NIST Consensus Builder
User's Manual

Amanda Koepke, Thomas Lafarge,
Antonio Possolo, Blaza Toman

Statistical Engineering Division
Information Technology Laboratory

May 13, 2017

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

Contents

Exhibits

1 Introduction

The NIST Consensus Builder (NICOB) serves to combine measurement results obtained by different laboratories, or by application of different measurement methods, into a consensus estimate of the value of a scalar measurand. The NICOB qualifies the consensus estimate with an evaluation of measurement uncertainty that captures not only the stated uncertainties associated with the individual measured values, but also any additional component of uncertainty that manifests itself only when these measured values are inter-compared.

The NICOB can also report the differences between individual measured values and the consensus value, and the differences between pairs of values measured by different laboratories or methods, in both cases qualifying these differences with evaluations of associated uncertainty. In the context of *Key Comparisons*, these differences and associated uncertainties are called (unilateral, and bilateral, respectively) *degrees of equivalence* (DoE) (Comité International des Poids et Mesures (CIPM), 1999).

When the reported measurement uncertainties associated with the individual measured values are qualified with the numbers of degrees of freedom that they are based on, these numbers are taken into account as well. In general, the numbers of degrees of freedom convey the reliability of the evaluations of measurement uncertainty, expressing the extent of the underlying evidentiary basis, be it the size of the experimental data or the strength of other information used when producing the evaluations.

According to the *Technical protocol for a key comparison* (Comité International des Poids et Mesures (CIPM), 2014, 4.4), reporting these numbers of degrees of freedom is required for Key Comparisons: "Uncertainties are evaluated at a level of one standard uncertainty and information must be given on the number of effective degrees of freedom required for a proper estimation of the level of confidence." However, the reports of many Key Comparisons do not list them.

Section 2 summarizes the steps that need to be taken to use the NICOB. Section 3 outlines several guiding principles that define the methods implemented in the NICOB, and illustrates and discusses these methods as they are applied to a set of measurement results for one of the measurands considered in key comparison CCQM-K25 (Schantz et al., 2003).

After reading Sections 2 and 3, users should be ready to make informed choices to apply the NICOB to their own data, and to interpret the results, without further

study of this manual. However, for the reader wishing to gain a more thorough appreciation for the technology implemented in the NICOB, the Appendix reviews details of the statistical methods, and Section 4 presents additional examples of application using data from the following studies, in all cases providing background information on the study, detailing the data that were used, and explaining the meaning of the results.

- *Carotid Artery Stenosis* (§4.1) reviews a meta-analysis in medicine comparing the performance of two alternative procedures for the treatment of carotid stenosis. Such retrospective comparisons of medical procedures, or of medical centers, which aim to strengthen conclusions by pooling data from multiple studies, account for the bulk of the inter-comparisons and collaborative trials conducted and published in any particular year and across all fields of application. In this example, the data are counts of cases of stroke or death, and require some pre-processing before they can be input into the NICOB.

- *Length of Gauge Blocks* (§4.2) uses the results of key comparison CCL-K1, carried out by the CIPM's Consultative Committee for Length, and addresses the issue of heterogeneity differently from a previously published reanalysis of the same data (Cox, 2007). In this example, numbers of degrees of freedom are available that qualify the standard uncertainties associated with the measured values.

- *Triple Point of Water* (§4.3) is part of key comparison CCT-K7, conducted by the CIPM's Consultative Committee for Thermometry, and concerns a comparison between national reference standards and a BIPM reference standard. A mixture model (*Linear Pool*) is fitted to a sample of simulated values of the consensus value that is produced by the NICOB, and compared with a similar statistical analysis described in the CCT-K7 Final Report.

- *Activity of Radionuclide* ^{60}Co (§4.4) is being measured in the ongoing key comparison BIPM.RI(II)-K1.Co-60, organized by the CIPM's Consultative Committee for Ionizing Radiation (Section II, Measurement of Radionuclides) that supports the International Reference System (SIR) maintained at the International Bureau of Weights and Measures (BIPM) in Sèvres, France. This inter-comparison involves a considerably larger number of participants than most key comparisons, and exhibits marked differences

both between the measured values and between the associated uncertainties.

- *Radiofrequency Power Sensor* (§4.5) reanalyzes measurement results from key comparison CCEM.RF-K25.W, and compares the key comparison reference value (KCRV) computed in the original study with its counterparts produced by the procedures implemented in the NICOB. It also discusses data selection based on a statistical criterion for outlier detection.

The accompanying graphical representations of the results obtained in these examples, any pre-processing that the data will have had to undergo in preparation for their use in the NICOB, and also alternative analyses that are presented in some cases, using methods not available in the NICOB, all were done using the R environment for statistical computing and graphics (R Core Team, 2015).

Section 5 emphasizes that the NICOB ought not be misconstrued as a toolbox capable of addressing all the needs of data reductions arising in the context of interlaboratory studies or inter-method comparisons, and discusses cases where either it simply cannot provide a satisfactory solution to the problem of consensus building, or where its application would be inappropriate.

For example, the NICOB does not offer means to address the challenge posed by measurands whose values may drift in the course of an inter-comparison. Neither is it suitable for the analysis of results from proficiency tests because it does not produce the performance metrics that typically are the focus of such tests (Thompson et al., 2006).

Section 6 summarizes technical details of the implementation and deployment of the NICOB as an application available in the World Wide Web.

2 Quick Start

2.1 Access

Access the NICOB via a Web browser by visiting `consensus.nist.gov`, which will display a page as illustrated in Exhibit 1 on Page 7. Clicking on About the NIST Consensus Builder brings up general information about the application. After inputting values for the fields displayed on the Enter data page, as described in §2.2, the user can verify that these inputs are valid by clicking the

button at the bottom of the page labeled Validate inputs. Then the user selects a data reduction method from among the three listed on the left hand-side of the page: DerSimonian-Laird, Hierarchical Bayes, or Linear Pool.

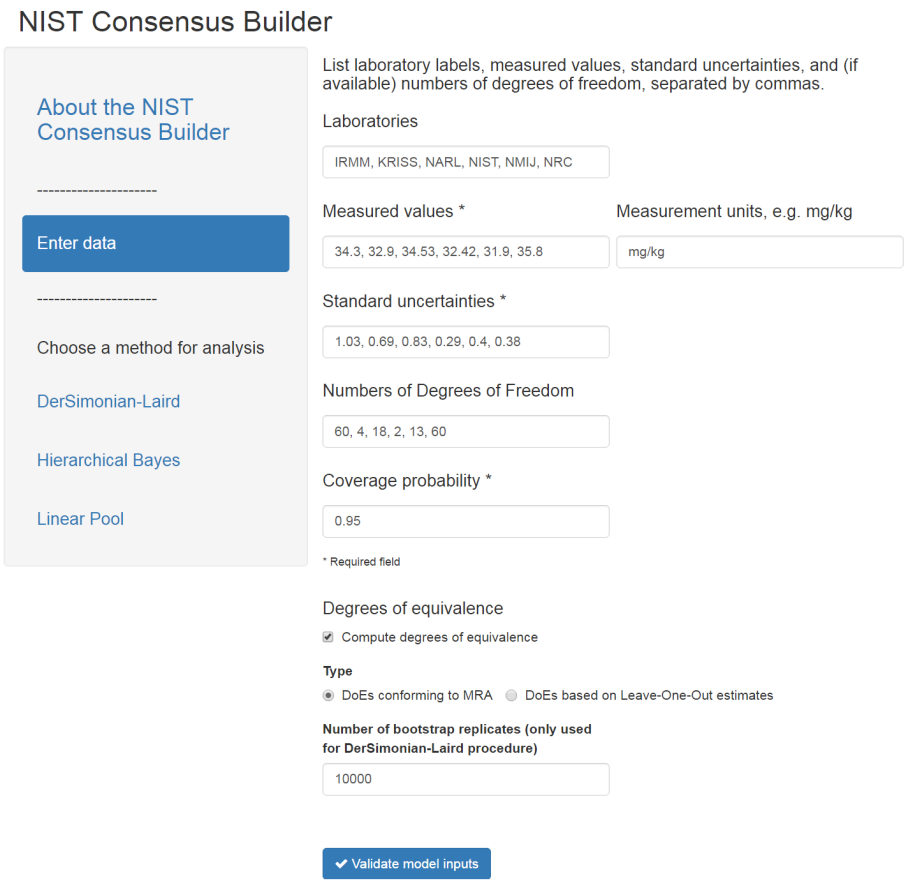

Exhibit 1: User interface for the NICOB presented by a Web browser when visiting consensus.nist.gov.

Buttons at the bottom of the Enter data page, shown in Exhibit 2, allow the user to load and save configuration files with inputs for the NICOB. Clicking the button labeled Save Configuration File downloads a plain text file named consensus.ncb to the local machine, which specifies the current inputs for the NICOB. To use a previously saved configuration file, search for and select the file using

the Browse button.

Alternatively, the NICOB also accepts configuration files with inputs specified as comma separated values and extensions `.ncb`, `.csv`, or `.txt`. Each row of the file designates data from a different laboratory or measurement method, and the file can have two, three, or four comma separated columns. For each row, data should by entered in the order: name (if available), measured value, standard uncertainty, and number of degrees of freedom (if available; missing or infinite degrees of freedom should be entered as `Inf`).

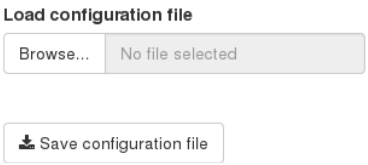

Exhibit 2: Buttons at the bottom of the Enter data page which allow the user to load or save configuration files with inputs for the NICOB.

2.2 General Inputs

- Labels designating the n participating laboratories (REQUIRED — character strings comprised of letters or numbers, separated from one another by commas).
- Measured values $x_1, \ldots, x_n$ produced by n different laboratories or measurement methods (REQUIRED — numbers separated by commas, which may be written in scientific notation as in 3.52e1 or 352e−1, both meaning 35.2).
- Measurement units to qualify the numerical values of the measured values, for example mg/kg, which are used to label axes of plots (OPTIONAL — character string).
- Standard uncertainties $u_1, \ldots, u_n$ associated with the measured values (REQUIRED — positive numbers separated by commas).
- Numbers of degrees of freedom $\nu_1, \ldots, \nu_n$ on which the standard uncertainties are based (OPTIONAL — positive numbers separated by commas). Missing or infinite degrees of freedom should be entered as `Inf`.

- Coverage probability (positive number between 0 and 1) desired for the coverage intervals (REQUIRED, DEFAULT: 0.95).
- Indication, by means of a check-box, of whether *degrees of equivalence* should be computed (DEFAULT: Not computed).

  If this box is checked, additional input fields appear and the user is prompted to enter the following:

  - Indication, by means of a radio button, of whether *degrees of equivalence* should be computed as defined in the MRA or based on *leave-one-out* estimates, as explained in Sections 3.2.8 and A.5.
  - Number of bootstrap replicates for *degrees of equivalence* uncertainty calculation. This is only used for the DerSimonian-Laird procedure (DEFAULT: 10 000); the Hierarchical Bayes and Linear Pool procedures use for this number the sample sizes of their method specific inputs.

2.3 Method Selection

Many different models for data from interlaboratory studies and meta-analysis, and many different ways of fitting them, have been proposed in the literature, for example by: Mandel and Paule (1970), Rocke (1983), Hedges and Olkin (1985), Mandel (1991), Whitehead and Whitehead (1991), Crowder (1992), Hunter et al. (1992), Searle et al. (1992), Vangel and Rukhin (1999), Cox (2002), Steele et al. (2002), Iyer et al. (2004), Toman (2007), Cooper et al. (2009), Rukhin (2009), Toman and Possolo (2009b), Elster and Toman (2010), Rukhin and Possolo (2011), and Bodnar et al. (2016), among many others.

The three methods implemented in the NICOB were selected deliberately to be very different from one another in several important ways. They are not meant to be interchangeable, and the user should consider their characteristics, including advantages and disadvantages indicated below, to determine which may be best for the intended purpose. Trying them all and selecting the procedure that produces the results that best match the user's preconceptions, or the user's notion of "ideal" results, would be statistical malpractice.

In the examples discussed in §3.2 and in §4, typically all three procedures implemented in the NICOB are applied, not because they are comparably adequate, but to provide an opportunity to assess the sensitivity of the conclusions to the choices of model and model-fitting procedure.

Several methods that are commonly used, for example, maximum likelihood estimation and cognates (ML and REML), are not offered in the NICOB. The omissions are not intended to suggest that models or model-fitting procedures different from those available in the NICOB are in any way inferior to those implemented in the NICOB.

DerSimonian-Laird The procedure most widely used for meta-analysis in medicine (DerSimonian and Laird, 1986) is generally recommended for the reduction of data from interlaboratory studies in measurement science. Jackson et al. (2010) point out that the procedure is "remarkably efficient" when estimating the consensus value.

The weighted mean, favored by Cox (2002) (where it is labeled *Procedure A*), is a particular case of the DerSimonian-Laird procedure. When the NICOB determines that the measurement results are homogeneous (that is, mutually consistent, as explained below), it reverts to the weighted average automatically.

An important advantage of the DerSimonian-Laird procedure is its ability to "dampen" the influence that measurement results with very small uncertainties have on the consensus value, particularly when such influence would be consequential: that is, when the measured values differ from one another considerably more than their associated uncertainties would suggest they should.

The computation of the DerSimonian-Laird consensus value does not use the numbers of degrees of freedom qualifying the uncertainty evaluations. However, when these numbers are available the NICOB uses them in the parametric statistical bootstrap evaluation of the uncertainty associated with the consensus value, and also in the characterization of the degrees of equivalence.

An important limitation of the DerSimonian-Laird procedure as originally defined by DerSimonian and Laird (1986) is the potential imprecision in the characterization of the dispersion of the measured values. The principal reason for such imprecision is that the uncertainty attributable to the typically small number of measurement results being inter-compared and combined is not recognized (Guolo and Varin, 2015; Hoaglin, 2016).

This limitation also impacts negatively the reliability of the evaluation of the uncertainty associated with the consensus value. However, the fairly

sophisticated uncertainty evaluation implemented in the NICOB, described in §3.2.4 and in §A.2, by and large mitigates this shortcoming. Still, the use of the DerSimonian-Laird procedure is most appropriate to combine measurement results from ten or more laboratories or measurement methods.

Hierarchical Bayes The three principal advantages of this Bayesian procedure are: (i) the ability to take the numbers of degrees of freedom into account, both for estimation of the consensus value and for the evaluation of the associated uncertainty; (ii) the proper recognition and propagation of the uncertainty associated with the dispersion of the measured values above and beyond what the laboratory-specific uncertainties $\{u_j\}$ already capture; and (iii) the opportunity to express some prior knowledge about both the $\{u_j\}$ and the standard deviation τ of the laboratory effects, described in §3.2.3.

The default value for the scale parameter (which is the median) of the prior distribution for τ is a robust estimate of the standard deviation of the measured values. If it is believed that the observed scatter of the measured values is unrealistically large (or small), then such belief may be injected into the analysis by assigning a value to that scale parameter smaller (or larger) than the default value.

It should be noted that the Bayesian procedure estimates the true value of the within-lab standard uncertainty σ_j only when the number of degrees of freedom ν_j associated with u_j is specified and is finite. Otherwise, $\sigma_j = u_j$ is treated as a known constant.

The ability to specify the medians of the prior distributions for τ and for $\{\sigma_j\}$ is crucial for two reasons: (i) it informs the procedure about the underlying measurement scale (the "right" value for the prior median for τ when the measured values are expressed in kilogram cannot be the same as when they are expressed in gram); (ii) it provides the means to convey some weak but relevant information about the expected whereabouts of the value of τ and of the $\{\sigma_j\}$, and in particular to "nudge" the posterior estimates of the latter toward larger or smaller values than the stated uncertainties suggest, if indeed there is credible information about their suffering from a shortcoming of this kind.

The hierarchical Bayesian procedure makes more assumptions about the probability distributions of the elements of the underlying model than the DerSimonian-Laird procedure. Furthermore, the prior distributions for τ

and for the $\{\sigma_j\}$, although only mildly informative, are likely to be influential, particularly when the number of participating laboratories is small (say, less than ten). Still, the hierarchical Bayesian procedure is recommended, especially if the sample size is less than ten. Note that for a sample size of two, the Bayesian procedure may run into computational difficulties.

This Bayesian procedure is intended to be of general purpose, hence the adoption of rather noncommittal prior distributions. When resources (consulting statisticians and suitable computational tools) are available, it is preferable to employ truly informative prior distributions that capture as completely as possible all the relevant preexisting knowledge about the value of the measurand and about other potentially influential aspects of the measurement, for example the typical size of the measurement errors to expect.

Linear Pool The oldest of the three methods implemented in the NICOB has been in use for a long time indeed, apparently dating back to Pierre Simon, Marquis de Laplace (Bacharach, 1979). It has been modified by many others, including Stone (1961), Lindley (1983), Genest et al. (1984), Clemen and Winkler (1999), and Toman (2007). The method has been "rediscovered" in different fields of measurement science, including in thermometry (Steele et al., 2002) and in chemistry (Duewer, 2004). The original purpose of the method was to aggregate expert opinions that were expressed in the form of probability distributions.

The Linear Pool relies on a model for the data that is structurally different from the laboratory effects model that underlies both the DerSimonian-Laird and Bayesian procedures (§3.2.3): it represents the probability distribution of the consensus value as a mixture of the probability distributions associated with the values measured by the participating laboratories. (Here, the word "mixture" is used in the technical sense of probability theory, for example as reviewed by McLachlan and Peel (2000), and more specifically in §A.4.)

The procedure can be explained very easily in non-technical terms, which is an important advantage. Furthermore, it makes only modest assumptions: either Student's t or Gaussian distributions are assigned to the individual measurement results, depending on whether numbers of degrees of freedom have, or have not been specified. The implementation in the

NICOB allows the user to specify weights for the values being pooled. The weights may express subjective assessments of either the quality of the participating laboratories, or the reliability of the measurement results being pooled.

The main downsides of the method are that it tends to produce multimodal distributions, and that it often leads to markedly larger uncertainty evaluations than the DerSimonian-Laird or the Bayesian procedures. However, in some cases the Linear Pool produces uncertainties that are quite comparable, or possibly even smaller, than alternative procedures. The procedure can be used when there are either a large number of measurement results or when there are only just a few.

2.4 Method-Specific Inputs

- For the DerSimonian-Laird procedure (Exhibit 3):
  - If the Knapp-Hartung adjustment (explained in A.2) is desired, then check the corresponding box in the graphical user interface;
  - To apply the parametric bootstrap for uncertainty evaluation, check the corresponding box. This reveals an input field for the desired number of bootstrap replicates (SUGGESTED VALUE: 10 000).
- For the hierarchical Bayesian procedure (Exhibit 4):
  - Positive numbers in the corresponding boxes in the graphical user interface, which are used as the medians of the prior distributions for the between-laboratory and for the within-laboratory (or, between-method and within-method) variance components (DEFAULT: robust indications of spread of the measured values for the between-laboratory variability, and for the laboratory-specific uncertainty).
  - Total number of iterations for the Markov Chain Monte Carlo (MCMC) sampler (DEFAULT: 250 000).
  - Length of burn-in for the Markov chain, which is the number of values discarded from the beginning of the realization of the chain (DEFAULT: 50 000).
  - Thinning rate for the Markov chain. The DEFAULT value is 25, meaning that only every 25th value generated in the chain should be kept.

When the NICOB determines that the total number of iterations may have been insufficient to ensure that the Markov chain has achieved its equilibrium distribution, the NICOB will suggest new values for the number of iterations, length of burn-in, and thinning rate.

- For the Linear Pool (Exhibit 5):
  - Weights (non-negative numbers separated by commas) to be associated with the different measurement results (DEFAULT: 1 for all).
  - Size of sample drawn from the mixture distribution of the measurand (DEFAULT: 100 000).

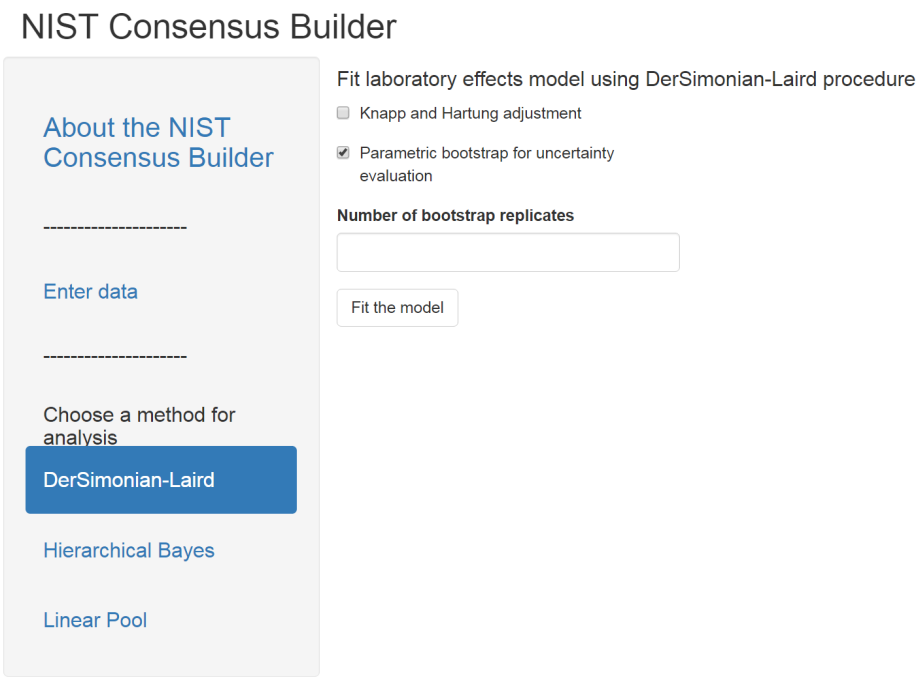

Exhibit 3: User interface for the NICOB corresponding to the specification of the DerSimonian-Laird procedure.

