NIST Decision Tree

User's Manual

— 3rd Edition —

David Newton, Amanda Koepke,
Antonio Possolo, & Michael Winchester

Statistical Engineering Division & Chemical Sciences Division
Information Technology Laboratory & Material Measurement Laboratory
National Institute of Standards and Technology

March 15, 2024

National Institute of
Standards and Technology
U.S. Department of Commerce

Contents

Exhibits

1 Introduction

The NIST Decision Tree (NDT) provides a recommendation for how to combine independent measurement results for the same scalar measurand, and then uses the method selected by the user (which may be the method the NDT recommends or any other that the user selects from among those implemented in the NDT) to produce:

(a) a consensus value — which will be the key comparison reference value (KCRV) in the context of CIPM key comparisons and of RMO supplementary comparisons (Comité International des Poids et Mesures (CIPM), 1999, 3.2,T10);

(b) an evaluation of the associated uncertainty;

(c) differences between the measured values and the consensus value, and their associated expanded uncertainties — which together are called *degrees of equivalence* (DoEs) in the context of key and supplementary comparisons (Comité International des Poids et Mesures (CIPM), 1999).

The inputs to the NDT are measured values, associated standard uncertainties, and (optionally) numbers of degrees of freedom that support these standard uncertainties.

Section 2 indicates the web address of the NDT and introduces its landing page. Sections 3-5 describe how to use the NDT and section 6 explains the meaning of the results that are displayed on-screen and that can also be saved as a PDF file. Section 7 illustrates the application of the NDT to select and fit a model to a set of measurement results obtained in key comparison CCQM-K25 (Schantz et al., 2003).

Possolo et al. (2021) provide details about the methods implemented in the NDT, and present several examples of application. Koepke et al. (2017) review consensus building for interlaboratory studies, key comparisons, and meta-analysis, and describe several methods in detail. Possolo and Meija (2022) and Meija et al. (2023) present several examples of application, with emphasis on Bayesian methods.

Borenstein et al. (2010) offer an introduction to different kinds of models used in meta-analysis, which are also relevant for interlaboratory studies, including key comparisons. The book-length treatments by Borenstein et al. (2009) and

Hartung et al. (2008) are highly recommended, the latter being more laden with mathematics than the former. The overview by Harrer et al. (2021) is particularly accessible.

2 Access

Access the NDT via a Web browser by visiting

`https://decisiontree.nist.gov`

The landing page has four tabs: **0. About**, which describes the application and includes a link to this manual, and **1. Data**, **2. Decision Tree**, and **3. Fit Model**: these last three should be visited in the order of their numbering.

3 Data

Measurement results are entered into a mini-spreadsheet in tab **1. Data**. Right-click anywhere on the mini-spreadsheet to add or remove rows, to undo or redo, and to choose an alignment option for the contents of the cells.

There should be no row of the mini-spreadsheet with all cells empty. However, it is acceptable to leave one or more entries in the column headed DegreesOfFreedom empty: this will be interpreted as indicating that the missing numbers of degrees of freedom are very large.

The number of columns of the mini-spreadsheet is fixed, and so are the column labels. Add or remove rows so that the mini-spreadsheet will have as many rows (excluding the header) as there are measurement results to be entered.

Each measurement result comprises an alphanumeric label, a measured value, the associated standard uncertainty, and (optionally) the number of degrees of freedom that support the standard uncertainty.

One or more cells intended to have numbers of degrees of freedom may be left empty when such numbers are not available. The valid entries for the cells in the column labeled Degrees of Freedom are either an empty cell or a number (not necessarily an integer) greater than or equal to 1.

The measurement results can be entered in any one of the following ways:

(a) Typing them directly into the cells of the mini-spreadsheet (cells in unused rows should be emptied, or those rows deleted);

(b) Copy a rectangular subset containing the measurement results, arranged according to how they shall be pasted onto the mini-spreadsheet, from an external spreadsheet, and paste them into the mini-spreadsheet;

(c) Click the button Browse... and select a comma-separated values (csv) file with the measurement results arranged according to how they are intended to be placed in the mini-spreadsheet, and whose first line must be

```
Laboratory,MeasuredValues,StdUnc,DegreesOfFreedom
```

The squares in the cells of the column of the mini-spreadsheet labeled Include can be clicked, which generates a check mark, to indicate that the measurement result in the same row shall be used in the calculation of the consensus value.

Any row whose square is left unchecked will not be considered by the Decision Tree, and it will not be used in the calculation of the consensus value or of its associated uncertainty. However, the corresponding degree of equivalence will be computed and displayed in **3. Fit Model**.

It is not necessary to add rows if one simply pastes a rectangle with the results copied from a spreadsheet external to the NDT, because the mini-spreadsheet expands as needed to accommodate the pasted data. However, if the pasted rectangle has fewer rows than the "default" mini-spreadsheet, then the extra rows will have to be deleted via right-clicks as mentioned above.

The contents of any cells in other columns of the mini-spreadsheet can be deleted in the same way.

To paste a rectangular subset of cells from an external spreadsheet (for example, LibreOffice Calc or Microsoft Excel):

(1) Copy a rectangular subset of cells from the spreadsheet, so that the first column in the rectangle has labels for the rows, the second column has measured values, the third column has standard uncertainties, and the fourth column (if present) has numbers of degrees of freedom;

(2) Click once on the topmost cell of the mini-spreadsheet in the column labeled Laboratory (it does not matter whether there is a value there already or not), so that it becomes selected;

(3) Press CTRL+V (or equivalent keystroke combination) to paste the clipboard onto the mini-spreadsheet.

(4) Remove any extraneous rows (to delete rows or columns, right click on any cell in the body of the mini-spreadsheet; to change the alignment of the entries in any column, right click on the column's header);

(5) Delete the contents of any cells that should be blank in the column labeled Degrees of Freedom of the mini-spreadsheet.

If some of the input data do not include numbers of degrees of freedom, then the corresponding cells in the column of the mini-spreadsheet labeled Degrees of Freedom will have to be blanked by selecting them and pressing BACKSPACE or DEL (or equivalent keys in the user's keyboard).

When the standard uncertainties are derived from expanded uncertainties based on *coverage factors* (JCGM 100:2008, §2.3.6) reported by the participants, then one can derive the corresponding, implied numbers of degrees of freedom by "inversion" of tables with percentiles of the Student's t distribution for different numbers of degrees of freedom. For example, $k = 2$ means that the effective number of degrees of freedom is 60.4.

