Accreditation and Quality Assurance
https://doi.org/10.1007/s00769-024-01614-w

RESEARCH

Value assignment and uncertainty evaluation for certified reference gas mixtures

Christina E. Cecelski[1] · Jennifer Carney[1] · Antonio Possolo[1]

Received: 17 January 2024 / Accepted: 22 July 2024

Abstract
The procedures used to assign values to certified reference gas mixtures and to evaluate their associated uncertainties, which are described in ISO 6143, and that were variously improved by Guenther and Possolo (Anal Bioanal Chem 399:489–500, 2011. 10.1007/s00216-010-4379-z), are further enhanced by the following developments: (i) evaluating and propagating uncertainty contributions derived from comparisons with historical reference gas mixtures of similar nominal composition; (ii) recognizing and quantifying mutual inconsistency (dark uncertainty) between primary standard gas mixtures used for calibration; (iii) employing Bayesian procedures for calibration, value assignment, and uncertainty evaluations; and (iv) employing state-of-the-art methods of meta-analysis to combine cylinder-specific measurement results. These developments are illustrated in examples of certification of two gas mixture Standard Reference Materials developed by the National Institute of Standards and Technology (NIST, USA). These examples serve only to demonstrate the methods described in this contribution and do not replace any official measurement results delivered in the certificates of any reference materials developed by NIST.

Keywords Calibration · Gas mixture · Certified reference material · Errors-in-variables regression · Maximum likelihood · Bayesian methods · Homogeneity · Dark uncertainty · Prediction

Introduction

The certification of Standard Reference Materials® (SRMs) developed by the National Institute of Standards and Technology (NIST) of the U.S. involves the careful and accurate characterization of a chemical or physical property (measurand) of a material, evaluation of the associated measurement uncertainty, and assessment of the stability of the measurement result during a specified period of validity of the certification [2].

NIST's portfolio of gas mixture SRMs [6] supports measurements made by government, industry, and academia, responding to international, national, and state regulations and agreements, addressing environmental concerns, characterizing global climate change, and ensuring fair trade of natural and other gases. The collection of NIST SRMs with reference gas mixtures includes SRMs in series 1600, 1700, 2600, and 2700, which are listed in the online NIST Store.

The Gas Sensing Metrology Group (of the Chemical Sciences Division in NIST's Material Measurement Laboratory) provides certified gas mixture SRMs, in aluminum cylinders, to customers worldwide. Since the compositions of these mixtures, usually expressed as amount fractions of specified analytes, are traceable to the international system of units (SI), these SRMs provide a link in the chain of comparisons that establishes traceability to the SI of the measurement results obtained by NIST customers that use these SRMs as calibrants.

A gas mixture SRM typically comprises a set of cylinders (*lot*) of nominally identical composition, filled by a specialty gas producer, and purchased and analyzed by

All authors have contributed equally to this work.

✉ Christina E. Cecelski
christina.cecelski@nist.gov

Jennifer Carney
jennifer.carney@nist.gov

Antonio Possolo
antonio.possolo@nist.gov

[1] National Institute of Standards and Technology, Gaithersburg, MD, USA

Published online: 09 September 2024

NIST. The certification of its composition includes the following key tasks:

(i) Assigning a value to each cylinder in the lot, using an analysis function derived from measurements of primary standard mixtures (PSMs) prepared gravimetrically at NIST, and evaluating the uncertainty that surrounds such value;
(ii) Determining whether the lot is sufficiently homogeneous to warrant assigning a single value to all of its cylinders, or whether different cylinders should be assigned different values; and
(iii) Assessing the stability of the lot, by comparing measurements of its composition made at different epochs over a suitably long period of time.

The analysis functions used to assign values to these SRMs have been built during calibration as described by Guenther and Possolo [12], who extended the method described in ISO 6143 [17] to recognize the typically small number of instrumental readings obtained for each PSM used as a calibrant. This contribution describes further enhancements and refinements that have been developed in the intervening thirteen years, of which the following are particularly noteworthy:

- Employing rigorous model selection criteria to choose the form of the analysis function that is used for value assignment (section Analysis function);
- Evaluating and propagating uncertainty contributions based on intercomparisons between the new SRM being developed and previously certified reference gas mixtures with the same nominal composition as the new SRM (section Analysis function);
- Assessing the homogeneity of the measurement results for the individual cylinders in a lot, using statistical methods that are commonly used to determine whether the results of independent measurements of the same measurand are mutually consistent (section Lot homogeneity);
- Applying state-of-the-art consensus-building techniques to combine cylinder-specific measurement results into a single measurement result for the whole lot, when such combination is warranted (section Consensus value and uncertainty evaluation).

