Errors-in-variables calibration with dark uncertainty

IOP Publishing | Bureau International des Poids et Mesures Metrologia

Metrologia **59** (2022) 045002 (17pp) https://doi.org/10.1088/1681-7575/ac711c

Errors-in-variables calibration with dark uncertainty

Christina E Cecelski[1,*], Blaza Toman[1], Fong-Ha Liu[2], Juris Meija[3] and Antonio Possolo[1]

[1] National Institute of Standards and Technology, Gaithersburg, MD, United States of America
[2] National Measurement Institute, North Ryde, New South Wales, Australia
[3] National Research Council, Ottawa, Ontario, Canada

E-mail: christina.cecelski@nist.gov

Received 1 December 2021, revised 16 May 2022
Accepted for publication 18 May 2022
Published 16 June 2022

Abstract

A model for errors-in-variables regression is described that can be used to overcome the challenge posed by mutually inconsistent calibration data. The model and its implementation are illustrated in applications to the measurement of the amount fraction of oxygen in nitrogen from key comparison CCQM-K53, and of carbon isotope delta values in steroids from human urine. These two examples clearly demonstrate that inconsistencies in measurement results can be addressed similarly to how laboratory effects are often invoked to deal with mutually inconsistent results from interlaboratory studies involving scalar measurands. Bayesian versions of errors-in-variables regression, fitted via Markov Chain Monte Carlo sampling, are employed, which yield estimates of the key comparison reference function in one example, and of the analysis function in the other. The fitting procedures also characterize the uncertainty associated with these functions, while quantifying and propagating the 'excess' dispersion that was unrecognized in the uncertainty budgets for the individual measurements, and that therefore is missing from the reported uncertainties. We regard this 'excess' dispersion as an expression of *dark uncertainty*, which we take into account in the context of calibrations that involve regression models. In one variant of the model the estimate of dark uncertainty is the same for all the participants in the comparison, while in another variant different amounts of dark uncertainty are assigned to different participants. We compare these models with the conventional errors-in-variables model underlying the procedure that ISO 6143 recommends for building analysis functions. Applications of this procedure are often preceded by the selection of a subset of the measurement results deemed to be mutually consistent, while the more discrepant ones are set aside. This new model is more inclusive than the conventional model, in that it easily accommodates measurement results that are mutually inconsistent. It produces results that take into account contributions from all apparent sources of uncertainty, regardless of whether these sources are already understood and their contributions have been included in the reported uncertainties, or still require investigation after they will have been detected and quantified.

Keywords: dark uncertainty, inconsistent results, Bayesian model, hierarchical model, Monte Carlo method, gas mixture, testosterone

(Some figures may appear in colour only in the online journal)

* Author to whom any correspondence should be addressed.

1. Introduction

This contribution proposes a new procedure to build calibration or analysis functions when the measurement results are mutually inconsistent, and describes two applications of this procedure in measurement science: to the construction of a key comparison reference function (KCRF) that plays the role of analysis function for gas mixtures; and to the calibration of carbon isotope delta values relative to the Vienna PDB (VPDB) standard.

Ordinary least squares (OLS) is still commonly employed to build functions that relate values of measurands delivered by standards and corresponding values of instrumental indications, for example in analytical chemistry and in the measurement of force, even in situations where all the values involved are surrounded by non-negligible uncertainties. The choice is then made to fit the regression curve that minimizes the errors in the direction (either vertical or horizontal) in which the uncertainties are smallest. This may yield either the calibration function or the analysis function, and if the other is needed, then the mathematical inverse is obtained via either analytical or numerical methods.

The calibration function translates values of the measurand into instrumental indications (or uncalibrated readings), and the analysis function does the opposite. In analytical chemistry, the analysis function is needed in practice for value assignment to material samples whose values of the measurand are unknown.

1.1. Gas mixtures

The value assignment to gas mixtures typically follows the guidelines in ISO 6143 [21], which is the internationally recognized standard used for determining the composition of calibration gas mixtures.

The value assignment is made using an analysis function that translates instrumental indications into values of amount fraction of the substance of interest. The analysis function is built from data acquired during calibration, which involves obtaining indications from primary standard gas mixtures that often are prepared gravimetrically. The statistical procedure used to build the analysis function is a so-called *errors-in-variables (EIV) regression* [2].

When EIV regression is used, one is at liberty to produce the analysis function directly, regardless of which quantity has the smallest uncertainties: that is what we do in this contribution, both for value assignments to selected properties of the chemical composition of a gas mixture (section 4), and to carbon isotope delta values in a sample containing steroid metabolites (section 6).

Both in key comparisons (KCs) and in multi-point calibrations to produce a calibration or analysis function, the situation often arises where some measurement results deviate from the consensus value or trend suggested by the bulk of the others.

In some KCs organized by the Gas Analysis Working Group of the Consultative Committee for Amount of Substance (CCQM), mixtures prepared gravimetrically by the participants are measured by the coordinating (or, pilot) laboratory using an instrumental method, and the results are then used to build a KCRF, which is akin to an analysis function used for value assignment in analytical chemistry, with the participants' mixtures playing the role of standards.

This was done in CCQM-K53 [25], for example, and the results turned out to be mutually inconsistent, yet there was no substantive reason to excluded any of them from the calculation of the KCRF. ('Mutual inconsistency' means that the absolute values of the differences between the values the participants assigned to their mixtures, and the corresponding values predicted by the KCRF, were appreciably larger than their uncertainties suggested that they should have been.)

Hein *et al* [14] used the term 'KCRF' in the same sense we give to it, and in a similar application. Cox and Harris [6] use a similar term, 'key comparison reference curve' (KCRC), but its meaning is very different from the meaning of KCRF as used here. In fact, KCRC refers to the result of blending individual curves determined by different laboratories, for example, curves that describe how the sensitivity of a hydrophone varies as a function of sound frequency, while in CCQM-K53 the participants determine no curves.

1.2. Isotope delta values

The assignment of isotope delta values to metabolites of anabolic androgenic steroids, for the detection of substances prohibited by the World Anti-Doping Agency (WADA) [43], also involves first the construction of an analysis function that relates determinations of such isotope delta values in calibration standards to the certified isotope delta values of these standards, and second its application to determinations made of a human urine sample whose isotope delta value is unknown [23, 30]. In the application discussed in subsection 6.1, the measurement results for the isotope delta values of the calibration standards also are mutually inconsistent.

The dispersion of measured values above and beyond their reported uncertainties is often described as heterogeneity or mutual inconsistency, and it is widely recognized and taken into account in meta-analysis [4, 13], and also in interlaboratory studies and KCs [24, 33, 39, 41, 44]. This 'excess' dispersion can be quantified using any one of several metrics [15], and also by a parameter in the statistical measurement model [22], whose value is estimated based on the measurement results, and that we call *dark uncertainty* in the sense defined by Thompson and Ellison [40].

1.3. Outline

Section 2 reviews the two sets of measurement results used in the illustrative applications: one from CCQM-K53 [25], the other from the Chemical Reference Values section of the National Measurement Institute, Australia (NMIA).

Section 3 defines three variants of the EIV model for the measurement results from CCQM-K53, and section 4 describes Bayesian procedures for fitting these variants to the measurement results, and compares the results. Section 5 explains how the corresponding degrees of equivalence (DoEs)

Table 1. Amount fractions, x, of oxygen in nitrogen in eleven mixtures prepared gravimetrically, associated uncertainties, $u(x)$, as reported by the participants, and ratios, r, of instrumental indications relative to a common, independently prepared reference mixture, and associated standard uncertainties, $u(r)$.

LAB	$x/(\mu\text{mol/mol})$	$u(x)/(\mu\text{mol/mol})$	r	$u(r)$
NMIJ	98.675	0.009	0.983 020	0.0006
NPL	99.002	0.048	0.986 981	0.0006
BAM	99.170	0.125	0.986 377	0.0006
NMIA	100.100	0.070	1.000 875	0.0006
NIST	100.410	0.060	0.998 881	0.0006
NMISA	100.585	0.011	1.005 079	0.0006
CENAM	100.970	0.150	1.006 298	0.0006
LNE	101.040	0.050	1.004 614	0.0006
KRISS	101.053	0.010	1.006 739	0.0006
VNIIM	101.080	0.070	1.009 951	0.0006
VSL	101.150	0.040	1.007 570	0.0006

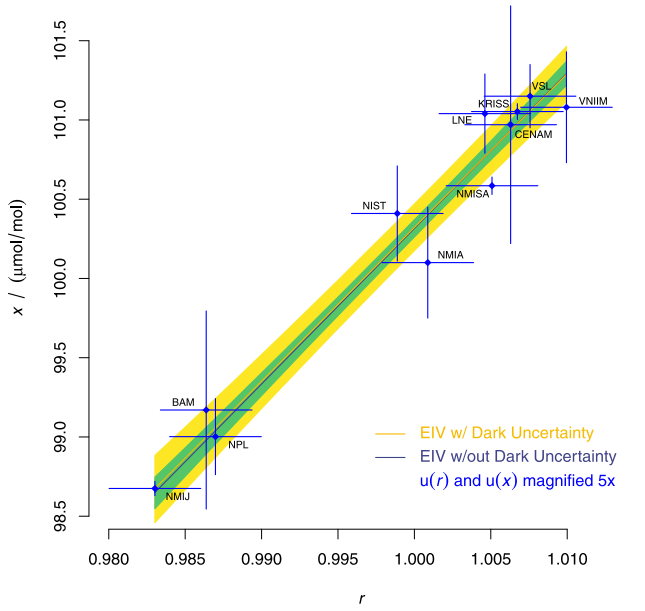

Figure 1. The (blue) diamonds represent the measured values, the horizontal line segments represent the $\{r_j \pm 5u(r_j)\}$, and the vertical line segments represent the $\{x_j \pm 5u(x_j)\}$. The light (yellow) sloping line represents the EIV regression line fitted taking dark uncertainty into account, and the wide, shaded (yellow) band is a simultaneous 95% coverage band for the whole line. The dark (purple) sloping line represents the EIV regression line fitted disregarding dark uncertainty, and the light (green) band is a simultaneous 95% coverage band for this whole line.

can be computed and discusses challenges that mutually inconsistent results pose when the EIV regression neglects the presence of dark uncertainty.

Section 6 extends the model to accommodate the presence of dark uncertainty in both variables related by the regression model, and also to take into account the fact that the reported uncertainties are based on specified, finite numbers of degrees of freedom, and then applies it to the measurements of carbon isotope delta values in metabolites of anabolic androgenic steroids. Section 7 presents some conclusions and lessons learned.

2. Measurement results

2.1. Amount fractions of oxygen in nitrogen

Table 1 transcribes the measurement results listed in table 6 of the Final Report for key comparison CCQM-K53 [25], except the results from CEM (Centro Español de Metrologia, Madrid, Spain), which CEM declared to be in error, and figure 1 depicts them. The measurand was the amount fraction of oxygen in nitrogen.