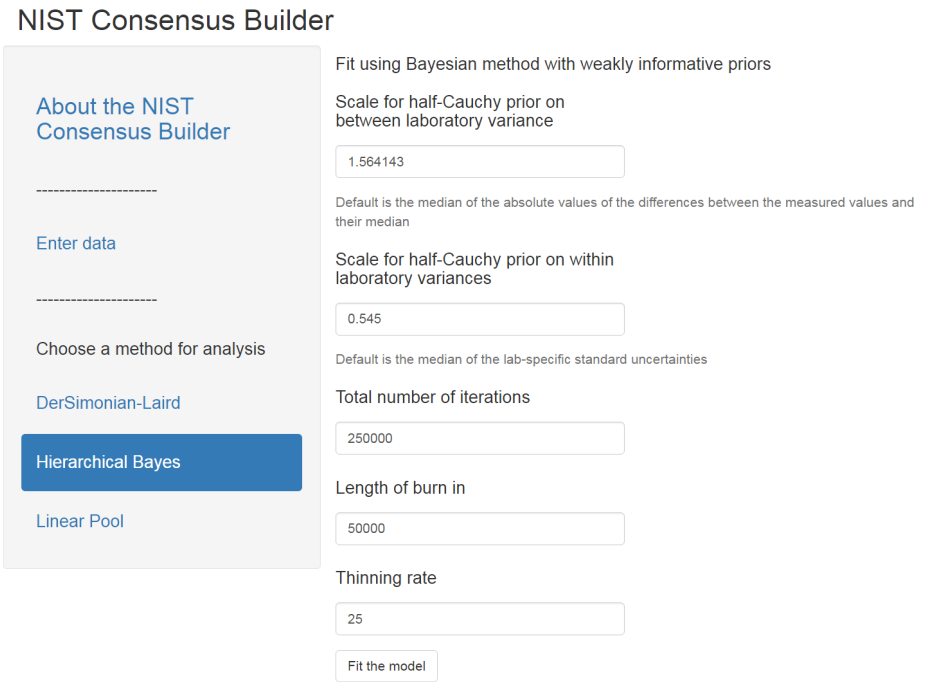

Exhibit 4: User interface for the NICOB corresponding to the specification of the hierarchical Bayesian procedure.

2.5 Output

The results appear on a refreshed web page under the section corresponding to the method selected, and include:

- Consensus estimate, associated standard uncertainty, and coverage interval for the true value of the measurand;
- If *degrees of equivalence* were requested, then estimates, standard uncertainties, and expanded uncertainties for differences between measured values and the consensus value, and between pairs of measured values are reported and depicted graphically.

Graphical outputs can be saved as PDF files by clicking their respective Download plot buttons. After fitting a model, if the user goes back to the Enter data page

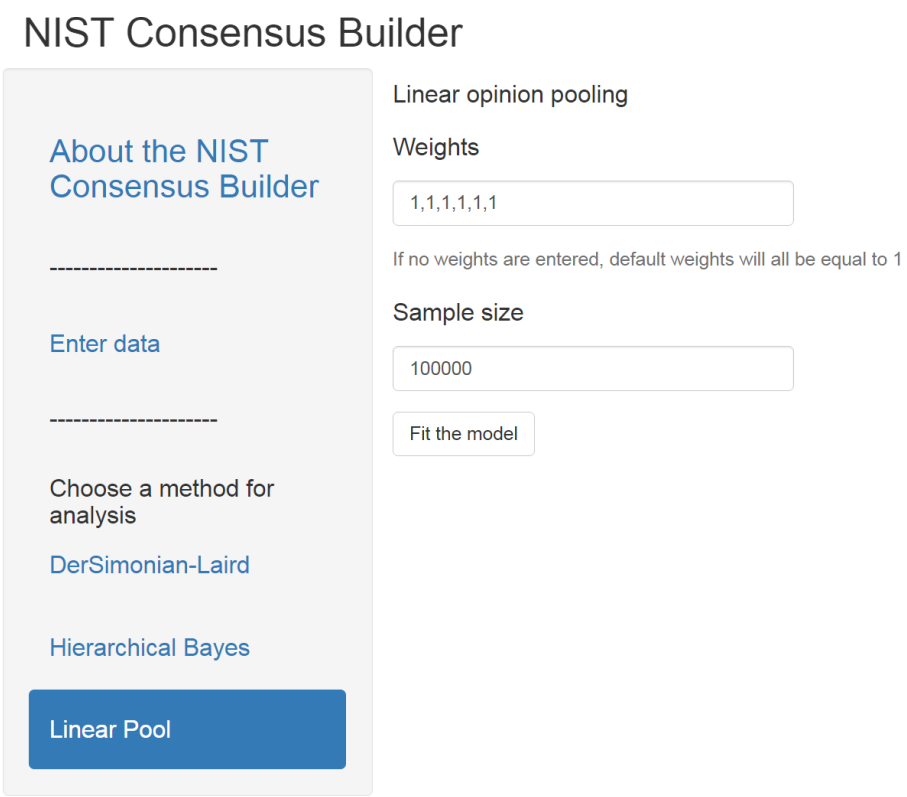

Exhibit 5: User interface for the NICOB corresponding to the specification of the Linear Pool.

and changes the data, then the user must click Fit the model again to update the results; the results are not updated automatically.

3 Orientation

The NICOB represents a compromise between practicability and best practices for the reduction of data from interlaboratory studies and inter-method comparisons, including collaborative trials. (In analytical chemistry, collaborative trials are interlaboratory studies to characterize the performance of a particular method of analysis when it is applied to a well-defined material (Thompson and Lowthian, 2011, Page 180).)

The NICOB is not suitable for the reduction of data from proficiency tests (Thompson et al., 2006) because these typically involve a reference value that is not a consensus value derived from all the participants' measurement results, and also because several performance metrics are usually evaluated that the NICOB does not produce.

On the one hand, practicability demands that scientists should have within easy reach (for example, in the World Wide Web) a toolkit to reduce measurement results obtained in the course of a comparison. This toolkit should implement widely accepted principles and methods for statistical data analysis, and be usable without specialized knowledge of statistics or of computer programming, and also without having to download and install any software. Practicability is further enhanced by minimizing the choices that need to be made when using the toolkit, and by making clear the meaning and potential impact of the different choices that the user needs to make.

On the other hand, best practices would require that each consensus building exercise be customized best to address the substantive needs and goals of the study and the peculiarities of the measurement results, and that this customization be the result of close collaboration between scientists and statisticians or applied mathematicians. However, a custom solution inevitably requires the development of customized computer codes for data analysis — an unrealistic requirement for most laboratories leading or otherwise participating in such interlaboratory studies.

The principles that drive the NICOB are outlined in §3.1, and §3.2 provides an extensive treatment of a particular example, in sufficient detail to provide an appreciation for the methods implemented in the NICOB, thus providing the minimal foundation that enables users to follow the steps described in the *Quick Start* (§2) and begin reducing their own data without further study of this manual.

The additional examples presented in §4 have been deliberately drawn from very different areas of application to increase the chances that users will find at least one example after which they may pattern their own analysis. Additional details about the statistical methods introduced in §3.2 are provided in Appendix A.

The all-important §5 emphasizes that the NICOB is not applicable universally: in fact, situations often arise when data reductions have to be done using statistical methods that the NICOB does not offer. At least one example in §4 indicates that occasionally some data preparation or analysis needs to be done before the NICOB can be used.

And even when the measurement results already are of a form that allows them to be entered into the NICOB directly, it may be preferable to re-express them into a scale of measurement different from the original, to improve compliance with assumptions that underlie the methods available in the NICOB.

For example, if there is a marked association between the measured values $\{x_j\}$ and their associated uncertainties $\{u_j\}$, whereby the latter tend to increase with increasing measured values, then it may be best to analyze instead $\{y_j\}$ with uncertainties $\{v_j\}$, where $y_j = \log(x_j)$, $v_j = u_j/|x_j|$, and $|x_j|$ denotes the absolute value of x_j, for $j = 1, \ldots, n$.

If the measured values $\{x_j\}$ have generally comparable associated uncertainties but are an implausible sample from a Gaussian distribution and the user wishes to employ the hierarchical Bayesian procedure (which assumes that the data are Gaussian), then it may be best to fit the model to $\{(\varphi(x_j), u_j|\varphi'(x_j)|\}$ instead, where φ is a suitable Box-Cox transformation (Box and Cox, 1964) and $|\varphi'(x_j)|$ denotes the absolute value of the first derivative of φ evaluated at x_j.

3.1 Principles

The NICOB is consistent with the following general principles for the combination of measurement results obtained independently by different laboratories or measurement methods.

(P1) No measurement result should be set aside except for substantive, documented cause. The mere fact that a measured value lies far from the bulk of the others, alone is insufficient reason to set it aside, even if a statistical test suggests that it is an outlier.

> Graphical and statistical detection of anomalous results, and examination of consistency indices like Cochran's Q (Cochran, 1954) or I^2 (Higgins and Thompson, 2002), are useful screening tools that may serve to draw the scientists' attention to measurement results deserving further scrutiny, but should be advisory, not decisional (Consultative Committee for Amount of Substance, 2013).
>
> In all cases, substantive considerations, rather than statistical tests, should drive the selection of the subset of the measurement results that should be combined into a consensus value.

(P2) No measured value should dominate the consensus value simply because the associated measurement uncertainty is much smaller than the uncertainties associated with the other measured values, especially when the measured values are markedly more dispersed than what their associated uncertainties alone would intimate.

> Methods consistent with this principle will therefore include a damping mechanism to limit the influence of unusually small laboratory-specific uncertainty evaluations.

(P3) Measurement methods should have been characterized sufficiently well to warrant the belief that the measured values, taken as a group, are roughly centered at the true value of the measurand.

> On the one hand, it is obvious that if all the measured values tend to be too low or too high, no statistical procedure that relies on the data alone will be able to detect this and "correct" the consensus estimate accordingly.
>
> For example, when immunoassays are used to measure the concentration of vitamin D, they may be persistently low or high, depending on the antibody that they use for targeting the vitamin, and on how the vitamin is bound to materials in the matrix of the sample (Tai et al., 2010; Farrell et al., 2012; Enko et al., 2015).
>
> On the other hand, it is desirable that the statistical procedures used for data reductions should be able to cope with situations where individual measured values lie far from the bulk of the others, and also with situations where there is some asymmetry in the apparent distribution of the measured values (Consultative Committee for Amount of Substance, 2013).
>
> For example, some methods for extracting polychlorinated biphenyls (PCBs) from riverine sediments, or for extracting arsenic from oyster tissue, may do so only incompletely. In such cases, the distribution of measured values may be markedly asymmetrical, showing a histogram whose left tail is longer than the right tail (Possolo, 2013).
>
> The models and methods implemented in the NICOB may be applicable only after the measured values will have been suitably transformed, for example using a Box-Cox transformation (Box and Cox, 1964), in which case the associated uncertainties should be transformed accordingly, typically by application of the Delta Method (Possolo and Toman,

2011, §5.2). The need for transformation arises often when there is a natural upper or lower bound for the value of the measurand and the measured values are close to a bound (e.g., mass fractions of rare elements in geochemical samples, or the amount-of-substance fraction of a chemical compound in a high-purity material).

(P4) A statistical model should be formulated that explicitly relates the measured values to the true value μ of the measurand, and the model should include elements representing contributions from the recognized sources of uncertainty. Furthermore, the estimation of μ, and the evaluation of associated uncertainty, should be consistent with the statistical model and with some principle of estimation whose general reliability is widely recognized.

> Typically, the required model will be a statistical model where μ appears as a parameter of the probability distribution of the measurement results. The measured values, and possibly also their associated uncertainties, are modeled as observed values of random variables.
>
> The principle suggests that a mere prescription or recipe for how the data should be reduced, without a clear description of how the data relate to the measurand and of implied assumptions, is not helpful.

(P5) The statistical model underlying data reductions should be able to detect, evaluate, and propagate uncertainty components that produce excess variance indicating significant heterogeneity of the measured values, which is an expression of so-called *dark uncertainty* (Thompson and Ellison, 2011), whereby the measured values are substantially more dispersed than is to be expected based on their stated, laboratory-specific uncertainties.

> Heterogeneity, or the presence of "excess" variance, is often the object of a statistical test, for example, a test based on Cochran's Q. However, the tests in common use have notoriously low power (probability of detecting heterogeneity when in fact the measurement results are heterogeneous) (Hoaglin, 2016). For this reason, it may be safest always to proceed on the assumption that there may be some heterogeneity, and then propagate it to all derivative quantities, including to the degrees of equivalence.
>
> The willingness of the participants in an interlaboratory study to engage in an inter-comparison includes a tacit agreement to abide by the resulting findings, in particular to recognize any component of uncertainty that their individual uncertainty evaluations would have missed

and that becomes apparent only once independent measurement results are inter-compared.

In particular, the detection of significant heterogeneity in cases where there is no reason to impugn the assumption of there being a common measurand, implies that the laboratory-specific uncertainty evaluations are too small, and that comparisons between individual measured values and the consensus value, or between pairs of individual measured values, should take dark uncertainty into account (Exhibit 6 provides an illustration of this situation).

(P6) Degrees of equivalence (differences between measured values and a consensus value, or between pairs of measured values, qualified with evaluations of associated uncertainty) should be computed consistently with their primary goal of identifying measured values that are significantly discrepant either from the consensus value or from one another.

3.2 Illustration and Overview of Methods

Key Comparison CCQM-K25 was carried out by the Consultative Committee for the Amount of Substance (*Metrology in Chemistry and Biology*), to compare measurement results for the mass fractions of five different polychlorinated biphenyl (PCB) congeners in sediment (Schantz and Wise, 2004). Exhibit 6 on Page 22 lists the measurement results for PCB 28 (2,4,4'-trichlorobiphenyl) that were selected based on the substantive reasons described by Schantz and Wise (2004), and depicts some of the results that the NICOB produced. To load these measurement results into the NICOB, click here.

The measurement results selected for further analysis were obtained by $n = 6$ national metrology institutes: Joint Research Centre Institute for Reference Materials and Measurements (IRMM, Geel, Belgium); Korea Research Institute of Standards and Science (KRISS, Daejeon, Republic of Korea); National Measurement Institute Australia (NARL, Sydney, Australia); National Institute of Standards and Technology (NIST, Gaithersburg, USA); National Metrology Institute of Japan (NMIJ, Tsukuba, Japan); and National Research Council Canada (NRC, Ottawa, Canada).

Exhibit 7 on Page 23 summarizes the results that the three procedures available in the NICOB produced when they were applied to the measurement results for PCB 28 listed in the top panel of Exhibit 6.

LAB	x (ng/g)	u (ng/g)	ν	LAB	x (ng/g)	u (ng/g)	ν
IRMM	34.30	1.03	60	NIST	32.42	0.29	2
KRISS	32.90	0.69	4	NMIJ	31.90	0.40	13
NARL	34.53	0.83	18	NRC	35.80	0.38	60

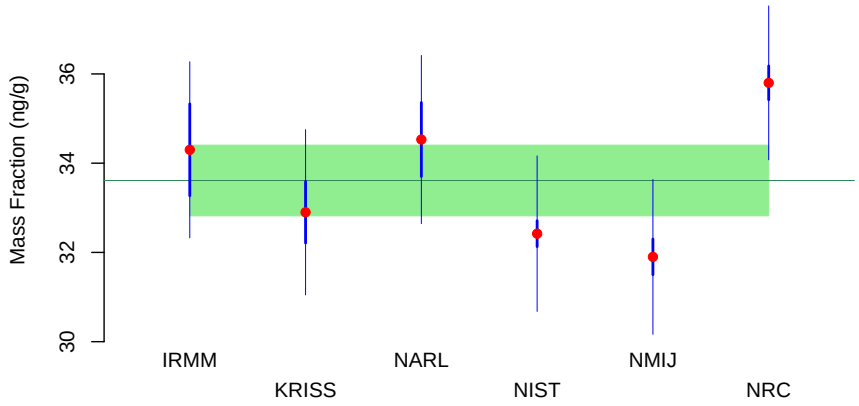

Exhibit 6: TOP PANEL: Measured values x of the mass fraction (ng/g) of PCB 28 in the material used in CCQM-K25, standard uncertainties u, and numbers of degrees of freedom ν on which the standard uncertainties are based. BOTTOM PANEL: Each large (red) dot represents the value x measured by a participating laboratory, and the thick, vertical (blue) line segment depicts $x \pm u$. The thin, vertical line segment depicts $x \pm (\hat{\tau}^2 + \widehat{\sigma}^2)^{1/2}$, where the uncertainty includes the contribution from *dark uncertainty*, estimated as the posterior mean $\hat{\tau} = 1.68$ ng/g, and the posterior mean $\widehat{\sigma}$ of the standard uncertainty u, as produced by the Bayesian procedure. The thin, horizontal (dark green) line marks the consensus value $\hat{\mu} = 33.6$ ng/g, which is the mean of the posterior distribution of μ obtained by application of the hierarchical Bayesian procedure. The shaded (light green) band represents $\hat{\mu} \pm u(\mu)$, where $u(\mu)$ is the standard deviation of the posterior distribution of μ.

PROCEDURE	CONSENSUS	STD. UNC.	95 % COV. INT.
DerSimonian-Laird	33.6	0.77	(32.0, 35.2)
Hierarchical Bayesian	33.6	0.79	(32.0, 35.2)
Linear Pool	33.6	1.55	(31.4, 36.3)

Exhibit 7: Results of the three consensus-building procedures implemented in the NICOB, for the mass fraction of PCB 28, all expressed in ng/g. The standard uncertainty and coverage interval for the DerSimonian-Laird procedure were computed using the version of the parametric statistical bootstrap described in §A.2, which includes consideration for the small number of measurement results that the estimate of dark uncertainty τ is based on.

3.2.1 Measurement Results

The measurement result from laboratory $j = 1, \ldots, n$ is a triplet (x_j, u_j, ν_j) comprising a measured value of the mass fraction x_j of PBC 28 in the material used in CCQM-K25, an evaluation of the associated standard uncertainty u_j, and the number of degrees of freedom ν_j on which the uncertainty evaluation is based. In many interlaboratory studies the $\{\nu_j\}$ are not reported. In such cases, the NICOB makes the (likely unrealistic) assumption that these numbers of degrees of freedom are very large, practically infinity, which may give rise to overoptimistic uncertainty evaluations for the consensus value.

The notion of degrees of freedom, as it is used in this context, often is a source of confusion and even acrimony. When u_j is the result of a Type A evaluation — discussed by Taylor and Kuyatt (1994, §3), Joint Committee for Guides in Metrology (2008a, §4.2), and Possolo (2015, §5) —, the corresponding ν_j is a function of the number of observations used to compute u_j. For example, when u_j is the standard deviation of the sampling distribution of the average of a set of m_j observations obtained under conditions of repeatability, then $\nu_j = m_j - 1$. But when, as is most often the case, u_j combines the results of Type A and Type B evaluations of several different components of uncertainty, any assignment of value to ν_j is likely to prove controversial.

Occasionally, one runs across suggestions to the effect that a Bayesian approach does away with the need to consider degrees of freedom (Bich, 2012; Kacker et al.,

2016), when in fact it does not. The paradigmatic example arises in the context of Bayesian estimation of the mean of a Gaussian distribution based on a sample of size m with average $\overline{x}$ and standard deviation s, when both the mean and the variance of this distribution are unknown, and the lack of *a priori* knowledge about their values is expressed using the invariant prior distribution suggested by Jeffreys (1946).

In these circumstances, the marginal posterior distribution of the mean is a re-scaled and shifted Student's t distribution with $\nu = m - 1$ degrees of freedom (Box and Tiao, 1973, Theorem 2.4.1). The associated standard uncertainty is the standard deviation of this posterior distribution. The probabilistic interpretation of every coverage interval (which is a Bayesian credible interval) of the form $\overline{x} \pm ks/\sqrt{m}$, and in particular the probabilistic interpretation of the standard uncertainty, depends on ν.