The measurement results entered into the mini-spreadsheet will be plotted automatically alongside. The dots represent the measured values, and each vertical line segment represents a measured value plus or minus one standard uncertainty. The data for the rows with a check-mark under Include are depicted in black, and the data for the rows without such check-mark are depicted in gray.

Once the data will have been entered and found to be satisfactory, and the check marks added as intended (in the column labeled Include), press the button labeled Validate Data to determine whether all inputs are valid. If so, then select the next tab, **2. Decision Tree**, by clicking on it at the top of the page.

4 Decision Tree

Exhibit 1 depicts the Decision Tree, which comprises four branching nodes (orange) and five leaves (blue). The leaves indicate different procedures for data reduction, each of which has an underlying statistical model.

To use the Decision Tree one answers a question at each node, and follows the course corresponding to the answer (YES or NO), until one reaches a leaf, which is the recommended procedure.

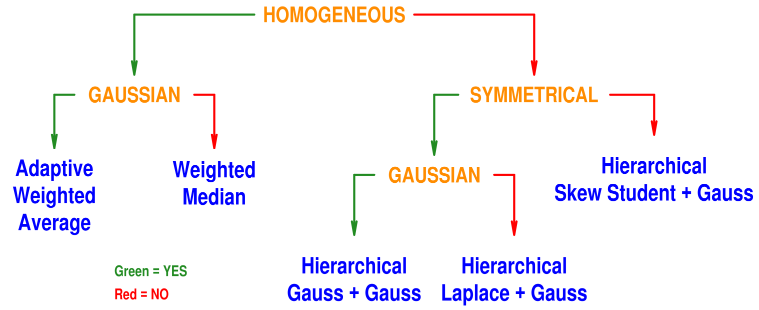

HOMOGENEOUS
GAUSSIAN
SYMMETRICAL
Adaptive Weighted Average
Weighted Median
GAUSSIAN
Hierarchical Skew Student + Gauss
Green = YES
Red = NO
Hierarchical Gauss + Gauss
Hierarchical Laplace + Gauss

Exhibit 1: The Decision Tree comprises four branching nodes (orange) and five leaves (blue) that suggest different models for the measurement results and corresponding procedures for data reduction. A question needs to be answered at each node: if the answer is YES, then one follows the green branch (toward the left); if the answer is NO, then one follows the red branch (toward the right), until one reaches a leaf.

Since the traversal of the Decision Tree is guided by classical statistical tests of hypotheses — for homogeneity (or mutual consistency), symmetry, and Gaussian shape —, it inherits all the shortcomings of such classical testing, including their limited power to detect heterogeneity, asymmetry, or deviations from the Gaussian shape.

For the test of Gaussian shape, which is the Shapiro-Wilk test (Shapiro and Wilk, 1965), other limitations derive from the fact that it is applied to the measured values standardized by centering them at their common median, and by dividing the centered values by the corresponding, reported standard uncertainties. However, the test does not take into account the uncertainty of the common median or the numbers of degrees of freedom that support the reported uncertainties.

A notable limitation of Cochran's Q test (Cochran, 1954) of homogeneity (§4.3), is its low power to detect heterogeneity when it exists. The test also assumes that the reported uncertainties all are based on very large (practically infinite) numbers of degrees of freedom, besides several other shortcomings (Hoaglin, 2016).

4.1 Measurement Model

Let $x_1, \dots, x_n$ denote the values measured by n laboratories or methods, $u(x_1)$, $\dots$, $u(x_n)$ denote their associated standard uncertainties, and $\nu_1, \dots, \nu_n$ denote the corresponding numbers of degrees of freedom.

The NDT aims to select and fit a specific version of the following model (except for the leaf labeled **Weighted Median**, whose corresponding model does not include λ_j) to the measurement results:

$$x_j = \mu + \lambda_j + \varepsilon_j, \quad \text{for } j = 1, \dots, n, \tag{1}$$

where μ denotes the true value of the measurand, the $\{\lambda_j\}$ denote laboratory or method effects (assumed to be a sample from a probability distribution with mean 0 and standard deviation τ), and the $\{\varepsilon_j\}$ denote measurement errors, all with mean 0 and possibly different standard deviations $\{\sigma_j\}$, of which the $\{u(x_j)\}$ are estimates.

The version of the model above that corresponds to the leaf of the NDT labeled **Weighted Median** has $\tau = 0$, therefore the model in fact reduces to the common median model, $x_j = \mu + \varepsilon_j$, where the measurement errors $\{\varepsilon_j\}$ are like outcomes of Laplace random variables with possibly different standard deviations, and μ is estimated by the weighted median with weights proportional to the reciprocals of the squared reported uncertainties, computed using R function `weighted.median` as defined in package `spatstat.geom` (Baddeley and Turner, 2005).

The NDT recommends the **Adaptive Weighted Average** (AWA) when the data are judged to be homogeneous (that is, mutually consistent), and the standardized measured values can reasonably be regarded as a sample from a Gaussian distribution. The estimation procedure is the version of the (DerSimonian and Laird, 1986) procedure that is described by Koepke et al. (2017).

In particular, if the DerSimonian-Laird estimate of τ is zero, then the AWA is the conventional weighted average of the measured values, with weights proportional to the reciprocal of the squared reported uncertainties. However, if the DerSimonian-Laird estimate of τ is not zero (even though heterogeneity will not have been deemed to be significant), then the estimate of μ is the DerSimonian-Laird estimate, and the uncertainty surrounding the estimate of τ is propagated as described by Koepke et al. (2017).

4.2 Statistical Tests and p-values

The NDT bases its recommendation, about the procedure to use for the data reductions, on the outcomes of three statistical tests of hypotheses, which are introduced in subsections 4.3, 4.4, and 4.5.

Each of these tests produces a p-value, not a decision: about whether the results are homogeneous, or whether the measured values are like a sample from a symmetrical distribution, or from a Gaussian (that is, normal) distribution.

It is the user's responsibility to make a decision, about each of the aforementioned issues, based on the p-value of each test. For this reason, here we review the classical concept of the kind of statistical tests of hypotheses that the NDT uses, explain the meaning of p-values, and suggest thresholds for decisions.

The tests of homogeneity, symmetry, and normality (Gaussian shape) all are *pure significance tests* pitting a rather general hypothesis against its negation: for example, that the measurement results are homogeneous, versus that they are not (regardless of the pattern of the heterogeneity).

Each such test is based on a criterion that gauges how far the data are from satisfying the hypothesis under test. For example, in Cochran's test, the criterion is the non-negative quantity Q that appears in the test's name: the smaller the Q, the more homogeneous the measurement results appear to be.