The coordinating laboratory KRISS measured all the mixtures using a gas chromatograph with a thermal conductivity detector (GC-TCD), and also an additional reference mixture before and after measuring each of the participants' mixtures. Since, in these types of comparisons, the coordinating laboratory is also a participant, its transfer standard typically is prepared by a staff member who does not become involved in any of the subsequent measurements for the key comparison.

Each value listed in the column headed r of table 1 is the ratio of the instrumental indication for a participant's mixture and the average of the two instrumental indications for the reference mixture obtained immediately before and after the indication for the participant's mixture. Both sets of standard uncertainties, $\{u(x_j)\}$ and $\{u(r_j)\}$, are assumed to be based on large numbers of degrees of freedom.

Lee *et al* [25, p 12] state that the measurement results from NMIA, NMISA, and VNIIM were excluded from the computation of the KCRF because 'they were inconsistent with the others owing to their ambiguity in impurity analysis'. For the present purposes of illustration of the model and method described in sections 3 and 4, we will use all the measurement results listed in table 1.

2.2. Carbon isotope delta values

Table 2 lists determinations of $\delta(^{13}\text{C})$ values made at NMIA, for eight isotopically different steroids in two calibration mixtures, and their assigned values relative to the VPDB standard of reference, which is a virtual material defined by NIST Reference Material 8544, NBS19 Limestone [18].

The two mixtures aforementioned were MX018-1 and MX018-3 from *NMIA MX018 Steroid Mixtures certified for Carbon Isotope Delta Value*, a certified reference material prepared by the National Measurement Institute of Australia [20]. The analysis was carried out by gas chromatography to allow separation of the different steroids in the mixtures, followed by combustion and isotope ratio mass spectrometry.

Adopting the notational conventions of Possolo *et al* [32], the isotope delta value for ^{13}C in material P relative to the standard of reference STD is

$$\delta_{\text{STD,P}}(^{13}\text{C}/^{12}\text{C}) = \frac{R_{\text{P}}(^{13}\text{C}/^{12}\text{C})}{R_{\text{STD}}(^{13}\text{C}/^{12}\text{C})} - 1, \qquad (1)$$

where $R_{\text{P}}(^{13}\text{C}/^{12}\text{C}) = N_{\text{P}}(^{13}\text{C})/N_{\text{P}}(^{12}\text{C})$ denotes the isotope ratio for ^{13}C and ^{12}C, where $N_{\text{P}}(^{13}\text{C})$ denotes the number of atoms of ^{13}C in the material, and similarly for $N_{\text{P}}(^{12}\text{C})$. Isotope

Table 2. Certified $\delta_{\mathrm{VPDB,S}}(^{13}\mathrm{C})$ values of 8 steroid calibrants (expressed in per mille relative to VPDB), and the corresponding measured values of $\delta_{\mathrm{NMIA,S}}(^{13}\mathrm{C})$ (relative to NMIA's in-house reference), reported alongside their associated standard uncertainties and numbers of degrees of freedom. The steroids are as follows: etiocholanolone (Etio); androsterone (A); 11-oxoetiocholanolone (11oxoEtio); testosterone (T_1 and T_2); 11β-hydroxyetiocholanolone (11OHEtio); 16-androstenol (16en); and dehydroepiandrosterone (DHEA).

SUBSTANCE	$\delta_{\mathrm{VPDB,S}}(^{13}\mathrm{C})$	$u(\delta_{\mathrm{VPDB,S}}(^{13}\mathrm{C}))$	ν_{V}	$\delta_{\mathrm{NMIA,S}}(^{13}\mathrm{C})$	$u(\delta_{\mathrm{NMIA,S}}(^{13}\mathrm{C}))$	ν_{N}
Etio	−27.94	0.120	41	−27.604	0.051	18
A	−27.79	0.105	15	−27.161	0.032	18
11oxoEtio	−13.58	0.115	28	−13.717	0.046	18
T_1	−27.87	0.120	24	−28.070	0.050	18
11OHEtio	−29.51	0.180	58	−29.584	0.045	18
16en	−30.96	0.185	47	−31.411	0.060	9
DHEA	−31.63	0.270	40	−31.454	0.051	9
T_2	−22.52	0.165	54	−21.972	0.044	9

delta values are conventionally expressed in *per mille* (parts per thousand). In the sequel we abbreviate them to $\delta_{\mathrm{STD,P}}(^{13}\mathrm{C})$ because they are always relative to $^{12}\mathrm{C}$.

Traceability of the $\delta(^{13}\mathrm{C})$ values to the VPDB was established via two-point linear calibration using the reference materials IAEA-CH-6 (sucrose, −10.45 ‰) and IAEA-CH-7 (polyethylene, −32.15 ‰), whose values of $\delta_{\mathrm{VPDB}}(^{13}\mathrm{C})$ bracket the $\delta(^{13}\mathrm{C})$ values in the standards listed in table 2 [10].

The fact that the same two reference materials were used to link all eight standards to the VPDB induces correlations between the calibrated isotope delta values, which, by application of a Monte Carlo method, we estimate to be all positive and in the range 0.1–0.5. If these correlations were taken into account, then, in subsection 6.2, the uncertainty associated with the isotope delta value assigned to a material sample with unknown isotopic composition would be slightly larger than what we report there. However, to avoid introducing an even more elaborate model than is described in subsection 6.1, we chose to disregard these correlations.

3. Errors-in-variables model for CCQM-K53

Since, in the case of CCQM-K53, a polynomial of the 1st degree provides an adequate model for the pairs of observations $\{(r_j, x_j)\}$, we will assume that the corresponding true values, $\{\xi_j\}$, of the amount fractions, and the corresponding true values, $\{\rho_j\}$, of the ratios of instrumental indications, are linearly related as $\xi_j = \beta_1 + \beta_2\rho_j$, for $j = 1, \ldots, n$, where $n = 11$ denotes the number of participants.

When depicting the ratios and amount fractions graphically, as in figure 1, the ratios are plotted against the horizontal axis, and the amount fractions are plotted against the vertical axis because the KCRF maps values of the ratio to values of the amount fraction of the analyte.

For this reason, we refer to the differences $\{r_j - \widehat{\rho}_j\}$ as *horizontal residuals*, and to the $\{x_j - \widehat{\xi}_j\}$ as *vertical residuals*, where the $\{\widehat{\rho}_j\}$ and the $\{\widehat{\xi}_j\}$ denote estimates of the true ratios and of the true amount fractions, respectively, obtained under the aforementioned constraint that $\xi = \beta_1 + \beta_2\rho$. Figures 2 and 5 illustrate these horizontal and vertical residuals variously, for one of the participants.

In section 4 we explain how we estimate the intercept, β_1, and the slope, β_2, of the analysis function, and also how we build a coverage region for the graph of this function, as depicted in figure 1.

We will present three variants of the EIV regression model for the results of CCQM-K53, according to whether (i) the model includes a single, common evaluation of dark uncertainty shared by all the participants (EIV-DARK), or (ii) entertains the possibility that different amounts of dark uncertainty, attributed to different participants, will suffice to achieve mutual consistency (EIV-SHADES), or (iii) no dark uncertainty is recognized at all (EIV-LIGHT). Versions (i) and (iii) are presented in subsection 3.1, and version (ii) in subsection 3.2.

Our main focus, however, will be on EIV-DARK, which has two sub-variants: one is used for the results of CCQM-K53, and involves consideration of dark uncertainty only for the amount fractions, not for the ratios of instrumental indications; the other is applied to the measurements of carbon isotope deltas in section 6, and involves consideration of dark uncertainty along both axes.

3.1. EIV model for CCQM-K53 with common dark uncertainty

The aforementioned EIV-DARK variant for the results of CCQM-K53 assumes that the measured values and corresponding true values for participant $j = 1, \ldots, n$, where $n = 11$ is the number of participants, are related as follows:

$$\xi_j = \beta_1 + \beta_2\rho_j, \qquad r_j = \rho_j + \gamma_j, \quad \text{and} \quad x_j = \xi_j + \lambda_j + \varepsilon_j. \tag{2}$$

The measurement errors for the amount fractions, $\{\varepsilon_j\}$, and for the ratios, $\{\gamma_j\}$, are assumed to be non-observable outcomes of independent Gaussian random variables, all with mean 0 and standard deviations equal to the corresponding standard uncertainties, $\{u(x_j)\}$ for the $\{\varepsilon_j\}$, and $\{u(r_j)\}$ for the $\{\gamma_j\}$.

When the standard uncertainties are based on specified, finite numbers of degrees of freedom, then it will be more appropriate to model either the $\{(x_j - \xi_j)/u(x_j)\}$ or the $\{(r_j - \rho_j)/u(r_j)\}$, or both, as outcomes of independent, Student's t random variables with the specified numbers of degrees of freedom [12].

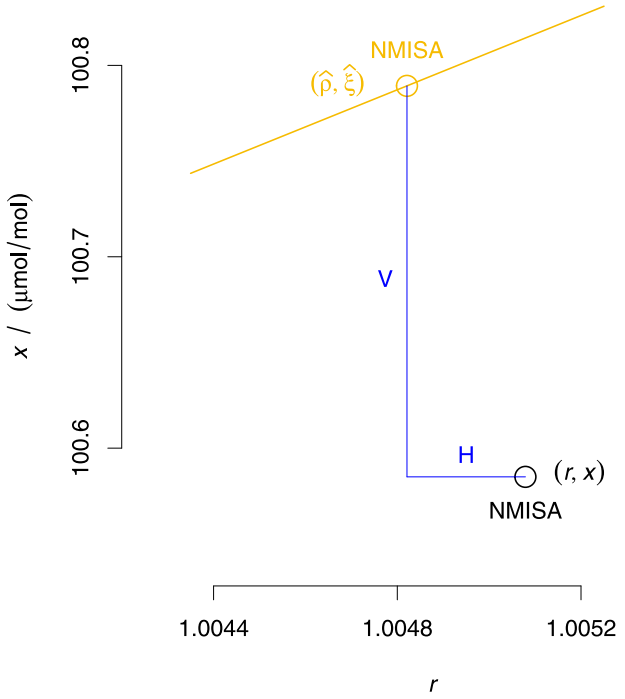

Figure 2. Magnified view of the portion of figure 1 around the values of r and x for NMISA, represented by the (black) circle, but without the line segments that represent the associated uncertainties, and without the EIV line (depicted in dark purple in figure 1) corresponding to the model without dark uncertainty (EIV-LIGHT). The (yellow) circle centered on the (yellow) EIV regression line, represents estimates of the true values of r and x corresponding to the EIV model that takes dark uncertainty into account (EIV-DARK). The absolute value of the horizontal residual, 0.000 26, is the length of the horizontal (blue) line segment labeled H, and the absolute value of the vertical residual, −0.204 µmol/mol, is the length of the vertical (blue) line segment labeled V, as listed in table 3. As will be explained in section 5, the vertical residual is the D in the degree of equivalence $(D, U_{95\%}(D))$ for NMISA. Note that the vertical residual is neither the vertical distance from the point with coordinates (r, x) to the EIV line, nor the distance from (r, x) to $(\widehat{\rho}, \widehat{\xi})$, which would be meaningless owing to the different units of the axes.

