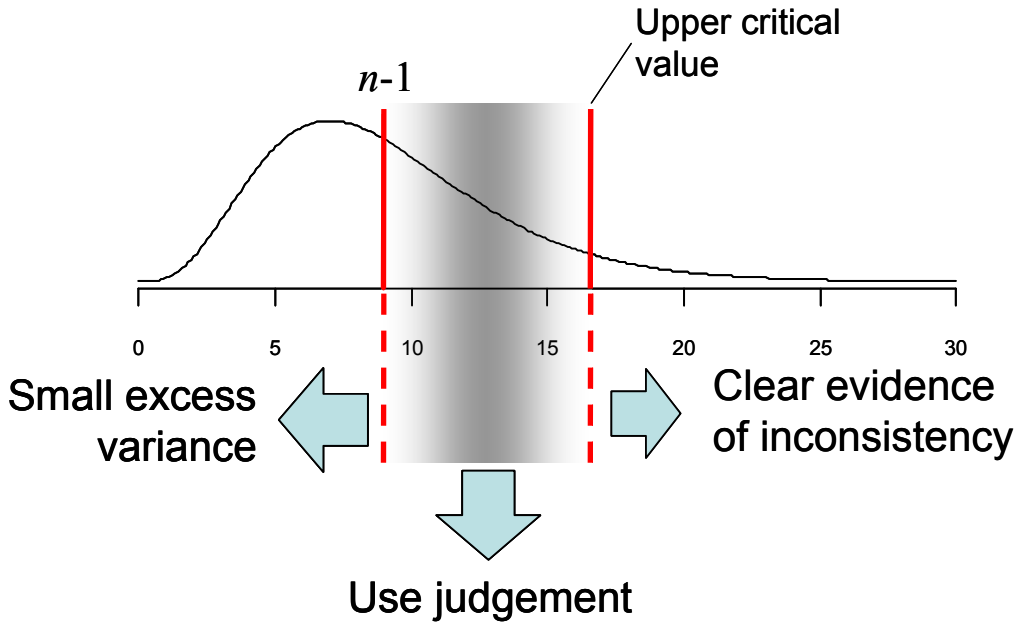
Note: The test given in Appendix 1 tests the hypothesis that laboratories share a common mean value and that the deviations from that value are normally distributed with mean 0 and standard deviation u_i.

5.2.2.2 It is important to treat consistency checks — including that in Appendix 1 — as advisory rather than as simple decision criteria. Figure 2 indicates a typical interpretation of the chi-squared test, taking n as the number of values included in the test. Below a chi-squared value of $n - 1$ there is no reason to believe there is inconsistency. Above the chosen critical value (usually the 5 % upper tail value) there is apparently clear evidence of inconsistency (but see Appendix 1 for additional comment). Between the two, there is no strong reason to believe there is appreciable inconsistency, but — particularly for a small data set — the chi-squared test result does not rule out the presence of potentially important over-dispersion. Particularly in the mid-range, therefore, it is important to consider all available information before reaching a decision about the method used to calculate the KCRV. In doing so, the working group should consider the following:

5.2.2.3 Evidence that would support a decision to treat the data as mutually consistent includes:

- Mutual consistency for most of the participating laboratories in at least three prior interlaboratory comparisons on closely similar materials.
- Evidence of consistently good performance across a broad range of prior comparisons within the working group's remit.

Figure 2: Indications from a chi-squared test



The figure shows a schematic illustration of a chi-squared distribution and the two most commonly used decision points. The particular distribution shown is for $n = 9$ measured values; the axis shows values of chi-squared.

5.2.2.4 The following circumstances make inconsistency more likely:

- Fewer than three prior interlaboratory comparisons involving the same measurand at similar levels in closely similar materials;
- Measurements on a substance not previously included in a Pilot or Key Comparison;
- Measurements on materials of unknown homogeneity;
- Measurements at concentrations not previously subjected to interlaboratory comparison by the working group;
- Application of new measurement methods by a substantial proportion of participants.

If any of these circumstances apply, it may be unsafe to treat the results as mutually consistent, and in particular to assume that the reported uncertainties can be taken as including all the effects that influence the measured values.

5.2.2.5 A working group may adopt a general policy in regard to the interpretation of consistency (or otherwise) based on experience of a range of materials over a period of time.

5.2.3 Validity of assumptions

5.2.3.1 Many of the KCRV calculation methods rely to a greater or lesser extent on assumptions of underlying normality (either for the data set as a whole or for the distributions associated with reported uncertainties) and large degrees of freedom for reported uncertainty. Where they are important to the working group's preferred method(s) of KCRV calculation, these assumptions should be checked to the extent possible.

5.2.3.2 Degrees of freedom should normally be reported. Where they are not immediately available, reported coverage factors should be reviewed. Any coverage factor greater than 2.0 for an assumed 95 % level of confidence should be regarded as evidence of limited degrees of freedom. Any coverage factors

over 2.2 (corresponding to approximately 11 degrees of freedom) should be regarded as evidence of low degrees of freedom.

Note: It is assumed that laboratories have considered and incorporated degrees of freedom for type B evaluations of uncertainty in accordance with the GUM. In particular GUM clause G.4.2 suggests a method for assigning finite degrees of freedom for Type B evaluations involving judgements which are themselves subject to doubt.

5.2.3.3 The distribution assumptions associated with reported uncertainties can be checked by contacting participants if necessary. Usually, however, provision of a single expanded uncertainty implies a symmetric distribution that can usually be assumed to be approximately normal or *t*-distributed. Asymmetric reported intervals indicate non-normality.

5.2.3.4 With the exception of the presence of outlying values, departures from normality for the data set as a whole are hard to detect using normality tests because the statistical power of such tests is inadequate for the relatively small data sets in metrology comparisons. Visual inspection is accordingly recommended.

Note: Where reported uncertainties differ appreciably, measured values are not generally expected to be normally distributed. Rather, the scaled deviations $(x_i - \hat{x})/u(x_i)$ where $\hat{x}$ is a candidate KCRV, should be approximately normally distributed.

5.2.3.5 Where serious departures from the assumptions above are detected, action should be taken to address the issue. In particular, KCRV calculation methods should be chosen for robustness to departures from these assumptions, and statistical tests (including chi-squared and outlier tests) should be used with caution.

5.2.3.6 If an initial consistency check shows evidence of remaining inconsistencies, or if there remains a risk that additional factors are contributing to the dispersion and it is considered unsafe to assume mutual consistency, the features of the data set responsible for failure of the consistency check should be determined. These features are usually identified on the basis of graphical inspection with the assistance of outlier checks, possibly supplemented by repetition of consistency checks after removal of clear outliers.

The two most relevant features are:

- The continued presence of a small number of extreme values.
- Apparently general over-dispersion; that is, evidence that most laboratory uncertainties are insufficient to account for the observed dispersion of results.

Figure 1 shows these features schematically.

Both features may be present. If outliers appear to be present, checking for consistency after removal of clear outliers will help to establish whether the majority of the results are mutually consistent.

5.2.3.7 If there is evidence of inconsistency other than the presence of a small number of extreme values, the working group should consider whether calculation of a KCRV remains justifiable. If there is evidence of severe inconsistency (for example, no overlap in reported expanded uncertainty intervals or several discrepant subsets) it is normally considered prudent to abandon the attempt to assign a KCRV and to undertake further investigations of the cause.

6 Selection of methods for KCRV calculation

6.1 General criteria for the selection of methods for KCRV calculation

6.1.1 Methods used for KCRV calculation should have the following characteristics:

i) Well-characterised theoretical performance. At least the following performance characteristics should have been established, preferably from theoretical considerations:

- Bias
- At least large-sample variance characteristics
- Asymptotic relative efficiency for a normal distribution (see Note below)
- Breakdown point (see Note below).

ii) Known performance on data typical of Key and Pilot Comparisons. For example:

- Performance on smaller data sets (of size 3 to 30) should have been established;
- Performance in the presence of extreme values should have been established, including performance on sets with extreme values regarded as drawn from an asymmetric distribution if appropriate;
- Comparison with any previously used methods meeting similar assumptions is useful.

iii) Broad scientific acceptance, usually shown by prior publication in appropriate refereed statistical journals.