The p-value of the test is the probability that the criterion used in the test be at least as deviant as it was observed to be, from the value the test criterion should have if the hypothesis under test were true. In other words, the p-value is like a surprise index: the smaller the p-value (a number between 0 and 1) the more surprising the data would be if the hypothesis under test in fact were true.

The user is then invited to subscribe to this reasoning: since surprises are surprising by definition, meaning that they occur only rarely, if a surprising event does occur then this challenges the hypothesis that renders the event surprising. For this reason, small p-values are interpreted as speaking against the validity of the hypothesis under test.

To make this concrete, consider these four measurement results for the mass fraction of arsenic in kudzu: 0.920(73) mg/kg, 0.916(6) mg/kg, 0.963(21) mg/kg, and 0.890(13) mg/kg. The criterion Q takes the value 8.95. If the measurement results in fact were homogeneous, then this value should be like a drawing from a chi-square distribution with 3 degrees of freedom. The corresponding p-value is 0.03 (pink area in Exhibit 2).

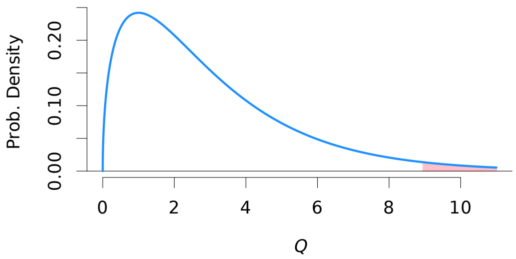


Exhibit 2: Reference distribution (chi-squared with 3 degrees of freedom) of Cochran's Q for quartets of mutually consistent measurement results, with pink area denoting the probability that Q is greater than 8.95.

The question then becomes: how small does the p-value need to be to warrant rejecting the hypothesis under test (which, in the context of this example, is that the results are homogeneous)? The answer depends on the user's appetite for risk: if the user rejects the hypothesis of homogeneity, then the user does so knowing that the probability is 0.03 of this being the wrong decision.

This risk needs to be weighed against another risk, quantified by the probability of failing to reject a hypothesis that is false. The complementary of this probability, which is the probability of rejecting a false hypothesis, is called the *power* of the test.

Ideally, a decision should minimize the probability of rejecting a true hypothesis, while maximizing the power of the test. Unfortunately, these are conflicting goals, and the user must resign herself to striking a balance between them. The choice is further complicated by the fact that, in significance tests such as are used in the NDT, the power of the tests is not a well-defined concept because there are so many different patterns of heterogeneity, and of asymmetry, and so many different shapes of non-Gaussian distributions.

Another consideration that plays a role when answering the question of how small the p-value needs to be to warrant rejecting the hypothesis under test, concerns the consequence of a wrong decision.

It just so happens that the NDT was designed to offer substantial protection against nefarious consequences of wrong decisions that the user can make when interpreting the p-values issuing from the three tests that drive the recommendation for how to reduce the data. In this light, we can suggest the following thresholds:

Cochran's Q Test It is safe to reject the hypothesis of homogeneity even for fairly high p-values. In this conformity, we recommend that the hypothesis be rejected rather freely, when the p-value is 10 % or smaller.

Miao-Gel-Gastwirth Test of Symmetry Since asymmetry can easily be confused with heaviness of the tails of the probability distribution of the laboratory or method effects (the $\{\lambda_j\}$ in Equation (1)), especially in small data sets as commonly arise in key comparisons, we recommend a very conservative stance on this count, hence that the hypothesis of symmetry be rejected only when the p-value is smaller than 1 %.

Shapiro-Wilk Test of Gaussian Shape The generally small numbers of participants in key comparisons call for a moderately conservative stance regarding distributional shape. Accordingly, we recommend that the hypothesis of Gaussian shape be rejected only when the p-value is smaller than 5 %.

4.3 Homogeneity and Dark Uncertainty

After the data will have been validated, and the user will have clicked on the tab for **2. Decision Tree**, the NDT will immediately perform and display the results of Cochran's Q test of homogeneity (Cochran, 1954) to determine whether the measured values are significantly more dispersed than their associated uncertainties suggest that they should be. *Homogeneity* thus means *mutual consistency*, not homogeneity of a material.

The p-value of the test is listed. A small p-value (less than 0.05 is conventionally regarded as being "small") suggests heterogeneity, hence that τ, the so-called *dark uncertainty* (Thompson and Ellison, 2011), is positive.

In this case, it is preferable to reject the hypothesis of homogeneity when it is true, than to accept it when it is false, hence, and in this conformity, it will not be harmful to regard p-values somewhat larger than 0.05 as being "small" for the purpose of the test.

The NDT also displays the size of τ relative to the median of the measured values and relative to the median of the standard uncertainties.

It is the user's responsibility to decide whether the NDT should assume that the measurement results are homogeneous, and to press either **Yes** or **No** accordingly.

The views differ, across the metrological community concerned with key comparisons, about whether the DoEs should, or should not express, in their uncertainty component, the amount of dark uncertainty when in fact dark uncertainty is detected and quantified in an estimate of the parameter τ. The issue, and the options that the NDT offers to include or exclude dark uncertainty from the DoEs, are introduced in section 6.

It should be noted that dark uncertainty, when it is detected and quantified, is always expressed in the uncertainty that surrounds the KCRV. The astute user may then note, and find the fact surprising that, for the example presented in section 7, the KCRV has associated standard uncertainty 0.55 ng/g, while the estimate of τ is 1.48 ng/g. Should not that uncertainty be larger than this estimate of τ, if such uncertainty indeed, and somehow, expresses this τ?

The solution of this apparent paradox is resolved once one realizes that the impact that τ has on the uncertainty associated with the KCRV is reduced through the averaging that takes place behind the scenes every time a model like that of Equation (1) is fitted to measurement results. The reason why is very much the same as the reason why, when m replicated determinations of a scalar quantity are combined into an equally weighted arithmetic average, the standard uncertainty of this average is $\sqrt{m}$ times smaller than the standard deviation of the replicates.

Remember that τ is the standard deviation of the laboratory effects, $\{\lambda_j\}$, in Equation (1), and that these laboratory effects can be construed as laboratory-specific biases that reflect the fact that some laboratories tend to measure high, while others tend to measure low.

The consensus value, which is the KCRV, the estimate of μ in Equation (1), is a kind of weighted average, with weights that depend on the particular model and on the measurement results themselves. Therefore, the magic of averages applies to it similarly to how it applies to averages of replicated determinations, the end-result being that the impact of the biases is reduced by averaging, and accordingly only a fraction of the dark uncertainty effectively becomes folded into the uncertainty associated with the KCRV.

