NIST Decision Tree
User's Manual

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

May 20, 2022

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

Contents

1 Introduction **4**

2 Access **4**

3 Data **5**

4 Decision Tree **7**

4.1 Measurement Model 8
4.2 Homogeneity 8
4.3 Symmetry 9
4.4 Normality 9
4.5 Model Selection 9

5 Model Fitting and Results **10**

6 Example **11**

Exhibits

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. 7

2 **1. Data** after filling the mini-spreadsheet with the measurement results for PCB 28 from CCQM-K25. 12

3 **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. . . 13

4 **3. Fit Model** showing default values of the parameters that need to be specified to fit the model selected previously. 13

5 **3. Fit Model** after fitting the Hierarchical Gauss+Gauss model to the measurement results specified in the Data tab. 14

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 key comparisons;

(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* in the context of key 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. Section 6 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.

2 Access

Access the NDT via a Web browser by visiting

https://decisiontree.nist.gov

The landing page has three tabs: **1. Data**, **2. Decision Tree**, and **3. Fit Model**, which 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.

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 (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.

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

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;

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

The measurement results entered into the mini-spreadsheet will be depicted automatically alongside. The dots represent the measured values, and each vertical line segment represents a measured value plus or minus one standard uncertainty.

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.

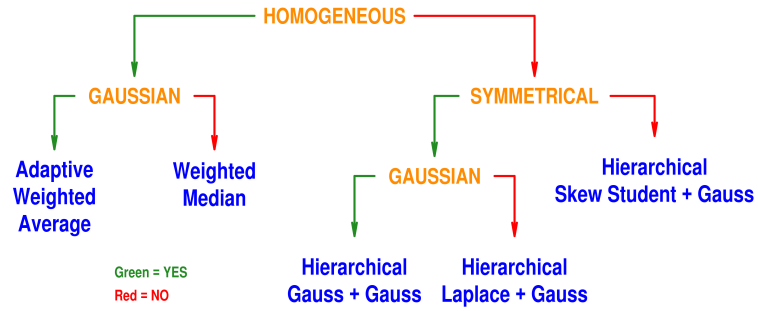


Exhibit 1: The Decision Tree comprises four branching nodes (orange) and five leaves (blue) that suggest different models for the measurement results and corresponding procedures for data reduction. A question needs to be answered at each node: if the answer is YES, then one follows the green branch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

4 Decision Tree

Figure 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.

Since the traversal of the Decision Tree is guided by classical statistical tests of hypotheses (for homogeneity, symmetry, and Gaussian shape), it inherits all the general and specific limitations of such classical testing.

As an example of the general limitations, consider the test of Gaussian shape for the probability distribution of the measured values: as the number of measurement results increases, it is increasingly likely that the test will reject the "null" hypothesis of Gaussian shape for the simple reason that the test becomes increasingly sensitive to departures from such idealized shape, regardless of whether such departures are consequential or not.

As an example of the specific limitations, consider the test of homogeneity (or, mutual consistency) discussed in §4.2. This particular test has low power to detect heterogeneity when it exists, and suffers from several other, well-known

shortcomings (Hoaglin, 2016).

4.1 Measurement Model

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

The NDT aims to select and fit a specific version of the following model to the measurement results:

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

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.

4.2 Homogeneity

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.

4.3 Symmetry

When needed, the NDT will perform the test proposed by Miao et al. (2006) to determine whether the measured values 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 (say, less than 0.05, or maybe preferably in this case, less than 0.01) 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.

4.4 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) indicates that the standardized measured values are unlikely to 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.

4.5 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 displayed item 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 and Results

Depending on the model selected in 4.5, additional inputs may be required in **3. Fit Model** before the selected model will be fitted to the data. The NDT will offer default values for these additional inputs, which the user is free to modify. We recommend that the number of bootstrap replicates be at least 5000.

The required inputs are as follows, for each of the five models implemented in the NDT:

(1) **Adaptive Weighted Average** The number of bootstrap samples for the evaluation of the expanded uncertainties that are part of the degrees of equivalence, using the parametric bootstrap.

(2) **Weighted Median** The number of bootstrap samples for the evaluation of the standard uncertainty associated with the consensus value. When there are at least 15 numerically different measurement results, the non-parametric 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** The median of the prior distribution for τ, and the median of the prior distributions for the $\{\sigma_j\}$ (the latter is relevant only for those laboratories or measurement methods whose numbers of degrees of freedom supporting the $\{u(x_j)\}$ are available).

(4) **Hierarchical Laplace + Gauss** The median of the prior distribution for τ, and the median of the prior distributions for the $\{\sigma_j\}$ (the latter is relevant only for those laboratories or measurement methods whose numbers of degrees of freedom supporting the $\{u(x_j)\}$ are available).

(5) **Hierarchical Skew Student + Gauss** The median of the prior distribution for τ, and the median of the prior distributions for the $\{\sigma_j\}$ (the latter is relevant only for those laboratories or measurement methods whose numbers of degrees of freedom supporting the $\{u(x_j)\}$ are available). Also the shape and rate parameters for the gamma prior distribution for the number of degrees of freedom of Student's skewed t distribution, and the prior standard deviation for its skewness parameter.

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.

Eventually, numerical results will be printed 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.

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 segments represents $x_j \pm u(x_j)$, and the thin vertical line segment represents $x_j \pm (\tau^2 + u^2(x_j))^{1/2}$.

The rightmost plot depicts the degrees of equivalence $\{(D_j, U_{95\,\%}(D_j)$, where $D_j = x_j - \widehat{\mu}$ for $j = 1, \ldots, n$.

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.