6.1.1 A selection of estimators currently considered to meet the above criteria is provided in Appendix 2.

Note: The asymptotic relative efficiency of a location estimator means, here, the large-sample variance of the estimator divided by the variance of the mean, which is the minimum variance estimator for the normal case. High efficiency indicates low variability and correspondingly low uncertainty. "Breakdown point" for an estimator describes the proportion of values that can move to infinity before affecting the estimate. High breakdown point corresponds to high resistance to outliers.

6.2 Factors affecting choice of a specific estimator

6.2.1 Location estimators fall into four broad classes, according to whether or not they are resistant to the effect of outliers (robust) and whether or not they use uncertainty information in the estimation of location. The appropriate class of estimator to use is therefore guided by two features of the data set identified during preliminary inspection:

i) The presence or likelihood of outliers. If outliers are present or likely, an estimation procedure that is robust to the presence of outliers should be used.

ii) The reliability of the reported uncertainties for the majority of participants, usually indicated by the degree of consistency found. If reported uncertainties are considered reliable and the majority of the data are mutually consistent, estimators that weight measured values according to their reported uncertainties should be used.

Estimators chosen to match the properties of the data set are regarded as valid estimators for the purpose of KCRV calculation.

Table 1 lists a selection of valid estimators, grouped by applicability based on the features above. Section 6.3 discusses the four different classes of estimator in more detail and compares estimators within each class. Appendix 2 gives details of calculations, uncertainty evaluation and properties of some useful estimators.

6.2.2 Additional factors that motivate the choice of a specific estimator include:

- The uncertainty associated with the estimator (related to the efficiency);
- The degree of resistance to extreme values (breakdown point and bias given extreme values regarded as drawn from an asymmetric distribution);
- Simplicity of application and presentation.

6.2.3 Preference should normally be given to valid estimator(s) having the smallest *theoretical* uncertainty (equivalent to best efficiency). Appendix 2 provides information on the efficiency of various estimation methods.

Note Calculated uncertainties derived from a given data set do not necessarily reflect the order of preference. By way of illustration, given a choice between the median and the mean and assuming both are otherwise equally valid choices for a particular data set, it is entirely possible that the scaled median absolute deviation of this particular data set (often used as a basis for the uncertainty associated with the median) might be appreciably smaller than the standard deviation used as the basis for the uncertainty of the mean. Despite this, the mean remains the preferred estimator because its theoretical variance (and therefore efficiency) is by far the smaller of the two.

6.2.4 Where different valid approaches provide similar values for location and similar values for the associated uncertainty, the simplest approach is normally preferred.

6.2.5 The choice of KCRV estimator should also permit the determination of degrees of equivalence. A desire for informative DoE uncertainties may restrict the choice of estimator. For example, KCRV estimators that do not use reported uncertainties often lead to degree-of-equivalence uncertainties that do not reflect the differing uncertainties reported by participants.

6.3 Recommended estimators for the KCRV

Recommended estimators for each of the four main scenarios in Table 1 are listed below. In these paragraphs, *m* denotes the number of laboratories accepted for KCRV calculation (the candidate set after any adjustments by the working group).

Note: These recommendations should not be taken as a requirement to apply one and only one estimator. It is often helpful to review the values and associated uncertainties returned by several nominally valid estimators operating on different principles. Appreciable differences between different estimators can help to identify the features of the data set that are responsible for these differences and inform the final choice of estimator. Where several estimators agree well, confidence in the KCRV calculation is improved. The recommended estimators indicated below, however, are expected to provide the best efficiency within their domain of applicability and should normally be preferred where there are differences between KCRVs.

6.3.1 Estimators for mutually consistent results

6.3.1.1 Essentially all estimators listed in Table 1 and Appendix 2 are applicable to mutually consistent results. However, some estimators are uniquely applicable to consistent data and should not normally be applied where anomalous values exist or where inconsistency is found or suspected. The uncertainty-weighted mean, also widely known as the Graybill-Deal estimator, is the most important of these.

6.3.1.2 The uncertainty-weighted mean weights the measured values by the reciprocals of the squared standard uncertainties. Where the reported uncertainties are consistent with the observed dispersion in the data set (that is, there is no over-dispersion), the uncertainty associated with the resulting KCRV involves only the reported uncertainties.

6.3.1.3 For validity, this procedure relies on large degrees of freedom associated with the reported uncertainties, and the absence of additional effects resulting in over-dispersion. Under these circumstances, the weighted mean is the recommended estimator.

6.3.1.4 Where reported uncertainties are closely similar, the arithmetic mean and the uncertainty-weighted mean are equivalent; the arithmetic mean may then be used as the location estimate.

Note: The standard uncertainty associated with the arithmetic mean is usually based on the standard deviation associated with the reported values. This uncertainty may differ from that of the weighted mean, which is based on reported uncertainties. For instance, the dispersion of reported values might by chance be substantially lower than expected from the reported uncertainties, resulting in too small a value for the KCRV uncertainty. For this reason, the uncertainty-weighted mean and its associated uncertainty should normally be used wherever the results are demonstrably consistent, and especially where $\chi^2_{\text{obs}} < m - 1$ (see section 5.2.2).

6.3.1.5 The uncertainty-weighted mean, with associated standard uncertainty set to $\sqrt{1/\sum_{i=1}^{m} w_i}$, should *not* be used if there is appreciable risk of significant over-dispersion (see section 5.2.2 and prior assumptions); that situation should be treated as in section 6.3.3.

Note: Zhang [6] recommends a modified estimate and associated uncertainty which are preferred where any participant uncertainties are associated with small degrees of freedom.

Recommended estimator: Uncertainty-weighted mean.

Note 1: Where reported uncertainties are very similar, and the arithmetic mean provides similar location and uncertainty estimates to the weighted mean, the arithmetic mean may be reported as the KCRV.

6.3.2 Mutually consistent results with some outliers present

6.3.2.1 CCQM-09-03 paragraph 5 notes that "… some values, through human error or unexpected chemical or sample effects, might be discrepant." This principle, which is based on prior experience in measurement and not on statistical considerations, is interpreted here as indicating that unexplained anomalous values must be regarded as possible erroneous results. Most identifiable anomalies appear as outliers in the final data set. Outliers cause deviations in location estimates and have severe effects on dispersion estimates, in turn affecting the resulting KCRV uncertainty. With an acknowledged risk of human or other error, unexplained extreme values should normally be given less weight in the calculation of the KCRV.

6.3.2.2 Working groups may direct that no measures may be taken to reduce the adverse effects of outlying values. In that case, the presence of outlying values should be taken as strong evidence of over-dispersion and the methods of section 6.3.3 should be applied.

6.3.2.3 Where the working group agrees that the provisions of CCQM-09-03 paragraph 5 should apply, two general types of procedure are available: i) Procedures based on automatic outlier detection and removal ('Outlier rejection procedures'), and ii) the use of robust statistics. The advantages and disadvantages of each are discussed below.

Note: At this stage, outlier testing and rejection or robust statistics are used to reduce the risk of *undue* influence from extreme values. By their nature, these procedures will inevitably remove or downweight some valid results; this is the price paid for increased resistance to possible error. For this reason, removal or downweighting of a value as part of a robust estimation procedure should not be taken to imply that the values so treated are excluded from calculation of the KCRV.

Outlier rejection

6.3.2.4 Automatic outlier rejection (that is, exclusion of extreme values based purely on statistical criteria) can be considered as a robust estimation technique, and for this reason as a valid procedure for calculation of a KCRV when mutual consistency is compromised by a small number of extreme values.

6.3.2.5 Outlier rejection procedures have the advantage of comparative simplicity once outliers are removed, because they permit the application of comparatively simple KCRV calculation methods. Disadvantages include: i) outlier rejection based on tests at low confidence levels (95 % and below) reject a substantial proportion of valid entries, causing estimates of dispersion and therefore the KCRV uncertainty to be biased low; ii) Repeated outlier testing and rejection (or some other methods for identifying multiple outliers) is often necessary to identify all extreme values, and this procedure may result in an unreasonably high rejection rate and ultimately conceal real inconsistency; iii) the most common outlier tests do not take account of uncertainty information. Working groups should therefore consider outlier rejection carefully before use.

6.3.2.6 Taking the above factors into account, if a working group adopts the use of outlier rejection for calculation of a KCRV, the following principles should be observed:

i) If used, automatic rejection should be applied at confidence levels of 99 % or above. This choice provides reasonable protection against gross error while minimising adverse effects on uncertainty evaluation.

ii) Repetitive outlier testing and rejection should be used if necessary to identify multiple extreme values, but no more than 20 % of the values in a data set should be rejected (note that this constrains the breakdown point to 20 %).

iii) Where reported uncertainties differ appreciably, outlier tests should take account of these uncertainties.