Exhibit 3 uses measurement results simulated according to the version of the model in Equation (1) that applies to Example 1 as described in section 7, to illustrate how the aforementioned averaging effect reduces the impact of τ on the uncertainty associated with the KCRV, the more so the larger the number of participating laboratories.

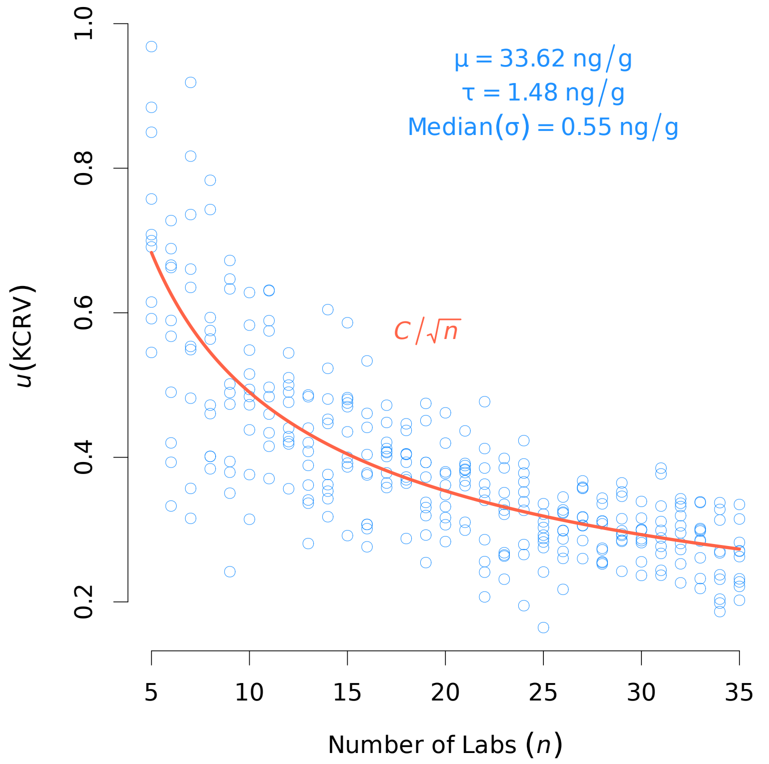


Exhibit 3: The impact upon u(KCRV), of the dark uncertainty, τ, which remained constant for the different numbers of laboratories considered in a simulation study, decreases with increasing n at a rate that is approximately inversely proportional to the square root of the number of laboratories (red curve).

For each of n laboratories, we simulated one measurement result and computed the corresponding KCRV and its associated uncertainty, u(KCRV). We did this for $n = 5, 6, \ldots, 35$ laboratories, and then ten times for each of these values of n, ending up with $(35 - 4) \times 10 = 310$ evaluations of u(KCRV).

The exhibit shows how u(KCRV) varies as a function of n. Since all these u(KCRV)s include the contribution from the same τ (which remained fixed at the value listed in the body of the plot for all 310 simulated datasets), the decrease of u(KCRV) with increasing n shows that the impact of τ upon the u(KCRV) decreases (through the averaging implied by procedure whereby the KCRV is estimated) with increasing n at a rate that is approximately inversely proportional to the square root of the number of laboratories.

4.4 Symmetry

When needed, the NDT will perform the test proposed by Miao et al. (2006) to determine whether the measured values (without any standardization) appear to be consistent with the hypothesis of their being a sample drawn from a symmetrical probability distribution.

The output of this test is its *p*-value, where, again, a small *p*-value (preferably less than 0.01 for this particular test) suggests that the data may not be a sample from a symmetrical distribution.

It is the user's responsibility to decide whether the NDT should assume that there is such symmetry, and to press either **Yes** or **No** accordingly.

4.5 Gaussian Shape (Normality)

Also when needed, the NDT will perform the Shapiro-Wilk test of Gaussian shape (Shapiro and Wilk, 1965), which is applied to the "roughly" standardized measured values: the differences between the measured values and their median, divided by the reported standard uncertainties.

The output of this test is its *p*-value: a small *p*-value (typically less than 0.05) suggests that the standardized measured values may not be a sample from a Gaussian distribution.

It is the user's responsibility to decide whether the NDT should assume that the standardized measured values are like a sample from a Gaussian distribution, and to press either **Yes** or **No** accordingly.

4.6 Model Selection

At this stage, the NDT displays the path that traverses the Decision Tree determined by the previous decisions, from its root (at the top) to a leaf (at the bottom), and offers the recommendation that corresponds to this leaf.

The recommendation is the item displayed in a drop-down menu from which the user will select the model to be fitted to the data: this can be the model that the NDT recommends, or any of the other four that the user may choose to adopt.

Once the user will have kept or modified the model selection, the user should proceed to **3. Fit Model**, by clicking this tab as displayed at the top of the page.

5 Model Fitting

Depending on the model selected in 4.6, the specification of model parameters will be required in **3. Fit Model** before the selected model will be fitted to the data, unless the user is willing to accept the default values suggested by the NDT, and merely clicks the button Run Method.

There are four classes of parameters that control model fitting and the presentation of results: General Parameters, Model Estimation Parameters, MCMC Parameters, and Prior Distribution Parameters. The last two categories are relevant only for Bayesian models, which are those that correspond to the three rightmost leaves of the NDT.

The General Parameters are the Random Number Seed and the Number of Significant Digits Reported. The Random Number Seed must be set prior to clicking the button labeled Run Method (bottom of this tab **3. Fit Model**).

If the same random number seed is used in two different runs, and the data, model, and the values of the other parameters also are the same, then exactly the same results will be obtained. Otherwise, if different random number seeds are used in different runs, then there may be slight differences between corresponding results.

The Number of Significant Digits Reported can be changed before or after the calculations will have been done: the displayed results will change accordingly without having to redo the calculations.

Values must also be specified for the following parameters (or the default values offered by the NDT accepted), for each of the five models implemented in the NDT:

(1) **Adaptive Weighted Average (AWA)** The number of bootstrap samples for the evaluation of the expanded uncertainties that are part of the degrees of equivalence, using the parametric bootstrap is entered into the box labeled Number of Bootstrap Replicates for Uncertainty Evaluations. We recommend that the number of bootstrap replicates be 5000 or larger.