Table 3. Horizontal and vertical residuals for the EIV regression line fitted taking dark uncertainty into account, and corresponding standard uncertainties for the ratios and for the amount fractions. The horizontal and vertical residuals for NMISA are depicted in figure 2. All the horizontal residuals have smaller absolute values than the standard uncertainties associated with the ratios, while nine of the eleven vertical residuals have larger absolute values than the standard uncertainties for the corresponding amount fractions.

LAB	$r - \widehat{\rho}$	$u(r)$	$x - \widehat{\xi}$/(µmol/mol)	$u(x)$
NMIJ	0.000 00	0.0006	0.000	0.009
NPL	0.000 05	0.0006	−0.052	0.048
BAM	−0.000 12	0.0006	0.158	0.125
NMIA	0.000 29	0.0006	−0.278	0.070
NIST	−0.000 20	0.0006	0.178	0.060
NMISA	0.000 26	0.0006	−0.204	0.011
CENAM	−0.000 02	0.0006	0.035	0.150
LNE	−0.000 28	0.0006	0.243	0.050
KRISS	−0.000 09	0.0006	0.069	0.010
VNIIM	0.000 20	0.0006	−0.187	0.070
VSL	−0.000 10	0.0006	0.084	0.040

Table 4. Bayes estimates of the intercept and slope of the EIV regression lines fitted to the measurement results from CCQM-K53, corresponding to the model variants with the same dark uncertainty for all the participants (EIV-DARK), with shades of dark uncertainty (EIV-SHADES), and that ignore dark uncertainty (EIV-LIGHT).

model variant	β_1/(µmol/mol)	$u(\beta_1)$/(µmol/mol)	β_2	$u(\beta_2)$
EIV-DARK	3.3	4	97	4
EIV-SHADES	1.4	4	99	4
EIV-LIGHT	2.2	2	98	2

The presence of measurement errors associated with observed values for the ratios and for the amount fractions is the reason why the model for the relationship between the $\{r_j\}$ and the $\{x_j\}$ is called an EIV regression model [2, 7]. But the model in equation (2) is not the conventional EIV model because it also recognizes the possible presence of one additional source of uncertainty, as we explain next.

The $\{\lambda_j\}$ in the part of the model for the amount fractions, $x_j = \xi_j + \lambda_j + \varepsilon_j$, are hypothetical laboratory effects assumed to be a sample from a Gaussian distribution with mean 0 and standard deviation τ. (Depending on the circumstances, distributions other than the Gaussian may be more appropriate, for example Student's t or Laplace's.) When $\tau = 0$ this model devolves into the common EIV model, which we refer to as EIV-LIGHT.

The EIV-DARK model variant for the amount fractions accommodates the fact that the vertical residuals, $\{x_j - \widehat{\xi}_j\}$, are more dispersed than the associated uncertainties $\{u(x_j)\}$ suggest that they should be, on the assumption that the relationship between the $\{\xi_j\}$ and the $\{\rho_j\}$ has been characterized correctly. In such case, the measurement results for the amount fractions appear to be mutually inconsistent in the context of the EIV regression model, in the same sense that this term is used for results from conventional interlaboratory studies and KCs [24].

The standard deviation, τ, quantifies the *dark uncertainty* for the amount fractions, a concept that Thompson and Ellison [40] introduced in the context of interlaboratory studies. The corresponding uncertainty component, or amalgam of components, is described as 'dark' because it does not appear in the uncertainty budgets that underlie the $\{u(x_j)\}$, and reveals itself only after a candidate analysis function has been fitted to all the measurement results together.

In the case of CCQM-K53, the estimate of τ is both statistically and substantively significant: in other words, it is confidently and sufficiently greater than zero to make a difference. The median of the Markov Chain Monte Carlo (MCMC) sample drawn from the posterior distribution of τ, as explained in section 4, is 0.165 µmol/mol, which is 3.3 times larger than the $\{u(x_j)\}$. A 95% credible interval for its true value ranges from 0.097 µmol/mol to 0.296 µmol/mol.

One may wonder whether a similar component of dark uncertainty might exist for the ratios. However, table 3 shows

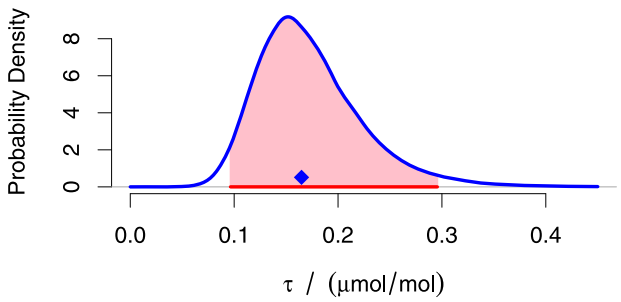

Figure 3. Estimate (thick, blue curve) of the posterior probability density of τ. The shaded (pink) area comprises 95% of the area under the curve, and its footprint on the horizontal axis, which does not straddle 0 μmol/mol, is a 95% credible interval for the true value of τ. The blue diamond marks the median of this distribution, which is 0.165 μmol/mol.

that such component is not warranted for the measurement results under consideration here. In fact, the standard deviation of the horizontal residuals is 3.2 times smaller than the median of the $\{u(r_j)\}$, while the standard deviation of the vertical residuals is 3.4 times larger than the median of the $\{u(x_j)\}$. Furthermore, and even more conclusively, if the EIV model is fitted entertaining components of dark uncertainty both for the amount fractions and for the ratios, the component for the latter does not differ significantly from zero.

3.2. EIV model for CCQM-K53 with shades of dark uncertainty

Merkatas *et al* [29] introduced a model for interlaboratory studies where the dark uncertainty is allowed to vary among participants, making the least contribution to those that are already in fair mutual agreement among themselves, and making the largest contributions to those that are in more marked disagreement with the bulk of the others, either because their measured values lie farther out, or because their reported uncertainties may be relatively much too small.

We extend this model to the context of EIV regression, and simplify it in the process, as follows: the relationship between the true values of the amount fractions and ratios of instrumental indications remains the same as in equation (2); however, the variances of the amount fractions, $\{x_j\}$, become

$$v_j^2 = u^2(x_j) + a_j\tau^2, \tag{3}$$

where a_j is a Bernoulli random variable with probability p_j of being 1 and probability $1 - p_j$ of being 0, for $j = 1, \ldots, n$. In other words, the EIV-SHADES model variant expresses the 'extra' variance component associated with x_j as fraction of the estimate of τ that corresponds to this model variant.

That is, this model variant can assign variance v_j^2 to the jth participant that ranges from $v_j^2 = u^2(x_j)$ to $v_j^2 = u^2(x_j) + \tau^2$, with all possible values in-between, with the $\{p_j\}$ representing mixing proportions of these two extremes, hence similar to the probabilities of inconsistency that Mana [28] introduced. Such mixtures can be modeled in still other ways, for example as in the original account of the shades model [29], or without involving the Bernoulli indicators $\{a_j\}$.

4. Bayesian procedure to estimate the analysis function for CCQM-K53

The different versions of the EIV model described above, in subsections 3.1 (EIV-DARK and EIV-LIGHT) and 3.2 (EIV-SHADES), were fitted to the data using Bayesian procedures. In subsection 4.1, we explain these procedures informally and intuitively, and in subsection 4.2 we provide additional, more technical details.

4.1. Bayesian approach

The measurement results used to determine the analysis function comprise n sets of measurement results $\{(r_j, u(r_j), x_j, u(x_j)) : j = 1, \ldots, n\}$. Therefore, we have n points $\{(r_j, x_j)\}$ to which we fit the regression line, plus $2n$ standard uncertainties that quantify reasonable ranges within which to locate the estimates of the true $\{(\rho_j, \xi_j)\}$.

The $\{\lambda_j\}$ in equation (2), which are assumed to be random effects, usually are not estimated. Their key property that needs to be recognized and estimated is their standard deviation, τ, which is then expressed in the uncertainty of the resulting KCRF and of derivative quantities [38].

Since the number of parameters, $n + 3$, (which are the intercept and slope of the EIV regression and the true values, $\{\rho_j\}$, of the ratios of instrumental indications) is of the same order of magnitude as the number of available data, this risks turning the optimization into an ill-posed problem: that is, a problem for which there may be multiple solutions, or where the solution may be not only hard to find but also very sensitive to the uncertainties associated with the observations.

More consequential still is the fact that, under commonly made assumptions, the classical EIV model is not identifiable [35]. This means that different combinations of values of the parameters can yield the same probability distribution for the data, and implies that the maximum likelihood estimate may not be unique unless additional assumptions are made or constraints are introduced.

The Bayesian approach can be conceived as a cautious way of solving an ill-posed optimization problem that also overcomes the issue of non-identifiability. Conventional, deterministic optimization aims to locate the solution by finding the most direct path towards an optimum. However, it may end-up in the wrong place because there may be multiple local optima, or because the global optimum may not be defined sharply.

The Bayesian approach thoroughly explores the set of possible values for the parameters, which in this case is a subset of a space of $n + 3$ dimensions. And this exploration is driven not only by the measurement results, but also by prior expectations about where the solution is likely to be. These expectations derive from the structure of the problem, from the design of the experiment, and from prior knowledge about fitting EIV models to data of this kind.

For example, we expect that the ratios of instrumental indications will be around 1, given how the composition of the standard used to form the ratios was selected. But we also expect that the ratios will exhibit appreciable dispersion around 1, considering that the participants were asked to

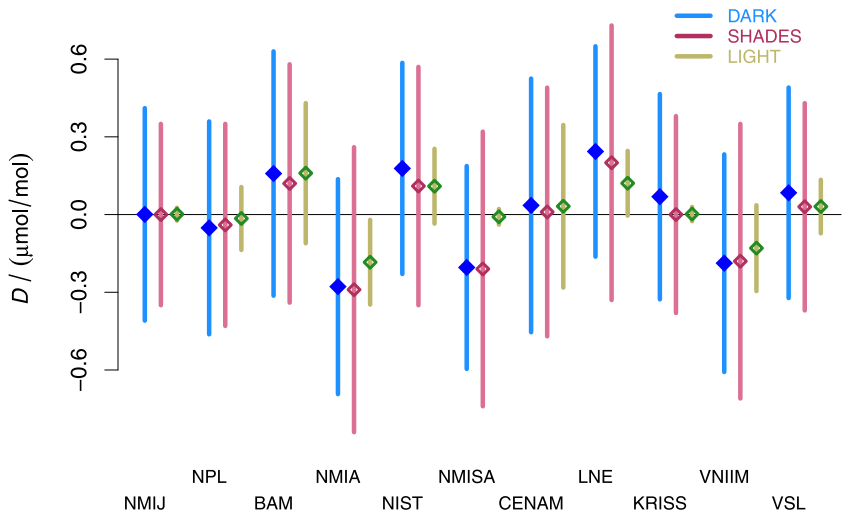

Figure 4. DoEs for the three variants of the EIV model that are listed in table 5: DARK, labeled $\{D_j\}$ in table 5, with the same τ for all the participants; SHADES, labeled $\{D_j^+\}$ in table 5, with shades of dark uncertainty that contribute different values of τ to different participants; LIGHT, labeled $\{D_j^-\}$ in table 5, ignoring dark uncertainty. The diamonds and the vertical line segments represent the corresponding DoEs.