Robust statistics

6.3.2.7 Robust statistics allow explicitly for the presence of extreme values and (usually) accommodate them by assigning weights that decrease with distance from the body of data ('down weighting'). They have been recommended for general application in analytical chemistry [7], particularly in connection with interlaboratory study [8, 9], and are in wide use. The early estimators proposed for analytical chemistry did not use uncertainty information, but the statistical literature has long treated this as a special case of estimators that do use uncertainty information [10]; there are therefore many well-characterised estimators that are both robust to the presence of outliers and use uncertainty information appropriately. The most useful class of robust statistics for calculation of a KCRV is the class known as M-estimators, which include the median and arithmetic mean as particular cases.

6.3.2.8 Robust estimators take due account of down-weighting when calculating dispersion and associated uncertainties, do not require removal of values from the candidate set, and accommodate marginal outliers appropriately, overcoming many of the disadvantage of excluding extreme values. They have well-characterised efficiency, which can be chosen explicitly for most M-estimators and is usually set to 85 % to 95 % depending on the degree of resistance required to extreme values. Their breakdown point is usually substantially higher than for outlier rejection procedures. Their principal disadvantages are i) comparative complexity in implementation and ii) sensitivity to estimates of dispersion, which can degrade their efficacy for very small data sets, particularly if several extreme values are present.

6.3.2.9 The sample median is a robust and simple estimator, often used in CCQM studies. Its breakdown point is high (50 %), but asymptotic efficiency for normally distributed data is very low (64 %) Uncertainties associated with the location estimate are therefore typically about 20 % to 25 % larger for the median than for other M-estimates. The sample median does not take reported uncertainties into account. Uncertainty estimation for the median usually uses a scaled median absolute deviation (MAD_E); this is very simple and robust, but again has very poor efficiency and is additionally biased low for data sets with $m < 10$. These factors make it hard to recommend the median over more sophisticated robust estimators given the priorities in paragraph 6.2.2.

Recommended estimators: Any well-characterised robust estimator with breakdown point of at least 20 % and efficiency of at least 85 % for the size of data set in question.

Note 1: Efficiency for the size of data set in question may differ substantially from the asymptotic efficiency usually available from theoretical considerations.

Note 2. Where the median and MAD_E provide closely similar values and uncertainties to the recommended estimators above, the reported KCRV may be based on the simpler estimates.

Note 3. Among robust estimators, the class known as MM-estimators offers high efficiency and breakdown point with minimal sensitivity to extreme values.

Note 4. Robust statistics should not normally be used on data sets of fewer than 7 values unless there is evidence to support their applicability. Below this, the median and MAD_E are likely to behave very nearly as well in the presence of moderate outliers, and outlier rejection followed by classical estimates can perform better than either for very small sets.

6.3.3 Lack of mutual consistency with no individual anomalous values

6.3.3.1 General inconsistency manifests as larger observed dispersion than can be accounted for by reported uncertainties – that is, over-dispersion or excess variance. Such dispersion can arise from understatement of uncertainties for most or all participants, or from the presence of genuine differences between items tested by each participant. For determining a KCRV, the principal effects are, first, that the validity of weighting based on reported uncertainties depends on the nature of the effect causing over-dispersion and, second, that reported uncertainties cannot be treated as appropriate for KCRV uncertainty evaluation.

6.3.3.2 The nature of over-dispersion affects the choice of estimator. The recommended approach is to decide, based on chemical knowledge and experience, the most likely form of the over-dispersion (proportional or fixed contribution) and act accordingly as set out below:

i) If the deviations resulting in over-dispersion are, to a reasonable approximation, proportional to reported uncertainty, weighting on the basis of reported uncertainty remains approximately valid, and it is sufficient to increase the calculated KCRV uncertainty by a simple scale factor. This is the basis of the correction for over-dispersion described in connection with the uncertainty-weighted mean in Appendix 2, which uses the Birge ratio. Under these circumstances, the uncertainty-weighted mean with KCRV uncertainty corrected for over-dispersion is appropriate. Note, however, that effects on chemical and biological measurements are rarely strictly proportional to laboratory uncertainties; simple scale factor increases in KCRV uncertainty are therefore best regarded as approximate adjustments only.

ii) If over-dispersion is attributable to a random factor, such as inhomogeneity of test materials, which operates on the same scale for all participants irrespective of their reported uncertainty, the combination of reported uncertainty and uncertainty associated with the additional random effect is no longer proportional to reported uncertainty. As the additional effect increases, the effective uncertainties of the reported values, taking the additional random effect into account, increase (and when the additional effect dominates they converge to the same uncertainty). Three situations can be distinguished:

   a) With small excess variance (for example, where the calculated Birge ratio is between 1.0 and 1.5), the uncertainty-weighted mean with a scale correction for over-dispersion is an appropriate estimator.

   b) With substantial inconsistency the effective weights are essentially equal, and the uncertainty-weighted mean converges to the arithmetic mean. The arithmetic mean is then the most appropriate estimator.

   c) At intermediate levels of over-dispersion, the most accurate representation involves estimation of the variance associated with over-dispersion, combination of that variance with the reported uncertainties, and recalculation based on the revised weights. This is usually an iterative process, but numerical methods with assured convergence exist. Implementations of this principle include, for example, the Mandel-Paule estimate [11], DerSimonian-Laird estimator [12] and the Ruhkin-Vangel restricted maximum likelihood estimate [13]. Note that the latter

is additionally valid for small degrees of freedom in the reported uncertainties and should be preferred where available. Toman and Possolo have provided an accessible illustration of the methodology [14].

Note: Approach c) is valid for all three situations, but a) and b) provide simpler and sufficient approximations at the extremes.

Recommended estimators:

Over-dispersion arising from unquantified effects proportional to reported uncertainty	Uncertainty-weighted mean with scale adjustment of the uncertainty
Over-dispersion arising from modest unquantified additive effect	Uncertainty-weighted mean
Over-dispersion arising from large unquantified additive effect	Arithmetic mean
Over-dispersion arising from unquantified additive effect	Mandel-Paule or Ruhkin-Vangel estimate

6.3.4 Lack of mutual consistency in addition to one or more anomalous values

6.3.4.1 The presence of a minority of extreme values together with either suspected or apparent mutual inconsistency of the remaining majority combines the features of the preceding two sections. The same considerations apply to the nature of the effect responsible for over-dispersion in the bulk of the data set as in section 6.3.3. Treatment, however, must accommodate outliers. Since there is no current implementation capable of modelling excess variance in the presence of outliers, only two scenarios are amenable to treatment. The recommended approaches are as follows:

i) When over-dispersion is, to a reasonable approximation, proportional to reported uncertainty, excluding extreme values at the 99 % level followed by uncertainty-weighted mean with correction for over-dispersion is applicable. For data sets of size 7 or greater, an uncertainty-weighted robust estimator, such as the MM-estimate, with uncertainty corrected for observed dispersion, is likely to perform at least as well, and will usually perform better for large sets. Note: The term 'MM-estimator' refers to a specific class of robust estimators suggested by Yohai et all [10] and should not be confused with the mixture model median, sometimes abbreviated as 'MM-median', which is not currently recommended for KCRV estimation pending a validated method of estimating uncertainty.

ii) When over-dispersion is attributable to a random additive effect, such as inhomogeneity of materials, which applies for all participants irrespective of their reported uncertainties, excluding extreme values at the 99 % level followed by the methods of section 6.3.3.2 ii) c) is the most generally applicable approach. A robust estimator applied without taking reported uncertainties into account is applicable for data sets of size 7 or larger. For data sets of size 6 or less, the median with uncertainty based on MAD_E provides similar or better overall performance.

Note: Where there is clear evidence of over-dispersion as well as a number of serious outliers, it is not usually appropriate to estimate a KCRV; rather, further investigations of the cause should be undertaken and the study repeated if necessary.

7 Degrees of equivalence

7.1 General

A degree of equivalence (DoE) is defined [1] as follows:

> The degree of equivalence of each national measurement standard is expressed quantitatively by two terms: its deviation from the key comparison reference value and the uncertainty of this deviation (at a 95 % level of confidence). The degree of equivalence between pairs of national measurement standards is expressed by the difference of their deviations from the reference value and the uncertainty of this difference (at a 95 % level of confidence).