The techniques described in this contribution are illustrated using data obtained during certification of two particular SRM lots: one, which we will refer to as SN, is a mixture of sulfur dioxide in nitrogen; the other, which we will refer to as PA, is a mixture of propane in air. The measurands are the amount fractions of the analyte: sulfur dioxide in the case of SN, and propane in the case of PA. SN is used to illustrate all the steps of the data reduction workflow as they are presented, while the particular challenges posed by PA, and how they were addressed, are described in section Lot PA's calibration challenge.

Over the years, NIST has developed several lots of mixtures similar to SN and PA, with various nominal amount fractions of the same analytes. For example, SRMs 1661–1664, 1693, 1694, and 1696, all deliver certified values for amount fractions of sulfur dioxide in nitrogen, and SRMs 1665–1669 do likewise for amount fractions of propane in air.

The results for SN and PA are presented for purposes of illustration only and shall not be used as replacement for the certified values listed in the certificates of any SRMs that NIST has developed and delivered to customers, not even for those whose nominal or actual compositions are very close to SN's or PA's, as presented here.

Section Data acquisition and data reduction workflow summarizes the acquisition of instrumental readings, and the subsequent data reduction workflow. Section Diagnostics and data selection describes diagnostic and data selection procedures used to ensure the quality of the experimental data and the reliability of the results both for the construction of the analysis function and for value assignment to individual cylinders.

Section Analysis function reviews the criteria employed to select a model for the analysis function, discusses *dark uncertainty* that can be uncovered during construction of this function, and presents a novel, Bayesian statistical procedure to build the analysis function and to evaluate the associated uncertainty, which includes consideration of historical information encapsulated in a prior distribution for what we call *historical uncertainty*.

Section Measurement results for individual cylinders explains how the analysis function is used to assign values of the measurand to individual cylinders, and the corresponding uncertainty evaluations. Section Lot homogeneity presents the technique used to assess the homogeneity of the lot. And if the lot indeed is deemed to be sufficiently homogeneous for the purpose that the SRM is intended to serve, then a single, consensus value is computed as presented in section Consensus value and uncertainty evaluation, and it is assigned to the whole lot. Also, a predictive interval is produced that, with specified probability, is believed to include the true amount fraction of the analyte in any cylinder in the lot that is shipped to a customer. Section Conclusions summarizes lessons learned and the corresponding best practices.

Notation and terminology: A symbol like $\{x_j\}$ is shorthand for $x_1, \ldots, x_n$, when the order of these x_js is immaterial. Ordered sets appear in boldface, as in $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)$, which denotes the coefficients of a polynomial used as analysis function. If G denotes a polynomial whose argument is r and whose coefficients are the elements of $\boldsymbol{\beta}$, then we write $G(r, \boldsymbol{\beta})$ to denote the value that the polynomial takes at r when the coefficients are the elements of $\boldsymbol{\beta}$. The maximum

likelihood estimate of a parameter θ is denoted $\widehat{\theta}$, and Bayesian estimates (posterior means, posterior medians, posterior standard deviations, etc.) are denoted $\widetilde{\theta}$. We use the term "standard error" as it is commonly understood in statistics: the standard deviation of a function of observations obtained under conditions of repeatability or of reproducibility, or, in the language of the Guide to the Expression of Uncertainty in Measurement (GUM) [19], as a Type A evaluation of the standard uncertainty of the same function.

Data acquisition and data reduction workflow

We employ a variety of analytical instruments to measure the amount fractions of different analytes in gas mixture SRMs: for example, a pulsed fluorescence analyzer to measure sulfur dioxide, or a gas chromatograph with flame ionization detection (GC-FID) to measure propane [6, Table 1].