Therefore, it is only a matter of semantics whether one says that this number ν of degrees of freedom pertains to the posterior distribution of the mean or to the standard uncertainty. In any case, ν summarizes the extent of the evidentiary basis on which evaluations and expressions of measurement uncertainty (say, coverage intervals) are based.

3.2.2 Dark Uncertainty

The standard deviation of the measured values $\{x_j\}$ listed in Exhibit 6 equals 1.48 ng/g, while the laboratory-specific standard uncertainties range from 0.29 ng/g to 1.03 ng/g, and their median equals 0.545 ng/g. Therefore, the measured values are almost three times more dispersed than the typical, within-laboratory standard uncertainty.

This "excess" variance is often interpreted as suggesting that the laboratories have failed to identify and evaluate one or more important sources of uncertainty, whose contribution Thompson and Ellison (2011) have called *dark uncertainty*.

Dark uncertainty is uncovered often in interlaboratory studies carried out in measurement science, both for physical and chemical quantities, and in meta-analytic studies of medical procedures and therapies. Most recently, measurements made using the watt balance revealed a surprisingly large dark uncertainty in mass determinations using state-of-the-art equipment (Stock, 2011). The measurement of the Newtonian constant of gravitation, G, affords another cogent illustration of the existence of dark uncertainty (Speake and Quinn, 2014).

Hundreds of experiments have been performed over time to measure the value of G, one of the earliest having been Henry Cavendish's (Cavendish, 1798). Even though much effort and great rigor have been applied to the characterization of the corresponding uncertainty budgets, the dispersion of the resulting measured values of G still is strikingly larger than what the stated uncertainties would lead one to expect.

This state of affairs has given great impetus to collective efforts to identify the underlying causes, which are expected to produce not only improved estimates of G, but also, and maybe most importantly, significant advances in the measurement of very weak forces generally.

These collective efforts include rather extreme measures in attempts to identify the underlying causes for such dispersion of values: for example, moving a torsion balance from a laboratory in Sèvres, France, to another in Gaithersburg, Maryland, USA. And also the exploration of techniques radically different from those that have traditionally been used for measuring G, for example using laser-cooled atoms and quantum interferometry (Rosi et al., 2014).

The documentation of the presence of dark uncertainty, achieved through its reliable detection and rigorous quantification, perforce ought to impact the computation of the consensus value, the evaluation of the associated uncertainty, and also the assessment of the significance of the differences between individual measured values and the consensus value (unilateral degrees of equivalence). The detection and quantification of dark uncertainty act as stimulants for research and discovery, and ultimately will lead to improvements in the quality of measurements.

The quantification of dark uncertainty, based on a set of measured values qualified with evaluations of measurement uncertainty, has been for quite some time, and continues to be, a very active area of research (Higgins and Thompson, 2002; Higgins et al., 2003; Viechtbauer, 2007; Rukhin, 2013; Turner et al., 2015). In general, by taking dark uncertainty properly into account:

(i) The influence that measured values accompanied by very small stated uncertainties have upon the consensus value is reduced, because a common dark uncertainty component acts as a damping or modulating factor;

(ii) The (statistical) significance of differences between individual measured values and the consensus value, also is reduced, and the proportion of discrepant laboratories is reduced accordingly, because the existence of dark

uncertainty blurs differences that might otherwise, but only illusorily, appear sharp.

Consideration of the contribution that dark uncertainty makes to the effective uncertainty surrounding each measured value enables a rigorous assessment of how realistic the stated uncertainties are, separately and distinctly from the assessment of apparent bias (that is, persistent deviation from the consensus value) of the measured values, when both criteria of performance are appraised by the community of participating laboratories.

Consistently with the spirit of the GUM (Joint Committee for Guides in Metrology, 2008a), although not possibly with its letter because the GUM does not consider interlaboratory studies, these downstream calculations must take into account contributions from all sources of uncertainty: both those that were captured in bottom-up evaluations performed by the participating laboratories individually, and those, whose joint effect is manifest in the dark uncertainty, that were the object of the top-down evaluation enabled by the inter-comparison of the measurement results (Possolo, 2015).

A laboratory random effects model is able to characterize dark uncertainty quantitatively (Toman and Possolo, 2009b, 2010; Borenstein et al., 2010). This model has many variants and a long history of use and proven utility, often being described as a variance components model for the analysis of variance (Searle et al., 2006).

3.2.3 Laboratory Random Effects Model

A laboratory random effects model represents the value of the mass fraction measured by laboratory j as $x_j = \mu + \lambda_j + \varepsilon_j$ for $j = 1, \ldots, n$, where n (which equals 6 in the case of PCB 28) is the number of laboratories, μ denotes the measurand that is estimated by the consensus value, $\lambda_1, \ldots, \lambda_n$ are the laboratory effects, and $\varepsilon_1, \ldots, \varepsilon_n$ represent measurement errors. This model underlies two of the procedures implemented in the NICOB: DerSimonian-Laird (§A.2) and hierarchical Bayes (§A.3).

If the data were only the $\{x_j\}$ then it would be impossible to distinguish the laboratory effects $\{\lambda_j\}$ from the measurement errors $\{\varepsilon_j\}$ because increasing λ_j and decreasing ε_j correspondingly, while keeping μ fixed, would still reproduce the same x_j.

However, since the $\{u_j\}$ also are part of the data, and we know that the absolute values of the $\{\varepsilon_j\}$ are generally comparable to the $\{u_j\}$, we can conclude that any "excess variance" exhibited by the $\{x_j\}$ is attributable to the $\{\lambda_j\}$, whose dispersion (or scatter) is gauged by τ.

The specific version of the laboratory random effects model is determined by the modeling choices made for the $\{\lambda_j\}$ and the $\{\varepsilon_j\}$. The DerSimonian-Laird procedure makes fewer assumptions about them than the hierarchical Bayes procedure, except when it comes to the uncertainty analysis, where the assumptions made by both methods are quite comparable.

The more specific (and more restrictive) assumptions that underlie the uncertainty evaluation for the DerSimonian-Laird procedure and the Bayesian model are: (i) the $\{\lambda_j\}$ are considered to be a sample from a Gaussian distribution with mean 0 and standard deviation τ; (ii) the $\{\varepsilon_j\}$ are assumed to be outcomes of independent Gaussian random variables with mean 0 and possibly different standard deviations $\{\sigma_j\}$; and (iii) the $\{\lambda_j\}$ and the $\{\varepsilon_j\}$ are mutually independent.

Since the random effects model "explains" the data as an additive superposition of effects, both the DerSimonian-Laird and hierarchical Bayesian procedures rely on the assumption that the underlying measurement scale is linear, in the sense that the same differences between measured values have the same meaning irrespective of whether the values being differenced are both low or both high. For some measurands, and in some measurement situations, this is not the case at all.

When there is a natural bound for the true value of the measurand, the measurement scale in fact may be "compressed" non-linearly in the vicinity of the bound. For example, when measuring purity of an alloy that comprises mostly a single metal with only minor amounts of impurities, the upper bound is 100 % and the measured values often appear to be a sample from a distribution that has a left tail heavier that the right tail (which terminates abruptly at 1).

In some cases there is a relationship between the measured values $\{x_j\}$ and the corresponding associated uncertainties $\{u_j\}$. This relationship is so prevalent in analytical chemistry, for example, that it has been exploited to produce approximate uncertainty evaluations in the absence of any other information (Horwitz, 1982, 2003). Even though the model can take into account different uncertainties for different laboratories, the existence of such relationship is a subtle hint of non-linearity in the underlying measurement scale.

For these and other reasons, it may be advantageous to carry out the analysis

after the data will have been re-expressed into another scale. Two generally useful classes of transformations are the Box-Cox transformations and the folded-power transformations.

The Box-Cox transformations map a measured value x onto $(x^p - 1)/p$ for some suitable value of p (Box and Cox, 1964; Mosteller and Tukey, 1977), with $p = 0$ indicating the logarithm.

The folded-power transformations map a proportion $0 \leqslant x \leqslant 1$ onto $x^p - (1-x)^p$ for some value of p usually selected from between 0 and 1. For $p = 0$, the transformation reduces to the folded logarithm, which is the familiar logit, $\log(w/(1-w))$ (Tukey, 1977; Mosteller and Tukey, 1977; Emerson, 1991).

If the measured values are re-expressed, then uncertainties need to be computed accordingly. This can be done by application of either the Delta Method (Possolo and Toman, 2011, §5.2), or a suitable Monte Carlo method. And similar care needs to be taken when transforming the results back onto the original measurement scale. The NICOB does not offer facilities to apply Box-Cox or folded-power transformations.

3.2.4 DerSimonian-Laird Procedure

The DerSimonian-Laird consensus value is a weighted average of the values measured by the participating laboratories, $\hat{\mu}_{\mathrm{DL}} = \sum_{j=1}^{n} w_j x_j \,/\, \sum_{j=1}^{n} w_j = 33.6$ ng/g, with weights $w_j = 1/(\tau^2 + \sigma_j^2)$ for $j = 1, \ldots, n$. The presence of τ in all the weights dampens the impact of very small values among the $\{\sigma_j\}$ (*cf.* principle P2).

Since both τ and the $\{\sigma_j\}$ are unknown, they are substituted by estimates, $\hat{\tau}_{\mathrm{DL}}$ and $\widehat{\sigma}_j = u_j$. The $\{u_j\}$ may be the result of either Type A or Type B evaluations, or combine results of both Type A and Type B evaluations, or they may be standard deviations of Bayesian posterior distributions.

In the DerSimonian-Laird procedure, the estimate of τ is obtained by equating observed and expected values of a particular function of the measurement results (a so-called *method of moments* estimate). The resulting $\hat{\tau}_{\mathrm{DL}}$ is then used in the definition of the weights $\{w_j\}$ as if it were known without uncertainty, which is obviously unsatisfactory, especially when the number of participating laboratories is as small as it is for PCB 28.

Even though there is a closed-form expression for the variance of a weighted mean, the reliability of such expression is questionable in this case because the weights depend on quantities (τ in particular) that have to be estimated from the

data. Similarly, treating the $\{u_j\}$ as known constants when they are based on small numbers of degrees of freedom also in unrealistic.

Alternatively, and preferably, $u_{\mathrm{DL}}(\mu)$ may be evaluated by application of the parametric statistical bootstrap, which offers the ability to take into account the numbers of degrees of freedom $\{\nu_j\}$ associated with the $\{u_j\}$, and also to recognize the impact that the typically small number n of measurement results has on the reliability with which τ is estimated (refer to §A.2 for details).

3.2.5 Hierarchical Bayesian Procedure

The Bayesian procedure is based on a principle of estimation whereby μ is modeled as a random variable and the consensus value is the mean of the probability distribution of μ given the measurement results $\{(x_j, u_j, \nu_j)\}$, computed according to Bayes's rule (DeGroot and Schervish, 2011; Possolo and Toman, 2011).

In a Bayesian analysis, unknown quantities (μ, $\{\lambda_j\}$, τ, $\{\sigma_j\}$) are modeled as outcomes of non-observable random variables, and the measurement results $\{(x_j, u_j, \nu_j)\}$ are modeled as actually observed outcomes of random variables.

The expression "random variable" does not imply that there is anything *chancy* about the value of the corresponding quantity. It simply indicates that there is a probability distribution associated with the quantity. This probability distribution recognizes that the value of the quantity is generally unknown, owing either to natural sampling variability, or to incomplete knowledge.

The probability distributions for the unknowns, reflecting *a priori* (that is, before acquiring any data) states of knowledge about their true values, are called *prior distributions*. These distributions may involve parameters (called *hyperparameters*) that must be assigned values at the outset.

In the NICOB, the prior distributions assigned to unknown quantities, including the measurand, are all weakly informative to give the data the broadest opportunity to influence the consensus value and all the other results of the data reductions. For example, the prior distribution for μ is Gaussian with mean 0 and a very large standard deviation (10^5). But even weakly informative priors can be rather influential, particularly when the number n of measured values is small (Lambert et al., 2005).

In many cases encountered in practice, there is considerable *a priori* information about the measurand (for example, that the amount-of-substance fraction of ethane, in a synthetic mixture designed to emulate natural gas, lies within a

fairly narrow interval). There may also exist *a priori* information about other parameters in the model (for example, about the measurement uncertainty σ_j associated with a gravimetric preparation). In general, taking such information into account requires a custom solution that the NICOB is unable to provide.

The only two prior distributions used in the NICOB that have adjustable hyper-parameters pertain to the between-laboratory dispersion of values τ, and to the laboratory-specific, true standard uncertainties $\{\sigma_j\}$. The prior distributions for τ and for the $\{\sigma_j\}$ are re-scaled Cauchy distributions truncated at zero, as suggested by Gelman (2006). The hyper-parameters are the corresponding medians. These half-Cauchy distributions are weakly informative because their variances are infinite (and their means are undefined).

In the example we have been considering concerning the mass fraction of PCB 28 in CCQM-K25, when we accept the default value, 1.56 ng/g, for the hyper-parameter corresponding to τ, we are expressing our *a priori* belief that the true value of τ is as likely to be smaller as it is to be larger than this value. The default value of this hyper-parameter is a robust estimate (produced by R function `mad`) of the standard deviation of the measured values.

And when we accept the default value (the median of the $\{u_j\}$) 0.545 ng/g for the hyper-parameter corresponding to the $\{\sigma_j\}$, we are indicating that the true values of the $\{\sigma_j\}$ are as likely to be smaller to be larger than this value.

These default hyper-parameter values are reasonable approximations for the medians of the half-Cauchy prior distributions, and allow the NICOB to be easily applied to a variety of datasets by being aware of the actual units of measurement. A more informative, and likely more useful Bayesian approach would require the user to examine and elicit their prior beliefs before looking at the data and choose hyper-parameters that incorporate actual prior knowledge.

Since the probability distribution that Bayes's rule produces for μ cannot be computed explicitly and analytically with the modeling choices just described, the consensus value produced by the Bayesian procedure is obtained as the average of a large sample of values drawn from the posterior distribution of μ via Markov Chain Monte Carlo (MCMC) sampling, and the associated standard uncertainty is the standard deviation of this MCMC sample (Gelman et al., 2013). For the mass fraction of PCB 28, the Bayesian procedure yields $\hat{\mu}_{HB} = 33.6$ ng/g and $u_{\mathrm{HB}}(\mu) = 0.79$ ng/g.

One 95 % coverage interval (usually called a 95 % *credible interval* in this Bayesian context) for μ is defined so that 2.5 % of the values in the MCMC sample are smaller

than its lower endpoint, and another 2.5 % are larger than its upper endpoint. This interval ranges from 32.0 ng/g to 35.2 ng/g. Exhibit 8 on Page 31 shows a smooth histogram of the MCMC sample and a Gaussian approximation to it.

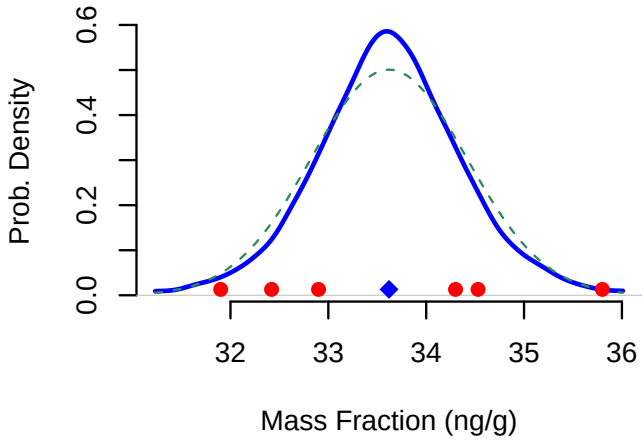

Exhibit 8: Posterior probability density of the consensus value, produced by the Bayesian procedure. The large (blue) diamond marks the estimate of the consensus value, and the (red) dots indicate the measured values. The thin, dashed (green), bell-shaped curve is a Gaussian probability density with the same mean and standard deviation as the posterior distribution of the consensus value: the latter has markedly heavier tails than this Gaussian approximation.

3.2.6 Sampling-Theoretic *vs.* Bayesian Procedures

Even though the DerSimonian-Laird procedure and the hierarchical Bayes procedure rely on the same model, they are fundamentally different in concept, in how they are fit to the data, and in the meaning that they implicitly ascribe to the respective results. These differences may be summarized by saying that the former is sampling-theoretic (or, frequentist), while the latter is Bayesian — which is not saying much unless one is already familiar with the meaning of these epithets, hence we explain.

The key differences between the sampling-theoretic and the Bayesian approaches are: (i) how they regard μ, and (ii) how they interpret the uncertainty surround-

ing the true value of μ. (The same differences apply to τ, $\{\lambda_j\}$, and the $\{\sigma_j\}$.) These approaches may also differ on how they regard the very measurement results $\{x_j, u_j, \nu_j\}$ that one wishes to combine. From the Bayesian perspective, the $\{x_j, u_j, \nu_j\}$ may already be attributes of posterior distributions, not of sampling distributions.

The sampling-theoretic viewpoint focuses on the variability of the consensus value (the estimate of μ) under hypothetical repetitions of the process that generated the data. The Bayesian viewpoint focuses on the information that the particular data in hand provide about μ, and uses this information to update the prior distribution for μ and produce a posterior distribution that typically will be appreciably less dispersed than the prior distribution.

The sampling-theoretic approach is concerned with (allegedly objective) fluctuations of the consensus value attributable to the vagaries of sampling a possibly hypothetical population, while the Bayesian approach updates a (subjective) state of knowledge about μ based on the data that have actually been observed.

The aforementioned fluctuations are only allegedly objective (and not categorically objective) because the sampling contemplated from the sampling-theoretic viewpoint much more often than not is from a hypothetical population — that is, from a population whose definition is a subjective construct.

The merits and demerits of these different approaches have been debated *ad infinitum*, not only within statistical circles but also in many areas of science, as well as within epistemology, which is the branch of philosophy concerned with knowledge in general and with scientific knowledge in particular (Steup, 2014). Within measurement science, some argue that "it is only the definition of probability as a degree of belief that is applicable" (O'Hagan, 2014), while others argue quite the opposite, "that the change from a frequentist treatment of measurement error to a Bayesian treatment of states of knowledge is misguided" (White, 2016).

These foundational issues aside, it should be noted that the Bayesian procedure implemented in the NICOB enjoys two very important, practical advantages: (i) it captures and propagates effectively the uncertainty surrounding the estimate of the between-laboratory dispersion of measured values (τ) without resorting to complex approximations, and (ii) it offers the means to express *a priori* knowledge about either the value of τ or about the reliability of the uncertainty evaluations produced by the participants in the study — a feature illustrated in Example 4.3.

3.2.7 Linear Pool

The Linear Pool makes the fewest assumptions about the data, and tends to produce the largest evaluation of $u(\mu)$ and the widest coverage intervals. As already noted above, in some cases the Linear Pool produces uncertainties that are quite comparable, or possibly even smaller, than alternative procedures.

Its starting point is a set of n probability distributions for the measurand, each of which describes a state of knowledge about the measurand. Specifically, the probability distribution for laboratory $j = 1, \ldots, n$ has mean x_j and standard deviation u_j. In the NICOB, this distribution is chosen to be either a re-scaled and shifted Student's t distribution (when the number of degrees of freedom ν_j has been stated and is finite), or a Gaussian distribution (otherwise).

The n probability distributions are aggregated by *mixing* using weights $\{w_j\}$, to produce a consensus distribution whose mean is the consensus value, and whose standard deviation is the standard uncertainty associated with the consensus value. The weights represent the quality or reliability of the participating laboratories or methods, as perceived by the person performing the aggregation.

Typically, a large sample of size K is drawn from the mixture distribution of the measurand by repeating the following process K times: select a laboratory at random, with probabilities proportional to the weights, and then draw a value from the corresponding distribution.

The K values obtained through this process are summarized in the same way as the bootstrap sample for the DerSimonian-Laird procedure, or as the MCMC sample drawn in the Bayesian procedure. Specifically, the mean of these K values is the consensus value, and the standard deviation is the associated standard uncertainty. Exhibit 9 on Page 34 shows a smooth histogram of the results for the default value $K = 10^6$, the corresponding estimate of the measurand, and a 95 % coverage interval for its true value.

3.2.8 Degrees of Equivalence

The principal goal of some interlaboratory studies, key comparisons in particular, is not so much to produce a consensus value as to detect and identify measurement results that deviate significantly from the consensus value, or from one another when considered in pairs.