Thus a DoE has two components: a value component and an uncertainty component.

7.2 Correlation associated with the KCRV and participants' measured values

7.2.1 When a consensus value is calculated, the consensus value and any participant value contributing to the estimate has correlation associated with them; if a participant value changes, so in general does the estimate. For small data sets, this effect can be large. CCQM-09-03 therefore requires that correlation associated with the KCRV and the reported participant values be taken into account in calculating DoEs using that KCRV.

7.2.2 Calculation of a particular participant's DoE generally involves the participant reported uncertainty, the KCRV uncertainty and a covariance term that takes account of the correlation associated with the KCRV and the participant value. The general calculations for DoEs are provided in Appendix 2.

7.2.3 The covariance term depends on the participant uncertainty and on the weight given to the participant value in calculating the KCRV. The appropriate weight depends on two factors. First, if the participant was not among the qualified participants used in obtaining the KCRV, the weight is set to zero. Second, the weight allocated depends on the estimator employed and, for robust estimators, on the reported value as well as on the reported uncertainty. Weighting functions for common estimators are given, together with covariance calculations, in Appendix 2. In general, the weights used are available from the software employed to calculate the KCRV. For members of the candidate set rejected using an outlier-rejection scheme, the weights should be set to zero for DoE calculation purposes.

7.3 Interpretation of degrees of equivalence and their uncertainties

7.3.1 The value components of the degrees of equivalence and, to a greater extent, their associated uncertainties, depend heavily upon the choice of KCRV calculation method. In all but the most straightforward case (perfect consistency among participants with large degrees of freedom and good evidence for the absence of undetected additional variance) the DoE uncertainties are not directly related to the reported laboratory uncertainties. Indeed, for KCRV estimation methods that do not make use of the reported uncertainties (including the simple mean and median) there is no relationship between the DoE uncertainty and an individual participant's reported uncertainty. For such KCRV calculations, therefore, the resulting DoEs and their associated uncertainties may be extremely unreliable indicators of laboratory performance even though the KCRV and its associated uncertainty may appear sensible.

7.3.2 In addition to the above, the KCRV uncertainty evaluation methods presented in Appendix 2 do not include any allowance for uncertainty associated with the particular choice of KCRV. Where several different KCRV calculation methods might reasonably be chosen for a given study, the interpretation of individual DoEs is inevitably more complex. In particular, a given DoE may show apparently significant inconsistency with one reasonable choice of KCRV calculation, whilst being apparently consistent with another.

7.3.3 Where the interpretation of the DoEs is appreciably affected by the issues in paragraphs 7.3.1 and 7.3.2, working groups should consider including appropriate cautionary notes in the KC report to discourage over-interpretation of the results.

7.4 Use of pairwise degrees of equivalence

7.4.1 The Technical Annex to the CIPM MRA [1] provides for the reporting and use of DoEs between pairs of institutes. Pairwise degrees of equivalence have the advantage that they do not depend on a particular choice of KCRV calculation; only on the laboratory results and reported uncertainties. Where the interpretation of DoEs with respect to the KCRV is difficult (for example where section 7.3 applies) it

may be useful to report pairwise degrees of equivalence in addition to, or instead of, the degrees of equivalence with respect to the KCRV.

8 Inhomogeneity

8.1 CCQM-09-03 states that:

> "Homogeneity effects should be taken into account by the coordinating NMI when determining a KCRV and its associated uncertainty. Homogeneity effects should be characterized by the coordinating NMI before distributing samples.
>
> NOTE 1. Homogeneity effects should be expressed in terms of a standard uncertainty u_{wb} for within-bottle effects and a standard uncertainty u_{bb} for between-bottle effects.
>
> NOTE 2. A similar statement applies to some other effects such as stability."

8.2 Where one unit per participant is distributed and inhomogeneity needs to be considered, the simplest approach is for the coordinator to include an additional term, equal to the estimated between-unit standard deviation s_{bb}, in the calculation of the standard uncertainty associated with the KCRV.

Note: This is equivalent to assigning an individual KCRV to each participant with a standard uncertainty appropriate for the unit(s) supplied. This standard uncertainty will be larger than that associated with the estimated true mean of the population of results.

8.3 Where multiple units are distributed to each participant and the measurand is the estimated average across units, treatment of inhomogeneity becomes far more complex. With multiple units per laboratory, the dispersion of measured values within the laboratory will normally increase, inflating uncertainties slightly. Some (but rarely all) participants may detect between-unit inhomogeneity and make explicit allowance for between-unit effects in their reported uncertainty. It then becomes impossible for the coordinator to make a reliable allowance for inhomogeneity for each laboratory without risking some double-counting of the effect of inhomogeneity.

8.4 In general, the complexity of possible inhomogeneity effects when multiple units are circulated makes it advisable to take all measures possible to minimise inhomogeneity and to design homogeneity tests with sufficient power to rule out significant inhomogeneity wherever possible.

9 Application to Pilot studies

9.1 The principles of this document apply both to Key Comparisons and to Pilot Studies, although the term 'KCRV' would normally be replaced by 'reference value' or 'assigned value' in a Pilot comparison. However, there are important differences between pilot and key comparisons that may affect the choice of estimators used:

- Pilot studies are usually exploratory, involving new measurement methods or more challenging measurement problems.
- Pilot studies often involve laboratories that are not NMIs.
- Pilot studies often involve participants in the early stages of implementing a measurement technique, or wishing to add a new technique to their established capabilities.
- Degrees of equivalence are not generally calculated in pilot studies.

These features make pilot studies much more likely to show unexpected results, either for individual participants or for all participants.

9.2 The increased likelihood of unexpected values coupled with the reduced need for degrees of equivalence makes it more appropriate to use robust estimation methods for pilot studies. Robust estimation methods as described in section 6.3.2 are therefore recommended for pilot studies.

Table 1: Choice of Estimators and Uncertainty Evaluation Methods for Consensus Assignment in CCQM Studies [a]

Consistency[b]	Presence of outliers[c]	
	No extreme values	**One or more extreme values**
There is no evidence of significant inconsistency in the bulk of the data set and the working group concludes that it is safe to assume that there is no source of over-dispersion. *Note: This situation is rare in practice; it is usually safer to assume some undetected between-laboratory effects.*	If uncertainties do not differ significantly: ▪ Arithmetic Mean ▪ Uncertainty-weighted mean (The two are comparable for consistent data with approximately equal uncertainties) If uncertainties differ appreciably: ▪ Uncertainty-weighted mean *Note: Approaches valid for data contaminated by outliers but otherwise consistent are also valid for this case, but will generally result in slightly or appreciably larger uncertainties.*	If uncertainties do not differ significantly: • Mean, after rejection of some outliers, q, say, in number, identified at the 99 % level, with standard uncertainty estimated as $s/\sqrt{(m-q)}$ • Huber (H15) or other M-estimates without prior weights • Median If uncertainties differ appreciably: • M-estimators, including those using Huber, Hampel and bisquare weighting functions, with accommodation for prior weights based on reported uncertainties • MM-estimates with prior weights based on reported uncertainties
There is insufficient evidence of mutual consistency, or evidence of significant inconsistency[d]	If uncertainties do not differ appreciably: • Arithmetic mean If uncertainties differ significantly: • Uncertainty-weighted mean with correction for over-dispersion • Mandel-Paule, Vangel-Ruhkin or equivalent weighted-mean estimates of location and uncertainty *Note: The arithmetic mean and Mandel-Paule approaches converge as inconsistency becomes large.*	If uncertainties do not differ significantly: • Arithmetic mean of outlier-rejected data • M- or MM-estimates (including Huber, Hampel and bisquare weighting functions) without prior weights • Median If uncertainties differ appreciably: • M-estimators, including Huber, Hampel and bisquare weighting functions, with prior weights based on reported uncertainties • MM-estimates with prior weights based on reported uncertainties • Mandel-Paule, Vangel-Ruhkin, or equivalent weighted-mean estimates of location and uncertainty after outlier rejection at the 99 % level

a Approaches known to CCQM at April 2013 and meeting the criteria of section 6.1 are listed. Section 6.3 includes recommendations and additional remarks on some estimators. For the small numbers of values typically found in the studies covered by this Table, alternative valid approaches often provide very similar performance.

b Checks for consistency are considered at section 5.2.2.

c "Outliers" in this Table refers both to values that are unusually distant from the bulk of the data set and to values that appear remote from a candidate KCRV when their reported uncertainty is taken into account. Inspection for outliers and other anomalies is discussed in section 5.2.1.

d Severe inconsistency is normally considered grounds for abandoning any attempt to calculate a KCRV (see paragraph 5.2.2.2).