Table 5. DoEs for the three variants of the EIV model that are depicted in figure 4: EIV-DARK ($\{D_j\}$, with the same τ for all the participants); EIV-SHADES ($\{D_j^+\}$, with shades of dark uncertainty that contribute different values of τ to different participants); and EIV-LIGHT ($\{D_j^-\}$, ignoring dark uncertainty).

LAB	$D/(\mu\text{mol/mol})$	$U_{95\%}(D)/(\mu\text{mol/mol})$	$D^+/(\mu\text{mol/mol})$	$U_{95\%}(D^+)$	D^-	$U_{95\%}(D^-)$
NMIJ	0.00	0.41	0.00	0.35	0.00	0.03
NPL	−0.05	0.42	−0.04	0.39	−0.02	0.12
BAM	0.16	0.47	0.12	0.46	0.16	0.27
NMIA	−0.28	0.42	−0.29	0.55	−0.18	0.16
NIST	0.18	0.41	0.11	0.46	0.11	0.14
NMISA	−0.20	0.39	−0.21	0.53	−0.01	0.03
CENAM	0.04	0.49	0.01	0.48	0.03	0.31
LNE	0.24	0.41	0.20	0.53	0.12	0.12
KRISS	0.07	0.40	0.00	0.38	0.00	0.03
VNIIM	−0.19	0.42	−0.18	0.53	−0.13	0.16
VSL	0.08	0.41	0.03	0.40	0.03	0.10

prepare mixtures with different amounts of oxygen [25, table 2] best to support the regression line.

Since the ratios should be around 1, and considering that the slope of the line must be positive, we also expect that β_2 will have a numerical value close to the average of the amount fractions of the participants' mixtures. Not only this, we also expect that both β_1 and β_2 will be close to their OLS estimates because the relative uncertainties for the ratios all are fairly small (around 0.06%).

The Bayesian approach uses probability distributions to encapsulate such prior knowledge about the values of the parameters. For example, Gaussian distributions are assigned to both the intercept and slope. The prior distribution for the slope is centered at the OLS estimate, and the prior distribution for the intercept is centered at zero because that's the expected response of the GC-TCD instrument in the absence of oxygen [25, p 4].

The prior relative standard deviations for the slope and intercept are either 10% or the relative uncertainties of the corresponding OLS estimates, whichever are largest. For the intercept, β_1, the prior relative standard deviation was 105%, and for the slope, β_2, it was 10%.

The prior distributions for the mean values of the ratios, $\{\rho_j\}$, are Gaussian distributions centered at the observed ratios with standard deviations three times larger than the standard uncertainties for the ratios.

Finally, we specify that τ is equally likely to be smaller or greater than the median of the standard uncertainties associated with the amount fractions. We impose the constraint that τ must have a non-negative numerical value by assigning a prior distribution to it that is concentrated on the positive numbers, and otherwise choose this distribution so that it rules out no positive values whatever, no matter how large, but with decreasing probability for values of τ that are increasingly higher than the median of the $\{u(x_j)\}$. We accomplish all this by assigning a Cauchy distribution truncated at zero to τ, with median equal to the median of the $\{u(x_j)\}$.

The Bayesian procedure does not deliver a particular solution. That is, it does not produce a specific set of 'optimal'

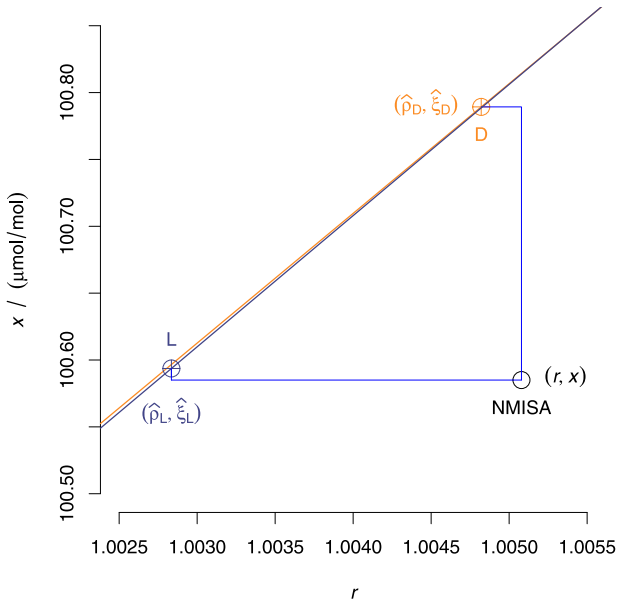

Figure 5. Magnified view of the portion of figure 1 around the values of r and x for NMISA, represented by the (black) circle, but without the line segments that represent the associated uncertainties. The (orange) crossed circle, labeled D, represents estimates of the true values of r and x corresponding to the EIV model that takes dark uncertainty into account. The (dark purple) crossed circle, labeled L, is its counterpart for the EIV model that disregards dark uncertainty. The DoE for the first model is the (signed) vertical residual, -0.20 µmol/mol, between the points labeled NMISA and D. The DoE for the second model is the (signed) vertical residual, -0.01 µmol/mol, between the points labeled NMISA and L. Figure 2 explains graphically the meaning of horizontal and vertical residuals.

values for β_1, β_2, τ, and the $\{\rho_j\}$. Instead, it samples the multidimensional space where the parameters live, guided by the prior expectations about these and by the statistical model for the data in equation (2). The resulting large sample is then used to locate the optimal values, in particular for the intercept and slope of the EIV line drawn in figure 1, and also for all the other parameters.

For the EIV-SHADES version of the model described in subsection 3.2, the intercept and slope of the regression line were assumed to be Gaussian and independent *a priori*, with means set equal to the OLS estimates, and standard deviations about 50 times larger than their least squares counterparts, hence rather noninformative. The true values of the ratios, $\{\rho_j\}$, also were assumed to be independent *a priori*, and had Gaussian prior distributions with mean 1 and standard deviation almost 100 times larger than the uncertainties reported for the ratios. The prior distribution for τ was the same half-Cauchy described above, and the $\{p_j\}$ were independent *a priori* and uniformly distributed between 0 and 1.

In addition to the thorough exploration of the set of possible parameter values, the Bayesian approach offers this great practical advantage relative to the classical approach: in a single stroke, it provides all the elements needed to evaluate the uncertainty surrounding the KCRF (represented by the shaded (yellow) band in figure 1), as well as the expanded uncertainties that are part and parcel of the DoEs, which are discussed in section 5.

4.2. Markov Chain Monte Carlo

The Bayesian version of the model defined in equation (2), which is the EIV-DARK variant, is a hierarchical model: given the values of β_1, β_2, τ and ρ_j, the ratio r_j is an outcome of a Gaussian random variable with mean ρ_j and standard deviation $u(r_j)$, and x_j is an outcome of a Gaussian random variable with mean $\xi_j = \beta_1 + \beta_2\rho_j$ and variance $\tau^2 + u^2(x_j)$, for $j = 1, \ldots, n$.

Fitting such model to the measurement results then becomes an exercise in Bayesian inference, which is done based on a sample of large size K drawn from the joint posterior distribution of β_1, β_2, τ, and $\rho_1, \ldots, \rho_n$, given the measurement results, by application of a device called MCMC sampling [8]. We denote the resulting samples of the values of the parameters $\{\widetilde{\beta}_{1,k}\}$, $\{\widetilde{\beta}_{2,k}\}$, $\{\widetilde{\tau}_k\}$, $\{\widetilde{\xi}_{j,k}\}$, and $\{\widetilde{\rho}_{j,k}\}$, for $j = 1, \ldots, n$ and for $k = 1, \ldots, K$, where K was 80 000.

This number, $K = 80\,000$, is only a small fraction (effectively, 10%) of the MCMC iterations that explored the range of all possible values of the parameters and that were stored to become a sample of the joint posterior distribution of the parameters, from which all subsequent inferences (estimates and uncertainty evaluations) were drawn.

The Bayesian procedure was implemented using the Stan language [1], in the code listed in figure A1, which is invoked for execution using the R [34] code listed in figure A2. Only 6% of the z-scores proposed by Geweke [9] to assess convergence of the MCMC chains to their equilibrium distribution, had absolute values greater than 2. Examination of the autocorrelation functions for the samples drawn from the posterior distribution of the parameters showed that the thinning that was employed was satisfactory. The diagnostics were computed and examined graphically using R functions `ggs_geweke` and `ggs_autocorrelation` defined in package `ggmcmc` [19].

After thinning, and all together, the four chains that were used (figure A1), produced 80 000 samples from the posterior joint distribution of the parameters given the data. The effective sample sizes for the parameters ranged from 77 000 to 80 000. The ratios of the total variability to the within-chain variability (so-called `Rhat`) were very close to 1 for all the parameters.

The EIV-SHADES variant was formulated in the BUGS language (as implemented in OpenBUGS [26]) because it provides an easier way to incorporate discrete latent parameters than Stan. This variant, too, was fitted to the data using MCMC.

Table 4 lists Bayes estimates of the EIV regression coefficients for the lines depicted in figure 1, and also for the model with shades of dark uncertainty. EIV-DARK and EIV-LIGHT correspond to the models described in subsection 3.1, and EIV-SHADES corresponds to the model described in subsection 3.2. Ignoring dark uncertainty when there is dark uncertainty that is both statistically and substantively significant, inevitably yields uncertainty evaluations (for the DoEs, in particular) that are unrealistically optimistic, and biased estimates of the regression coefficients [2].

Table 6. Reported standard uncertainties, $\{u(x_j)\}$, for the amount fractions of oxygen listed in table 1, and posterior medians of the $\{v_j\}$ for the model variant EIV-SHADES, which are defined in equation (3).

LAB	$u(x_j)$	v_j
NMIJ	0.009	0.009
NPL	0.048	0.048
BAM	0.125	0.125
NMIA	0.070	0.231
NIST	0.060	0.186
NMISA	0.011	0.216
CENAM	0.150	0.150
LNE	0.050	0.218
KRISS	0.010	0.010
VNIIM	0.070	0.216
VSL	0.040	0.040

Figure 3 shows an estimate of the posterior probability density of the dark uncertainty, τ, for the model with a common value of the dark uncertainty for all the participants, and a 95% credible interval for its true value, which substantiates the conclusion that τ indeed is significantly greater than zero.

5. Degrees of equivalence for CCQM-K53

The DoEs are pairs $\{(D_j, U_{95\%}(D_j))\}$, where D_j is the difference between the amount fraction measured by participant j and the corresponding value predicted by the KCRF, and $U_{95\%}(D_j)$ is the expanded uncertainty associated with this difference, for 95% coverage.

The $\{\widehat{\rho}_j\}$ are estimates of the true values of the ratios, and $\{\widehat{\xi}_j\}$ are estimates of the true values of the amount fractions, computed under the assumption that $\xi_j = \beta_1 + \beta_2\rho_j$ for $j = 1, \ldots, n$ stated in equation (2).