In some interlaboratory studies, the consensus value used for these evaluations is not derived from the measurement results of the participants. For example,

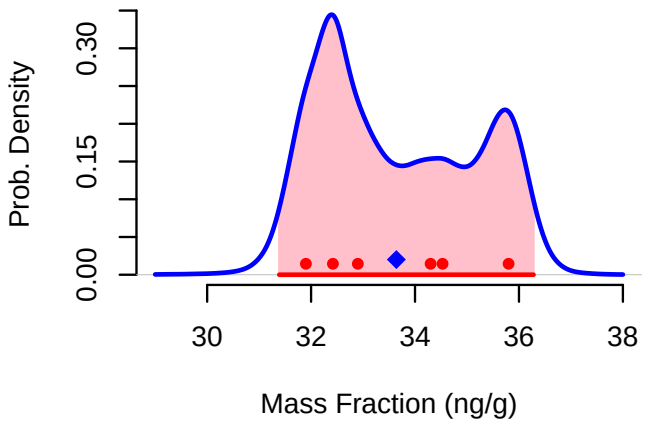

Exhibit 9: Probability density of the consensus value based on a sample of size $K = 10^6$ drawn from the equally weighted mixture of re-scaled and shifted Student's t distributions assigned to the $n = 6$ participating laboratories. The large (blue) diamond marks the estimate of the consensus value, and the (red) dots indicate the measured values. The (pink) shaded region under the curve comprises 95 % of the area under the curve: its projection onto the horizontal axis (red line segment) is a 95 % coverage interval for the measurand.

in key comparison CCQM-K1 (Alink et al., 1999), the consensus value was the amount-of-substance fraction of each of several gas species in nitrogen that had been determined gravimetrically by the pilot laboratory as it prepared the gas mixtures for distribution to the participants.

In other cases, the consensus value is not even a meaningful physical quantity, defined only as a convenient baseline for the inter-comparisons between laboratories. For example, in key comparison CCM.FF-K6.2011 (Benková et al., 2014) the consensus value was a weighted average of the relative errors in the measurement of the transfer standard, defined as relative differences between volumes of gas indicated by the transfer standard and the corresponding volumes measured by the reference (national) standard.

The goal of detecting and identifying measurement results that are significantly inconsistent with a reference or consensus value is paramount in *proficiency tests* (Thompson et al., 2006), and in performance rankings of medical centers

(MacKenzie et al., 2015).

Key comparisons (KCs), which are defined in the Mutual Recognition Arrangement (MRA) (Comité International des Poids et Mesures (CIPM), 1999), include both this goal and the goal of identifying pairs of measured values that differ significantly from one another, because KCs serve to evaluate the intercomparability of measurement results produced by different national metrology institutes.

In the context of KCs, the relevant comparisons are based on unilateral and bilateral *degrees of equivalence* (DoE). Owing to the legal force of the MRA that frames KCs, there is less latitude for KCs than there is for studies done in other contexts regarding how to define the DoE. In fact, the MRA is quite specific, stating in the Technical Supplement to the Arrangement:

> *For the purposes of this arrangement, the term degree of equivalence of measurement standards is taken to mean the degree to which a standard is consistent with the key comparison reference value. 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).*
>
> — Comité International des Poids et Mesures (CIPM) (1999, T.3)

That is, the unilateral DoE are the pairs $\{(D_j, U_{95\,\%}(D_j))\}$, where $D_j = x_j - \hat{\mu}$ and $U_{95\,\%}(D_j)$ denotes the associated expanded uncertainty for 95 % coverage of the true difference, and the bilateral DoE are pairs $\{(B_{ij}, U_{95\,\%}(B_{ij}))\}$, where $B_{ij} = D_i - D_j$, and the $\{U_{95\,\%}(B_{ij})\}$ are the counterparts of the $\{U_{95\,\%}(D_j)\}$. Note that B_{ij} need not equal $x_i - x_j$: this will happen when the $\{D_j\}$ are replaced by the $\{D_j^*\}$ that are defined below.

Occasionally, the measurement results for a participant in a KC are not used in the computation of the consensus value. This may happen when a participant is traceable to the SI via another participant, or when the participant deviated from the protocol agreed for the comparison. But even in such cases the unilateral DoE for the participant may be computed.

This suggests an alternative definition for the unilateral DoE that could be applied generally: based on the difference $D_j^* = x_j - \hat{\mu}_{-j}$, where $\hat{\mu}_{-j}$ denotes an estimate of the consensus value derived from the measurement results produced by all the participants but leaving out the results from participant j, for $j = 1, \ldots, n$. Viechtbauer and Cheung (2010) have used this idea to identify outliers in meta-analysis, and Duewer et al. (2014) have promoted it specifically for the evaluation of degrees of equivalence in key comparisons.

This alternative definition may then be carried forward and lead to an alternative definition of the bilateral DoEs based on $B_{ij}^* = D_i^* - D_j^*$ (which generally differs from $x_i - x_j$), for $1 \leqslant i \neq j \leqslant n$.

It may be argued that D_j^* is more accurate than D_j as an assessment of the "distance" between the value measured by laboratory j and the values measured by the other laboratories. In fact, $x_j - \hat{\mu}$ may be too small in absolute value because $\hat{\mu}$ incorporates ("tracks") x_j. Furthermore, since x_j and $\hat{\mu}_{-j}$ are uncorrelated, this alternative definition greatly simplifies the evaluation of $U_{95\,\%}(D_j^*)$.

The NICOB offers the user the possibility of computing DoEs as defined in the MRA, or according to the *leave-one-out* strategy just described, both for unilateral and bilateral DoEs corresponding to the three procedures available.

The question must also be answered as to whether, once their associated uncertainties are taken into account, the $\{D_j\}$ and the $\{B_{ij}\}$ (or the $\{D_j^*\}$ and the $\{B_{ij}^*\}$) differ significantly from 0. The NICOB follows guidance from Jones and Spiegelhalter (2011) about how to identify participants with "unusual" results in an interlaboratory study, in the sense that their measured values lie "beyond the range allowed by the model", as described in §A.5.

Exhibit 10 on Page 37 compares several versions of the unilateral DoE for the PCB 28 measurement results that correspond to the three statistical procedures implemented in the NICOB. Exhibit 22 on Page 51 displays significant bilateral DoE in an interlaboratory study of the activity of radionuclide ^{60}Co.

4 Examples

4.1 Carotid Artery Stenosis

Carotid stenosis is the narrowing of the carotid artery that conveys freshly oxygenated blood to the brain, usually caused by the build up of plaque. When

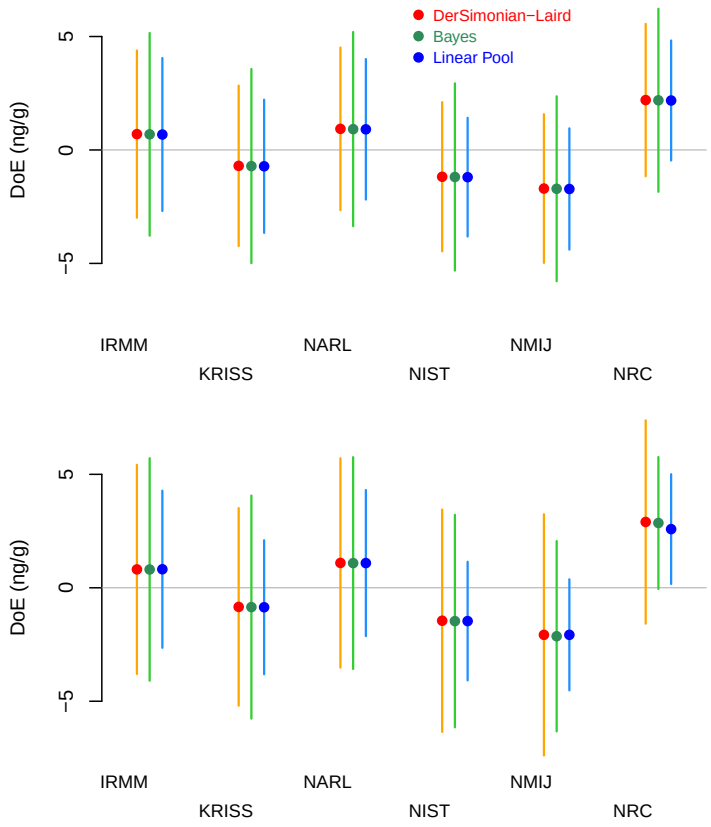

Exhibit 10: The red (DerSimonian-Laird), green (Bayes), and blue (Linear Pool) dots, and line segments of matching colors, depict the unilateral DoE that correspond to the three methods of data reduction implemented in the NICOB: the MRA version in the top panel, and the Leave-One-Out version in the bottom panel. The dots (in either panel), representing the $\{D_j\}$ (upper panel) and the $\{D_j^*\}$ (in the lower panel), are not all at the same heights exactly, even if they appear to be: in the MRA version, their differences are attributable to differences in the consensus values produced by the three procedures for data reduction. The $\{D_j^*\}$ (in the lower panel) are generally larger in absolute value than the corresponding $\{D_j\}$ (in the upper panel), and similarly for the expanded uncertainties.

fragments of plaque break off and blood flow carries them into the brain, they may block smaller arteries and cause a stroke.

Carotid endarterectomy and carotid stenting are two procedures that may be used to treat this condition. The former is the surgical removal of the plaque deposits. The latter involves deployment of an expansible tube (*stent*) inside the artery that mechanically widens the artery's inner diameter.

The upper panel of Exhibit 11 on Page 39 lists the results of nine randomized controlled clinical trials that Meier et al. (2010) selected and combined in a meta-analysis, indicating the numbers of patients involved in each study, and the numbers of these that either suffered a stroke or died within 30 days following the procedure.

4.1.1 Log-Odds Ratios and Standard Uncertainties

The data are a set of nine 2 × 2 tables of counts: for example, for Naylor-1998, the table is displayed in Exhibit 12 on Page 40. Before these data can be used in the NICOB, each 2 × 2 table of counts needs to be reduced to a scalar summary of the relationship between the probability of stroke or death and the two alternative treatments. The log odds ratio is the summary that we shall use in this example, and that we explain next. But first we point out that, regarded as a function of random variables with binomial distributions (the counts k_E and k_S listed in Exhibit 11 on Page 39, assuming that the counts n_E and n_S have been fixed by design) the log-odds ratio has approximately a Gaussian distribution, hence the models implemented in the NICOB are adequate in principle.

Again for Naylor-1998, a naive estimate of the probability of stroke or death among those who underwent endarterectomy is $p_E = 0/12 = 0$. Its counterpart for the group that underwent stenting is $p_S = 5/11$. Therefore, the odds of stroke or death are $p_E/(1 - p_E) = 0$ in the endarterectomy group, and $p_S/(1 - p_S) = 5/6$ in the stenting group. The corresponding *odds ratio* is the ratio of these two odds, which equals 0, whose logarithm (*log-odds ratio*) is negative infinity — obviously problematic as potential input to the NICOB.

Bayesian estimates of p_E and p_S are much more reasonable than the naive estimates above, especially when the number of cases of stroke or death is zero, being of the form $p_E = (k_E + ½)/(n_E + 1)$, and similarly for p_S. These estimates are the means of the posterior distributions that correspond to the Jeffreys prior distribution for the probability of success in n_E and n_S binomial trials, respectively. The posterior distribution for the probability of stroke or death in the

STUDY	n_E	k_E	n_S	k_S	log(OR)	u(log(OR))	ν_{OR}
Naylor-1998	12	0	11	5	−4.2670	2.3209	14
CAVATAS-2001	253	21	251	18	0.1585	0.3342	499
Brooks-2001	51	0	53	0	0.0393	3.1485	102
Brooks-2004	42	0	43	0	0.0211	3.1492	83
SAPPHIRE-2004/8	167	5	167	6	−0.1883	0.6150	330
EVA-3S-2006/8	262	9	265	24	−1.0290	0.4009	445
SPACE-2006	584	36	599	45	−0.2123	0.2316	1165
BACASS-2007	10	1	10	0	2.1059	2.4654	15
ICSS-2009	857	34	853	65	−0.6915	0.2175	1575

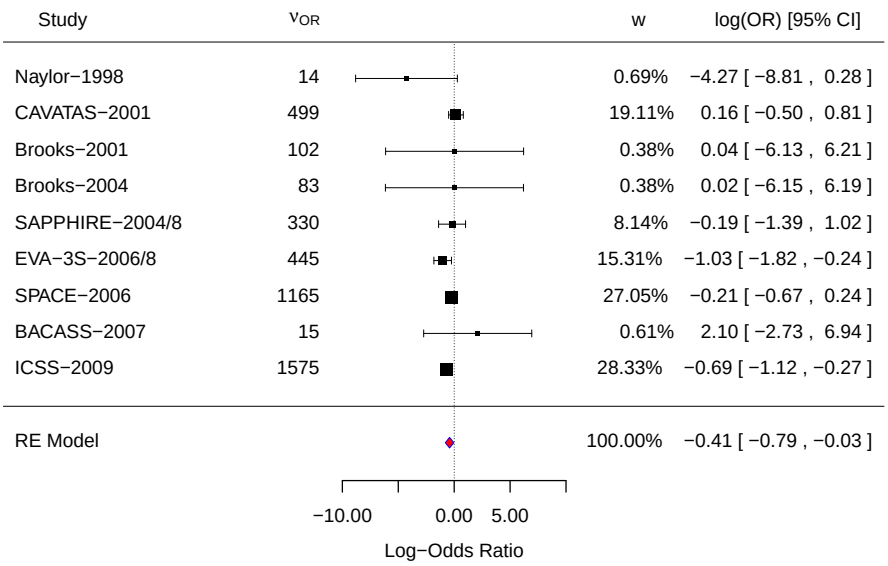

Exhibit 11: TOP PANEL: Results of nine randomized, controlled trials comparing the incidence of strokes among patients suffering from carotid stenosis, and estimates of the corresponding log-odds ratios, associated uncertainties, and degrees of freedom (last three columns in the table). For each study, n_E denotes the total number of patients that underwent carotid endarterectomy, and k_E denotes the number of these who suffered a stroke or died within 30 days following the procedure; n_S and k_S have similar meaning for carotid stenting. BOTTOM PANEL: The *forest plot* shows, for Naylor-1998, for example: the number of degrees of freedom (14) underlying the standard uncertainty of the corresponding log-odds ratio; the weight that the DerSimonian-Laird procedure assigned to the result (2.84 %); the estimate of the log-odds ratio (–4.27); and an approximate, 95 % coverage interval for the true log-odds ratio. The (red) diamond at the bottom indicates the consensus value (–0.41) and associated uncertainty.

	STROKE/DEATH	NONE	
ENDARTERECTOMY	0	12	*12*
STENTING	5	6	*11*
	5	*18*	***23***

Exhibit 12: Counts of outcomes of two alternative procedures to treat carotid artery stenosis corresponding to the measurement results reported in Exhibit 11 on Page 39 for Naylor-1998.

endarterectomy group is a beta distribution with shape parameters $k_E + ½$ and $n_E - k_E + ½$, and similarly for the stenting group.

To evaluate the log-odds ratio and its standard uncertainty computed using Bayes estimates of the relevant probabilities, we make a large number K of draws from these beta distributions (with n_E and n_S kept fixed at the values given), form the corresponding log-odds ratios, and finally compute the mean and standard deviation of the resulting K values of the log-odds ratio. The means and standard deviations obtained in this way, based on samples of size $K = 10^7$ drawn from the appropriate posterior distributions, are listed under log(OR) and u(log(OR)) in Exhibit 11.

The last column in the upper panel of the same Exhibit lists the effective numbers of degrees of freedom ν_{OR} that the values of u(log(OR)) are based on, computed using the Welch-Satterthwaite formula (Miller, 1986, Equation (2.44)), as suggested by Taylor and Kuyatt (1994, §B.3). To load the data for this example into the NICOB, click here.

4.1.2 Results

The results listed in Exhibit 13 on Page 41 correspond to the estimates of the log-odds ratios, and associated standard uncertainties and numbers of degrees of freedom, listed in the last three columns of the upper panel of Exhibit 11.

PROCEDURE	CONSENSUS	STD. UNC.	95 % COV. INT.
DerSimonian-Laird	−0.41	0.21	(−0.83, +0.012)
Hierarchical Bayesian	−0.41	0.24	(−0.88, +0.066)
Linear Pool	−0.46	2.35	(−6.34, +5.01)

Exhibit 13: Results of the three consensus building procedures implemented in the NICOB, for the log-odds ratio that compares the performance of carotid endarterectomy and carotid stenting using the data compiled by Meier et al. (2010). The standard uncertainty and coverage interval for DerSimonian-Laird were computed using the version of the parametric statistical bootstrap described in §A.2, which includes consideration for the small number of measurement results that the estimate of dark uncertainty τ is based on.

4.2 Length of Gauge Blocks

Exhibit 14 on Page 42 lists the measurement results that were used in data reductions for tungsten carbide block 20-23289, of nominal length 1 mm, in key comparison CCL-K1 conducted by the Consultative Committee for Length (CCL). To load these results into the NICOB, click here. The measurement results from VNIIM and NIM are not listed in this Exhibit because they were not used by CCL to compute the consensus value, for the substantive reasons explained by Thalmann (2001, A2.4).

We do, in this manner, comply here with provision (P1) in §3. This should be contrasted with the treatment that Cox (2007, 6.1) makes of the same data that leads to the exclusion of the results from CENAM and VNIM, but not of the results from NIM, based on purely statistical considerations. Cox (2007) also disregarded the numbers of degrees of freedom supporting the $\{u_j\}$. Toman and Possolo (2009a,b, 2010) provide alternative analyses for the data from the same key comparison.

4.2.1 Results

Exhibit 15 on Page 43 lists the results produced by the NICOB for the measurement results in Exhibit 14.

LAB	L	$u(L)$	ν
OFMET	15.0	9.0	500
NPL	15.0	14.0	119
LNE	30.0	10.0	94
NRC	18.0	13.0	9
NIST	24.0	9.0	50
CENAM	−9.0	7.0	72
CSIRO	33.0	9.0	205
NRLM	12.5	8.6	5
KRISS	8.8	10.0	55

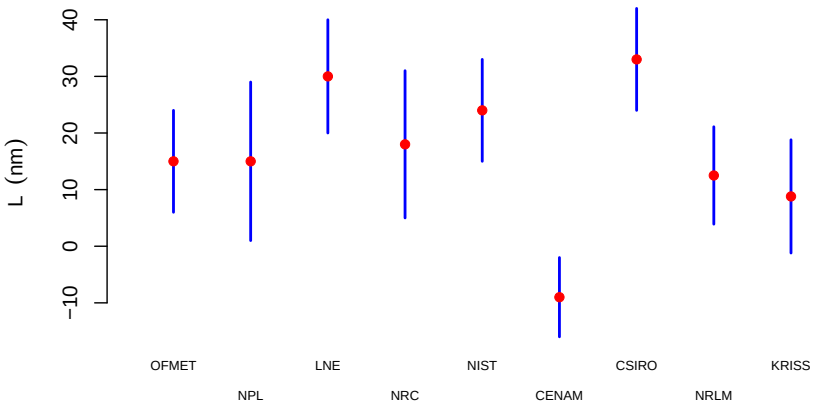

Exhibit 14: TOP PANEL: Differences L between the measured length of titanium carbide block 20-23289, and its nominal length of 1 mm, and associated standard uncertainties $u(L)$, all expressed in nm. Rudolf Thalmann (METAS, Switzerland) kindly shared the numbers of degrees of freedom associated with the laboratory-specific standard uncertainties. BOTTOM PANEL: Measured differences $\{L_j\}$ represented by the red dots, and $\{L_j \pm u(L_j)\}$ represented by the vertical blue line segments, for $j = 1, \ldots, 9$.

PROCEDURE	CONSENSUS	STD. UNC.	95 % COV. INT.
DerSimonian-Laird	15.7	5.1	(5.6, 25.8)
Hierarchical Bayesian	15.5	5.0	(6.1, 25.6)
Linear Pool	16.3	15.5	(−15.2, 44.7)

Exhibit 15: Results of the three consensus building procedures implemented in the NICOB, for the measurement results accepted for use in CCL-K1, pertaining to titanium carbide block 20-23289, of nominal length 1 mm. All values are expressed in nm. The standard uncertainty and coverage interval for DerSimonian-Laird were computed using the version of the parametric statistical bootstrap described in §A.2, which includes consideration for the small number of measurement results that the estimate of dark uncertainty τ is based on.

4.3 Triple point of water

Exhibit 16 on Page 45 lists values of the difference Δ (defined more precisely below), between a national reference cell for the triple point of water and the BIPM reference cell, and associated standard uncertainties, determined in the context of key comparison CCT-K7 conducted by the Consultative Committee for Thermometry (Stock et al., 2006, Table 19). To load the data for this example into the NICOB, click here.

This comparison aimed to achieve (i) "a direct comparison of high-quality water triple point cells to quantify differences between cells", and (ii) "a comparison of the national realizations of the water triple point which served to calibrate the transfer cells" (Stock et al., 2006, Pages 5–6). The value of the difference Δ_j for laboratory j was determined as a difference between two differences: $\Delta_j = (T_{\text{BIPM}}(\text{TRANSTD}_j) - T_{\text{BIPM}}(\text{REF})) - (T_j(\text{TRANSTD}_j) - T_j(\text{NTLSTD}_j))$, where $T_{\text{BIPM}}(\text{TRANSTD}_j)$ denotes the value that the BIPM measured for the transfer standard sent by laboratory j, $T_{\text{BIPM}}(\text{REF})$ denotes "the temperature attributed to the BIPM reference group", $T_j(\text{TRANSTD}_j)$ denotes the average of the two values of temperature of the transfer standard that laboratory j measured before and after the transfer standard went to the BIPM, and $T_j(\text{NTLSTD}_j)$ denotes the temperature of the national standard used by laboratory j.