10 References

1 BIPM, Mutual recognition of national measurement standards and of calibration and measurement certificates issued by national metrology institutes, Bureau International des Poids et Mesures, 1999

2 M Cox (2008) Data Evaluation Principles for CCQM Key Comparisons. CCQM reference CCQM-09-03 (http://www.bipm.org/cc/CCQM/Restricted/15/CCQM09_03.pdf).

3 ISO, Guide to the expression of uncertainty in measurement. International Organization for Standardization (1995). Also available as JCGM-100:2008 (http://www.bipm.org/utils/common/documents/jcgm/JCGM_100_2008_E.pdf)

4 Guidelines for CIPM Key Comparisons. CIPM (1999) amended 2003. Available at http://www.bipm.org/en/cipm-mra/guidelines_kcs/

5 See, for example, J Chambers, W Cleveland, B Kleiner, and P Tukey, Graphical Methods for Data Analysis, Wadsworth. (1983).

6 Nien-Fan Zhang, Metrologia, (2006), 43, 195-204

7 Analytical Methods Committee, Analyst, (1989), 114, 1693

8 Analytical Methods Committee, Analyst, (1989), 114, 1699

9 ISO 13528:2005 Statistical methods for use in proficiency testing by interlaboratory comparisons. International Organization for Standardization, Geneva, 2005.

10 RA Maronna, RD Martin, VJ Yohai, (2006) Robust Statistics: Theory and methods. John Wiley & Sons Ltd, Chichester, England.

11 RC Paule, J Mandel (1982) Consensus values and weighting factors. J Res Nat Bur Std 87:377–385

12 R DerSimonian, N Laird. (1986). Meta-analysis in clinical trials. Controlled Clinical Trials, 7, 177–188.

13 A Ruhkin, M Vangel, Journal of the American Statistical Association, 1998, Vol. 93, No. 441

14 B Toman, A Possolo, Accred. Qual. Assur. (2009) 14, 553-563

Appendix 1: A common consistency check

A simple consistency check meeting the requirements of section 5.2.2 is as follows:

i) Calculate the uncertainty-weighted mean* $\bar{x}_u$ of the candidate set:

$$\bar{x}_u = \frac{\sum_{i=1}^{m} x_i / u^2(x_i)}{\sum_{i=1}^{m} 1/u^2(x_i)} \qquad \text{(A1.1)}$$

ii) Calculate

$$\chi^2_{\text{obs}} = \sum_{i=1}^{m} \left(\frac{x_i - \bar{x}_u}{u(x_i)} \right)^2 \qquad \text{(A1.2)}$$

iii) Compare χ^2_{obs} with $m - 1$ and with $\chi^2_{0.05,\, m\text{-}1}$, the 95 percentile of χ^2 with $m - 1$ degrees of freedom.

iv) If $\chi^2_{\text{obs}} < m - 1$, it is normally safe to proceed with the assumption that the results are mutually consistent and that the uncertainties account fully for the observed dispersion of values.

v) If $m - 1 \le \chi^2_{\text{obs}} \le \chi^2_{0.05,m-1}$ the data provide no strong evidence that the reported uncertainties are inappropriate, but there remains a risk that additional factors are contributing to the dispersion. Refer to the prior working group decision on presumptive consistency and proceed accordingly.

vi) If $\chi^2_{\text{obs}} > \chi^2_{0.05,m-1}$, the data should be considered mutually inconsistent.

A worked example is given overleaf.

Notes:

i) Any other consistency check may be used if it has equivalent power of detecting over-dispersion.

ii) The consistency check above depends on a reliable location estimate. Replacing the weighted mean in the test above with a robust estimator is a useful precaution against undue influence from extreme values. High breakdown point is more important than efficiency at this stage. The median is therefore a useful simple estimator where reported uncertainties are reasonably similar. High-breakdown robust estimates that additionally take account of reported uncertainties (including, for this purpose, the mixture model median and largest consistent subset method) are useful if reported uncertainties differ appreciably (see Appendix 2).

iii) The chi-squared test assumes approximately normally distributed error. This test is inappropriate if any of the results included in the test have small degrees of freedom. Use of the critical values for chi-squared then leads to a higher probability of rejection of the null hypothesis. For this and other reasons, section 5.2.2 recommends that the result of a chi-squared test be used as a guide and not as a simple decision criterion.

* This is often referred to as the Graybill-Deal estimator.

Example: A consistent data set

Reported data on lead in a test material from 6 laboratories, with standard uncertainties, are given in Table 2 and plotted in Figure 3.

Consistency test:

i) The uncertainty-weighted mean $\bar{x}_u$ is calculated as follows:

$$\sum_{i=1}^{6} x_i / u_i^2 = 29135.941$$

$$\sum_{i=1}^{6} 1/u_i^2 = 9906.024$$

$$\bar{x}_u = \frac{\sum_{i=1}^{6} x_i / u^2(x_i)}{\sum_{i=1}^{6} 1/u^2(x_i)} = \frac{29135.941}{9906.024} = 2.941\ mg\,l^{-1}$$

ii) The chi-squared statistic is calculated as

$$\chi^2_{\text{obs}} = \sum_{i=1}^{6}\left(\frac{x_i - \bar{x}_u}{u(x_i)}\right)^2 = \left[\left(\frac{2.938 - 2.941}{0.018}\right)^2 + \left(\frac{2.917 - 2.941}{0.035}\right)^2 + \ldots\right] = 1.499$$

(only the first two summed terms are shown)

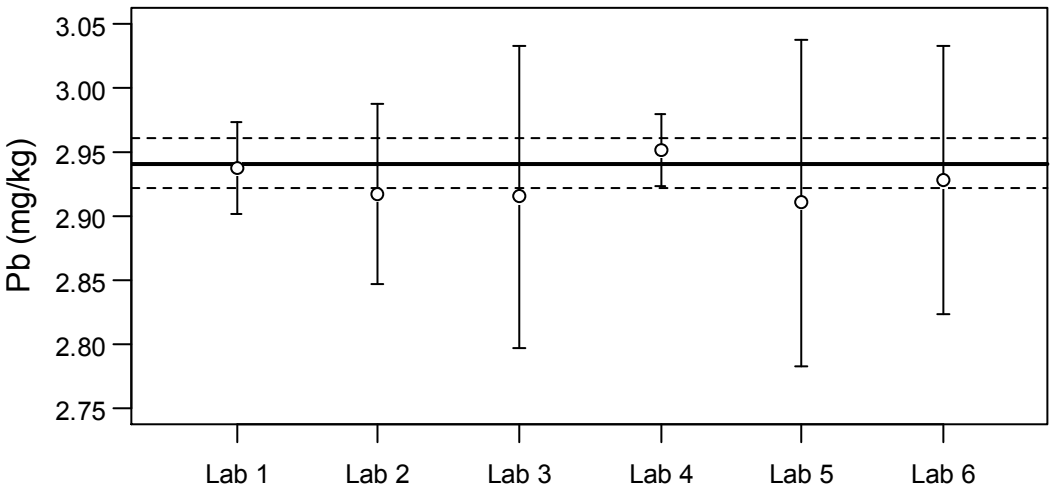
iii) The chi-squared statistic is compared with m–1 and with $\chi^2_{0.05,\,m\text{-}1}$. For m = 6, $\chi^2_{0.05,\,m\text{-}1}$ is 11.07. The calculated value of 1.499 is considerably below m-1 = 5 and the upper 95% critical value of 11.07. There is therefore no evidence of important excess dispersion.

Table 2: Data for consistency check example[Note 1]

Laboratory	**1**	**2**	**3**	**4**	**5**	**6**
x (mg kg^{-1})	2.938	2.917	2.915	2.951	2.910	2.928
u (mg kg^{-1})	0.018	0.035	0.059	0.014	0.064	0.052

Note 1: The data are simulated, with mean and uncertainties based on the KCRV and median reported uncertainty for CCQM K-30 (lead in wine).

Figure 3: Consistency test example



The figure shows values from Table 2. Error bars show expanded uncertainties using coverage factors of k=2. The solid and dashed horizontal lines are the weighted mean and the upper and lower limits of the corresponding expanded uncertainty interval (again with k = 2) , respectively.