6 Example

Exhibit 2 shows the **1. Data** tab after pasting the measurement results for PCB 28 from CCQM-K25 into the mini-spreadsheet. The NDT creates their graphical representation automatically.

Exhibit 3 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

	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

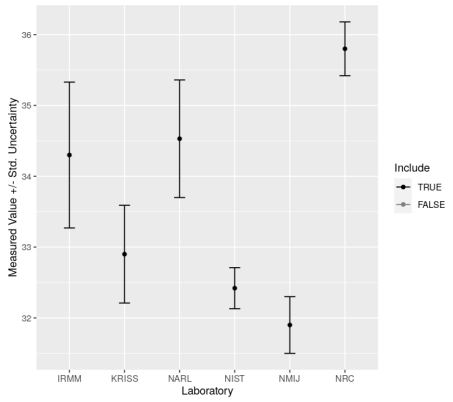


Exhibit 2: **1. Data** after filling the mini-spreadsheet with the measurement results for PCB 28 from CCQM-K25.

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 4 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 5 shows **3. Fit Model** after fitting the selected model to the measurement 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 at the top left corner.

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 at the bottom.

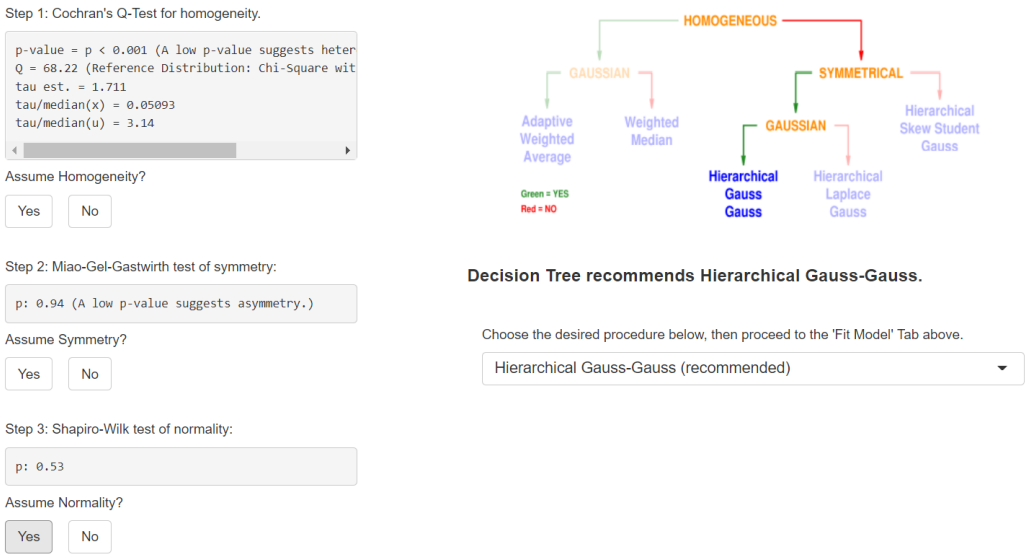


Exhibit 3: **2. Decision Tree** with recommendation offered by the NDT after the user's decisions about homogeneity of the measurement results, and symmetry and normality of the measured values.

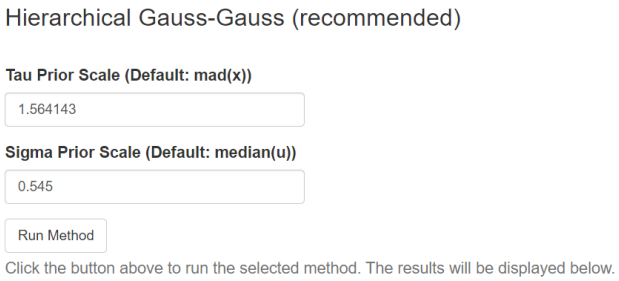
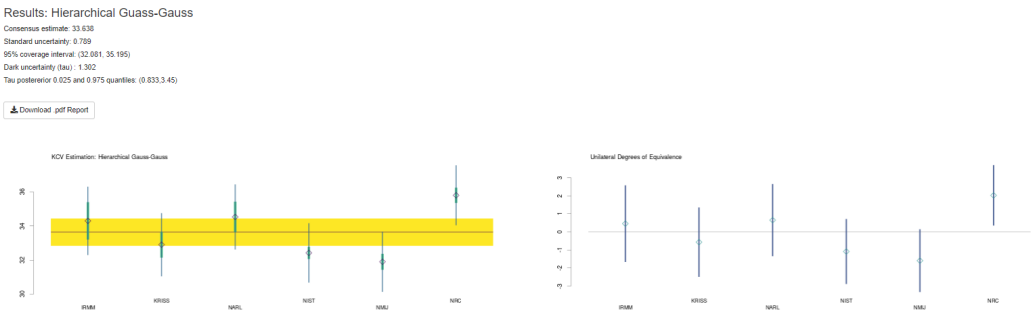


Exhibit 4: **3. Fit Model** showing default values of the parameters that need to be specified to fit the model selected previously.



Unilateral Degrees of Equivalence Table

Lab	DoE.x	DoE.U95	DoE.Lwr	DoE.Upr
IRMM	0.455062	1.06795	-1.63491	2.54503
KRISS	-0.569318	0.974913	-2.45829	1.31966
NARL	0.650742	0.991453	-1.3149	2.61638
NIST	-1.08663	0.882166	-2.85351	0.680257
NMIJ	-1.592	0.860267	-3.2975	0.113504
NRC	2.02318	0.84111	0.382453	3.6639

Showing 1 to 6 of 6 entries

Exhibit 5: **3. Fit Model** after fitting the Hierarchical Gauss+Gauss model to the measurement results specified in the 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.

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.

References

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.

B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, London, UK, 1993.

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.

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, 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, 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.