All the models offered in the NICOB assume that, given the values of the underly-

ing parameters, the measured values are like observed outcomes of independent random variables. The fact that $T_{\text{BIPM}}(\text{REF})$ figures in all of them challenges the validity of such assumption in this case. The concern may be alleviated to some extent by the understanding that the uncertainty associated with $T_{\text{BIPM}}(\text{REF})$ may be negligible by comparison with the uncertainties associated with all the other measured values.

The situation (and the pattern of possibly induced correlations) is further complicated by the fact that a least squares adjustment was applied to all the measurement results made at the BIPM (Stock et al., 2006, 4.3). We will ignore these complications and proceed on the still questionable, aforementioned assumption of independence.

It should be noted that all three data reduction procedures available in the NICOB can be generalized to accommodate correlations between values measured by different studies, even though they are not currently implemented in the NICOB. Chen et al. (2016) describes such generalization for the DerSimonian-Laird procedure. The modification of the hierarchical Bayesian procedure involves using a multivariate Gaussian distribution for the likelihood function, and application of a copula (Nelsen, 1999; Possolo, 2010) would accomplish the same purpose for the Linear Pool.

4.3.1 Results

The NICOB results listed in Exhibit 17 on Page 46 correspond to the estimates of the difference $\Delta = T_{\text{LAB}} - T_{\text{BIPM}}$ between a national reference cell for the triple point of water and the BIPM reference, and associated standard uncertainties, listed in the upper panel of Exhibit 16.

Stock et al. (2006, Page 68) point out that "*For this key comparison, the KCRV is based on the mean value of the results from all of the participants, including some laboratories who made corrections for the influence of chemical impurities and isotopic composition, and some who did not. The uncertainty of the KCRV is taken to be the standard deviation of the mean of the data set. Because the distribution of the pooled data is multimodal, care should be taken when using this quantity for calculating confidence intervals.*"

The authors then explain that "*it was decided not to use the propagated uncertainties of the participants' results because many of them are underestimated.*" However, they do not offer suggestions for how severe this underestimation may be. The Bayesian procedure implemented in the NICOB facilitates taking this type of

LAB	Δ	$u(\Delta)$	LAB	Δ	$u(\Delta)$
BIPM	0	44	NIST	−40	33
BNM	−54	66	NMIJ	54	151
CEM	−14	41	NMi-VSL	16	55
CENAM	−5	27	NPL	45	39
CSIR	105	74	NRC	85	23
CSIRO	−29	34	PTB	−14	56
IMGC	−15	27	SMU	69	53
IPQ	40	160	SPRING	34	71
KRISS	69	56	UME	−53	91
MSL	117	16	VNIIM	22	46
NIM	33	61			

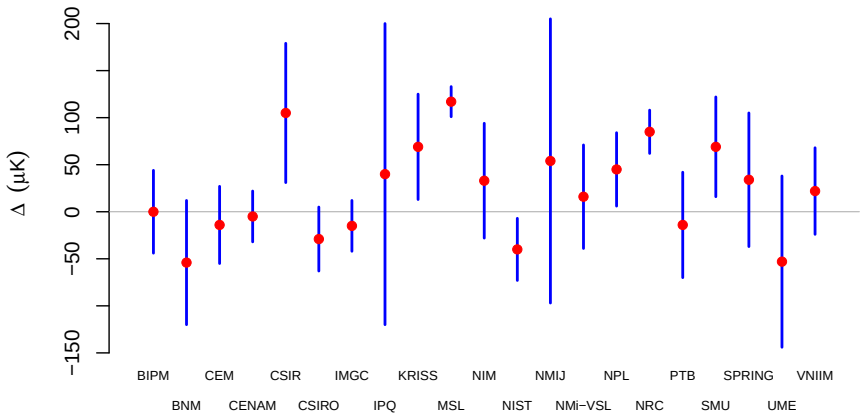

Exhibit 16: TOP PANEL: Values of the difference $\Delta = T_{\text{LAB}} - T_{\text{BIPM}}$ between a national reference cell for the triple point of water and the BIPM reference cell, and associated standard uncertainties, all expressed in μK. BOTTOM PANEL: Measured differences $\{\Delta_j\}$ represented by the red dots, and $\{\Delta_j \pm u(\Delta_j)\}$ represented by the vertical blue line segments.

external information into account. Under the belief that the uncertainty evaluations produce values that are too small, the median of the prior distribution for the corresponding $\{\sigma_j\}$ may be increased, and multiple values tried, to determine how influential this choice may be upon the final results.

We have considered three scenarios: first where the scale was set equal to the median, 53 µK, of the reported laboratory-specific standard uncertainties $\{u_j\}$ (which is the default choice in the NICOB); second where the scale was set equal to the largest of the $\{u_j\}$, which is 160 µK; and third where it was set equal to five times the median of the $\{u_j\}$, at 265 µK. The corresponding estimates of the consensus value, and the associated standard uncertainties, varied by about 0.1 µK in consequence, which suggests that, in this case at least, the results are rather resilient to substantial misreporting of the laboratory-specific standard uncertainties.

PROCEDURE	CONSENSUS	STD. UNC.	95 % COV. INT.
DerSimonian-Laird	23	15	(−8, 53)
Hierarchical Bayesian	24	14	(−4, 52)
Linear Pool	22	84	(−140, 191)

Exhibit 17: Results of the three consensus building procedures implemented in the NICOB, for the difference $\Delta = T_{\text{LAB}} - T_{\text{BIPM}}$ between the national reference cells for the triple point of water and the BIPM reference, all expressed in µK. The standard uncertainty and coverage interval for DerSimonian-Laird were computed using the version of the parametric statistical bootstrap described in §A.2, which includes consideration for the small number of measurement results that the estimate of dark uncertainty τ is based on.

For these data, Stock et al. (2006) note: "*It is instructive to look at the joint or pooled probability distribution, calculated as the sum of the individual probability distributions (Figure 29). The individual distributions were assumed as Gaussian. The joint distribution looks like the superposition of a broader distribution centered at* −5 µK *and a narrower distribution centered at* 110 µK." Exhibit 18 on Page 47 is the counterpart of Figure 29 of Stock et al. (2006), who suggest that the secondary peak at higher temperatures is attributable to measurements made by laboratories that were able to apply corrections for deviations between the iso-

topic composition of the water in their cells and the "*isotopic composition of the cell water from V-SMOW (Vienna Standard Mean Ocean Water, prepared by the Atomic Energy Commission in Vienna).*"

We imported into R the sample of 10 000 values drawn from the consensus value by application of the Linear Pool, and fitted a mixture of Gaussian distributions to it, using R function `normalmixEM` defined in package `mixtools` (Benaglia et al., 2009). The Bayesian Information Criterion (BIC) for model selection suggested that the best model should have three components (Burnham and Anderson, 2002), which are depicted in Exhibit 18 on Page 47. The two components with smallest standard deviations are centered at −2 µK (accounting for 59 % of the unit of probability) and 102 µK (accounting for 17 % of the unit of probability), and they determine the more salient characteristics of the distribution.

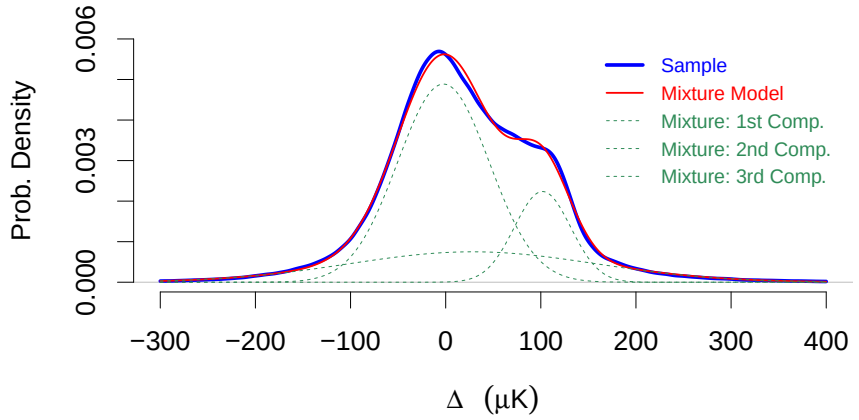

Exhibit 18: The thick (blue) line delineates the probability density of the result of linearly pooling Gaussian probability distributions with means $\{A_j\}$ and standard deviations $\{u(A_j)\}$. The thin (blue) lines are scaled to one-half the scale of the vertical axis.

4.4 Activity of Radionuclide ^{60}Co

^{60}Co is a synthetic radioactive isotope of cobalt whose atoms have nuclei with 33 neutrons and 27 protons. Its half-life is 5.2711 years, decaying to stable ^{60}Ni by emission of a beta particle and highly energetic gamma rays (releasing about

20 W per gram of the radioisotope). ^{60}Co is widely used for sterilization of medical instruments, fruit, etc., and for cancer radiotherapy.

The upper panel of Exhibit 19 on Page 49 lists measurement results for the International Reference System (SIR) equivalent activity (Michotte, 2002) of ^{60}Co produced by 19 laboratories, as reported in the BIPM Key Comparison Database (KCDB, http://kcdb.bipm.org) for ongoing key comparison BIPM.RI(II)-K1.Co-60 (Ratel et al., 2003a,b; Ratel and Michotte, 2003; Ratel et al., 2006; Michotte et al., 2010). To load these measurement results into the NICOB, click here.

4.4.1 Results

The results listed in Exhibit 21 on Page 50 correspond to the measurement results listed in the top panel of Exhibit 19. Exhibit 20 shows the probability density that represents the result of linearly pooling (that is, mixing) distributions assigned to the participating laboratories.

4.5 Radiofrequency Power Sensor

Exhibit 23 on Page 52 lists measurement results from key comparison CCEM.RF-K25.W (Judaschke, 2014, 2015), of the calibration factor η_{CAL} at 33 GHz, for a commercial, temperature-compensated thermistor power sensor that circulated among the participants. To load these results into the NICOB, click here. The calibration factor is the proportion of the true power of the radiofrequency signal that the sensor actually measures, hence it is a dimensionless quantity with values typically between 0 and 1.

In the original study, the measurement result from NMIA (Australia), which is not listed in Exhibit 23, was not used in the computation of the key comparison reference value (KCRV) because the value measured by this institute is traceable to the international system of units (SI) (BIPM, 2006) through another participant.

The original study also excluded the measurement result from NRC (Canada) because a statistical criterion described by Randa (2005) identifies it as an outlier, and computed the KCRV as the arithmetic average 0.8184 of the seven remaining measured values, with associated standard uncertainty 0.0028. We have included the Canadian result in our analysis, and in the end conclude that the corresponding unilateral degree of equivalence, evaluated by all three procedures implemented in the NICOB, does not differ significantly from zero (Exhibit 23).

LAB	A_j	$u(A_j)$	DATE	LAB	A_j	$u(A_j)$	DATE
	kBq				kBq		
LNMRI	7077	8	1984-11-21	PTB	7057	16	2001-07-02
ENEA	7065	26	1991-01-22	NMISA	7098	16	2002-05-30
ANSTO	7056	10	1992-05-13	CNEA	7050	15	2003-01-17
KRISS	7047	22	1995-01-18	RC	7040	40	2003-06-17
MKEH	7051	18	1999-06-11	NMIJ	7050	8	2004-03-17
LNE-LNHB	7060	4	1999-10-20	IRMM	7039	17	2005-01-27
CIEMAT	7090	11	1999-11-30	IFIN-HH	7101	24	2007-05-10
NPL	7053	21	2000-06-30	NIST	7083	14	2007-08-07
IRA	7037	8	2000-12-06	BEV	7057	17	2007-09-28
BARC	7099	46	2001-01-10				

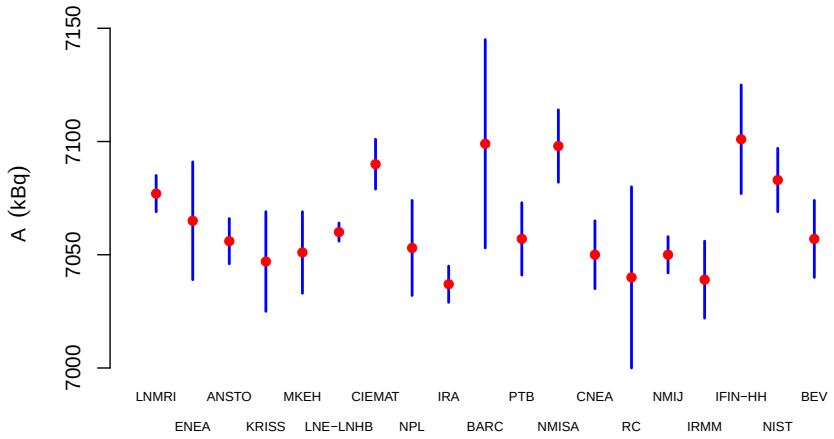

Exhibit 19: TOP PANEL: Measurement results for the SIR equivalent activity of ^{60}Co. BOTTOM PANEL: The large (red) dots represent the values of activity $\{A_j\}$ measured by the participating laboratories; the vertical (blue) line segments depict the intervals $\{A_j \pm u(A_j)\}$.

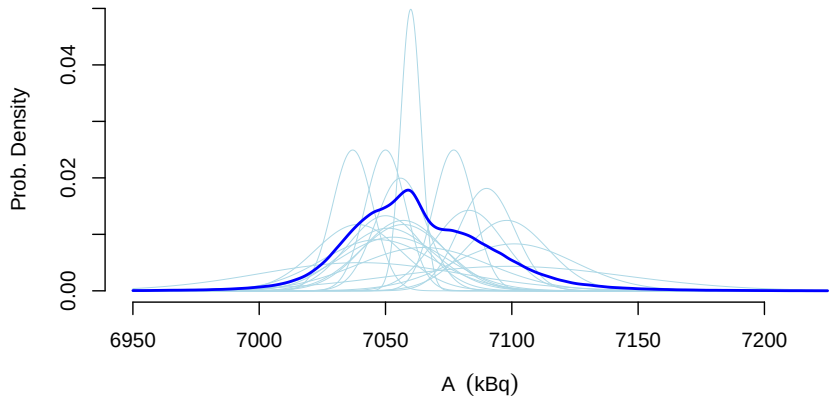

Exhibit 20: The thick (blue) line delineates the probability density that represents the result of linearly pooling Gaussian probability distributions with means $\{A_j\}$ and standard deviations $\{u(A_j)\}$ that represent the states of knowledge of the 19 participating laboratories about the true SIR equivalent activity of ^{60}Co. The thin (blue) lines are scaled to one-half the scale of the vertical axis.

PROCEDURE	CONSENSUS	STD. UNC.	95 % COV. INT.
DerSimonian-Laird	7062	4	(7053, 7071)
Hierarchical Bayesian	7062	5	(7053, 7072)
Linear Pool	7064	29	(7012, 7127)

Exhibit 21: Results of the three consensus building procedures implemented in the NICOB, for the SIR equivalent activity of ^{60}Co. The standard uncertainty and coverage interval for DerSimonian-Laird were computed using the version of the parametric statistical bootstrap described in §A.2, which includes consideration for the small number of measurement results that the estimate of dark uncertainty τ is based on.

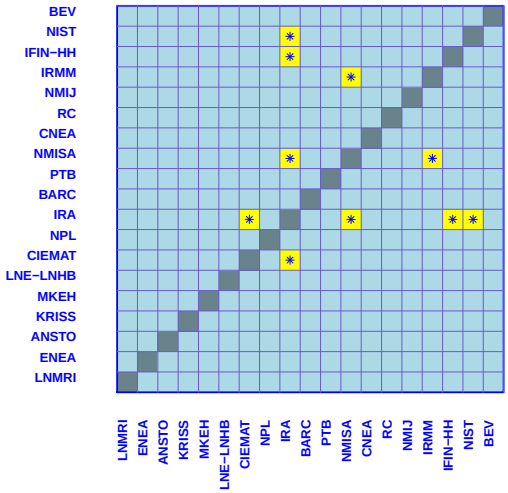

Exhibit 22: The yellow squares with an asterisk inside indicate B_{ij} that differ significantly from 0 in the sense that the interval $B_{ij} \pm U_{95\,\%}(B_{ij})$ does not include 0.

Cochran's Q-test of homogeneity yields a p-value of 0.59, therefore not suggesting heterogeneity. Hoaglin (2016) points out several shortcomings of this test. However, the criterion described by Randa (2005) can be criticized because it neglects the uncertainties associated with the measured values. In any case, principle (P1) from §3 rules out the practice of excluding measured values from the calculation of the KCRV based on statistical criteria alone.

The DerSimonian-Laird procedure, as implemented in the NICOB and applied to all eight measurement results listed in Exhibit 23, estimates the KCRV as 0.8192, with associated standard uncertainty 0.0022, and a 95 % coverage interval ranging from 0.8147 to 0.8235 (where the standard uncertainty and the coverage interval were obtained by application of the parametric bootstrap).

Exhibit 24 on Page 53 lists the results of data reductions from the final report of the key comparison, and those produced by the NICOB. The close agreement between the unilateral degrees of equivalence (computed as defined in the MRA) produced by the DerSimonian-Laird and Bayesian procedures is particularly striking. The Linear Pool produces considerably larger uncertainty evaluations than the other two procedures. Still, none of the three procedures suggests

LAB	η_{CAL}	$u(\eta_{\text{CAL}})$
KRISS	0.8247	0.0095
LNE	0.8184	0.0112
NIM	0.8196	0.0033
NIST	0.8170	0.0070
NPL	0.8069	0.0072
NRC	0.8355	0.0130
PTB	0.8186	0.0038
VNIIFTRI	0.8236	0.0058

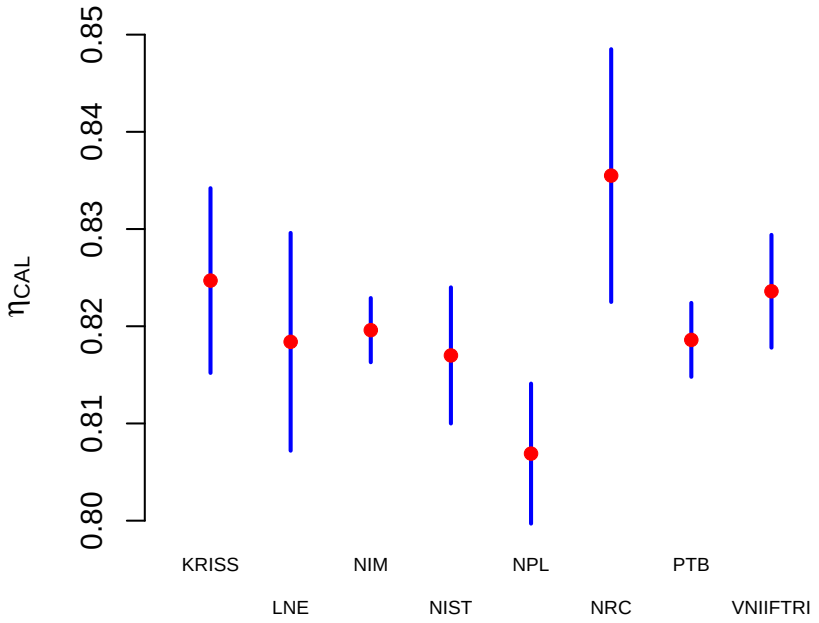

Exhibit 23: TOP PANEL: Measurement results for the calibration factor η_{CAL} of traveling standard SN 216 used in key comparison CCEM.RF-K25.W. BOTTOM PANEL: The large (red) dots represent the values measured by the participating laboratories; the vertical (blue) line segments depict the intervals $\{\eta_{\text{CAL},j} \pm u(\eta_{\text{CAL},j})\}$.

a significant discrepancy for the measurement results from NRC.