(2) **Weighted Median (WM)** We recommend that the Number Bootstrap Runs be 5000 or larger. When there are at least 15 numerically different measured values, the nonparametric bootstrap is used; otherwise the parametric bootstrap is used (Efron and Tibshirani, 1993). In either case, the standard uncertainty associated with the weighted median is half the length

of the interval centered at the weighted median that includes 68 % of the bootstrap replicates.

(3) **Hierarchical Gauss + Gauss (HGG)** The Total Number of MCMC Steps should be twice as large as the Number of MCMC Warm-Up Steps. The positive integer assigned to, Keep an MCMC Draw Every ____ Steps specifies the thinning rate to apply to the MCMC samplers so as to reduce autocorrelations between draws from the posterior distribution of the model parameters given the data. If the value is 15, then these means that only every 15th element in the samples produced by the MCMC samplers will be retained, while the others are discarded.

The prior mean and prior standard deviation for μ in Equation (1) are specified in the boxes labeled Mu Prior Location (Default: `mean(x)`) and Mu Prior Scale (Default: `sd(x)/sqrt(3)`). The user should change these only if there is credible prior information about the measurand, for example as there can be when the measurement results are from a proficiency test, and the user wishes to use a consensus value that is the value determined by an expert laboratory, which may not even be a participant in the study.

The prior median for the dark uncertainty, τ, goes in the box labeled Tau Prior Median (Default: mad(x)), where `mad` denotes the rescaled median absolute deviation from the median (of the values in the vector `x`). The prior median for the standard deviations of the measurement errors, the $\{\epsilon_j\}$ in Equation (1), goes in Sigma Prior Median (Default: med(u)), where `med(u)` denotes the median of the standard uncertainties reported by the participants. However, this value is relevant only for those laboratories that have provided numbers of degrees of freedom supporting the $\{u(x_j)\}$.

(4) **Hierarchical Laplace + Gauss (HLG)** The control parameters are the same as for HGG.

(5) **Hierarchical Skew Student + Gauss** In addition to the control parameters required for HGG and HLG, the shape and scale of the prior gamma distribution for the number of degrees of freedom need to be specified in the boxes labeled Gamma Shape for Nu Prior and Gamma Scale for Nu Prior, and the standard deviation of the Gaussian prior distribution for the skewness parameter needs to be specified in the box labeled Alpha (Skewness) Prior Scale (the corresponding prior mean is zero). Refer to Possolo et al. (2021, 3.2.5) for details.

Once all the required choices will have been made, click the tab heading **Fit Model** and click the button Run Method. A progress bar on the lower left corner of the browser page indicates how close to completion the calculations are.

6 Results

Eventually, numerical results will be shown on-screen and two plots will be drawn. In addition, clicking the button labeled Download Report (PDF File) downloads an Adobe PDF with the data, choices made, and both numerical and graphical results.

Two buttons under Type of DoEs to Display serve to choose the kind of DoEs to present: these can be either DoEs Recognizing Dark Uncertainty or DoEs Ignoring Dark Uncertainty. In this manual, we refer to the former as $\{U_{95\,\%}(D_j)\}$, and to the latter as $\{U^*_{95\,\%}(D_j)\}$. The estimates $\{D_j\}$, of the DoEs are the same in both cases; only the associated expanded uncertainties for 95 % coverage differ.

The expanded uncertainties, $\{U_{95\,\%}(D_j)\}$, that recognize dark uncertainty, τ, include the contribution from τ (and from the uncertainty that surrounds the estimate of τ). In general, these uncertainties are larger than the corresponding uncertainties, $\{U^*_{95\,\%}(D_j)\}$, that ignore dark uncertainty. The choice between one kind and the other should reflect the purpose that the DoEs are intended to serve.

The random effects model in Equation (1) regards the participants in an interlaboratory study or key comparison, as being representative of the community of comparable laboratories even if they will not have been drawn at random from such community. In addition, determining which laboratories are "comparable" to the participating laboratories is a subjective judgment.

For this reason, the expanded uncertainties, $\{U_{95\,\%}(D_j)\}$, that recognize dark uncertainty, characterize the performance of such community as a collective, not the performance of the individual laboratories that the $\{U_{95\,\%}(D_j)\}$ are associated with. The typical $\{U_{95\,\%}(D_j)\}$ (defined, for example, as the median or the geometric average of these expanded uncertainties) quantifies the closeness to the KCRV that a laboratory, randomly selected from that community, would have been expected to have achieved if this laboratory would have participated in the comparison.

The expanded uncertainties, $\{U^*_{95\,\%}(D_j)\}$, that ignore dark uncertainty, are inconsistent with the model in Equation (1), and reflect only the uncertainty claimed (and reported) by the individual participants, as well as the uncertainty associated with the KCRV, and the correlation that exists between the error affecting the KCRV and the errors affecting to measured values. The $\{U^*_{95\,\%}(D_j)\}$ can play the role of uncorroborated claims made by the participants, which the participants feel reasonably entitled to make when they advertise their measurement services.

The CIPM's Mutual Recognition Arrangement (Comité International des Poids et Mesures (CIPM), 1999, T.3) points out that while "a key comparison reference value is normally a close approximation to the corresponding SI value, it is possible that some of the values submitted by individual participants may be even closer." However, and by the same token, it is conceivable that the KCRV, being no more than an estimate of the true value of the measurand, can deviate markedly from this true value. Therefore, no degree of equivalence, regardless of its kind or size, can, with full confidence, be interpreted as documenting actual measurement performance.

The situation is rather different in a proficiency test where the true value of the measurand is known to within a margin of uncertainty sufficiently narrow for the purpose the test is intended to serve. In such case, it is possible to gauge absolute performance to at least within this margin of uncertainty.

The leftmost plot depicts the measurement results, the consensus value, and a horizontal (yellow) band whose height represents $\widehat{\mu} \pm u(\widehat{\mu})$. For each measurement result, an open diamond marks the measured value, x_j, a thick vertical line segment represents $x_j \pm u(x_j)$, and the thin vertical line segment represents $x_j \pm (\tau^2 + u^2(x_j))^{½}$.

The rightmost plot depicts the degrees of equivalence, which will be $\{(D_j, U_{95\,\%}(D_j))\}$ or $\{(D_j, U^*_{95\,\%}(D_j))\}$, depending on the choice made using the buttons under Type of DoEs to Display, where $D_j = x_j - \widehat{\mu}$ for $j = 1, \ldots, n$.