For our measurements, neither do we use the instrumental readings directly, nor do we calibrate the instrument once and for all and then rely on the amount fractions that it outputs. Instead, we begin by designating a particular cylinder, in the lot of cylinders containing the reference mixture, as the *lot standard* (LS), which will serve as the analytical control, and will be sampled repeatedly by the instrument during the data acquisition phase of the certification process.

During data acquisition, we sample and obtain instrumental readings for the PSMs, and for the cylinders that will become the units in an SRM lot, in alternation with instrumental readings for the LS. Then, we form ratios between the readings for the PSMs and approximately contemporaneous readings for the LS and do the same with the readings for the SRM cylinders. Utilizing these ratios enables compensation for any instrumental drift that may have occurred throughout the period when the PSMs and the cylinders in the lot are sampled [6].

Typically, one to three cylinders are sampled and analyzed between consecutive samplings and instrumental analyses of the LS, in a random order determined by a computer-operated gas analysis system (COGAS). The corresponding ratios are formed using a cylinder's or a PSM's instrumental reading as numerator, and the drift-corrected, approximately contemporaneous reading of the LS as denominator. Subsection Correlations between Ratios quantifies the correlations between these ratios, induced by their sharing instrumental indications obtained for the LS, and discusses the potential impact of such correlations.

The raw data comprise all ratios of instrumental indications obtained from multiple samplings of each PSM and of each cylinder in the SRM lot. For example, for SN, 11 replicated determinations of the ratio were made for each of four PSMs, and either 12 or 18 replicated determinations of the ratio were made for each of 32 cylinders. (However, typically, we make between 6 and 10 replicated determinations of each ratio.)

Both the calibration that yields the analysis function used for value assignment, and the actual value assignment and uncertainty evaluation are done offline, after all the relevant raw data have been collected.

The workflow to produce an SRM with either a single certified value of the measurand for the whole lot, or with an individually certified value for each cylinder in the lot, involves the following steps of data reduction applied to the ratios described above:

W1 Identification and possible removal of apparently anomalous ratios of instrumental indications, either because they are deemed outliers, or because they correspond to a setting of a controllable, experimental factor (for example, day of data acquisition, or port of the COGAS system) that seems to be compromised (subsection Identifying anomalous ratios);

W2 Selection of the form for the analysis function, usually a polynomial of low degree, estimation of its coefficients, and evaluation of the associated uncertainty, which includes a contribution from historical uncertainty when relevant historical lot standards are available (section Analysis function);

W3 Value assignment and uncertainty evaluation for the individual cylinders in the lot, recognizing the contributions from all identified sources of uncertainty, including the contributions from any dark uncertainty (subsection Two sources of dark uncertainty) that may be detected, and from historical uncertainty (section Measurement results for individual cylinders);

W4 Assessment of the homogeneity of the lot (section Lot homogeneity) according to whether:

- The probability distribution of values of the measurand assigned to the different cylinders is unimodal (that is, the histogram of these values has a single maximum, or "peak") or multimodal;
- The measurement results (cylinder-specific values of the measurand and their associated uncertainties) are mutually consistent as judged by Cochran's Q [8] and Welch's F [40] tests of homogeneity.

W5 Value assignment and uncertainty evaluation for the whole lot provided the lot has been found to be sufficiently homogeneous (section Consensus value and uncertainty evaluation); and

W6 Value assignment and uncertainty evaluation for one or several lot standards (section Analysis function), which will be preserved and used in future stability tests and also for evaluations of historical uncertainty (subsection Two sources of dark uncertainty).

Diagnostics and data selection

Identifying anomalous ratios

Figure 1 shows boxplots of the ratios for the PSMs used to build the analysis function for lot SN, and for the cylinders in the same lot, exposing the few, apparently anomalous ratios that are candidates for removal as outliers.

For each boxplot: the thick, horizontal line segment across the box indicates the median of the ratios for the corresponding cylinder; half of the ratios lie between the top and bottom of the box; the height of the box is the *interquartile range* (IQR); the whiskers extend from the top and bottom of the box to the outermost ratio not more than $1.5 \times$ IQR away from the bottom or top of the box; the red dots indicate ratios more than $1.5 \times$ IQR away from the bottom or top of the box.