PROCEDURE	CONSENSUS	STD. UNC.	95 % COV. INT.
CCEM.RF-K25.W	0.8184	0.0028	
DerSimonian-Laird	0.8192	0.0022	(0.8147, 0.8235)
Hierarchical Bayesian	0.8192	0.0022	(0.8192, 0.8239)
Linear Pool	0.8205	0.0112	(0.7993, 0.8471)

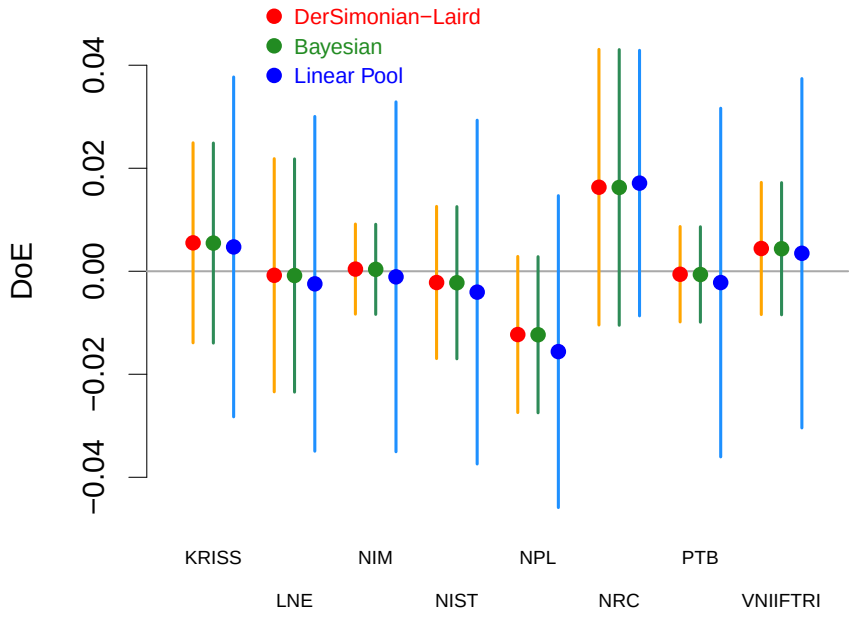

Exhibit 24: TOP PANEL: Results of the three consensus building procedures implemented in the NICOB for the calibration factor η_{CAL} of traveling standard SN 216 used in key comparison CCEM.RF-K25.W. The standard uncertainty and coverage interval for the DerSimonian-Laird procedure were computed using the version of the parametric statistical bootstrap described in §A.2, which includes consideration for the small number of measurement results that the estimate of dark uncertainty τ is based on. BOTTOM PANEL: Unilateral degrees of equivalence as defined in the MRA corresponding to the DerSimonian-Laird, Bayesian, and Linear Pool procedures.

5 Advisory

The NICOB ought not be misconstrued as a toolbox capable of addressing all the needs of data reductions arising in the context of interlaboratory studies or inter-method comparisons.

Quite the contrary: even a cursory examination of Final Reports of key comparisons available in the BIPM Key Comparison Database (KCDB, http://kcdb.bipm.org), or of articles in medical journals describing meta-analyses, reveals that most data reductions done in these contexts require customized treatments that only a statistician, biostatistician, or applied mathematician can give, who is committed to learning about and understanding the substantive issues, and willing to work collaboratively with scientists or medical doctors.

A common occurrence that renders the NICOB inapplicable is temporal drift in the value of the measurand in the course of the interlaboratory study (Zhang, 2012). This often arises when the study requires that a physical artifact be circulated and some attribute of it measured by the participating laboratories in turn. Unless this drift can be reliably corrected for, and the resulting uncertainty quantified, the NICOB affords no built-in means to address this type of situation.

We have pointed out already (in §3) that the NICOB is not suitable for the reduction of data from proficiency tests (Thompson et al., 2006) because these typically involve a reference value that is not a consensus value derived from the participants' measurement results, and also because the data reductions for such tests usually produce several performance metrics that the NICOB does not evaluate.

We have also pointed out in §3.2.5 that in many cases encountered in practice there is considerable *a priori* information about the measurand, or about other parameters in the model, but that taking such information into account requires a custom solution that the NICOB is unable to provide, except to the limited extent that it affords to tune the hyper-parameters governing the prior distributions for τ and for the $\{\sigma_j\}$.

In some key comparisons there is a different reference value for each laboratory: for example, when different laboratories measure amount-of-substance fractions of the same gas in mixtures nominally of the same composition but prepared separately from one another, or degrading over time differently from one another, as in CCQM-K90 (formaldehyde in nitrogen). The NICOB cannot address the challenges posed by these, either.

In many cases, the least that needs to happen before the NICOB can be used is

some suitable pre-processing of the data observed in an interlaboratory study or gathered for a meta-analysis. The example presented in §4.1 illustrates an instance of this need.

In other cases, the NICOB should simply not be used at all because there is a fundamental mismatch between the structure of the experimental data and the assumptions underlying the statistical methods implemented in the NICOB. The mere fact that a table in a journal article or technical report lists measurement results in a form that appears suitable for use in the NICOB, is not sufficient reason to employ any of the statistical models and methods that it offers.

One such case is key comparison CCT-K4, conducted by the Consultative Committee for Thermometry (CCT), to compare local realizations of the ITS-90 temperature scale using aluminum and silver freezing-point cells (Nubbemeyer and Fischer, 2002). Table 8 in the corresponding Final Report lists 12 differences $T_{\text{LAB}} - T_{\text{MC}}$, and associated standard uncertainties, between the temperature of the freezing point of silver measured by a participating laboratory, and the corresponding temperature measured by the pilot laboratory on a circulating, master fixed-point cell.

Cox (2007) used this data set to illustrate the concept of largest consistent subset. At first blush, these data, depicted in Exhibit 25 on Page 56, seem to be begging for analysis using the NICOB or some comparable other means. However, yielding to such temptation would be inappropriate because none of the models available in the NICOB are suitable.

The reason is that each of the differences is a linear combination (with coefficients that vary from laboratory to laboratory) of the same set of other differences (between temperatures of two fixed-point cells) that were measured by the participating laboratories on different occasions in the course of the experiment. Thus, the random variables used to model the differences in Table 8 are correlated to an extent that the Final Report does not consider, and that the NICOB is not equipped to recognize or address. There are, however, versions of the random effects model that can take the web of correlations pervading the data in this example into account, for example as implemented in R package `mvmeta` (Gasparrini, 2012).

The Final Report reveals that the CCT did not reach unanimity about whether a consensus value (KCRV) should be computed, and that a vote was taken to adjudicate the matter. The majority were inclined to compute the KCRV (as a particular weighted average), but the Final Report notes that "The KCRV has no physical

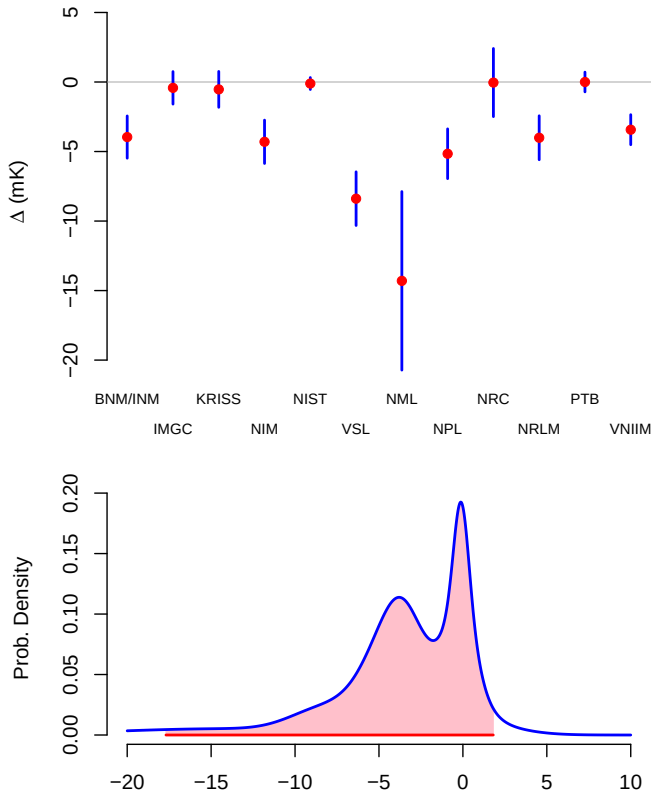

Exhibit 25: TOP PANEL: Measurement results from CCT-K4, given in Nubbemeyer and Fischer (2002, Table 8), where $\Delta = T_{\text{LAB}} - T_{\text{MC}}$ is the difference between the temperature of the freezing point of silver measured by a participating laboratory, and the corresponding temperature measured by the pilot laboratory on a circulating, master fixed-point cell. The value for PTB, which was the pilot, is 0 mK by construction. BOTTOM PANEL: Probability distribution for the consensus value, obtained by linearly pooling, and interval of values of Δ (thick, horizontal red line) that is the projection of the (pink) region that encompasses 95 % of the area under the (blue) curve. As noted in the text, none of the methods implemented in the NICOB are suitable for reducing these data.

meaning and is used only as a notational shorthand for presenting a common baseline against which all laboratory values can be compared" (Nubbemeyer and Fischer, 2002, Page 25).

Members in the minority offered explanations for their opposition to computing a KCRV, which the same Final Report also records, and that provide informative insights into some of the challenges in statistical modeling and data analysis that no set of fixed "recipes" (for example as are implemented in the NICOB) is likely to be able to address successfully.

This was the objection raised by NIST (Nubbemeyer and Fischer, 2002, 5.3.2):

> *In order to make valid comparisons in a key comparison in which different laboratories have made measurements using non-identical transfer standards, the data must first be normalized to a common basis. The normalization of the data usually introduces different levels of uncertainty and correlations between different laboratories' results, which complicates the computation of the uncertainties of the pair-wise differences between the laboratories. Under these conditions, which are present in KC4 (and KC3), determination of the uncertainties usually requires the use of a variance-covariance matrix (or equivalent non-matrix calculations) to be used in determining the bilateral difference uncertainties. Therefore, unless the uncertainties and correlations introduced by the normalization of the data to accommodate the use of different transfer instruments happens to be the same for all comparisons between laboratories, the KCRV approach to expressing the degree of equivalence cannot be easily implemented in the usual way.*

Another dissenting voice (on the appropriateness of computing a consensus value for CCT-K4) arose from NRC (Canada), expressing a different reason for concern about the computation of a consensus value:

> *Why am I against the calculation of a KCRV? Basically, I see the K4 results as a failure to demonstrate compatibility amongst the world's best laboratories. Clearly, much work needs to be done if we wish to achieve such interoperability. It is disconcerting that the identification of the failing laboratories is so sensitive to the method used to calculate the KCRV. This represents, to me, a sufficient technical reason to avoid defining a KCRV.*

Albeit indirectly, the remark relates to the presence of *dark uncertainty*, already mentioned in §3, whereby the dispersion or scatter of measured values is much larger than what the stated, laboratory-specific uncertainties can account for.

The laboratory random effects model discussed in Appendix A provides an effective framework to address this challenge, even if it does not, by itself, solve the fundamental scientific problem that induces the observed heterogeneity. However, by identifying and quantifying this heterogeneity (in an estimate of the between-laboratory standard deviation that we have been calling τ), a flag is raised signaling the need for a solution of the underlying scientific or technical problem.

Interestingly, NRC qualified their objections by presenting the probability distribution for the difference $\Delta = T_{\text{LAB}} - T_{\text{MC}}$ that results from equally-weighted mixing of Gaussian probability distributions with means equal to the measured differences $\{\Delta_j\}$, and standard deviations equal to the corresponding $\{u(\Delta_j)\}$. This is the same as linearly pooling (§A.4), even though in this case it would neglect those correlations. Since the mixture distribution is bimodal (Exhibit 25), NRC concluded that a consensus value would be meaningless. When the Linear Pool in the NICOB is applied to this data, it produces a 95 % coverage interval (supposedly for the true value of the consensus value), that is so wide that it covers all 12 measured values of Δ. If the correlations were taken into account, then this interval likely would be wider still.

6 Implementation

The implementation of the NICOB has been done mostly in R (R Core Team, 2015), given R's fitness for purpose, wealth of specialized functionality, and universal availability. We leveraged these resources as deployed in packages `metafor` (Viechtbauer, 2010) for the DerSimonian-Laird procedure, and in `R2jags` (Su and Yajima, 2015) for the Bayesian procedure, which employs JAGS, a computer program for the analysis of Bayesian hierarchical models using Markov Chain Monte Carlo sampling (Plummer, 2015).

The codes for the parametric bootstrap, for drawing samples from the approximate sampling distribution of the estimate of τ, and for the degrees of equivalence, were developed specifically for the NICOB.

To make the application accessible to users with no knowledge of R we have created an easy-to-use graphical user interface displayed in a web browser em-

ploying facilities provided by the R package `shiny` (Chang et al., 2016). NIST hosts the NICOB at `consensus.nist.gov`, with several instances of the Shiny app running concurrently for load balancing.

A Appendix: Statistical Procedures

This section provides details of the statistical models and methods implemented in the NICOB, and discusses the conditions under which they are expected to produce valid results. The three procedures made available in this version of the NICOB have a long history of usage and a proven track record of reliable performance. They are not, however, interchangeable, and generally one should be chosen that seems most adequate for the data in hand, and best fit for the purpose that the analysis is intended to serve.

A large collection of statistical procedures for the analysis of results from interlaboratory studies is available. In March 2016, there were about one dozen R packages available on the Comprehensive R Archive Network (CRAN) offering functions for such analyses. Therefore, our selection of only three among this multitude of procedures may be easily challenged.

We selected no more than three because offering the user a large set of alternatives risks devolving model selection into something akin to tasting the 48 flavors of *Rick's Rather Rich Ice Cream* (Palo Alto, CA), until one finds the results that one likes best — a deplorable, if common, manifestation of statistical malpractice.

We selected no fewer than three because we wish to encourage the user to entertain models and sets of assumptions that are clearly different from one another, in hopes that at least one of the three will be appropriate for the data in hand and will produce results that are fit for purpose.

- We selected the DerSimonian-Laird procedure (§A.2) because it is used most often in practice, in particular in medical meta-analysis, which accounts for the bulk of the interlaboratory studies performed each year, and because it makes fewer assumptions about the nature of the data than the hierarchical Bayesian procedure.
- We selected a hierarchical Bayesian procedure (§A.3) because it allows the expression of a modicum of prior knowledge about uncertainty components that typically are difficult to evaluate reliably, and because it takes

into account, without special "add-ons", the uncertainty surrounding the estimate of τ.

- We selected the Linear Pool (§A.4) due to its longevity and because it makes the fewest assumptions about the nature of the data.

Furthermore, the key motivating ideas behind all three are fairly easy to explain in non-technical terms. Since random effects models underlie both the DerSimonian-Laird and the hierarchical Bayesian procedure, §A.1 explains the difference between random and fixed effects models, and reviews the reasons why we favor the former over the latter, for general use in the reduction of data from interlaboratory studies and from inter-comparisons of alternative measurement methods. The approaches implemented in the NICOB for the evaluation of degrees of equivalence are explained in §A.5.

A.1 Random *versus* Fixed Effects

Three different models are frequently used (and often confused) for values measured in interlaboratory studies: (i) *random effects* models; (ii) *fixed effects* models; and (iii) *common mean* or *fixed effect* (note the singular) models, defined as follows.

Random Effects $x_j = \mu + \lambda_j + \varepsilon_j$, where μ denotes an unknown constant, and λ_j and ε_j are values of non-observable random variables, for $j = 1, \ldots, n$, the former with mean 0 and common variance τ^2, the latter with mean 0 and possibly different variances $\sigma_1^2, \ldots, \sigma_n^2$.

Fixed Effects $x_j = \theta_j + \varepsilon_j$, where θ_j denotes an unknown constant and ε_j denotes the value of a non-observable random variable with mean 0 and variance σ_j^2, for $j = 1, \ldots, n$.

Common Mean (Fixed Effect) $x_j = \mu + \varepsilon_j$, where μ denotes an unknown constant and ε_j denotes the value of a non-observable random variable with mean 0 and variance σ_j^2 for $j = 1, \ldots, n$.

Both the DerSimonian-Laird procedure and the hierarchical Bayesian procedure are based on random effects models. These models are appropriate when one wishes to derive lessons from an interlaboratory study that will be applicable

to laboratories similar to those that have participated in the study, or when one must recognize between-laboratory differences as a source of measurement uncertainty, hence very broadly indeed (Mandel, 1964; Mandel and Paule, 1970).

Entertaining a fixed effects model where the $\{\theta_j\}$ are different from one another is tantamount to admitting that the different laboratories are measuring different quantities owing to persistent effects (biases) that do not average out as each laboratory replicates its measurements. In such circumstances, no consensus value can be meaningful.

Only when the $\{x_j\}$ have a common expected value μ is there a meaningful consensus value, and this happens only for the random effects model and for the common mean model. And among these two, the former should be preferred, as Borenstein et al. (2010, Page 107) argues:

> "If we were going to use one model as the default, then the random-effects model is the better candidate because it makes less stringent assumptions about the consistency of effects. Mathematically, the fixed-effect model is really a special case of the random-effects model with the additional constraint that all studies share a common effect size. To impose this constraint is to impose a restriction that is not needed, not proven, and often implausible."

A.2 DerSimonian-Laird

The DerSimonian-Laird (§A.2) and the Bayesian hierarchical (§A.3) procedures implement a *random effects* model that expresses each measured value as an additive superposition of three elements: $x_j = \mu + \lambda_j + \varepsilon_j$, for each of $j = 1, \ldots, n$ laboratories, where μ is the measurand, the $\{\lambda_j\}$ denote laboratory (or, method) effects, and the $\{\varepsilon_j\}$ represent measurement errors. (The Linear Pool is based on a different model, which will be discussed in §A.4.)

The random variables that the $\{\lambda_j\}$ represent are assumed to be independent and to have mean 0 and standard deviation $\tau \geqslant 0$. These assumptions mean that, taken collectively, the measured values are unbiased (that is, are centered on μ).

When $\tau = 0$, the measurement results are said to be *homogeneous* (or, mutually *consistent*) — in such case, the DerSimonian-Laird procedure reduces to Procedure A of Cox (2002). Even if the results are homogeneous, the evaluation of the

uncertainties associated with the consensus value and the degrees of equivalence (§A.5) performed by the NICOB is far more sophisticated and realistic than what Cox (2002) describes for the reasons given below.

The measurement errors are assumed to be realized values of independent random variables with mean 0, but possibly different standard deviations $\{\sigma_j\}$. The standard measurement uncertainties $\{u_j\}$ are assumed to be equal to the $\{\sigma_j\}$ only when they are based on infinitely many degrees of freedom. The DerSimonian-Laird procedure (DerSimonian and Laird, 1986; Whitehead and Whitehead, 1991) treats the uncertainties associated with the measured values as known constants (that is, based on infinitely many degrees of freedom). This assumption is unwarranted in most cases.

However, when the standard uncertainties that are used as inputs are qualified with numbers of degrees of freedom, then the NICOB takes them into account when it evaluates uncertainty using the parametric statistical bootstrap. The Bayesian procedure does likewise to some extent even when numbers of degrees of freedom are not reported.

The consensus value, which estimates μ, is a weighted average of the values measured by the participating laboratories, $\hat{\mu} = \sum_{j=1}^{n} w_j x_j / \sum_{j=1}^{n} w_j$, with weights $w_j = 1/(\tau^2 + \sigma_j^2)$ for $j = 1, \ldots, n$. Since τ is unknown, it is replaced by a method-of-moments estimate: $\hat{\tau}^2_{\mathrm{DL}} = \max\{0, \hat{\tau}^2_{\mathrm{M}}\}$, where $\hat{\tau}^2_{\mathrm{M}} = (Q - n + 1)/(\sum_{j=1}^{n} u_j^{-2} - \sum_{j=1}^{n} u_j^{-4} / \sum_{j=1}^{n} u_j^{-2})$, and $Q = \sum u_j^{-2}(x_j - \hat{\mu})^2$. The $\{u_j\}$ replace the $\{\sigma_j\}$ that usually figure in the expression for $\hat{\tau}$.

The approximate standard uncertainty associated with the consensus value is $u_{\mathrm{DL}}(\mu) = \sqrt{1/\sum_{j=1}^{n} w_j}$ (Higgins et al., 2009). The presence of τ^2 in the denominator of the weights $\{w_j\}$ acts as a moderating influence, preventing very small laboratory-specific uncertainties from influencing the consensus value to an extent that is often found to be objectionable when conventional weighted averages are used, as in Procedure A of Cox (2002).

The NICOB offers two, more accurate alternatives to this approximation: one suggested by Knapp and Hartung (2003), the other made possible by technology developed by Biggerstaff and Tweedie (1997) and Biggerstaff and Jackson (2008). However, Hoaglin (2016) warns that, although improving on naive evaluations, these alternatives still rely on the typically unrealistic assumption that $u_j = \sigma_j$, which is particularly harmful when the number of participating laboratories is small (say, less than 10). These concerns are mitigated by the parametric bootstrap option that the NICOB offers for uncertainty evaluation, even though this

involves additional assumptions.

In the DerSimonian-Laird procedure, neither the estimate of the measurand, nor the (naive) corresponding uncertainty evaluation $u_{\mathrm{DL}}(\mu)$, depend on any specific distributional assumption about the laboratory effects or the measurement errors, provided that they are independent and all have finite variances. In particular, the $\{x_j\}$ are not assumed to be outcomes of Gaussian random variables.

The DerSimonian-Laird procedure, and the Knapp-Hartung adjustment, are implemented in the NICOB via R function `rma` defined in package `metafor` (Viechtbauer, 2010). The version of the parametric bootstrap that we have developed and implemented recognizes the typically small number of degrees of freedom supporting the estimate of τ along lines suggested by Biggerstaff and Tweedie (1997) and Biggerstaff and Jackson (2008).