Appendix 2: Calculations for consensus KCRV estimators

The following tables provide calculations for some common estimators used for KCRVs. The list is not intended to be either comprehensive or restrictive; any estimator that meets the criteria of section 6 and is appropriate to the assumptions in force may be used.

It is assumed that the study includes a total of n laboratories of which the first m ($m \leq n$) are included in the calculation of the KCRV (but all n laboratories are used in the calculation of degrees of equivalence).

1 General

1.1 Degrees of equivalence

For a KCRV $\hat{\mu}$ and an individual reported value x_i, the degree of equivalence or DoE is (d_i , $U(d_i)$), where

$$U(d_i) = k\mathrm{u}(d_i).$$

When normality can be assumed, k can be taken as 2. When normality cannot be assumed, k is chosen based on knowledge of the distribution. $u(d_i)$ is given by

$$d_i = x_i - \hat{\mu}, \qquad u^2(d_i) = u^2(x_i) + u^2(\hat{\mu}) - 2\mathrm{cov}(x_i, \hat{\mu}). \tag{A2.1}$$

1.2 Uncertainty and Covariance

Many estimators can be expressed as a linear combination of values x_i with associated weights w_i, so that the KCRV $\hat{\mu}$ can be expressed as

$$\hat{\mu} = \sum_{i=1}^{m} w_i x_i. \tag{A2.2}$$

For example, the arithmetic mean can be treated as an instance of equation (A2.2) with all w_i set to $1/m$. Although there are specific formulae for many estimators, the general form of equation (A2.2) is particularly useful for obtaining uncertainties and covariances in key comparisons, because it is applicable not only to common classical estimators, but also to many robust estimators.

Where the $u(x_i)$ are compatible with the observed dispersion of the values x_i (for example, where a chi-squared test shows no evidence of over-dispersion), the standard uncertainty $u(\hat{\mu})$ associated with $\hat{\mu}$ can, in the absence of correlation associated with the x_i, be calculated from

$$u^2(\hat{\mu}) = \sum_{\mathrm{i}=1}^{m} w_i^2 u^2(x_i). \tag{A2.3}$$

Note: Strictly, this expression applies only to exact weights; however, it is likely to be a sufficient approximation for the purposes of this document.

Under the same conditions, the covariance associated with x_i and $\hat{\mu}$ is

$$\mathrm{cov}(x_i, \hat{\mu}) = w_i u^2(x_i), \qquad i = 1, \ldots, m. \tag{A2.4}$$

Note 1: When the standard uncertainties $u(x_i)$ are not compatible with the dispersion of the x_i, Equation (A2.3) can underestimate $u(\hat{\mu})$. A compensating adjustment to equation (A2.3) that is sometimes used is to modify $u(\hat{\mu})$ by a scaling factor based on the observed dispersion of (scaled) deviations

(the Birge ratio does this explicitly; many robust estimates implicitly choose scale based on dispersion). This aspect will be considered further below in relation to particular estimators.

Note 2: In some cases – including the arithmetic mean and the median – $u(\hat{\mu})$ as usually evaluated is formally *unrelated* to the reported standard uncertainties $u(x_i)$. Equations (A2.3) and (A2.4) *do not then apply correctly* unless $s(x)$, the standard deviation of $x_1, \ldots, x_m$, is used in place of all the $u(x_i)$. An important exception is the case where $s(x)$ is largely attributable to an additional random effect outside laboratory control; this aspect is discussed in connection with the calculations for classical estimators below.

Note 3: Many robust estimators (including the median) can be expressed in the form of equation (A2.2) using 'posterior weights' that become available following, or in the process of, estimation.

Note 4: Values excluded from the calculation of $\hat{\mu}$ can be treated as having zero values for the corresponding weights w_i and consequently there is zero covariance associated with these values and $\hat{\mu}$. Correlation may, however, arise for other reasons, such as use of a calibrant common to all study participants.

1.3 Uncertainty component of the degree of equivalence

The uncertainty component

$$U_i(d_i) = ku(d_i)$$

of the DoE for the ith participant is given by expression (A2.1).

1.4 Efficiency and breakdown point

The performance of robust (and other) estimators can usefully be described in terms of two properties; *efficiency* and *breakdown point*; these properties are included in the tables below.

Efficiency describes the dispersion properties of an estimator when applied to well-behaved data; it is usually given as asymptotic relative efficiency, defined as the inverse of the ratio of estimator variance to the variance of the corresponding minimum variance estimator when applied to the normal distribution. High efficiencies are desirable, as higher efficiency leads to smaller KCRV uncertainty.

Note: CCQM 08-08 principle 7 requires that "The most efficient approach [that is giving the smallest value of $u(x_{\text{ref}})$] of those consistent with the applicable assumptions is preferred."

The breakdown point (or simply 'breakdown') can be thought of as the proportion of the data set that can go to infinity while keeping the estimate finite. It gives an indication of outlier resistance. A high breakdown point is desirable for outlier resistance.

Note: No useful estimator has a breakdown point higher than 0.5, but many approach 0.5. The mean and weighted mean have a breakdown point of zero, indicating that they have essentially no resistance to the presence of outliers.

2 Classical estimators

2.1 Arithmetic mean

Value $\bar{x}$	$$\bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i .$$
Standard uncertainty $u(\bar{x})$	$$u^2(\bar{x}) = \frac{1}{m} s^2(x),$$ where $s(x)$ is the standard deviation of the measured values $x_1, \ldots, x_m$, given by $$s^2(x) = \frac{1}{m-1}\sum_{i=1}^{m} (x_i - \bar{x})^2 .$$
Covariance $\mathrm{cov}(x_i, \bar{x})$	$$\mathrm{cov}(x_i, \bar{x}) = \begin{cases} \frac{1}{m} s^2(x), & i = 1,...,m, \\ 0, & \text{otherwise.} \end{cases}$$
DoE uncertainty component:	$$U(d_i) = ku(d_i),$$ where $$u^2(d_i) = \begin{cases} \left(1 - \frac{1}{m}\right) s^2(x), & i = 1, ..., m, \\ u^2(x_i) + u^2(\bar{x}), & \text{otherwise.} \end{cases}$$
Breakdown point	Zero.
Efficiency	One (for identically equal and reliable uncertainties $u(x_i)$, $i = 1, \ldots, m$).
Special cases	Where the $u(x_i)$ are considered to be reliably determined and an additional random effect increases the dispersion of values x_i, so that $s^2(x)$ is greater than $$\frac{1}{m^2}\sum_{i=1}^{m} u^2(x_i) :$$ $$\mathrm{cov}(x_i, \bar{x}) = \frac{1}{m} u^2(x_i)$$ and $$u^2(d_i) = \left(1 - \frac{2}{m}\right) u^2(x_i) + u^2(\bar{x}).$$
Software	The arithmetic mean and standard deviation are routinely included in spread sheets and statistical software.
Additional remarks	The arithmetic mean is not a minimum-variance estimator unless *all* $u(x_i)$ are identical or an additional random term dominates the dispersion so that $s^2(x)$ is very much greater than $$\frac{1}{m^2}\sum_{i=1}^{m} u^2(x_i).$$

2.2 Uncertainty-weighted mean

Note: This is also referred to as the 'Graybill-Deal' estimator.

Value $\bar{x}_u$	$$\bar{x}_u = \sum_{i=1}^{m} w_i x_i ,$$ where $$w_i = \frac{1/u(x_i)^2}{\sum_{j=1,m} 1/u(x_j)^2}$$
Standard uncertainty $u(\bar{x}_u)$ a) Uncorrected for observed dispersion b) Corrected for observed dispersion	$$\frac{1}{u^2(\bar{x}_u)} = \sum_{i=1}^{m} \frac{1}{u^2(x_i)}.$$ $$u^2_{\text{corr}}(\bar{x}_u) = \frac{\chi^2_{\text{obs}}}{m-1} u^2(\bar{x}_u),$$ where $$\chi^2_{\text{obs}} = \sum_{i=1}^{m} \frac{(x_i - \bar{x}_u)^2}{u^2(x_i)}.$$
Covariance $\text{cov}(x_i, \bar{x})$	$$\text{cov}(x_i, \bar{x}) = \begin{cases} w_i u^2(x_i), & i = 1, \ldots m, \\ 0, & \text{otherwise.} \end{cases}$$
Uncertainty component of degree of equivalence: i) Result included in calculation of $\bar{x}$ a) Uncorrected for observed dispersion b) Corrected for observed dispersion ii) Result not included in calculation of $\bar{x}$	$$U(d_i) = ku(d_i),$$ where $$u^2(d_i) = u^2(x_i) - u^2(\bar{x}_u),$$ $$u^2_{\text{corr}}(d_i) = u^2_{\text{corr}}(\bar{x}_u) + (1 - 2w_i)u^2(x_i).$$ (see Additional remark ii) below). $$u^2(d_i) = u^2(x_i) + u^2(\bar{x})$$ or $$u^2(x_i) + u^2_{\text{corr}}(\bar{x}),$$ as appropriate.
Breakdown point	Zero
Efficiency	One (with reliably reported uncertainties with large degrees of freedom)
Special cases	Over-dispersion requires a scale correction to the reported uncertainties; see above.