According to Tukey's rules [36] for the construction of boxplots, if the replicates are a sample from a Gaussian distribution, then the corresponding boxplot will (incorrectly) flag about 0.7 % of the replicates as apparent outliers. If the replicates are a sample from a probability distribution whose tails are heavier than Gaussian tails, then the proportion of replicates thus flagged will be greater than that.

We set apparent outliers aside only for substantive cause, based on scientific judgment and concrete findings of such causes. Erring on the side of caution, by not setting aside all the apparent outliers that might warrant such action, is counterbalanced by our reliance on robust statistical methods that are part of the process of value assignment (item (L2) in section Measurement results for individual cylinders), which provide safeguards against the undue influence of any apparently anomalous ratio that will not have been set aside.

Correlations between ratios

As mentioned in section Data acquisition and data reduction workflow, correlations can arise when the same instrumental reading of the LS is used to form consecutive ratios for the same or different cylinders (i.e., PSMs or SRM units). The acquisition of data for the construction of the analysis function used for lot SN followed this repeating pattern, $L_1P_1L_2P_2L_3$ …, where the $\{L_i\}$ denote replicated instrumental readings for the LS, and the $\{P_i\}$ denote replicated instrumental readings for the same PSM. Since the ratios corresponding to P_1 and P_2 are based on interpolated readings of the adjacent readings for the LS, they share the same measurement error that affects L_2, hence they are correlated. For lot PA, the repeating pattern is $L_1A_1B_1L_2$ …, where the $\{A_i\}$ and the $\{B_i\}$ pertain to two different PSMs.

As explained in step (M6) of subsection Characterization of the analysis function, we use the averages of the replicates of the ratios to build the analysis function. For lot SN, each such average summarizes either 10 or 11 replicates. For lot

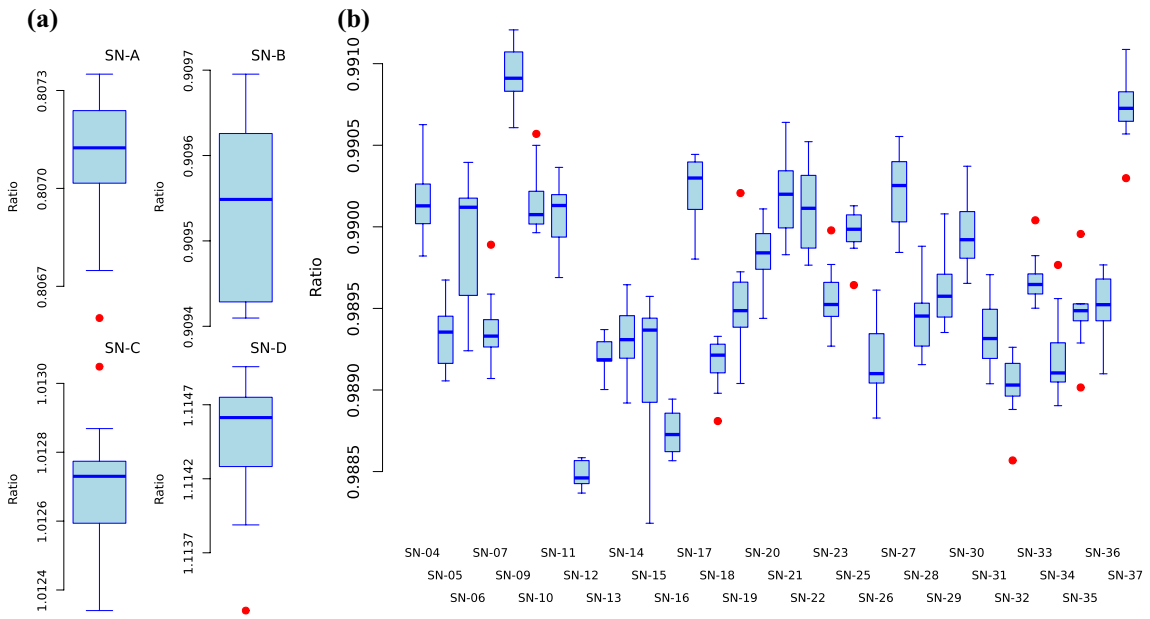

Fig. 1 LEFT PANEL: Boxplots (explained in the text) of ratios for PSMs used to build the analysis function for lot SN. RIGHT PANEL: Boxplots of ratios for the cylinders in the same lot. Each boxplot in the left panel summarizes 11 replicates of the ratios, while the boxplots in the right panel summarize either 12 or 18 replicates of the ratios for the cylinders. Red dots indicate apparently anomalous values in each batch (color figure online)

PA, each average ratio summarizes either 7 or 8 replicates. In this conformity, we have evaluated the correlations between averages of replicated ratios for the four PSMs used to build the analysis function for lot SN and done the same for the five PSMs used for lot PA, by application of a Monte Carlo method.