Figure 4 depicts the DoEs that are listed in table 5. Three sets of DoEs are presented: $\{(D_j, U_{95\%}(D_j))\}$, which pertain to the EIV model variant (EIV-DARK) that assigns the same amount of dark uncertainty to all the participants; $\{(D_j^+, U_{95\%}(D_j^+))\}$ for the model with shades of dark uncertainty (EIV-SHADES); and $\{(D_j^-, U_{95\%}(D_j^-))\}$ when dark uncertainty is ignored (EIV-LIGHT).

The $\{(D_j, U_{95\%}(D_j))\}$ were computed as follows:

(a) D_j is the difference $x_j - \widehat{\xi}_j$, for $j = 1 \ldots, n = 11$;

(b) $U_{95\%}(D_j)$ is such that the interval $D_j \pm U_{95\%}(D_j)$ includes 95% of the following K differences: $x_j - (\widehat{\xi}_{j,1} + y_{j,1} + z_{j,1})$, …, $x_j - (\widehat{\xi}_{j,K} + y_{j,K} + z_{j,K})$, where $y_{j,k}$ is drawn from a Gaussian distribution with mean 0 and standard deviation $u(x_j)$, and $z_{j,k}$ is drawn from a Gaussian distribution with mean 0 and standard deviation $\widehat{\tau}_k$, for $j = 1 \ldots, n = 11$, and for $k = 1, \ldots, K$.

The $\{y_{j,k}\}$ reflect the fact that the $\{D_j\}$ are predicted values, not expected values, and the $\{z_{j,k}\}$ put into play the 'extra' uncertainty expressed in the dark uncertainty. The $\{(D_j^-, U_{95\%}(D_j^-))\}$ are computed similarly, except that, in step (b), there are no $\{z_{j,k}\}$.

For the model with shades of dark uncertainty, $U(D_j^+)$ is twice the square root of the estimated variance of the predictive distribution, $2\sqrt{u^2(\xi_j) + v_j^2}$, for $j = 1, \ldots, n$.

Considering that the EIV regression lines corresponding to the models with and without dark uncertainty, which are depicted in figure 1, are just about coincident, the differences between corresponding $\{(D_j, U_{95\%}(D_j))\}$ and $\{(D_j^-, U_{95\%}(D_j^-))\}$, apparent both in figure 4 and in table 5, certainly are surprising. This difference is most pronounced for NMIJ, NMISA and KRISS, but its counterparts for NMIA and LNE are also striking.

Figure 5 provides a close-up of the differences D_6 and D_6^-, which pertain to NMISA. Given the explanations given above, and illustrated in figure 2, of how horizontal and vertical residuals are defined, here we focus on the wide discrepancy between the points labeled D and L in figure 5.

Since both EIV lines are practically coincident immediately above the point with coordinates (r_6, x_6), which are the ratio and the amount fraction measured for NMISA, it could be surprising that the corresponding D_6 and D_6^- should be so very different: the first -0.20 μmol/mol, the second -0.01 μmol/mol. This is a consequence of how the residuals are defined in equation (2), which is the same for both models, except that the $\{\lambda_j\}$ are missing from the model that ignores dark uncertainty. Furthermore, the estimates $(\widehat{\rho}_6, \widehat{\xi}_6)$ for either model, represented by the aforementioned points labeled D and L, are estimates of true values computed under the assumption that $\xi = \beta_1 + \beta_2\rho$.

This very strong assumption is just another way of saying that the measurement results are assumed to be mutually consistent, in the sense that the deviations of the observations $\{(r_j, x_j)\}$ from a straight line are commensurate with the reported uncertainties associated with the ratios and with the amount fractions.

However, this is not the case here: the observations and their associated uncertainties are not consistent with their lying on a straight line. The estimation procedure for the EIV model that neglects dark uncertainty does its best under such assumption but can do no better than provide an inadequate solution. And the best it manages to achieve for NMISA involves placing $(\widehat{\rho}_6, \widehat{\xi}_6)$ at location L in figure 5. The procedure feels at liberty to do so given the ample horizontal uncertainty that surrounds (r_6, x_6), which is depicted in figure 1.

The EIV model that recognizes dark uncertainty does not need to go through the same gyrations owing to the freedom it has to accommodate a large vertical deviation of (r_6, x_6) from the corresponding regression line: a freedom that the incorporation of dark uncertainty gives it. For this reason, the vertical residual for NMISA is almost the vertical distance between (r_6, x_6) and the regression line.

The moral of this story is that the substantial differences between corresponding $\{D_j\}$ and $\{D_j^-\}$ are merely a consequence of the fact that the EIV model that disregards dark uncertainty is unable to cope with the mutual inconsistency of the measurement results from CCQM-K53, and the constraint $\xi = \beta_1 + \beta_2\rho$ forces the estimation procedure to place some of the (bivariate) fitted values at locations that may be surprising.

The original data reductions in Lee *et al* [25] overcome this challenge by setting aside the results from NMIA, NMISA, and VNIIM. In this contribution we retain them because the model we have developed accommodates them easily.

Table 6 compares the reported uncertainties with their 'adjusted' counterparts produced by the EIV-SHADES model variant, where v_j is as defined in equation (3). The posterior distributions of the $\{v_j\}$ were markedly skewed, and for this reason the table lists their medians. For NMIA, NIST, NMISA, LNE, and VNIIM the $\{v_j\}$ had posterior medians appreciably larger than their $\{u(x_j)\}$. For the other participants, the posterior medians of the $\{v_j\}$ were essentially equal to the corresponding $\{u(x_j)\}$.

6. Errors-in-variables model for $\delta(^{13}C)$

The model we shall use to relate measured and calibrated values of $\delta(^{13}C)$ for the standards listed in table 2, stated below in equation (4), involves components of dark uncertainty for both quantities related by the EIV regression.

6.1. Analysis function for $\delta(^{13}C)$

An analysis function is built based on the measurement results listed in table 2, and this analysis function will be used subsequently to assign a calibrated isotope delta value to a material sample. The novelties in this case are: (i) the reported uncertainties are qualified with finite (and not very large) numbers of degrees of freedom; (ii) there are uncertainty components attributable to dark uncertainty for both quantities related by the EIV regression model; and (iii) the EIV regression is used to assign a value to a material sample, and the uncertainty associated with this prediction is evaluated.

Fact (i) implies that the $\{u\,(\delta_{\mathrm{NMIA,S}}(^{13}\mathrm{C}))\}$ and the $\{u\,(\delta_{\mathrm{VPDB,S}}(^{13}\mathrm{C}))\}$ are estimates of true but unknown standard deviations of the measurement errors $\{\gamma_j\}$ and $\{\varepsilon_j\}$, and these $2n$ standard deviations are additional parameters in the version of the EIV model to be entertained in this case.

Need (ii) involves the inclusion of two dark uncertainty parameters in the model, which now becomes:

$$\theta_j = \beta_1 + \beta_2\eta_j, \qquad \delta_{\mathrm{NMIA,S}_j}(^{13}\mathrm{C}) = \eta_j + \varphi_j + \gamma_j, \quad \text{and}$$
$$\delta_{\mathrm{VPDB,S}_j}(^{13}\mathrm{C}) = \theta_j + \psi_j + \varepsilon_j, \tag{4}$$

for $j = 1, \ldots, n$, where j refers to the substance among the $n = 8$ listed in table 2. Here, ψ_j has the role that λ_j had in equation (2), and φ_j is its counterpart for $\delta_{\mathrm{NMIA,S}_j}(^{13}\mathrm{C})$, which was not needed in the case of CCQM-K53. The $\{\varphi_j\}$ and the $\{\psi_j\}$ are modeled as non-observable values of independent Gaussian random variables with mean 0, the former with standard deviation τ_φ, the latter with standard deviation τ_ψ, which are the two components of dark uncertainty operating along the two axes relative to which the analysis function is depicted in figure 6. The $\{\gamma_j\}$ and the $\{\varepsilon_j\}$ represent measurement errors whose uncertainty contributions are captured in the reported uncertainties associated with the measured isotope delta values.

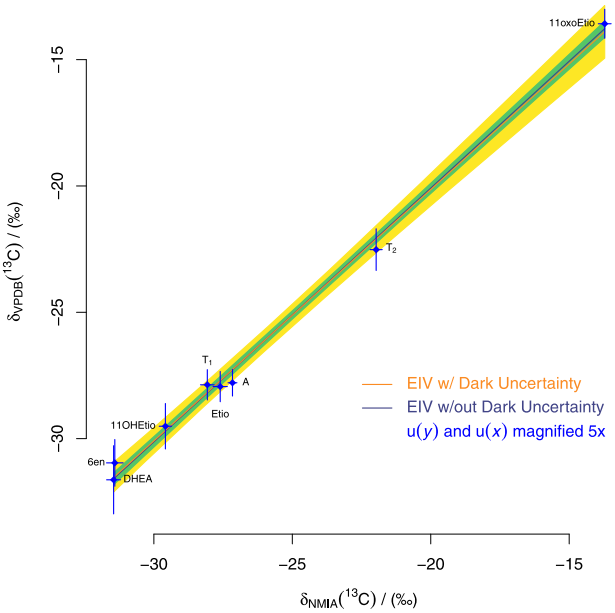

Figure 6. The (blue) diamonds represent the measured values, the horizontal line segments represent the $\{\delta_{\mathrm{NMIA,S}_j}(^{13}\mathrm{C}) \pm 5u\,(\delta_{\mathrm{NMIA,S}_j}(^{13}\mathrm{C}))\}$, and the vertical line segments represent the $\{\delta_{\mathrm{VPDB,S}_j}(^{13}\mathrm{C}) \pm 5u\,(\delta_{\mathrm{VPDB,S}_j}(^{13}\mathrm{C}))\}$. The light (orange) sloping line represents the EIV regression line fitted taking dark uncertainty into account, and the wide, shaded (yellow) band is a simultaneous 95% coverage band for the whole line. The dark (purple) sloping line represents the EIV regression line fitted disregarding dark uncertainty, and the light (green) band is a simultaneous 95% coverage band for this whole line.

This model was fitted to the calibration data using a Bayesian procedure whose rationale has already been explained in section 4. MCMC was employed similarly to how it was employed in subsection 4.2, with differences in implementation (figure A1 versus A3) that take into account the finite numbers of degrees of freedom supporting the reported uncertainties, and the need for considering dark uncertainty along both axes. Figures A3 and A4 list the corresponding Stan and R codes.

Table 7 lists the estimates of the intercept and slope, and associated uncertainties, for the model that recognizes dark uncertainty, and for the model that ignores it. Figure 6 depicts the EIV regression lines corresponding to these two models, and the associated uncertainty bands. Figure 7 shows an estimate of the density of the joint posterior probability distribution of τ_φ and τ_ψ. Both τ_φ and τ_ψ are strictly greater than zero with 99% probability.