Software	Weighted means are implemented in some statistical software. General-purpose linear modelling software usually implements weighting, and will return the weighted mean if instructed to fit an intercept-only regression model; the standard error will usually be corrected for dispersion by default.
Additional remarks	i) The uncertainty associated with the uncertainty-weighted mean should normally be corrected for observed dispersion where that dispersion is greater than can be accounted for by the reported standard uncertainties. ii) The correction given for observed dispersion is equivalent to an assumption that excess variance affects each laboratory to an extent proportional to its reported uncertainty. This circumstance is unlikely in practice, and should consequently be regarded as an approximation. iii) The calculation for $u_{\text{corr}}(d_i)$ assumes that $u_{\text{corr}}(\bar{x}_u)$ is greater than $u(\bar{x}_u)$ due to excess variance arising from effects outside the control of the laboratories. iv) The uncertainty-weighted mean described above is often referred to as simply 'the weighted mean'. v) Zhang [1] recommends modified estimates and associated uncertainty which are preferred where any reported uncertainties are associated with small degrees of freedom.

2.3 Median

Value med(x)	$$\text{med}(x) = \begin{cases} \frac{1}{2}(x'_{m/2} + x'_{m/2+1}), & m \text{ even}, \\ x'_{m/2}, & m \text{ odd}, \end{cases}$$ where $x'_1, \ldots, x'_m$ denote the participant values arranged in increasing order (or if there are ties in non-decreasing order).
Standard uncertainty $u(\text{med}(x))$	$$u^2(\text{med}(x)) = \frac{\pi}{2m}\hat{\sigma}^2,$$ where $\hat{\sigma}$ is a robust estimate of standard deviation, usually based on the median absolute deviation mad(x) multiplied by 1.483. (This corrected estimate is sometimes called MAD_E.)
Covariance $\text{cov}(x_i, \text{med}(x))$	$$\text{cov}(x_i, \text{med}(x)) = \begin{cases} \hat{\sigma}^2/m, & i = 1, \ldots, m, \\ 0, & \text{otherwise}. \end{cases}$$
Uncertainty component of degree of equivalence: i) Result included in calculation of med(x) ii) Result not included in calculation of med(x)	$$U(d_i) = ku(d_i),$$ where $$u^2(d_i) = \left(1 + \frac{\pi - 4}{2m}\right)\hat{\sigma}^2.$$ $$u^2(d_i) = u^2(x_i) + u^2(\text{med}(x)).$$

Breakdown point	1/2.
Efficiency	0.637.
Special cases	Where the $u(x_i)$ are considered to be reliably determined and an additional random effect increases the dispersion of values x_i, $$\mathrm{cov}(x_i, \mathrm{med}(x)) = \frac{1}{m} u^2(x_i)$$ and $$u^2(d_i) = \left(1 - \frac{2}{m}\right) u^2(x_i) + u^2(\mathrm{med}(x)).$$
Software	The median is implemented routinely in spreadsheets and statistical software. MAD_E is implemented in most statistical software. In the free open-source package R [2] MAD_E is implemented as mad() and in the AMC Excel add-in Robstat.xla [3] as MADE().
Additional remarks	i) The median takes no account of the reported laboratory uncertainties. ii) MAD_E is inefficient and is negatively biased for small m. At $m = 5$, MAD_E^2 underestimates σ^2 by approximately 10 %, which may be acceptable, but at $m = 4$, MAD_E^2 underestimates σ^2 by approximately 30 %. MAD_E is therefore not recommended for use with $m < 5$.

3 Additional-variance estimators

3.1 Overview

These estimators arise from the assumption of a random between-laboratory effect in addition to the effects accounted for by the laboratory uncertainties. They calculate an additional variance component to model over-dispersion in addition to the reported standard uncertainties. They are most appropriate when an effect such as test material inhomogeneity or instability unexpectedly affects all laboratories in a similar manner. They are also appropriate when there is evidence of appreciable inconsistency and where robust methods are not considered appropriate. Note that in the summaries below, it is assumed that the excess variance arises from an effect such as material inhomogeneity which is outside the control of the participants and is therefore properly included in the uncertainty of any calculated degrees of equivalence.

3.2 Mandel-Paule and Vangel-Ruhkin estimators

This section points to implementations of the Mandel-Paule (M-P) and Vangel-Ruhkin (V-R) estimates and presents formulae for calculating *u*(KCRV) and degree-of-equivalence uncertainties.

Estimator	Recommended software implementation(s)
Vangel-Ruhkin	Dataplot [4] **Remarks**: i) The Vangel-Ruhkin method provides an iterative restricted maximum likelihood estimate as the KCRV and the associated uncertainty; when software is available for its computation it is recommended over the M-P procedure. ii) The efficiency is high, but breakdown point zero unless supported by outlier rejection
Mandel-Paule [5]	Dataplot [4] **Remarks**: i) An additional free implementation is available in the experimental R Packege 'metRology' available at http://sourceforge.net/projects/metrology/ ii) The iterative algorithm converges reasonably fast and can consequently be implemented in a spreadsheet. See reference [5] for details. iii) The efficiency is high, but breakdown point zero

3.3 Uncertainty and DoE calculations

Both the Mandel-Paule and Vangel-Ruhkin methods effectively estimate an additional component of variance and combine this with reported uncertainties, with weights based on the resulting combined uncertainties. The following calculations apply to either of these methods where degrees of freedom are large and the estimated additional variance component is available.

Value $\bar{x}_{\mathrm{AV}}$	$$\bar{x}_{\mathrm{AV}} = \sum_{i=1}^{m} w_i x_i ,$$ where

	$$w_i = \frac{\dfrac{1}{u^2(x_i) + u^2(q)}}{\displaystyle\sum_{j=1,m} \frac{1}{\left(u(x_j)^2 + u^2(q)\right)}}$$ and $u^2(q)$ is the estimated additional variance from the iterative V-R or M-P procedure
Uncertainty component of degree of equivalence:	$U(d_i) = ku(d_i)$, where
Value x_i included in calculation	$u^2(d_i) = u^2(x_i) + u^2(q) - u^2(\bar{x}_{\mathrm{AV}})$
Value x_i not included in calculation	$u^2(d_i) = u^2(x_i) + u^2(q) + u^2(\bar{x}_{\mathrm{AV}})$

3.4 The DerSimonian-Laird procedure

The DerSimonian-Laird estimator is a non-iterative method that includes a calculated excess variance term. It can therefore be implemented reasonably easily in a spreadsheet. It provides very similar results to the V-R and Mandel-Paule estimators above and has been suggested (CCQM-11-18) as a preferred calculation where calculation simplicity is desired and where an excess-variance estimator is appropriate (see above).

The method starts with an initial calculation of the Graybill-Deal uncertainty-weighted mean identical to $\bar{x}_u$ in 2.2 and then calculates an estimate of excess variance, denoted by λ below.