For lot SN, all of these correlations are between −0.019 and 0.024, and for lot PA they are between −0.020 and 0.022. When either set of correlations is propagated to the values assigned to the lot — which involves replacing the n univariate, rescaled and shifted Student's t distributions specified in (M6) of subsection Characterization of the analysis function, with a multivariate (rescaled and shifted) Student's t distribution with the appropriate correlation matrix — the resulting uncertainty associated with the lot value, which we call x_{SRM} in step (S1) of section Consensus value and uncertainty evaluation, is just about identical to its counterpart obtained assuming that all the correlations are zero.

For the same reasons, the replicates of the ratios for the individual cylinders in the lots also are correlated. These correlations can impact not only the uncertainty surrounding the lot value but also reduce the effective numbers of degrees of freedom supporting the evaluations of the uncertainty associated with the averages of the ratios.

Since there are so many more cylinders in either lot than there are PSMs underlying the corresponding analysis functions, the proportion of pairs of ratios that share the same instrumental reading of the LS is much smaller than for the PSMs, and in consequence, the correlation matrices for the cylinder-specific averages of the ratios have many zero entries. And the nonzero entries are of similarly small magnitudes as those, aforementioned, that were computed for the PSMs.

Furthermore, the impact of any nonzero correlations between average ratios for the cylinders in the lot is even smaller than the impact of nonzero correlations between average ratios for the PSMs because historical uncertainty (subsection Two sources of dark uncertainty) and between-cylinder differences make the largest contributions by far to the uncertainty associated with the lot value. For the same reasons, some over-estimation of the effective numbers of degrees of freedom is inconsequential in practice.

Gauging influence of experimental factors

The controllable experimental factors affecting the determinations of the ratios for the cylinders are, for each ratio: (a) the identity, Cyl, of the cylinder the ratio pertains to; (b) the Port of the COGAS system that the cylinder was connected to when the ratio was determined; (c) the Day when the ratio was determined; (d) its Break-Set, which is the set of instrumental indications obtained between consecutive changes of cylinder connections to the COGAS manifold; (e) its LS-Set, which is the set of ratios corresponding to instrumental indications obtained between consecutive readings for the LS; and (f) the specific sequence of acquisition of the replicated instrumental readings, which impacts the correlations between individual replicates of the ratios, and the correlations between the averages of these replicates, as already discussed in subsection Correlations between ratios. The LS-Set is nested within the Break-Set, and the Break-Set is either nested within or is identical to Day.

These factors are modeled as random effects, and their impact can be gauged in terms of the variance components that they contribute to the overall dispersion of the ratios. The variance component attributable to differences between cylinders provides a first indication about the homogeneity of the lot, which is addressed more rigorously as described in section Lot homogeneity. The variance components attributable to the other factors serve as diagnostics indicating potential anomalies during data acquisition: for example, whether a particular port of the COGAS manifold may have been malfunctioning, or whether the formation of the ratios may have been insufficient to correct for instrumental drift.

To evaluate the random effects aforementioned, we use a conventional, linear, Gaussian mixed effects model [25] fitted to the replicates of the cylinder values using the method of restricted maximum likelihood (REML) [31] as implemented in R function `lmer` defined in package `lme4` [1, 28].

The 384 replicates of the ratios from 32 cylinders in lot SN were arranged into 214 LS-Sets grouped into 36 Break-Sets, which were measured in the course of 9 Days, using two COGAS Ports. The QQ-plots in Fig. 2 suggest that this model is adequate for the ratios obtained during certification of lot SN, and Table 1 lists the standard deviations of the random effects. Only differences between cylinders seem to make a contribution to the overall dispersion of the replicates that is significantly larger than the dispersion of the residuals, thus not raising concerns about the influence of the experimental factors, but suggesting that the lot may not be homogeneous.