The choice of values for the parameters of the prior distributions was driven by the following considerations:

(a) The prior means for the $\{\eta_j\}$ are set equal to the measured values $\{\delta_{\mathrm{NMIA,S}_j}(^{13}\mathrm{C})\}$, and the prior standard deviations are set equal to $\{3u\,(\delta_{\mathrm{NMIA,S}_j}(^{13}\mathrm{C}))\}$;

JUSTIFICATION: the choice for the means is based on the typical repeatability achieved in mass spectrometry [11], and the adoption of extravagantly large values for the standard deviations renders the prior distribution very weakly informative.

Table 7. Bayes estimates of the intercept and slope of the EIV regression lines fitted to the measurements of $\delta(^{13}\mathrm{C})$ listed in table 2, for the models that recognize (EIV-DARK) or ignore (EIV-LIGHT) dark uncertainty. The estimator used for the intercept and slope is as implemented in R function `huberM` defined in package `robustbase` [27, 36], and the estimator used for their standard uncertainties is as implemented in R function `Qn` defined in the same package.

model	β_1	$u(\beta_1)$	β_2	$u(\beta_2)$
EIV-DARK	−0.240 ‰	0.741 ‰	0.995	0.028
EIV-LIGHT	0.003 ‰	0.265 ‰	1.004	0.010

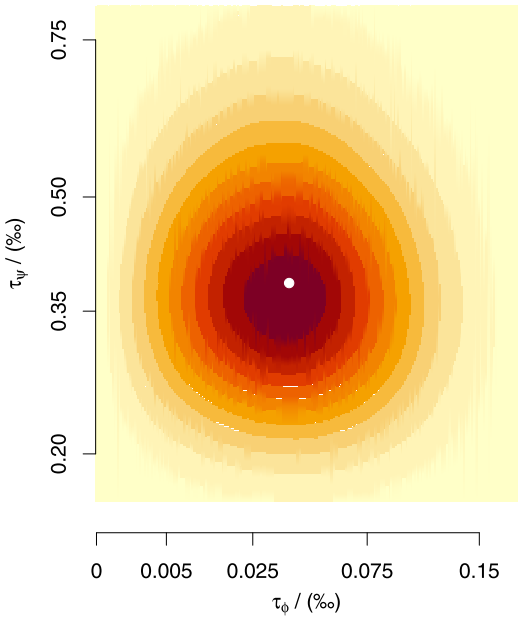

Figure 7. Density of the joint posterior distribution of τ_φ and τ_ψ in the model underlying the analysis function for $\delta(^{13}\mathrm{C})$, where the darker the hue the higher the density. Note that the axes have non-linear scales. The coordinates of the white dot are the median of the posterior distribution of τ_φ, 0.038 ‰, and the median of the posterior distribution of τ_ψ, 0.38 ‰ (the fact that the latter is ten times larger than the former is merely a numerical coincidence).

(a) The intercept, β_1, has prior mean 0 ‰ and prior standard deviation 10 ‰, and the slope, β_2, has prior mean 1 and prior standard deviation 0.5;

JUSTIFICATION: the choice of prior mean for β_2 is based on the fact that corresponding true values of $\{\theta_j\}$ and $\{\eta_j\}$ are expected to be nearly identical, unless NMIA's in-house reference were grossly miscalibrated, which it is not considering that a robust estimate of the slope is close to 1: this value was obtained using R function `lmrob`, defined in package `robustbase` [27] disregarding the reported uncertainties and neglecting the presence of components of dark uncertainty.

The relative prior standard uncertainty of $0.5/1 = 50\%$ for β_2 is much larger than what experience suggests would be reasonable to entertain, and thus delivers a largely uninformative prior.

The intercept, β_1, is expected to be zero owing to considerations of mass balance. In the extreme, inconceivable case that the slope should be zero, the uncertainty surrounding the value of the intercept still should not exceed the variability of $\delta(^{13}\mathrm{C})$ values in plants (−10 ‰ to −34 ‰ [5]) or in steroid metabolites in the urine of human populations (−16 ‰ to −26 ‰ [3]).

Therefore, the prior standard deviation of 10 ‰ for the intercept generously makes allowance for such far-fetched possibility, effectively rendering this prior distribution non-informative.

(c) The prior median for τ_ψ is set equal to the median of the absolute values of the residuals from a line fitted to the $\{\delta_{\mathrm{VPDB,S}_j}(^{13}\mathrm{C})\}$ as a function of the $\{\delta_{\mathrm{NMIA,S}_j}(^{13}\mathrm{C})\}$, and similarly for τ_φ but for the residuals from a line fitted to the $\{\delta_{\mathrm{NMIA,S}_j}(^{13}\mathrm{C})\}$ as a function of the $\{\delta_{\mathrm{VPDB,S}_j}(^{13}\mathrm{C})\}$;

JUSTIFICATION: the prior expectation is neutral about whether dark uncertainty is or is not needed, best to model the variability of the measured values (separately for the $\{\delta_{\mathrm{VPDB,S}_j}(^{13}\mathrm{C})\}$ and for the $\{\delta_{\mathrm{NMIA,S}_j}(^{13}\mathrm{C})\}$). The choice for these prior medians captures that 'neutrality' by recognizing that the absolute values of the vertical or horizontal residuals are the benchmarks for deciding whether components of dark uncertainty are called for.

(d) The prior median for the true variances corresponding to the reported standard uncertainties $\{u\,(\delta_{\mathrm{VPDB,S}_j}(^{13}\mathrm{C}))\}$ is set equal to the squared median of these standard uncertainties, and similarly for the true variances corresponding to the $\{u\,(\delta_{\mathrm{NMIA,S}_j}(^{13}\mathrm{C}))\}$.

JUSTIFICATION: the established, typical repeatability of mass spectrometry implies that this choice gives a very wide margin to the true variances, and in fact makes the prior only very weakly informative.

6.2. Prediction of $\delta(^{13}\mathrm{C})$ in urine sample

The measurement campaign was seeking the carbon isotope delta value of testosterone, $\delta_{\mathrm{VPDB,T}}(^{13}\mathrm{C})$, in a human urine sample denoted T. The measurement results were $\delta_{\mathrm{NMIA,T}}(^{13}\mathrm{C}) = -26.87$ ‰ with $u\,(\delta_{\mathrm{NMIA,T}}(^{13}\mathrm{C})) = 0.124\ ‰/\sqrt{8}$, based on 7 degrees of freedom.

The predicted value is $\widehat{\theta}_\mathrm{T} = -0.240‰ + 0.995 \times (-26.87\,‰) = -26.98\,‰$ where −0.240 ‰ and 0.995 (table 7) are the intercept and slope of the EIV regression line depicted in figure 6 that corresponds to the model with dark uncertainty along both axes, and θ_T is the counterpart of the $\{\theta_j\}$ in equation (4).

The uncertainty associated with the predicted value is evaluated using a Monte Carlo method that produces a sample of size K from the probability distribution of $\delta_{\mathrm{VPDB,T}}(^{13}\mathrm{C})$, whose

implementation in R code is listed in figure A5. The values in this sample are of the following form, for $k = 1, \ldots, K$:

$$\eta_{T,k} = \delta_{NMIA,T}(^{13}C) + (\widetilde{\tau}_{\varphi,k} \times a_k) + \left(u(\delta_{NMIA,T}(^{13}C)) \times t_k / \sqrt{\nu_T/(\nu_T - 2)}\right), \quad \text{and}$$
$$\theta_{T,k} = \widetilde{\beta}_{1,k} + \left(\widetilde{\beta}_{2,k} \times \eta_{T,k}\right) + (\widetilde{\tau}_{\psi,k} \times b_k), \tag{5}$$

where a_k and b_k denote independent drawings from a Gaussian distribution with mean 0 and standard deviation 1, $\widetilde{\tau}_{\varphi,k}$ and $\widetilde{\tau}_{\psi,k}$ denote drawings from the posterior distributions of the two components of dark uncertainty (which were generated during MCMC sampling using the Stan and R codes in figures A3 and A4), ν_T denotes the number of degrees of freedom supporting $u\,(\delta_{NMIA,T}(^{13}C))$, $\widetilde{\beta}_{1,k}$ and $\widetilde{\beta}_{2,k}$ denote drawings from the posterior distributions of the intercept and slope of the analysis function (also generated during MCMC sampling), and t_k denotes a value drawn from a Student's t distribution with ν_T degrees of freedom.

The uncertainty evaluation for $\delta_{VPDB,T}(^{13}C)$ is based on the $\{\theta_{T,k}\}$, and takes into account the fairly small number of degrees of freedom supporting $u\,(\delta_{NMIA,T}(^{13}C))$, as well as both components of dark uncertainty. Since the probability distribution of the $\{\theta_{T,k}\}$ has fairly heavy tails, the standard uncertainty is evaluated employing a robust, highly-efficient alternative to the standard deviation that is implemented in R function `Qn` defined in package `robustbase` [27, 36], which yields standard uncertainty 0.44 ‰. Its counterpart when dark uncertainty is neglected amounts to 0.08 ‰.

The standard uncertainty is appreciably smaller than the WADA requirement that it should not exceed 0.7 ‰ [42, table 1]. A 95% prediction interval for the true isotope delta value of the sample ranges from −27.93 ‰ to −26.01 ‰, and the corresponding expanded uncertainty for 95% coverage is 0.96 ‰. Since $0.96/0.44 = 2.18$, this uncertainty evaluation is effectively based on 12 degrees of freedom.

7. Conclusions

The modified EIV regression models described in sections 3 and 6 serve to reconcile measurement results that may be found to be mutually inconsistent when building a KCRF or an analysis function for value assignment either to a gas mixture or to a steroid $\delta(^{13}C)$ value in a urine sample.

By entertaining a yet unrecognized, possibly composite source of uncertainty, whose contribution is quantified by the dark uncertainty, along one or both axes of the regression, mutual consistency is achieved. However, accomplishing this enacts a cost in uncertainty surrounding this function, which becomes larger than it would have been if the 'extra', dark uncertainty had been disregarded. Both figures 1 and 6 depict this fact.

Concerning KCs, the inference our model and analysis substantiate, about the participants' ability to deliver measurement services of quality consistent with the uncertainty claims submitted to the KC, is meaningful only when the dark uncertainty is taken into account.

Whether dark uncertainty should be quantified using a single estimate of τ that applies to all measurement results equally (EIV-DARK), or using different estimates of τ for different subsets of the results (EIV-SHADES), depends on the roles that the evaluation of dark uncertainty is intended to serve.

Before we discuss these roles, below, we will note that all meta-analyses conducted in medicine that we have had a chance to examine over the years, entertain a single, common estimate of dark uncertainty, which there is named *heterogeneity* [17] and is quantified in sundry ways [16]. This fact is relevant because meta-analysis in medicine is closely comparable, methodologically, to the interlaboratory studies, including KCs, that are carried out in measurement science [37].

A single τ, applicable to all participants generally, is generalizable to characterize the state-of-the-art in the community of laboratories that the participants represent, even if the participants are not a simple random sample drawn from such community. Furthermore, a single τ also functions as an index that allows tracking the performance of that community over time, and succinctly summarizes their collective progress.

Shades of dark uncertainty, on the one hand describe more precisely and pointedly the degrees to which individual participants seem to be consistent with the consensus value, and with one another, which are principal goals of a KC, differently from a typical meta-analysis in medicine. On the other hand, because shades of dark uncertainty are specific to the individual participants, they do not lend themselves to generalization to the wider community that the participants originate from.