While $u^2(D_j)$, the squared standard uncertainty associated with the difference D_j could, in principle, be computed as $u^2(x_j) + u^2(\widehat{\mu}) - 2u(x_j)u(\widehat{\mu})r$, where r denotes the correlation between x_j and $\widehat{\mu}$, this is not how the NDT does it in practice because, in general, there are no closed-form expressions for $u(\widehat{\mu})$, for r, or for the coverage factor k that would yield $U_{95\,\%}(D_j)$ as $ku(D_j)$. Instead, the NDT uses the Monte Carlo methods that Koepke et al. (2017, §6.1,§6.2) describe, for the **Adaptive Weighted Average** and for the three Hierarchical Bayesian procedures.

Neither are closed-form expressions generally available to compute $u(\widehat{\mu})$ or $U_{95\,\%}(D_j)$

when $\widehat{\mu}$ is a weighted median. The corresponding uncertainty evaluations are done using either the nonparametric or the parametric version of the statistical bootstrap (Possolo et al., 2021, §3.2.2), depending on the number of laboratories involved whose results contribute to the calculation of the KCRV.

The Unilateral Degrees of Equivalence Table and the Table of Uncertainties for Each Lab show comprehensive collections of numerical results from fitting the model. The Table of Uncertainties for Each Lab can be downloaded by clicking the button labeled Download Lab Uncertainties Table (.csv File). The headers of these tables are explained under Column Name Descriptions, farther down the web page with results.

Finally, under MCMC Sampler Diagnostics the user can find numerical diagnostics indicative of the performance of the Markov Chain Monte Carlo sampler used to fit a Bayesian model to the data, if a Bayesian model indeed was selected by the user.

After selecting and fitting one model, the user can return to **2. Decision Tree**, and select and fit another model. However, we believe that trying several or all models implemented in the NDT for the purpose of choosing the procedure that produces what best matches the user's notion of "ideal" results would be statistical malpractice.

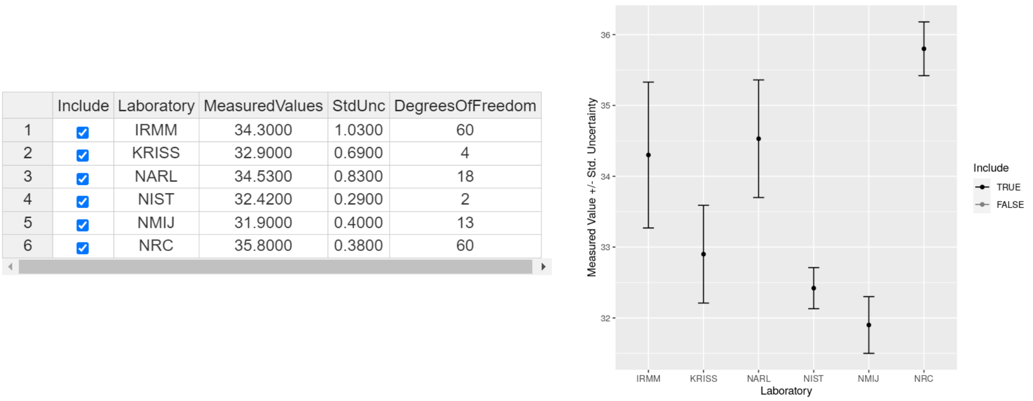
7 Example

Exhibit 4 shows the **1. Data** tab after pasting the measurement results for PCB 28 from CCQM-K25 (Schantz and Wise, 2004) into the mini-spreadsheet. The NDT creates their graphical representation automatically.

Exhibit 5 shows the results of the statistical tests of homogeneity, symmetry, and normality from the **2. Decision Tree** tab, after the user will have decided that the measurement results are heterogeneous and that the standardized measured values can reasonably be regarded as a sample from a probability distribution that is symmetrical and Gaussian. Note that the NDT has highlighted the path traversed from the root (at the top) to a leaf (at the bottom) of the *Decision Tree*.

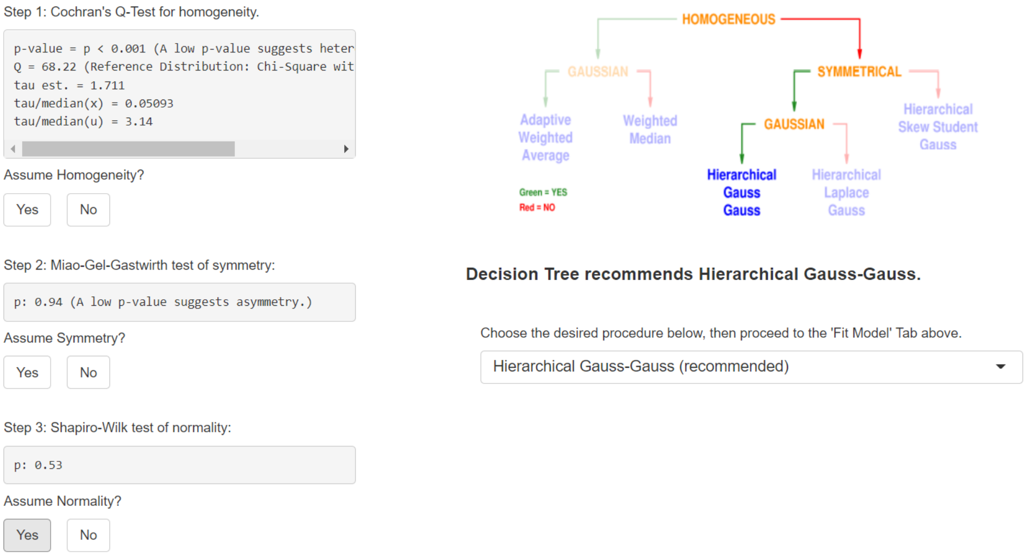
Exhibit 6 shows the **3. Fit Model** tab before fitting the model to the measurement results in **1. Data**. It lists the default values of the parameters that need to be specified in order to fit the selected model.

Exhibit 7 shows **3. Fit Model** after fitting the selected model to the measurement



	Include	Laboratory	MeasuredValues	StdUnc	DegreesOfFreedom
1	☑	IRMM	34.3000	1.0300	60
2	☑	KRISS	32.9000	0.6900	4
3	☑	NARL	34.5300	0.8300	18
4	☑	NIST	32.4200	0.2900	2
5	☑	NMIJ	31.9000	0.4000	13
6	☑	NRC	35.8000	0.3800	60

Exhibit 4: **1. Data** after filling the mini-spreadsheet with the measurement results for PCB 28 from CCQM-K25.