Analysis function

An analysis function, G, as defined in ISO 6143 [17], translates an instrumental indication, or a ratio r of instrumental indications in our case, into a value of the measurand, which is the amount fraction $x = G(r)$ of sulfur dioxide in lot SN. The analysis function used for lot SN was based on 10 or 11 replicates of each ratio obtained for each PSM (after removing any replicate that has been found to be unreliable, among those depicted in red in the left panel of Fig. 1) for each of $n = 4$ PSMs. Table 2 lists the calibration data used to build

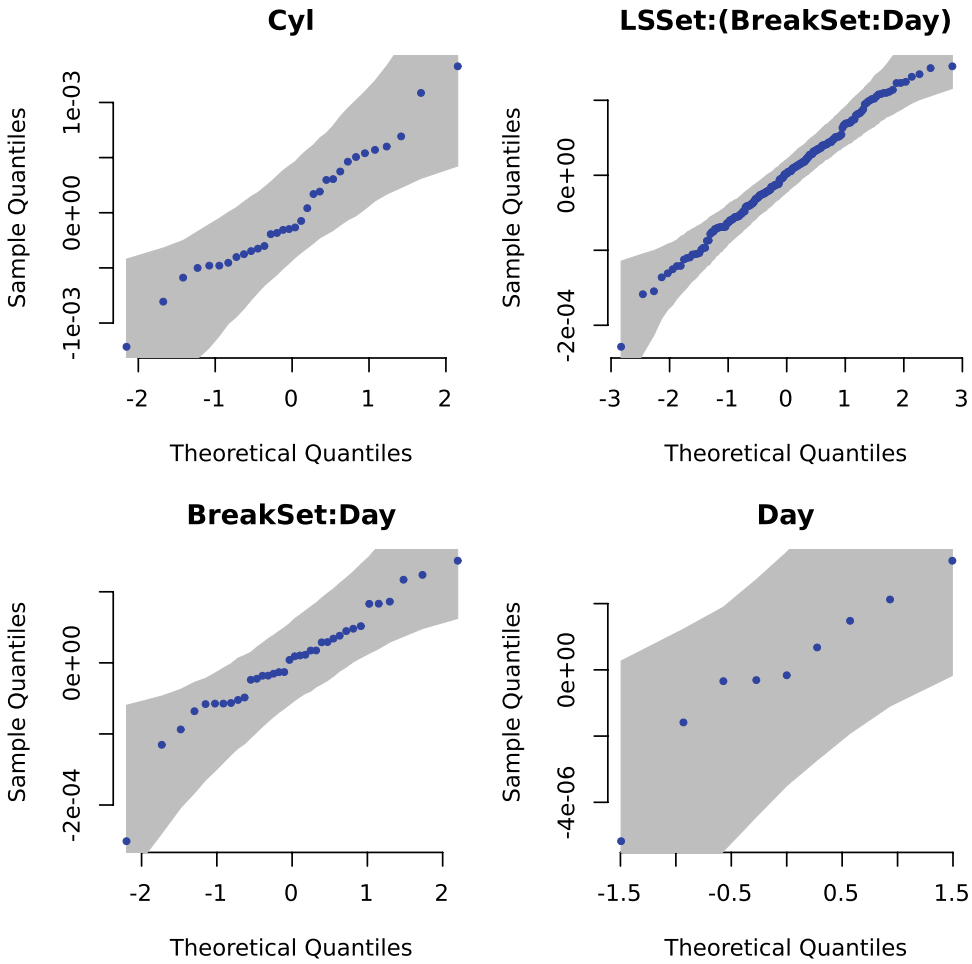

Fig. 2 QQ-plots [41] for the random effects and residuals corresponding to the linear, Gaussian mixed effects model fitted to 384 replicates of the ratios determined for the cylinders in lot SN. The fact that all the blue dots lie inside the 95% probability gray bands suggests that the assumptions are met that validate the model (color figure online)

the analysis function used for value assignment to the cylinders in lot SN.