Acknowledging the presence of dark uncertainty and evaluating and propagating it, ultimately contributes to a realistic indication of the dispersion of values to be expected when this type of measurement is made by multiple laboratories whose capabilities are similar to those demonstrated in CCQM-K53. When dark uncertainty is both statistically and substantively significant, it may be unrealistic for these laboratories subsequently to claim measurement capabilities at the levels of their reported uncertainties.

However, the extent to which the dark uncertainty that is detected and quantified in a key comparison should impact claims of calibration and measurement capabilities (CMCs), involves both technical and policy considerations that only the overseeing Consultative Committee can determine.

Concerning determinations of isotope delta values, and similar tasks in analytical chemistry as well as in other fields where EIV should be employed routinely: properly taking into account contributions from all apparent sources of uncertainty, including those that can be evaluated only in a top-down manner [31, (3f)] and whose contributions were not accounted for in the reported uncertainties, yields more realistic uncertainty evaluations than when perceptible contributions are disregarded.

The EIV model that disregards dark uncertainty predicts the same isotope delta value for the urine sample considered in subsection 6.2, but the associated standard uncertainty is only 0.08 ‰, which is 5.5 times smaller than its counterpart for the EIV model that takes dark uncertainty into account. In this

```
data {
  int n;                    // Number of measurement results
  vector<lower=0>[n] x;     // Amount fractions
  vector<lower=0>[n] ux;    // Std. unc. for amount fractions
  vector<lower=0>[n] r;     // Ratios
  vector<lower=0>[n] ur;    // Std. unc. for ratios
  vector[2] betaM;          // Prior means for betas
  vector[2] betaS;          // Prior SDs for betas
  vector<lower=0>[n] rhoM;  // Prior means for rhos
  vector<lower=0>[n] rhoS;  // Prior SDs for rhos
  real<lower=0> gammax; }   // Prior median for dark uncertainty

parameters {
  vector[2] beta;           // Regression coefficients
  vector[n] rho;            // Mean values of ratios
  real<lower=0> taux; }     // Dark uncertainty

transformed parameters {
  vector[n] xi;             // Mean values of amount fractions
  for (j in 1:n) {xi[j] = beta[1] + beta[2]*rho[j];} }

model {
  beta ~ normal(betaM, betaS); // Prior for betas
  taux ~ cauchy(0, gammax);    // Prior for dark uncertainty
  rho ~ normal(rhoM, rhoS);    // Prior for rhos
  r ~ normal(rho, ur);         // Likelihood for ratios
  // Likelihood for amount fractions
  x ~ normal(xi, sqrt(square(ux) + square(taux))); }
```

Figure A1. Stan code that implements MCMC for the model in equation (2) fitted to the measurement results from CCQM-K53. The variable `taux` represents the dark uncertainty τ for the amount fractions of oxygen in nitrogen.

```
require(rstan)

psi = lm(x ~ poly(r, 1, raw=TRUE))
psi.summary = summary(psi)
beta.pi.mean = psi.summary$coefficients[,"Estimate"]
beta.pi.cv = c(max(0.1, abs(psi.summary$coefficients[1,"Std. Error"] /
                            psi.summary$coefficients[1,"Estimate"])),
               max(0.1, abs(psi.summary$coefficients[2,"Std. Error"] /
                            psi.summary$coefficients[2,"Estimate"])))
beta.pi.sd = abs(c(beta.pi.mean*beta.pi.cv))
beta.pi.mean[1] = 0

eivModel.data = list(n=length(x), x=x, ux=ux, r=r, ur=ur,
                     betaM=beta.pi.mean, betaS=beta.pi.sd,
                     rhoM=r, rhoS=3*ur,
                     gammax=median(ux))

eivModel.fit = stan(model_code = eivModel,
                    data = eivModel.data,
                    control=list(adapt_delta=0.985, max_treedepth=12),
                    warmup=50000, iter=250000,
                    chains=4, cores=4, thin=10)

print(eivModel.fit, digits=5)
```

Figure A2. R code that executes the Stan code listed in figure A1, which is assumed to have been assigned to `eivModel` as a character string. The vectors `x` and `ux` have the amount fractions and associated standard uncertainties listed in the second and third columns of table 1, and `r` and `ur` have the corresponding values for the ratios of instrumental indications listed in the same table.

```
data {
  int n;                    // Number of observations
  vector[n] y;              // Measured deltas for standards
  vector<lower=0>[n] uy;    // Standard uncertainties for y
  vector[n] nuy;            // No. of DFs supporting uy
  vector[n] x;              // Calibrated deltas for standards
  vector<lower=0>[n] ux;    // Standard uncertainties for x
  vector[n] nux;            // No. of DFs supporting ux
  vector[2] betaM;          // Prior means for beta
  vector[2] betaS;          // Prior SDs for beta
  vector[n] etaM;           // Prior means for eta
  vector<lower=0>[n] etaS; // Prior SDs for eta
  real<lower=0> gammax;     // Prior median for taux
  real<lower=0> gammay;     // Prior median for tauy
  real<lower=0> alphax2;    // Prior median for sigmax2
  real<lower=0> alphay2; } // Prior median for sigmay2

transformed data {
  vector<lower=0>[n] ux2;
  vector<lower=0>[n] uy2;
  ux2 = square(ux);
  uy2 = square(uy); }

parameters {
  vector[2] beta;      // Regression coefficients
  vector[n] eta;       // True values for measured deltas
  real<lower=0> taux; // Dark uncertainty for calibrated deltas
  real<lower=0> tauy; // Dark uncertainty for measured deltas
  vector<lower=0>[n] sigmax2;   // Variances of errors in x
  vector<lower=0>[n] sigmay2; } // Variances of errors in y

transformed parameters {
  vector[n] xi; // True values of calibrated deltas
  vector<lower=0>[n] lambdax; // Rates of gamma likelihoods for ux2
  vector<lower=0>[n] lambday; // Rates of gamma likelihoods for uy2
  for (j in 1:n) {xi[j] = beta[1] + beta[2]*eta[j];} // True x
  for (j in 1:n) {lambdax[j] = nux[j]/(2*sigmax2[j]);};
  for (j in 1:n) {lambday[j] = nuy[j]/(2*sigmay2[j]);}; }

model {
  beta ~ normal(betaM, betaS);  // Prior for beta
  taux ~ cauchy(0, gammax);     // Prior for taux
  tauy ~ cauchy(0, gammay);     // Prior for tauy
  eta ~ normal(etaM, etaS);     // Prior for eta
  sigmax2 ~ cauchy(0, alphax2); // Prior for sigmax2
  sigmay2 ~ cauchy(0, alphay2); // Prior for sigmay2
  x ~ normal(xi, sqrt(sigmax2 + square(taux)));  // Likelihood for x
  y ~ normal(eta, sqrt(sigmay2 + square(tauy))); // Likelihood for y
  ux2 ~ gamma(nux/2, lambdax);                   // Likelihood for ux2
  uy2 ~ gamma(nuy/2, lambday); }                 // Likelihood for uy2
```

Figure A3. Stan code that implements MCMC for the model in equation (4) that was fitted to the isotope delta values in table 2. The vectors `x`, `ux` and `nux` have the values listed in the second, third, and fourth columns of this table. The vectors `y`, `uy` and `nuy` have the values listed in the fourth, fifth, and sixth columns. The values of `x` are plotted along the vertical axis in figure 6, and the values of `y` are plotted along the horizontal axis. `taux` represents the standard deviation, τ_ψ, of the $\{\psi_j\}$ in equation (4), and `tauy` represents τ_φ.

situation, a smaller uncertainty does not signify higher quality, being the result of ignoring contributions from important sources of uncertainty.

The measurement discussed in section 6 employs multipoint calibration involving eight standards, a much larger number than most laboratories use in routine work supporting verification of compliance with WADA rules for athletes participating in sports events where such rules apply. And it also demonstrates a higher level of sophistication, both in analytical chemistry and in data reductions, than are common in such routine work, which focuses on differences of isotope delta values between paired target and endogenous steroids, rather than aim to produce a determination that is traceable to a suitable standard of reference, VPDB in this case.

```
require(rstan)

beta.pi.mean = c(0, 1)
beta.pi.sd = c(10, 0.5)

eivModel.data = list(n=length(x), x=x, ux=ux, y=y, uy=uy,
                     nux=dfux, nuy=dfuy,
                     betaM=beta.pi.mean, betaS=beta.pi.sd,
                     etaM=y, etaS=3*uy,
                     alphax2=median(ux^2),
                     alphay2=median(uy^2),
                     gammax=median(abs(residuals(lm(x~y)))),
                     gammay=median(abs(residuals(lm(y~x)))))

eivModel.fit = stan(model_code = eivModel,
                    data = eivModel.data,
                    control=list(adapt_delta=0.975),
                    warmup=1000000, iter=2000000,
                    chains=4, cores=4, thin=15)

print(eivModel.fit, digits=5)
```

Figure A4. R code that executes the Stan code listed in figure A3, which is assumed to have been assigned to `eivModel` as a character string.

```
eivModel.post = rstan::extract(eivModel.fit)
betaTILDE = eivModel.post$beta
tauxTILDE = eivModel.post$taux
tauyTILDE = eivModel.post$tauy

deltaNMIA.T   = -26.87
udeltaNMIA.T =    0.124/sqrt(8)
nu.T          =   7

K = nrow(betaTILDE)
etaPRED = deltaNMIA.T + rnorm(K, mean=0, sd=tauyTILDE) +
          udeltaNMIA.T * rt(K, df=nu.T) / sqrt(nu.T/(nu.T-2))
thetaPRED = rnorm(K, mean=0, sd=tauxTILDE)
for (k in 1:K) {
  thetaPRED[k] = thetaPRED[k] +
                 betaTILDE[k,1] + betaTILDE[k,2]*etaPRED[k] }

require(robustbase)
round(c(thetaPRED=huberM(thetaPRED)$mu, "u(thetaPRED)"=Qn(thetaPRED),
      quantile(thetaPRED, probs=c(0.025, 0.975))), 2)
```

Figure A5. R code that predicts the value of $\delta_{\mathrm{VPDB}}(^{13}\mathrm{C})$ in the urine sample and that evaluates the associated uncertainty.

The modified EIV model is most conveniently and safely fitted using a Bayesian formulation and MCMC sampling, because this addresses the challenge posed by the typically small number of data points per parameter being estimated, and ensures a thorough exploration of the space of possible values for the parameters before developing specific estimates for them.

The Bayesian approach offers the additional advantage of producing all the ingredients necessary for the subsequent uncertainty evaluations, using the results from MCMC sampling, not only for the EIV regression line, but also for the DoEs in the case of CCQM-K53, and for the isotope delta value assigned to the testosterone in the urine sample.

When DoEs are in scope, as they are for the illustrative reanalysis of the measurement results from CCQM-K53, and they are determined by a KCRF computed using the EIV methods proposed in this contribution, then they are estimated by the 'vertical' residuals defined above and illustrated in figure 2.