Graybill-Deal mean	$\bar{x}_u = \frac{1}{W_1}\sum_{i=1}^{p} w_i x_i, \quad w_i = 1/u_i^2, \ i = 1,\ldots,p, \qquad W_1 = \sum_{i=1}^{p} w_i.$
Interlaboratory variance	$\lambda = \max\left[0, \ \frac{\sum_{i=1}^{p} w_i (x_i - \bar{x}_u)^2 - p + 1}{W_1 - W_2/W_1}\right], \quad W_2 = \sum_{i=1}^{p} w_i^2.$
DerSimonian-Laird mean x_{DL}	$x_{\mathrm{DL}} = \sum_{i=1}^{N} \tilde{w}_i x_i, \qquad \tilde{w}_i = \frac{(u_i^2 + \lambda)^{-1}}{\sum_{j=1}^{p} (u_j^2 + \lambda)^{-1}}.$
Standard uncertainty $u(x_{\mathrm{DL}})$	$u(x_{\mathrm{DL}}) = \left[\sum_{i=1}^{p} \tilde{w}_i^2 (x_i - x_{\mathrm{DL}})^2 / (1 - \tilde{w}_i)\right]^{1/2}.$
Uncertainty component of degree of equivalence:	$U(d_i) = ku(d_i)$, where
Value x_i included in calculation	$d_i = x_i - x_{\mathrm{DL}}, \qquad u^2(d_i) = u_i^2 + \lambda - u^2(x_{\mathrm{DL}}).$
Value x_i not included in calculation	$d_i = x_i - x_{\mathrm{DL}}, \qquad u^2(d_i) = u_i^2 + \lambda + u^2(x_{\mathrm{DL}}).$

Software	An implementation of the DerSimonian-Laird procedure is provided in the R package ‘metRology’ currently available at http://sourceforge.net/projects/metrology/ and at https://r-forge.r-project.org/projects/metrology/

4 Robust estimators (unweighted)

4.1 Huber estimate 2 (H15)

Value $\hat{\mu}_{\mathrm{H15}}$	$$\hat{\mu}_{\mathrm{H15}} = \frac{1}{W}\sum_{i=1}^{m} W_i x_i ,$$ where $$W_i = \min\left(1,\ \frac{k\hat{\sigma}}{\lvert x_i - \hat{\mu}_{\mathrm{H15}}\rvert}\right), \qquad W = \sum_{i=1}^{m} W_i ,$$ $\hat{\sigma}$ is a robust scale estimate (often $\mathrm{MAD_E}$ or a value $\hat{\sigma}_{\mathrm{H15}}$ determined during iteration) and k is a tuning constant, usually 1.345 or 1.5. For 95 % efficiency, 1.345 is recommended.
Standard uncertainty $u(\hat{\mu}_{\mathrm{H15}})$	$$u^2(\hat{\mu}_{\mathrm{H15}}) = \frac{1}{e}\hat{\sigma}^2_{\mathrm{H15}},$$ where $\hat{\sigma}_{\mathrm{H15}}$ is the robust estimate of standard deviation delivered simultaneously in the iterative estimation of $\hat{\mu}_{\mathrm{H15}}$ and e is the efficiency (0.95 for k = 1.345).
Covariance $\mathrm{cov}(x_i, \hat{\mu}_{\mathrm{H15}})$	$$\mathrm{cov}(x_i, \hat{\mu}_{\mathrm{H15}}) = \begin{cases} \frac{W_i}{W}\hat{\sigma}^2_{\mathrm{H15}}, & i = 1, ..., m, \\ 0, & \text{otherwise} \end{cases}$$ (see Additional Remarks ii)).
Uncertainty component of DoE:	$$U(d_i) = ku(d_i),$$ where
i) Result included in calculation of $\hat{\mu}_{\mathrm{H15}}$	$$u^2(d_i) = \hat{\sigma}^2_{\mathrm{H15}} + u^2(\hat{\mu}_{\mathrm{H15}}) - 2\mathrm{cov}(x_i, \hat{\mu}_{\mathrm{H15}})$$ (see Additional Remarks ii).
ii) Result not included in estimation of $\hat{\mu}_{\mathrm{H15}}$	$$u^2(d_i) = u^2(x_i) + u^2(\hat{\mu}_{\mathrm{H15}}).$$
Breakdown point	1/2 with prior scale estimate: 0.33 with simultaneous determination of $\hat{\sigma}_{\mathrm{H15}}$ using an absolute deviation basis [6].
Efficiency	0.95 for k = 1.345.
Special cases	Where the $u(x_i)$ are considered to be reliably determined and an additional random effect increases the dispersion of values x_i, $$\mathrm{cov}(x_i, \hat{\mu}_{\mathrm{H15}}) = w_i u^2(x_i),$$ where $$u^2(d_i) = u^2(x_i) + u^2(\hat{\mu}_{\mathrm{H15}}) - 2\mathrm{cov}(x_i, \hat{\mu}_{\mathrm{H15}}).$$
Software	Huber's estimate is implemented in some statistical software (in R as hubers() and rlm() in the MASS package; in S-plus in the robust package, also available for R) and in the AMC Excel add-in Robstat.xla [3].
Additional remarks	i) Because the weights depend on the estimate $\hat{\mu}_{\mathrm{H15}}$,

	Huber's estimate is, in common with many robust estimators, usually calculated by iterative re-weighting. ii) Huber's estimate is typically applied without regard to reported uncertainties; the uncertainty for an individual value is then implicitly assumed to be equal to the estimated standard deviation for the data (usually $\hat{\sigma}_{\mathrm{H15}}$) as it would be for the mean and median. The case for which the $u(x_i)$ are reliably estimated and an additional random effect is operating is covered above. iii) ISO 13528 suggests a value of $\hat{\sigma}_{\mathrm{H15}}\sqrt{\frac{\pi}{2m}}$ for the uncertainty associated with the Huber estimate. This value is unnecessarily conservative when the efficiency is known (and not equal to $\pi/2$).

5 Robust estimators (weighted)

Robust estimators that use prior weights (whether based on reported uncertainties or otherwise) are not described in detail here. Rather, software implementations are listed for information. Additional detail is under consideration for future releases of this guidance; in particular, the present software implementations do not provide ready access to DoE uncertainty components. Additional software is currently under development to provide this feature.

These estimators are appropriate when outliers are likely and reported uncertainties vary substantially for good reason.

Estimator	Recommended software implementation(s)
Huber estimate (H15) (weighted)	R [2], using package MASS[7]. The appropriate call is summary(rlm(x~1, method="huber", weights=1/u^2)). The return value includes the estimate and its associated standard error, which should be used as the KCRV and associated standard uncertainty respectively. **Remarks**: i) Posterior weights w_p can be obtained from the return value as the component '*w*'. These can be used to calculate effective weights w_i (above) using $w_i = w_p/u^2(x_i)$. ii) The implementation estimates the scale from the data.
MM- estimate (MM) (weighted)	R [2], using package MASS[7]. The appropriate call is summary(rlm(x~1, method="MM", weights=1/u^2)). The return value includes the estimate and its associated standard error, which should be used as the KCRV and associated standard uncertainty respectively. **Remarks**: i) Posterior weights can be obtained from the return value as the component 'w'. ii) The implementation estimates the scale from the data.

References for Appendix 2

1 Nien-Fan Zhang, Metrologia, (2006), 43, 195-204

2 R Development Core Team (2009). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

3 Robstat.xla. Available at http://www.rsc.org/amc/.

4 James J. Filliben and Alan Heckert (1978) with additional continuous contributions since. Available from http://www.itl.nist.gov/div898/software/dataplot.

5 RC Paule, J Mandel (1982) Consensus values and weighting factors. J Res Nat Bur Std 87:377–385

6 RA Maronna, RD Martin, VJ Yohai, (2006) Robust Statistics: Theory and methods. John Wiley & Sons Ltd, Chichester, England

7 Venables, W. N. & Ripley, B. D. (2002) Modern Applied Statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0

Appendix 3: Symbols and notation

The following symbols and notation are used in this guidance document:

$cov(x,y)$	covariance between x and y.
d_i	value component $x_i - x_{ref}$ of the degree of equivalence (DoE) for laboratory i ($i = 1, \ldots, N$)
k	coverage factor
m	number of qualified participants (see section 5)
n	number of participating laboratories
$u(d_i)$	standard uncertainty associated with d_i
$U(d_i)$	uncertainty component of the DoE for laboratory i ($i = 1, \ldots, N$). Note: The Technical Annex to the MRA states that this uncertainty is expressed at 95 % confidence.
$u(x_i)$	standard uncertainty associated with x_i ($i = 1, \ldots, N$)
$u(x_{ref})$	standard uncertainty associated with x_{ref}
$u^2(x)$	$u(x)$ expressed as a variance (x can be x_i, x_{ref}, etc.)
w_i	weighting factor applied to x_i in the calculation of a KCRV
x_i	value submitted by ith participating laboratory ($i = 1, \ldots, n$)
x_{ref}	key comparison reference value (KCRV)