Step 1: Cochran's Q-Test for homogeneity.

```
p-value = p < 0.001 (A low p-value suggests heter
Q = 68.22 (Reference Distribution: Chi-Square wit
tau est. = 1.711
tau/median(x) = 0.05093
tau/median(u) = 3.14
```

Assume Homogeneity?

Step 2: Miao-Gel-Gastwirth test of symmetry:

```
p: 0.94 (A low p-value suggests asymmetry.)
```

Assume Symmetry?

Step 3: Shapiro-Wilk test of normality:

```
p: 0.53
```

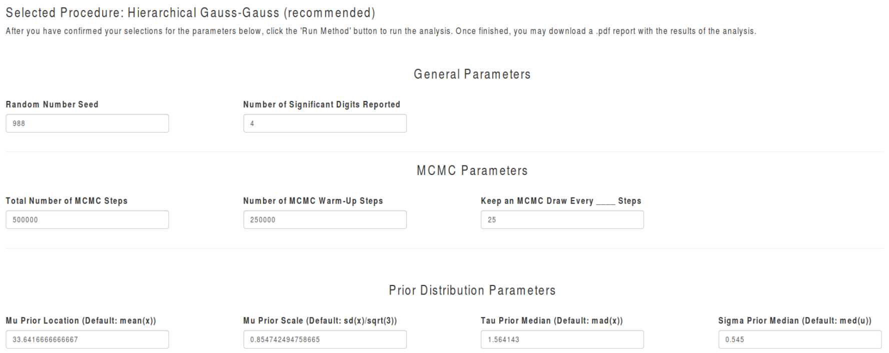
Assume Normality?

Decision Tree recommends Hierarchical Gauss-Gauss.

Choose the desired procedure below, then proceed to the 'Fit Model' Tab above.

Hierarchical Gauss-Gauss (recommended)

Exhibit 5: **2. Decision Tree** with recommendation offered by the NDT after the user's decisions about homogeneity of the measurement results, and symmetry and normality of the measured values.



Selected Procedure: Hierarchical Gauss-Gauss (recommended)

After you have confirmed your selections for the parameters below, click the 'Run Method' button to run the analysis. Once finished, you may download a .pdf report with the results of the analysis.

General Parameters

Random Number Seed: 988

Number of Significant Digits Reported: 4

MCMC Parameters

Total Number of MCMC Steps: 500000

Number of MCMC Warm-Up Steps: 250000

Keep an MCMC Draw Every ____ Steps: 25

Prior Distribution Parameters

Mu Prior Location (Default: mean(x)): 33.6416666666667

Mu Prior Scale (Default: sd(x)/sqrt(3)): 0.854742494758665

Tau Prior Median (Default: mad(x)): 1.564143

Sigma Prior Median (Default: med(u)): 0.545

Exhibit 6: **3. Fit Model** showing default values of the parameters that need to be specified to fit the model selected previously.

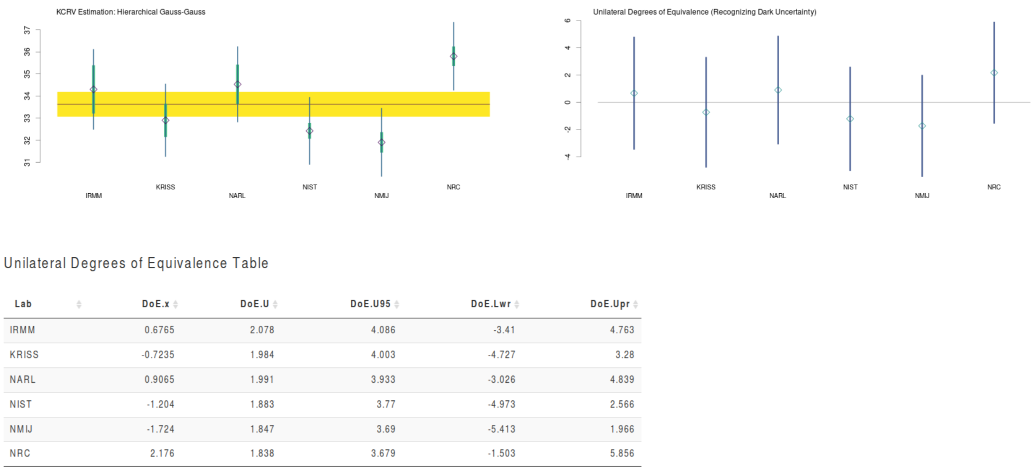
results in **1. Data**. The consensus value, its associated standard uncertainty, a 95 % coverage interval for its true value, and an estimate of the dark uncertainty τ, and the endpoints of a 95 % credible interval for τ are listed on the web page with the output, and are also transcribed into the report produced by clicking the button labeled Download Report (PDF File).

The measurement results are displayed graphically on the left panel, now also including the consensus value (horizontal brown line) and a 1-σ uncertainty band surrounding it, as well as the reported uncertainties associated with the measured values and the result of adding the contribution from τ in quadrature, to the reported uncertainties. The degrees of equivalence are displayed graphically on the right panel, and listed in a table toward the bottom of the page.

If one of the Bayesian models is run (Hierarchical Gauss-Gauss, Hierarchical Laplace-Gauss, or Hierarchical Skew-Student-t), then diagnostics for the MCMC sampler are also listed.

As a general recommendation, if any of the R-hat values are greater than 1.05, then the sampler may not have reached equilibrium, and the Total Number of MCMC Steps should be increased, and the run repeated. In any case, the Number of MCMC Warm-Up Steps should be about half of the Total Number of MCMC Steps.

The effective sample size, n.eff, takes into account autocorrelations between values of the parameters sampled by the Markov Chain(s), and expresses the equivalent number of independent values that the results are based on. The effective sample sizes can be different for different parameters.


KCRV Estimation: Hierarchical Gauss-Gauss
37
36
35
34
33
32
31
IRMM
KRISS
NARL
NIST
NMIJ
NRC
Unilateral Degrees of Equivalence (Recognizing Dark Uncertainty)
6
4
2
0
-2
-4
IRMM
KRISS
NARL
NIST
NMIJ
NRC

Unilateral Degrees of Equivalence Table

Lab	DoE.x	DoE.U	DoE.U95	DoE.Lwr	DoE.Upr
IRMM	0.6765	2.078	4.086	-3.41	4.763
KRISS	-0.7235	1.984	4.003	-4.727	3.28
NARL	0.9065	1.991	3.933	-3.026	4.839
NIST	-1.204	1.883	3.77	-4.973	2.566
NMIJ	-1.724	1.847	3.69	-5.413	1.966
NRC	2.176	1.838	3.679	-1.503	5.856

Exhibit 7: **3. Fit Model** after fitting the Hierarchical Gauss+Gauss model to the measurement results specified in the **1. Data** tab.