Classical EIV regression models can be non-identifiable [35], a shortcoming that the Bayesian approach overcomes with superlative ease and elegance when it uses proper prior distributions, as we have done throughout. A Bayesian formulation also effectively regularizes what can be an ill-posed

problem, or a problem with multiple solutions, and delivers the solution that is most likely in light of the data and with due consideration for relevant, prior knowledge about the problem.

Reliance on statistical models that are able to quantify excess variability, which throughout we call 'dark uncertainty' but that is also known as heterogeneity or mutual inconsistency, makes interlaboratory studies, KCs in particular, more inclusive than when the only way of coping with excess variability involved setting some measurement results aside, often for no better reason than their appearing not to belong with the bulk of the others. Proper accounting for excess dispersion makes these comparisons also more rigorous and realistic. The same enhanced inclusivity also facilitates reliance on multi-point calibration in challenging applications in analytical chemistry like those illustrated in section 6.

Acknowledgments

The authors are much indebted to their NIST colleagues Dr Michael Nelson and Dr Adam Pintar for their incisive reviews of this contribution, and for their generous and copious suggestions for improvement. The authors also express their appreciation to Dr Joële Viallon (BIPM, Sèvres, France) for asking many perceptive questions about the aspects of this proposal that impact gas metrology, which greatly stimulated its development and the explanations we offer in this contribution. We thank Dr Mark Lewin (National Measurement Institute, Australia) for facilitating the development of the example involving measurements of isotope delta values. Finally, the authors thank the members of the Gas Analysis Working Group of the CIPM Consultative Committee for Amount of Substance (Metrology in Chemistry and Biology), for considering and discussing this new method. The measurement results for testosterone used in subsection 6.2 are deliverables from a research project (20000R317) funded by the Partnership for Clean Competition Research Collaborative (Colorado Springs, CO, USA) and awarded to the National Measurement Institute, Australia. The content of this publication does not necessarily reflect the views or policies of the Research Collaborative.

Appendix A. Stan and R computer codes

The Stan code listed in figure A1, and the R code listed in figure A2, were used to fit the model described in sections 3 and 4 to the measurement results from CCQM-K53 that are listed in table 1. The R code in figure A2 assumes that the Stan code in figure A1 has been assigned to the R variable `eivModel` as a character string, including the line breaks.

The Stan code listed in figure A3, and the R code listed in figure A4, were used to fit the model described in subsection 6.1 to the measurement results for $\delta(^{13}\mathrm{C})$ that are listed in table 2. The R code in figure A4 assumes that the Stan code in figure A3 has been assigned to the R variable `eivModel` as a character string, including the line breaks.

The R code listed in figure A5 is used to produce the results described in subsection 6.2: the value of $\delta_{\mathrm{VPDB}}(^{13}\mathrm{C})$ for the sample of interest, and the associated uncertainty, expressed both as the standard measurement uncertainty and as a 95% credible interval for the corresponding true value.

ORCID iDs

Christina E Cecelski https://orcid.org/0000-0001-6782-3106
Fong-Ha Liu https://orcid.org/0000-0002-6572-251X
Juris Meija https://orcid.org/0000-0002-3349-5535
Antonio Possolo https://orcid.org/0000-0002-8691-4190

References

[1] Carpenter B *et al* 2017 Stan: a probabilistic programming language *J. Stat. Softw.* **76** 1–32
[2] Carroll R J, Ruppert D, Stefanski L A and Crainiceanu C M 2006 *Measurement Error in Nonlinear Models—A Modern Perspective* 2nd edn (Boca Raton, FL: CRC Press)
[3] Cawley A, Trout G, Kazlauskas R, Howe C and George A 2008 Carbon isotope ratio (δ^{13}C) values of urinary steroids for doping control in sport *Steroids* **74** 379–92
[4] Cooper H, Hedges L V and Valentine J C (ed) 2019 *The Handbook of Research Synthesis and Meta-Analysis* 3rd edn (New York: Russell Sage Foundation Publications)
[5] Coplen T and Shrestha Y 2017 Isotope-abundance variations and atomic weights of selected elements: 2016 (IUPAC Technical Report) *Pure Appl. Chem.* **88** 1203–24
[6] Cox M G and Harris P M 2012 The evaluation of key comparison data using key comparison reference curves *Metrologia* **49** 437–45
[7] Fuller W A 1987 *Measurement Error Models* (New York: Wiley)
[8] Gelman A, Carlin J B, Stern H S and Rubin D B 2003 *Bayesian Data Analysis* 2nd edn (Boca Raton, FL: CRC)
[9] Geweke J 1992 Evaluating the accuracy of sampling-based approaches to calculating posterior moments *Bayesian Statistics 4: Proceedings of the Fourth Valencia International Meeting* ed J M Bernado, J O Berger, A P Dawid and A F M Smith (Oxford: Clarendon) pp 169–93
[10] Gröning M 2007 *Reference Sheet for Reference Materials NBS 22, IAEA-CH-3, IAEA-CH-6, IAEA-CH-7, USGS24* Vienna, AustriaInternational Atomic Energy Agency (IAEA) nucleus.iaea.org/sites/ReferenceMaterials/
[11] Gross J 2017 *Mass Spectrometry: A Textbook* 3rd edn (Berlin: Springer)
[12] Guenther F R and Possolo A 2011 Calibration and uncertainty assessment for certified reference gas mixtures *Anal. Bioanal. Chem.* **399** 489–500
[13] Hartung J, Knapp G and Sinha B K 2008 *Statistical Meta-Analysis with Applications* (New York: Wiley)
[14] Hein S, Philipp R, Duewer D, Lippa K and Toman B 2013 Final report on CCQM-K79: comparison of value-assigned CRMs and PT materials: ethanol in aqueous matrix *Metrologia* **50** 08005
[15] Higgins J P T and Thompson S G 2002 Quantifying heterogeneity in a meta-analysis *Stat. Med.* **21** 1539–58
[16] Higgins J P T, Thompson S G, Deeks J J and Altman D G 2003 Measuring inconsistency in meta-analyses *Br. Med. J.* **327** 557–60
[17] Higgins J P T, Thomas J, Chandler J, Cumpston M, Li T, Page M J and Welch V A (ed) 2019 *Cochrane Handbook for Systematic Reviews of Interventions* 2nd edn (New York: Wiley)

[18] Hélie J-F, Adamowicz-Walczak A, Middlestead P, Chartrand M M G, Mester Z and Meija J 2021 Discontinuity in the realization of the Vienna Peedee Belemnite carbon isotope ratio scale *Anal. Chem.* **93** 10740–3
[19] Fernández i Marín X 2016 ggmcmc: analysis of MCMC samples and Bayesian inference *J. Stat. Softw.* **70** 1–20
[20] Iavetz R 2021 *NMIA MX018: Steroid Mixtures Certified for Carbon Isotope Delta Value* (North Ryde, NSW, Australia: National Measurement Institute) measurement.gov.au
[21] International Standard ISO 6143:2001(E) 2001 *Gas Analysis—Comparison Methods for Determining and Checking the Composition of Calibration Gas Mixtures* (Geneva, Switzerland: International Organization for Standardization (ISO))
[22] Joint Committee for Guides in Metrology 2020 *Guide to the Expression of Uncertainty in Measurement: VI. Developing and Using Measurement Models* (Sèvres, France: International Bureau of Weights and Measures (BIPM)) 2020. BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP and OIML, JCGM GUM-6
[23] Kioussi M K, Angelis Y S, Cawley A T, Koupparis M, Kazlauskas R, Brenna J T and Georgakopoulos C G 2011 External calibration in gas chromatography-combustion-isotope ratio mass spectrometry measurements of endogenous androgenic anabolic steroids in sports doping control *J. Chromatogr.* A **1218** 5675–82
[24] Koepke A, Lafarge T, Possolo A and Toman B 2017 Consensus building for interlaboratory studies, key comparisons, and meta-analysis *Metrologia* **54** S34–62
[25] Lee J *et al* 2010 Final report on international key comparison CCQM-K53: oxygen in nitrogen *Metrologia* **47** 08005
[26] Lunn D, Spiegelhalter D, Thomas A and Best N 2009 The BUGS project: evolution, critique and future directions *Stat. Med.* **28** 3049–67
[27] Maechler M *et al* 2021 *robustbase: Basic Robust Statistics* http://CRAN.R-project.org/package=robustbase R package version 0.93-9
[28] Mana G 2021 Interlaboratory consensus building *Metrologia* **58** 055002
[29] Merkatas C, Toman B, Possolo A and Schlamminger S 2019 Shades of dark uncertainty and consensus value for the Newtonian constant of gravitation *Metrologia* **56** 054001
[30] Munton E, Liu F-H, John Murby E and Brynn Hibbert D 2012 Certification of steroid carbon isotope ratios in a freeze-dried human urine reference material *Drug Test. Anal.* **4** 928–33
[31] Possolo A 2015 *Simple Guide for Evaluating and Expressing the Uncertainty of NIST Measurement Results* (Gaithersburg, MD: National Institute of Standards and Technology) NIST Technical Note 1900
[32] Possolo A, van der Veen A M H, Meija J and Hibbert D B 2018 Interpreting and propagating the uncertainty of the standard atomic weights (IUPAC Technical Report) *Pure Appl. Chem.* **90** 395–424 IUPAC Technical Report
[33] Possolo A, Koepke A, Newton D and Winchester M R 2021 Decision tree for key comparisons *J. Res. Natl. Inst. Stand.* **126** 126007
[34] R Core Team 2021 *R: A Language and Environment for Statistical Computing* (Vienna, Austria: R Foundation for Statistical Computing) https://R-project.org/
[35] Reiersøl O 1950 Identifiability of a linear relation between variables which are subject to error *Econometrica* **18** 375–89
[36] Rousseeuw P J and Croux C 1993 Alternatives to the median absolute deviation *J. Am. Stat. Assoc.* **88** 1273–83
[37] Rukhin A L 2003 Two procedures of meta-analysis in clinical trials and interlaboratory studies *Tatra Mt. Math. Publ.* **26** 155–68
[38] Rukhin A L 2009 Weighted means statistics in interlaboratory studies *Metrologia* **46** 323–31
[39] Rukhin A L 2013 Estimating heterogeneity variance in meta-analysis *J. R. Stat. Soc.* B **75** 451–69
[40] Thompson M and Ellison S L R 2011 Dark uncertainty *Accred. Qual. Assur.* **16** 483–7
[41] Toman B and Possolo A 2009 Laboratory effects models for interlaboratory comparisons *Accred. Qual. Assur.* **14** 553–63
[42] WADA Science/IRMS Working Group 2022 *Detection of Synthetic Forms of Prohibited Substances by GC/C/IRMS* (Montreal, Canada: World Anti-Doping Agency) WADA Technical Document TD2022IRMS
[43] WADA 2021 *Prohibited List: World Anti-Doping Code International Standard* (Montreal, Quebec, Canada: World Anti-Doping Agency)
[44] Westwood S *et al* 2021 Mass fraction assignment of bisphenol-A high purity material *Metrologia* **58** 08015