To approximate the distribution of τ^2, Biggerstaff and Tweedie (1997) derive the exact mean and variance of Cochran's Q statistic. From these, the distribution of Q is approximated using a gamma distribution. Since $\tau^2_{\mathrm{M}} = (Q - n + 1)/(\sum_{j=1}^{n} u_j^{-2} - \sum_{j=1}^{n} u_j^{-4} / \sum_{j=1}^{n} u_j^{-2})$, the distribution of τ^2_{M} can be approximated using a location-shifted, scaled gamma distribution. Thus to simulate from the approximate distribution of $\tau^2_{\mathrm{DL}} = \max\{0, \tau^2_{\mathrm{M}}\}$, we simulate τ^2_{M} from the appropriate gamma distribution and take the maximum of that simulated sample and 0.

The uncertainty evaluation via the parametric statistical bootstrap (Efron and Tibshirani, 1993) is consistent with the GUM Supplement 1 (Joint Committee for Guides in Metrology, 2008b). The raw materials for this evaluation are obtained by repeating the following steps a large number (K) of times, for $k = 1, \ldots, K$:

(a) Draw τ_k from the approximate probability distribution of $\hat{\tau}_{\mathrm{DL}}$;

(b) Draw x_{jk} from a Gaussian distribution with mean $\hat{\mu}$ and variance $\tau_k^2 + u_j^2$, for $j = 1, \ldots, n$;

(c) If ν_j is either infinity or unspecified, $u_{jk} = u_j$, otherwise $u_{jk} = u_j \sqrt{\nu_j / \chi^2_{\nu_j}}$ where $\chi^2_{\nu_j}$ denotes a value drawn from a chi-squared distribution with ν_j degrees of freedom, for $j = 1, \ldots, n$;

(d) Compute the DerSimonian-Laird consensus value μ_k corresponding to the triplets $(x_{1k}, u_{1k}, \nu_1), \ldots, (x_{nk}, u_{nk}, \nu_n)$.

The standard uncertainty associated with the DerSimonian-Laird consensus value is the standard deviation of the $\{\mu_k\}$, and one (among many alternatives) 95 % coverage interval for μ ranges from the 2.5th to the 97.5th percentile of the $\{\mu_k\}$.

Alternatively, a Bayesian treatment (§A.3) can also remedy the defect that the conventional DerSimonian-Laird uncertainty evaluation suffers from: not recognizing the uncertainty surrounding the estimate of τ. The Bayesian treatment offers additional advantages that will become apparent in §A.3 and §A.5.

A.3 Hierarchical Bayesian

The distinctive traits of a Bayesian treatment are these: (i) all quantities whose values are unknown are modeled as non-observable random variables; (ii) data are modeled as observed values of random variables; (iii) estimates and uncertainty evaluations for unknown values are derived from the conditional probability distribution of the unknowns given the data (the so-called *posterior distribution*) computed by application of Bayes rule (Gelman et al., 2013).

Enacting (i) and (ii) involves specifying probability distributions for all the quantities in play (unknowns as well as data), and (iii) typically involves Markov Chain Monte Carlo sampling to produce an arbitrarily large sample from the posterior distribution, standing as a proxy for its analytical characterization (which is impracticable in most cases) (Gelman et al., 2013). The NICOB uses the implementation of MCMC in JAGS (Plummer, 2015), via R package `R2jags` (Su and Yajima, 2015).

The distributions selected for the Bayesian analysis are these:

- μ has a prior Gaussian distribution with mean 0 and a very large standard deviation (10^5);
- τ and the $\{\sigma_j\}$ have prior half-Cauchy distributions as suggested by Gelman (2006) and further supported by Polson and Scott (2012). We have chosen the default values of the medians of these prior distributions as follows: for τ, equal to the median of the absolute values of the differences between the measured values and their median; and for the $\{\sigma_j\}$, equal to the median of the $\{u_j\}$. The user has the freedom to change both these values;
- Given τ, the $\{\lambda_j\}$ are Gaussian with mean 0 and standard deviation τ;
- Given μ, $\{\lambda_j\}$, and $\{\sigma_j\}$, the measured values $\{x_j\}$ are modeled as outcomes of Gaussian random variables with means $\{\mu + \lambda_j\}$ and standard deviations $\{\sigma_j\}$;

- When the standard uncertainties associated with the measured values are based on finitely many numbers of degrees of freedom $\{\nu_j\}$, then $\{\nu_j u_j^2/\sigma_j^2\}$ are modeled as outcomes of chi-squared random variables with $\{\nu_j\}$ degrees of freedom; when they are based on infinitely many numbers of degrees of freedom (that is, are regarded as known), $\sigma_j = u_j$.

The estimate of the consensus value μ is the mean of the corresponding posterior distribution, and the associated standard uncertainty $u(\mu)$ is the standard deviation of the same distribution. Since this distribution is not derived analytically, and instead we base our inferences on the sample that MCMC draws from it, the consensus value is the average of this sample, and $u(\mu)$ is its standard deviation.

The NICOB verifies that the MCMC sampling process has reached equilibrium by applying the convergence diagnostic test suggested by Geweke (1998) to the samples drawn from the distribution of all of the unknown quantities. If the NICOB concludes that equilibrium has not been reached, then it issues a message inviting the user to re-run the analysis using a larger MCMC sample, with suggested values for the new sample size, number of initial values to discard, and the subsampling (thinning) rate.

A.4 Linear Pool

The Linear Pool was suggested by Stone (1961) to aggregate the opinions or states of knowledge of several experts on a particular matter expressed as probability distributions, thereby producing a consensus. However, Bacharach (1979) attributes the idea to Pierre Simon, Marquis de Laplace.

In our case, the "experts" are the laboratories or measurement methods involved in an inter-comparison, and their opinions or states of knowledge are expressed in the form of probability distributions (whose means are the measured values $\{x_j\}$ and whose standard deviations are the associated standard uncertainties $\{u_j\}$). The Linear Pool produces a sample from a mixture of these probability distributions, which may then be suitably summarized to produce a consensus value and an evaluation of the associated uncertainty.

The distributions aforementioned are taken to be either Gaussian (when the number of degrees of freedom is infinite or unspecified), or re-scaled and shifted Student's t distributions (when the number of degrees of freedom is finite), in both cases with means and standard deviations as described above.

The NICOB gives the user the option of specifying weights for the different measurement results: the default is to weigh all the results equally. The weights represent the quality or reliability of the participating laboratories or methods, as perceived by the person performing the aggregation. However, this does not imply that the measured values themselves will end-up being equally weighed, for they are also weighed according to the reported measurement uncertainties.

If $\{\phi_j\}$ denote the probability densities of the distributions assigned to the participants as described above (Gaussian or re-scaled and shifted Student's t), and $\{w_j\}$ denote the corresponding weights (non-negative, which the NICOB normalizes to sum to 1), then the mixture distribution has probability density $f = \sum_{j=1}^{n} w_j \phi_j$, where n denotes the number of participants.

The average of the sample drawn from this mixture distribution is the consensus value, and its standard deviation is the corresponding standard uncertainty. Toman (2007, Equation (20)) provides an analytical expression for this standard uncertainty for a common implementation of the Linear Pool. Coverage intervals are built by selecting suitable percentiles from this sample, for example the 2.5th and the 97.5th for a 95 % coverage interval (which generally need not be centered at the consensus value).

The Linear Pool is but one of several ways in which the opinions of multiple experts, expressed as probability distributions, may be merged into a consensus distribution. Clemen and Winkler (2007) review some of the alternatives, and detail and compare the underlying assumptions and their properties.

The consensus distribution obtained by the Linear Pool may be multimodal (meaning that the mixture density f mentioned above may have multiple peaks): in such cases its mean and standard deviation may be poor indications of its typical value and spread (for example, as illustrated in Exhibit 25 on Page 56). To facilitate a critical evaluation of the fitness-for-purpose of the mean and standard deviation of the distribution, the NICOB also depicts f graphically.

A.5 Degrees of Equivalence

The NICOB follows guidance from Jones and Spiegelhalter (2011) about how to identify participants with "unusual" results in an interlaboratory study, in the sense that their measured values lie "beyond the range allowed by the model", and implements their Approach 2 to *Identify Outliers to the Random Effects Distribution*. Since the MRA does not specify how the expanded uncertainties that

are part of the degrees of equivalence ought to be computed, not only is our choice consistent with the MRA, it also reflects the state-of-the-art while upholding a measure of circumspection that we believe to be appropriate when flagging results as "unusual."

The perspective in this endeavor is one of testing, rather than of estimation: for the unilateral DoE, the goal is to identify measured values that, as Jones and Spiegelhalter (2011) put it, "lie beyond the range allowed by the model", and that effectively are outliers relative to the random effects distribution. Both Bayesian and sampling-theoretic approaches lead to the same criterion to identify significant discrepancies.

In §3.2.8 we introduced an alternative computation of the unilateral DoE for laboratory j as $D_j^* = x_j - \hat{\mu}_{-j}$, where $\hat{\mu}_{-j}$ denotes an estimate of the consensus value derived from the measurement results produced by all the participants but leaving-out the results from participant j, for $j = 1, \ldots, n$.

We also noted that Viechtbauer and Cheung (2010) and Duewer et al. (2014), possibly among others, have used this idea previously. The NICOB offers the user the possibility of computing DoEs using the conventional version, as defined in the MRA, or according to the *leave-one-out* strategy just described, both for unilateral and bilateral DoEs corresponding to the three procedures available.

A.5.1 DerSimonian-Laird

For the DerSimonian-Laird procedure, the conventional version of the unilateral DoE is $D_j = x_j - \hat{\mu}$ and the bilateral DoE is $B_{ij} = D_i - D_j$ for $i, j = 1, \ldots, n$. Both the $\{U_{95\,\%}(D_j)\}$ and the $\{U_{95\,\%}(B_{ij})\}$ are evaluated using the parametric statistical bootstrap, from raw materials computed by repeating the following steps a large number K of times, for $k = 1, \ldots, K$:

(a) Draw τ_k from the approximate sampling distribution of $\hat{\tau}_{\mathrm{DL}}$ described in §A.2;

(b) Draw $x_{j,k}$ from a Gaussian distribution with mean $\hat{\mu}$ and variance $\tau_k^2 + u_j^2$, for $j = 1, \ldots, n$;

(c) If ν_j is either infinity or unspecified, $u_{j,k} = u_j$, otherwise $u_{j,k} = u_j\sqrt{\nu_j/\chi^2_{\nu_j}}$ where $\chi^2_{\nu_j}$ denotes a value drawn from a chi-squared distribution with ν_j degrees of freedom, for $j = 1, \ldots, n$;

(d) Compute the DerSimonian-Laird consensus value μ_k corresponding to the triplets $(x_{1k}, u_{1k}, \nu_1), \ldots, (x_{nk}, u_{nk}, \nu_n)$;

(e) Compute $D_{j,k} = x_{j,k} - \mu_k$, for $j = 1, \ldots, n$.

Compute $U_{95\,\%}(D_j)$ as one half of the length of the shortest interval centered at the average of $D_{j,1}, \ldots, D_{j,K}$ and that includes 95 % of these $\{D_{j,k}\}$. The value of $U_{95\,\%}(B_{ij})$ is computed similarly, based on $B_{ij,k} = D_{i,k} - D_{j,k}$ for $i, j = 1, \ldots, n$.

For the leave-one-out version, $D^*_j = x_j - \hat{\mu}_{-j}$ and the bilateral DoE is $B^*_{ij} = D^*_i - D^*_j$ for $i, j = 1, \ldots, n$. Instead of performing a parametric bootstrap evaluation of $u(\hat{\mu}_{-j})$, we simply draw K samples from the Student's t approximation suggested by Knapp and Hartung (2003). The deviations that are used for the evaluation of the expanded uncertainty associated with D^*_j are of the form $D^*_{j,k} = x_j + e_{j,k} - \mu_{-j,k}$, where $e_{j,k}$ is an outcome of either a Student's t or a Gaussian distribution with mean 0 and variance $\tau^2_{-j,k} + u^2_j$ (depending on whether degrees of freedom have, or have not been specified as inputs to the NICOB), and $\tau^2_{-j,k}$ is drawn from the approximate distribution mentioned in §A.2.

A.5.2 Hierarchical Bayesian

The Bayesian approach that Jones and Spiegelhalter (2011) describe is based on the posterior predictive distribution for measured values, whose probability density is g such that

$$g(\xi_j | x_1, \ldots, x_n) = \iiint \phi(\xi_j | \mu, \tau^2 + \sigma_j^2)\, q(\mu, \tau, \sigma_j | x_1, u_1, v_1, \ldots, x_n, u_n, v_n)\, \mathrm{d}\mu\, \mathrm{d}\tau\, \mathrm{d}\sigma_j,$$

where ξ_j denotes a prediction for a value that laboratory j may measure, $\phi(\cdot|\mu, \tau^2 + \sigma_j^2)$ denotes the probability density of a Gaussian distribution with mean μ and variance $\tau^2 + \sigma_j^2$, and q denotes the probability density of the joint posterior distribution of μ, τ, and σ_j given the measurement results.

The unilateral degree of equivalence for laboratory $j = 1, \ldots, n$ comprises $D_j = x_j - \hat{\mu}$, where $\hat{\mu}$ denotes the average of an MCMC sample drawn from the posterior distribution of μ, and $U_{95\,\%}(D_j)$, which is derived from a sample $\xi_{j,1}, \ldots, \xi_{j,K}$ drawn from the aforementioned predictive distribution by repeating the following steps a large number (K) of times, for $k = 1, \ldots, K$:

(a) Draw μ_k, τ_k, and $\sigma_{j,k}$ from the corresponding posterior distributions via MCMC sampling;

(b) Draw $\xi_{j,k}$ from a Gaussian distribution with mean μ_k and variance $\tau_k^2 + \sigma_{j,k}^2$;

(c) Compute $D_{j,k} = x_j - \xi_{j,k}$.

The value of $U_{95\,\%}(D_j)$ is one half of the length of the shortest interval that is symmetrical around the average of the $\{D_{j,k}\}$ and includes 95 % of them.

Under the leave-one-out approach, we leave out the results from participant j, for $j = 1, \ldots, n$. This means there is no posterior distribution for σ_j, so in the leave-one-out versions of the unilateral and bilateral DoE we modify the algorithm. For k in $1, \ldots, K$, $\mu_{-j,k}$ and $\tau_{-j,k}$ are drawn from the corresponding posterior distributions, via MCMC sampling. Similar to the DerSimonian-Laird procedure, $D^*_{j,k} = x_j + e_{j,k} - \mu_{-j,k}$ where $e_{j,k}$ are drawn from either a Student's t or a Gaussian distribution, depending on whether the degrees of freedom have been specified, with mean 0 and variance $\tau^2_{-j,k} + u_j^2$.

A.5.3 Linear Pool

For the Linear Pool procedure, the conventional version of the unilateral DoE is $D_j = x_j - \hat{\mu}$, where $\hat{\mu}$ is the mean of the sample drawn from the mixture distribution, described in Section A.4. Again, the bilateral DoE is $B_{ij} = D_i - D_j$ for $i, j = 1, \ldots, n$. The $\{U_{95\,\%}(D_j)\}$ and $\{U_{95\,\%}(B_{ij})\}$ are evaluated using $D_{j,k} = x_j + e_{j,k} - \hat{\mu}$, where $e_{j,k}$ is drawn from a Student's t (or Gaussian) distribution with mean 0 and variance u_j^2.

For the leave-one-out version, let $\{\tilde{x}_{-j,k}\}$ for $k = 1, \ldots, K$ denote the sample of size K produced when the Linear Pool is applied to all the measurements excluding those from laboratory j for each $j = 1, \ldots, n$, and define $\hat{\mu}_{-j}$ as their average. The unilateral DoE have $D^*_j = x_j - \hat{\mu}_{-j}$, and the bilateral DoE have $B^*_{ij} = D^*_i - D^*_j$ for $i, j = 1, \ldots, n$. The $\{U_{95\,\%}(D^*_j)\}$ and $\{U_{95\,\%}(B^*_{ij})\}$ are computed similarly to how they are evaluated in the other two leave-one-out procedures, based on $\{D^*_{j,k}\}$ and $\{B^*_{ij,k}\}$ such that $D^*_{j,k} = x_j + e_{j,k} - \tilde{x}_{-j,k}$ and $B^*_{ij,k} = D^*_{i,k} - D^*_{j,k}$. Here $e_{j,k}$ are drawn from either a Student's t or a Gaussian distribution, depending on whether the degrees of freedom have been specified, with mean 0 and variance u_j^2.

Acknowledgments

The authors are greatly indebted to their NIST colleagues David Duewer, Adam Pintar, Andrew Rukhin, and Jolene Splett, who generously spent much time and effort reviewing and preparing a large and most helpful collection of suggestions for improving a draft of this contribution. David Duewer, in addition, did an extensive hands-on evaluation of a prototype of the NICOB, and offered such

a wealth of suggestions for improvement that we will continue to incorporate them into future versions of the NICOB. He also uncovered many errors and inconsistencies that we have corrected to the best of our ability.

References

A. Alink, M.J.T. Milton, F. Guenther, E.W.B. de Leer, H.J. Heine, A. Marschal, G. S. Heo, C. Takahashi, W. L. Zhen, Y. Kustikov, and E. Deak. Final report of Key Comparison CCQM-K1. Technical report, Nederlands Meetinstituut, Van Swinden Laboratory, Delft, The Netherlands, January 1999. URL kcdb.bipm.org/appendixB/appbresults/ccqm-k1.b/ccqm-k1_final_report.pdf. Corrected version, September 2001.

M. Bacharach. Normal Bayesian dialogues. *Journal of the American Statistical Association*, 74(368):837–846, December 1979. doi: 10.1080/01621459.1979.10481039.

T. Benaglia, D. Chauveau, D. R. Hunter, and D. Young. mixtools: An R package for analyzing finite mixture models. *Journal of Statistical Software*, 32(6):1–29, 2009. URL www.jstatsoft.org/v32/i06/.

M. Benková, S. Makovnik, B. Mickan, R. Arias, K. Chahine, T. Funaki, C. Li, H. M. Choi, D. Seredyuk, C.-M. Su, C. Windenberg, and J. Wright. CIPM key comparison CCM.FF-K6.2011: Comparison of the primary (national) standards of low-pressure gas flow — final report. Technical report, Consultative Committee for Mass and Related Quantities, Bureau International des Poids et Mesures, Sèvres, France, February 2014. URL kcdb.bipm.org/appendixB/appbresults/ccm.ff-k6/ccm.ff-k6_final_report.pdf.

W. Bich. From errors to probability density functions. evolution of the concept of measurement uncertainty. *IEEE Transactions on Instrumentation and Measurement*, 61(8): 2153–2159, August 2012. doi: 10.1109/TIM.2012.2193696.

B. J. Biggerstaff and D. Jackson. The exact distribution of Cochran's heterogeneity statistic in one-way random effects meta-analysis. *Statistics in Medicine*, 27:6093–6110, 2008. doi: 10.1002/sim.3428.

B. J. Biggerstaff and R. L. Tweedie. Incorporating variability in estimates of heterogeneity in the random effects model in meta-analysis. *Statistics in Medicine*, 16:753–768, 1997. doi: 10.1002/(SICI)1097-0258(19970415)16:7<753::AID-SIM494>3.0.CO;2-G.

BIPM. *The International System of Units (SI)*. International Bureau of Weights and Measures (BIPM), Sèvres, France, 8th edition, 2006.

O. Bodnar, A. Link, and C. Elster. Objective Bayesian inference for a generalized marginal random effects model. *Bayesian Analysis*, 11(1):25–45, March 2016. doi: 10.1214/14-BA933.

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.

G. E. P. Box and D. R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society, Series B (Methodological)*, 26(2):211–252, 1964.

G. E. P. Box and G. C. Tiao. *Bayesian Inference in Statistical Analysis*. Addison-Wesley, Reading, Massachusetts, 1973.

K.P. Burnham and D.R. Anderson. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. Springer-Verlag, New York, NY, 2nd edition, 2002.

H. Cavendish. Experiments to determine the density of the earth. by Henry Cavendish, Esq. F. R. S. and A. S. *Philosophical Transactions of the Royal Society of London*, 88: 469–526, 1798. doi: 10.1098/rstl.1798.0022.

W. Chang, J. Cheng, J. J. Allaire, Y. Xie, and J. McPherson. *shiny: Web Application Framework for R*, 2016. URL https://CRAN.R-project.org/package=shiny. R package version 0.13.2.

Y. Chen, Y. Cai, C. Hong, and D. Jackson. Inference for correlated effect sizes using multiple univariate meta-analyses. *Statistics in Medicine*, 35(9):1405–1422, 2016. doi: 10.1002/sim.6789.

R. T. Clemen and R. L. Winkler. Combining probability distributions from experts in risk analysis. *Risk Analysis*, 19:187–203, 1999.

R. T. Clemen and R. L. Winkler. Aggregating probability distributions. In W. Edwards, R. F. Miles Jr., and D. von Winterfeldt, editors, *Advances in Decision Analysis: From Foundations to Applications*, chapter 9, pages 154–176. Cambridge University Press, Cambridge, UK, 2007. ISBN 978-0-521-68230-5.

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

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.