Acknowledgments

Michal Mariassy (Slovak Institute of Metrology, SMU), Andrew Rukhin (National Institute of Standards and Technology, NIST), Blaza Toman (National Institute of Standards and Technology, NIST), and Jochen Vogl (Bundesanstalt für Materialforschung und -prüfung, BAM), reviewed a draft of Possolo et al. (2021), provided valuable comments and suggestions, and asked challenging questions, which induced numerous improvements. David Duewer and Donald Windover (both from NIST), have, time and again, provided excellent suggestions for improvement of both the NIST Decision Tree and the NIST Consensus Builder.

The authors are very grateful to the members of the Inorganic Analysis Working Group (IAWG) of the CCQM, for their interest in the *Decision Tree*, and for the questions and comments they offered about it during recent meetings of the CCQM-IAWG. The authors also thank Andrea Peruzzi (NRC Canada) and other members of the Consultative Committee for Thermometry (CCT) for expressing a similar interest in the Decision Tree. Finally, we are highly appreciative of Claudia Swart's (PTB, Germany) probing questions that prompted the discussion of the impact of dark uncertainty on the uncertainty associated with the KCRV that is presented in section 4.3.

References

A. Baddeley and R. Turner. spatstat: An R package for analyzing spatial point patterns. *Journal of Statistical Software*, 12:1–42, 2005. URL www.jstatsoft.org/v12/i06/.

M. Borenstein, L. V. Hedges, J. P. T. Higgins, and H. R. Rothstein. *Introduction to Meta-Analysis*. John Wiley & Sons, 2009. ISBN 978-0-470-05724-7. doi: 10.1002/9780470743386.refs.

M. Borenstein, L. V. Hedges, J. P. T. Higgins, and H. R. Rothstein. A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods*, 1:97–111, 2010. doi: 10.1002/jrsm.12.

W. G. Cochran. The combination of estimates from different experiments. *Biometrics*, 10(1):101–129, March 1954. doi: 10.2307/3001666.

Comité International des Poids et Mesures (CIPM). *Mutual Recognition of National Measurement Standards and of Calibration and Measurement Certificates Issued by National Metrology Institutes*. Bureau International des Poids et Mesures (BIPM), Pavillon de Breteuil, Sèvres, France, October 14th 1999. URL www.bipm.org/en/cipm-mra/. Technical Supplement revised in October 2003.

R. DerSimonian and N. Laird. Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7(3):177–188, September 1986. doi: 10.1016/0197-2456(86)90046-2.

B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Springer-Science+Business Media, Dordrecht, The Netherlands, 1993. ISBN 978-0-412-04231-7. doi: 10.1201/978-0-4292-4659-3.

M. Harrer, P. Cuijpers, T. A. Furukawa, and D. D. Ebert. *Doing Meta-Analysis With R: A Hands-On Guide*. Chapman & Hall/CRC Press, Boca Raton, FL, 2021. ISBN 978-0-367-61007-4.

J. Hartung, G. Knapp, and B. K. Sinha. *Statistical Meta-Analysis with Applications*. John Wiley & Sons, Hoboken, NJ, 2008. ISBN 978-0-470-29089-7.

D. C. Hoaglin. Misunderstandings about Q and 'Cochran's Q test' in meta-analysis. *Statistics in Medicine*, 35:485–495, 2016. doi: 10.1002/sim.6632.

Joint Committee for Guides in Metrology (JCGM). *Evaluation of Measurement Data — Guide to the Expression of Uncertainty in Measurement*. International Bureau of Weights and Measures (BIPM), Sèvres, France, 2008. URL https://www.bipm.org/documents/20126/2071204/JCGM_100_2008_E.pdf/cb0ef43f-baa5-11cf-3f85-4dcd86f77bd6. BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP and OIML, JCGM 100:2008, GUM 1995 with minor corrections.

A. Koepke, T. Lafarge, A. Possolo, and B. Toman. Consensus building for interlaboratory studies, key comparisons, and meta-analysis. *Metrologia*, 54(3): S34–S62, 2017. doi: 10.1088/1681-7575/aa6c0e.

J. Meija, O. Bodnar, and A. Possolo. Ode to Bayesian Methods in Metrology. *Metrologia*, 60:052001, 2023. doi: 10.1088/1681-7575/acf66b.

W. Miao, Y. R. Gel, and J. L. Gastwirth. A new test of symmetry about an unknown median. In A. C. Hsiung, Z. Ying, and C.-H. Zhang, editors, *Random Walk, Sequential Analysis and Related Topics: A Festschrift in Honor of Yuan-Shih Chow*, pages 199–214. World Scientific Publishing Company, Singapore, 2006. doi: 10.1142/9789812772558_0013.

A. Possolo and J. Meija. *Measurement Uncertainty: A Reintroduction*. Sistema Interamericano de Metrologia (SIM), Montevideo, Uruguay, 2nd edition, 2022. ISBN 978-0-660-42866-6. doi: 10.4224/1tqz-b038.

A. Possolo, A. Koepke, D. Newton, and M. R. Winchester. Decision Tree for Key Comparisons. *Journal of Research of the National Institute of Standards and Technology*, 126:126007, 2021. doi: 10.6028/jres.126.007.

M. Schantz and S. Wise. CCQM–K25: Determination of PCB congeners in sediment. *Metrologia*, 41(*Technical Supplement*):08001, 2004. doi: 10.1088/0026-1394/41/1A/08001.

M. Schantz, S. Wise, G. Gardner, C. Fraser, J. McLaren, P. Lehnik-Habrink, E. C. Galván, H. Schimmel, D.-H. Kim, G. S. Heo, D. Carter, P. Taylor, and T. Yarita. CCQM-K25: Key Comparison — determination of PCB congeners in sediment — Final Report 12 Dec 2003. Technical report, Consultative Committee for Amount of Substance — Metrology in Chemistry, Bureau International des Poids et Mesures, Sèvres, France, December 2003. URL kcdb.bipm.org/appendixB/appbresults/ccqm-k25/ccqm-k25_final_report.pdf. CCQM-K25.

S. S. Shapiro and M. B. Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3,4):591–611, 1965. doi: 10.2307/2333709.

M. Thompson and S. L. R. Ellison. Dark uncertainty. *Accreditation and Quality Assurance*, 16:483–487, October 2011. doi: 10.1007/s00769-011-0803-0.