Comité International des Poids et Mesures (CIPM). *Measurement comparisons in the CIPM MRA*. Bureau International des Poids et Mesures (BIPM), Pavillon de Breteuil, Sèvres, France, March 2014. URL www.bipm.org/en/cipm-mra/cipm-mra-documents/. CIPM MRA-D-05, Version 1.5.

Consultative Committee for Amount of Substance. CCQM guidance note: Estimation of a consensus KCRV and associated degrees of equivalence. Technical report, Bureau International des Poids et Mesures (BIPM), Sèvres, France, April 12th 2013. URL www.bipm.org/cc/CCQM/Allowed/19/CCQM13-22_Consensus_KCRV_v10.pdf. Version 10.

H. Cooper, L. V. Hedges, and J. C. Valentine, editors. *The Handbook of Research Synthesis and Meta-Analysis*. Russell Sage Foundation Publications, New York, NY, 2nd edition, 2009.

M. G. Cox. The evaluation of key comparison data. *Metrologia*, 39:589–595, 2002. doi: 10.1088/0026-1394/39/6/10.

M. G. Cox. The evaluation of key comparison data: determining the largest consistent subset. *Metrologia*, 44:187–200, 2007. doi: 10.1088/0026-1394/44/3/005.

M. Crowder. Interlaboratory comparisons: Round robins with random effects. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 41:409–425, 1992. doi: 10.2307/2347571.

M. H. DeGroot and M. J. Schervish. *Probability and Statistics*. Addison-Wesley, 4th edition, 2011.

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.

D. L. Duewer. A robust approach for the determination of CCQM key comparison reference values and uncertainties. Technical report, Consultative Committee for Amount of Substance: Metrology in Chemistry (CCQM), International Bureau of Weights and Measures (BIPM), Sèvres, France, 2004. URL www.bipm.info/cc/CCQM/Allowed/10/CCQM04-15.pdf. 9th Annual Meeting, Working Document CCQM/04-15.

D. L. Duewer, K. W. Pratt, C. Cherdchu, N. Tangpaisarnkul, A. Hioki, M. Ohata, P. Spitzer, M. Máriássy, and L. Vyskočil. "Degrees of equivalence" for chemical measurement capabilities: primary pH. *Accreditation and Quality Assurance*, 19:329–342, 2014. doi: 10.1007/s00769-014-1076-1.

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

C. Elster and B. Toman. Analysis of key comparisons: estimating laboratories' biases by a fixed effects model using Bayesian model averaging. *Metrologia*, 47:113–119, 2010.

J. D. Emerson. Introduction to transformation. In D. C. Hoaglin, F. Mosteller, and J. W. Tukey, editors, *Fundamentals of Exploratory Analysis of Variance*. John Wiley & Sons, New York, NY, 1991.

D. Enko, G. Kriegshäuser, R. Stolba, E. Worf, and G. Halwachs-Baumann. Method evaluation study of a new generation of vitamin D assays. *Biochemia Medica*, 25(2):203–212, 2015. doi: 10.11613/BM.2015.020.

C.-J. L. Farrell, S. Martin, B. McWhinney, I. Straub, P. Williams, and M. Herrmann. State-of-the-art vitamin D assays: A comparison of automated immunoassays with liquid chromatography-tandem mass spectrometry methods. *Clinical Chemistry*, 58(3):531–542, 2012. doi: 10.1373/clinchem.2011.172155.

A. Gasparrini. *mvmeta: multivariate meta-analysis and meta-regression*, 2012. R package version 0.2.4.

A. Gelman. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, 1(3):515–533, 2006. doi: doi: 10.1214/06-BA117A.

A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin. *Bayesian Data Analysis*. Chapman & Hall / CRC, Boca Raton, FL, 3rd edition, 2013.

C. Genest, S. Weerahandi, and J. V. Zidek. Aggregating opinions through logarithmic pooling. *Theory and Decision*, 17:61–70, 1984. doi: 10.1007/BF00140056.

J. Geweke. Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In J. M. Bernado, J. O. Berger, A. P. Dawid, and A. F. M. Smith, editors, *Bayesian Statistics 4: Proceedings of the Fourth Valencia International Meeting*. Clarendon Press, Oxford, UK, 1998.

A. Guolo and C. Varin. Random-effects meta-analysis: the number of studies matters. *Statistical Methods in Medical Research*, 2015. doi: 10.1177/0962280215583568. Published online before print May 7, 2015.

L. V. Hedges and I. Olkin. *Statistical Methods for Meta-Analysis*. Academic Press, San Diego, CA, 1985.

J. P. T. Higgins and S. G. Thompson. Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, 21:1539–1558, 2002. doi: 10.1002/sim.1186.

J. P. T. Higgins, S. G. Thompson, J. J. Deeks, and D. G. Altman. Measuring inconsistency in meta-analyses. *British Medical Journal*, 327(7414):557–560, September 2003. doi: 10.1136/bmj.327.7414.557.

J. P. T. Higgins, S. G. Thompson, and D. J. Spiegelhalter. A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society, Series A (Statistics in Society)*, 172 (1):137–159, January 2009.

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. Horwitz. Evaluation of analytical methods used for regulation of foods and drugs. *Analytical Chemistry*, 54(1):67A–76A, 1982. doi: 10.1021/ac00238a765.

W. Horwitz. The certainty of uncertainty. *Journal of AOAC International*, 86(1):109–111, 2003.

J. E. Hunter, F. L. Schmidt, and G. B. Jackson. *Meta-analysis: cumulating research findings across studies*. Sage, Beverly Hills, CA, 1992.

H. K. Iyer, C. M. J. Wang, and T. Mathew. Models and confidence intervals for true values in interlaboratory trials. *Journal of the American Statistical Association*, 99(468):1060–1071, December 2004. doi: 10.1198/016214504000001682.

D. Jackson, J. Bowden, and R. Baker. How does the dersimonian and laird procedure for random effects meta-analysis compare with its more efficient but harder to compute counterparts? *Journal of Statistical Planning and Inference*, 140(4):961–970, 2010. doi: 10.1016/j.jspi.2009.09.017.

H. Jeffreys. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London*, 186(1007):453–461, 1946.

Joint Committee for Guides in Metrology. *Evaluation of measurement data — Guide to the expression of uncertainty in measurement*. International Bureau of Weights and Measures (BIPM), Sèvres, France, 2008a. URL www.bipm.org/en/publications/guides/gum.html. BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP and OIML, JCGM 100:2008, GUM 1995 with minor corrections.

Joint Committee for Guides in Metrology. *Evaluation of measurement data — Supplement 1 to the "Guide to the expression of uncertainty in measurement" — Propagation of distributions using a Monte Carlo method*. International Bureau of Weights and Measures

(BIPM), Sèvres, France, 2008b. URL www.bipm.org/en/publications/guides/gum.html. BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP and OIML, JCGM 101:2008.

H. E. Jones and D. J. Spiegelhalter. The identification of "unusual" health-care providers from a hierarchical model. *The American Statistician*, 65(3):154–163, 2011. doi: 10.1198/tast.2011.10190.

R. Judaschke. CCEM key comparison CCEM.RF-K25.W, RF power from 33 GHz to 50 GHz in waveguide, Final Report of the Pilot Laboratory. Technical report, Physikalisch-Technische Bundesanstalt, Braunschweig, Germany, October 2014. URL www.bipm.org/utils/common/pdf/final_reports/EM/RF-K25/CCEM.RF-K25.W.pdf.

R. Judaschke. CCEM Key comparison CCEM.RF-K25.W. RF power from 33 ghz to 50 ghz in waveguide. final report of the pilot laboratory. *Metrologia*, 52(1A):01001, 2015. doi: 10.1088/0026-1394/52/1A/01001.

R. Kacker, R. Kessel, and J. F. Lawrence. Removing divergence of JCGM documents from the GUM (1993) and repairing other defects. *Measurement*, 88:194–201, 2016.

G. Knapp and J. Hartung. Improved tests for a random effects meta-regression with a single covariate. *Statistics in Medicine*, 22:2693–2710, 2003. doi: 10.1002/sim.1482.

P. C. Lambert, A. J. Sutton, P. R. Burton, K. R. Abrams, and D. R. Jones. How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. *Statistics in Medicine*, 24:2401–2428, 2005. doi: 10.1002/sim.2112.

D. V. Lindley. Reconciliation of probability distributions. *Operations Research*, 31(5):866–880, September-October 1983.

T. A. MacKenzie, G. L. Grunkemeier, G. K. Grunwald, A. J. O'Malley, C. Bohn, Y. X. Wu, and D. J. Malenka. A primer on using shrinkage to compare in-hospital mortality between centers. *The Annals of Thoracic Surgery*, 99:757–761, March 2015. doi: 10.1016/j.athoracsur.2014.11.039.

J. Mandel. *The Statistical Analysis of Experimental Data*. Interscience Publishers (John Wiley & Sons), New York, NY, 1964.

J. Mandel. The validation of measurement through interlaboratory studies. *Chemometrics and Intelligent Laboratory Systems*, 11(2):109–119, 1991. doi: 10.1016/0169-7439(91)80058-X.

J. Mandel and R. Paule. Interlaboratory evaluation of a material with unequal numbers of replicates. *Analytical Chemistry*, 42(11):1194–1197, September 1970. doi: 10.1021/ac60293a019.

G. McLachlan and D. Peel. *Finite Mixture Models*. John Wiley & Sons, New York, NY, 2000.

P. Meier, G. Knapp, U Tamhane, S Chaturvedi, and H. S. Gurm. Short term and intermediate term comparison of endarterectomy versus stenting for carotid artery stenosis: systematic review and meta-analysis of randomised controlled clinical trials. *BMJ*, 340:c467, 2010. doi: 10.1136/bmj.c467.

C. Michotte. Efficiency curve of the ionization chamber of the SIR. *Applied Radiation and Isotopes*, 56:15–20, 2002.

C. Michotte, S. Courte, G. Ratel, M. Sahagia, A. C. Wätjen, R. Fitzgerald, and F.-J. Maringer. Update of the ongoing comparison BIPM.RI(II)-K1.Co-60 including activity measurements of the radionuclide ^{60}Co for the IFIN-HH (Romania), NIST (USA) and the BEV (Austria). *Metrologia*, 47(1A):06010, 2010.

R. G. Miller. *Beyond ANOVA, Basics of Applied Statistics*. John Wiley & Sons, New York, NY, 1986.

F. Mosteller and J. W. Tukey. *Data Analysis and Regression*. Addison-Wesley Publishing Company, Reading, Massachusetts, 1977.

R. B. Nelsen. *An Introduction to Copulas*. Lecture Notes in Statistics, 139. Springer, New York, NY, 1999.

H. G. Nubbemeyer and J. Fischer. Report to the CCT on Key Comparison 4: Comparison of Local Realisations of Aluminium and Silver Freezing-Point Temperatures. Technical report, Physikalisch-Technische Bundesanstalt, Germany, January 29th 2002.

A. O'Hagan. Eliciting and using expert knowledge in metrology. *Metrologia*, 51(4):S237–S244, 2014. doi: 10.1088/0026-1394/51/4/S237.

M. Plummer. *JAGS Version 4.0.0 user manual*, October 2015. URL mcmc-jags.sourceforge.net/.

N. G. Polson and J. G. Scott. On the half-Cauchy prior for a global scale parameter. *Bayesian Analysis*, 7(4):887–902, 2012. doi: 10.1214/12-BA730.

A. Possolo. Copulas for uncertainty analysis. *Metrologia*, 47:262–271, 2010. doi: 10.1088/0026-1394/47/3/017.

A. Possolo. Five examples of assessment and expression of measurement uncertainty. *Applied Stochastic Models in Business and Industry*, 29:1–18, January/February 2013. doi: 10.1002/asmb.1947. Discussion and Rejoinder pp. 19–30.

A. Possolo. *Simple Guide for Evaluating and Expressing the Uncertainty of NIST Measurement Results*. NIST Technical Note 1900. National Institute of Standards and Technology, Gaithersburg, MD, 2015. doi: 10.6028/NIST.TN.1900.

A. Possolo and B. Toman. *Tutorial for metrologists on the probabilistic and statistical apparatus underlying the GUM and related documents*. National Institute of Standards and Technology, Gaithersburg, MD, November 2011. doi: 10.13140/RG.2.1.2256.8482. URL www.itl.nist.gov/div898/possolo/TutorialWEBServer/TutorialMetrologists2011Nov09.xht.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. URL www.R-project.org/.

J. Randa. Update to Proposal for KCRV and Degree of Equivalence for GTRF Key Comparisons. National Institute of Standards and Technology, Boulder, CO, February 2005. Consultative Committee for Electricity and Magnetism, Working Group on Radiofrequency Quantities (GT-RF).

G. Ratel and C. Michotte. BIPM comparison BIPM.RI(II)-K1.Co-60 of the activity measurements of the radionuclide ^{60}Co. *Metrologia*, 40(1A):06007, 2003.

G. Ratel, C. Michotte, R. Broda, and A. Listkowska. Activity measurements of the radionuclide ^{60}Co for the RC, Poland in the ongoing comparison BIPM.RI(II)-K1.Co-60. *Metrologia*, 40(1A):06033, 2003a.

G. Ratel, C. Michotte, B. R. S. Simpson, and A. Iglicki. Activity measurements of the radionuclide ^{60}Co for the CSIR-NML and the CNEA in the BIPM comparison BIPM.RI(II)-K1.Co-60. *Metrologia*, 40(1A):06010, 2003b.

G. Ratel, C. Michotte, Y. Hino, J. Keightley, and U. Wätjen. Update of the ongoing comparison BIPM.RI(II)-K1.Co-60 including activity measurements of the radionuclide ^{60}Co for the NMIJ, Japan and the IRMM (Geel). *Metrologia*, 43(1A):06003, 2006.

D. M. Rocke. Robust statistical analysis of interlaboratory studies. *Biometrika*, 70(2): 421–431, 1983. doi: 10.1093/biomet/70.2.421.

G. Rosi, F. Sorrentino, L. Cacciapuoti, M. Prevedelli, and G. M. Tino. Precision measurement of the newtonian gravitational constant using cold atoms. *Nature*, 510:518–521, June 26 2014. doi: 10.1038/nature13433.

A. L. Rukhin. Weighted means statistics in interlaboratory studies. *Metrologia*, 46(3): 323–331, 2009. doi: 10.1088/0026-1394/46/3/021.

A. L. Rukhin. Estimating heterogeneity variance in meta-analysis. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 75(3):451–469, 2013. doi: 10.1111/j.1467-9868.2012.01047.x.

A. L. Rukhin and A. Possolo. Laplace random effects models for interlaboratory studies. *Computational Statistics and Data Analysis*, 55:1815–1827, 2011. doi: 10.1016/j.csda.2010.11.016.

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. R. Searle, G. Casella, and C. E. McCulloch. *Variance Components*. John Wiley & Sons, New York, NY, 1992.

S. R. Searle, G. Casella, and C. E. McCulloch. *Variance Components*. John Wiley & Sons, Hoboken, NJ, 2006. ISBN 0-470-00959-4.

C. Speake and T. Quinn. The search for newton's constant. *Physics Today*, 67(7):27–33, 2014. doi: 10.1063/PT.3.2447.

A. G. Steele, K. D. Hill, and R. J. Douglas. Data pooling and key comparison reference values. *Metrologia*, 39(3):269–277, 2002.

M. Steup. Epistemology. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. The Metaphysics Research Lab, Center for the Study of Language and Information, Stanford University, Stanford, California, spring 2014 edition, 2014. URL plato.stanford.edu/archives/spr2014/entries/epistemology/.

M. Stock. The watt balance: determination of the planck constant and redefinition of the kilogram. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 369(1953):3936–3953, 2011. doi: 10.1098/rsta.2011.0184.

M. Stock, S. Solve, D. del Campo, V. Chimenti, E. Méndez-Lango, H. Liedberg, P.P.M. Steur, P. Marcarino, R. Dematteis, E. Filipe, I. Lobo, K.H. Kang, K.S. Gam, Y.-G. Kim, E. Renaot, G. Bonnier, M. Valin, R. White, T. D. Dransfield, Y. Duan, Y. Xiaoke, G. Strouse, M. Ballico, D. Sukkar, M. Arai, A. Mans, M. de Groot, O. Kerkhof, R. Rusby, J. Gray, D. Head, K. Hill, E. Tegeler, U. Noatsch, S. Duris, H.Y. Kho, S. Ugur, A. Pokhodun, and S.F. Gerasimov. Final report on CCT-K7 — key comparison of water triple point cells. Technical report, International Bureau of Weights and Measures (BIPM), Sèvres, France, January 2006. URL kcdb.bipm.org/appendixB/appbresults/cct-k7/cct-k7_final_report.pdf.

M. Stone. The opinion pool. *The Annals of Mathematical Statistics*, 32:1339–1342, December 1961. doi: 10.1214/aoms/1177704873.

Y.-S. Su and M. Yajima. *R2jags: Using R to Run 'JAGS'*, 2015. URL https://CRAN.R-project.org/package=R2jags. R package version 0.5-7.

S. S.-C. Tai, M. Bedner, and K. W. Phinney. Development of a candidate reference measurement procedure for the determination of 25-hydroxyvitamin D3 and 25-hydroxyvitamin D2 in human serum using isotope-dilution liquid chromatography-tandem mass spectrometry. *Analytical Chemistry*, 82(5):1942–1948, 2010. doi: 10.1021/ac9026862.

B. N. Taylor and C. E. Kuyatt. *Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results*. NIST Technical Note 1297. National Institute of Standards and Technology, Gaithersburg, MD, 1994. URL physics.nist.gov/Pubs/guidelines/TN1297/tn1297s.pdf.

R. Thalmann. CCL Key Comparison CCL-K1: Calibration of gauge blocks by interferometry — Final report. Technical report, Swiss Federal Office of Metrology (METAS), Wabern, Switzerland, January 2001. URL kcdb.bipm.org/appendixB/appbresults/ccl-k1/ccl-k1_final_report.pdf.

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.

M. Thompson and P. J. Lowthian. *Notes on Statistics and Data Quality for Analytical Chemists*. Imperial College Press, London, UK, 2011.

M. Thompson, S. L. R. Ellison, and R. Wood. The International Harmonized Protocol for the proficiency testing of analytical chemistry laboratories (IUPAC Technical Report). *Pure and Applied Chemistry*, 78(1):145–196, 2006. doi: 10.1351/pac200678010145.

B. Toman. Bayesian approaches to calculating a reference value in key comparison experiments. *Technometrics*, 49(1):81–87, February 2007. doi: 10.1198/004017006000000273.

B. Toman and A. Possolo. Model-based uncertainty analysis in inter-laboratory studies. In F. Pavese, M. Bär, A. B. Forbes, J. M. Linares, C. Perruchet, and N. F. Zhang, editors, *Advanced Mathematical and Computational Tools in Metrology and Testing: AMCTM VIII*, volume 78 of *Series on Advances in Mathematics for Applied Sciences*, pages 330–343. World Scientific Publishing Company, Singapore, 2009a. ISBN 981-283-951-8.

B. Toman and A. Possolo. Laboratory effects models for interlaboratory comparisons. *Accreditation and Quality Assurance*, 14:553–563, October 2009b. doi: 10.1007/s00769-009-0547-2.

B. Toman and A. Possolo. Erratum to: Laboratory effects models for interlaboratory comparisons. *Accreditation and Quality Assurance*, 15:653–654, 2010. doi: 10.1007/s00769-010-0707-4.

J. W. Tukey. *Exploratory Data Analysis*. Addison-Wesley, Reading, MA, 1977.

R. M. Turner, D. Jackson, Y. Wei, S. G. Thompson, and J. P. T. Higgins. Predictive distributions for between-study heterogeneity and simple methods for their application in bayesian meta-analysis. *Statistics in Medicine*, 34:984–998, 2015.

M. G. Vangel and A. L. Rukhin. Maximum likelihood analysis for heteroscedastic one-way random effects ANOVA in interlaboratory studies. *Biometrics*, 55:129–136, March 1999. doi: 10.1111/j.0006-341X.1999.00129.x.

W. Viechtbauer. Confidence intervals for the amount of heterogeneity in meta-analysis. *Statistics in Medicine*, 26:37–52, 2007.

W. Viechtbauer. Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, 36(3):1–48, 2010. doi: 10.18637/jss.v036.i03.

W. Viechtbauer and M. W.-L. Cheung. Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods*, 1:112–125, 2010. doi: 10.1002/jrsm.11.

D. R. White. In pursuit of a fit-for-purpose uncertainty guide. *Metrologia*, 53:S107–S124, 2016. doi: 10.1088/0026-1394/53/4/S107.

A. Whitehead and J. Whitehead. A general parametric approach to the meta-analysis of randomized clinical trials. *Statistics in Medicine*, 10(11):1665–1677, 1991. doi: 10.1002/sim.4780101105.

N. F. Zhang. Statistical analysis for interlaboratory comparisons with linear trends in multiple loops. *Metrologia*, 49(3):390–394, 2012. doi: 10.1088/0026-1394/49/3/390.