We model the analysis function for our gas mixture SRMs as a polynomial of low degree $p-1$ for some integer $p>1$, whose coefficients are $\boldsymbol{\beta}=(\beta_1,\ldots,\beta_p)$. In general, however, the analysis function need not be a polynomial, in which case $\boldsymbol{\beta}$ denotes the parameters that identify the particular function that has been chosen to play the role of analysis function. The analysis function is fitted to the calibration data using an errors-in-variables (EIV) regression procedure [5, 11] comprising these simultaneous relations for each of $j=1,\ldots,n$ PSMs:

$$\xi_j = G(\rho_j,\boldsymbol{\beta}),\quad x_j=\xi_j+\lambda_j+\delta_j,\quad r_{i,j}=\rho_j+\epsilon_{i,j}, \tag{1}$$

where ξ_j is the true value of the amount fraction of the analyte in PSM j, ρ_j is the true value of the corresponding ratio of instrumental indications, x_j is the measured value of ξ_j, and $r_{1,j},\ldots,r_{m_j,j}$ are the m_j replicates of the measured ratio for PSM j. The model also involves assumptions about the analysis function, and about the measurement errors $\{\delta_j\}$ and $\{\epsilon_{i,j}\}$, which we describe next.

It is usually assumed that the true analysis function can be closely approximated by a polynomial G, whose estimate, $\widehat{G}$, is the analysis function used in practice. In other words, G is defined as $G(\rho,\boldsymbol{\beta})=\beta_1+\beta_2\rho+\cdots+\beta_p\rho^{p-1}$ where $\rho\geqslant 0$ denotes the true value of a ratio of instrumental indications. Typically, all p terms of a polynomial of degree $p-1$ are included in the model, but it is conceivable that, in some applications, not all powers of the argument will be included: for example, p can be 4 and the quadratic term may be excluded from the polynomial during the process of model selection described in subsection Selection of degree for analysis function polynomial. Clause 5.1.c of ISO 6143 [17] lists other functional forms that can be selected for G.

The errors $\{\delta_j\}$ and $\{\epsilon_{i,j}\}$ are usually assumed to be non-observable outcomes of independent random variables centered at 0. An assumption commonly made, which is implicit in the least squares criterion proposed in ISO 6143 [17] for fitting G to the calibration data, is that all these errors are like outcomes of mutually independent Gaussian random variables with possibly different standard deviations. We make the same assumption throughout this contribution

Table 1 Standard deviations of the random effects in the linear, Gaussian, mixed effects model fitted to the ratios obtained for the certification of lot SN.

Random effect	Std. dev.	Lwr95	Upr95	($/10^{-4}$µmol/mol)
Cyl	5.56	4.34	7.18	
LS-Set:(Break-Set:Day)	1.06	0.71	1.35	
Break-Set:Day	1.05	0.49	1.64	
Day	0.14	0.00	1.25	
Port	0.18	0.00	1.63	
Residual	1.61	1.44	1.83	

The columns headed Lwr95 and Upr95 list the endpoints of approximate 95 % confidence intervals for the standard deviations of the variance components, computed using the profile likelihood method [10]. Colons denote nesting, for example, Break-Set within Day. Only Cyl makes a contribution to the overall dispersion of the ratios that is significantly larger than the residual dispersion

and verify its adequacy for the calibration data that we use in each application, after removing any observations whose reliability is questionable.

$\{\lambda_j\}$ are PSM effects, assumed to be a sample from a probability distribution with mean 0 µmol/mol, and standard deviation τ_C, which is the component of dark uncertainty uncovered during calibration (hence the subscript "C"). If τ_C is greater than zero, then it means that the residuals $\{x_j - \xi_j\}$ are more dispersed than the uncertainties $\{u(x_j)\}$ intimate that they should be. The next subsection Two sources of dark uncertainty, discusses this τ_C and yet another source of dark uncertainty, both of which arise fairly commonly during the development of gas mixture SRMs.

Two sources for dark uncertainty

The term "dark uncertainty" was introduced by Thompson and Ellison [35] in the context of interlaboratory studies involving scalar measurands. Cecelski et al [7] extended the concept to make it meaningful also in the context of EIV regression, and both Cecelski et al [7] and Viallon et al [38] applied it to quantify "excessive" dispersion of measurement results used to build analysis functions in gas metrology.

Our justification for using the term in this contribution is the following: First, we select an analysis function that is adequate for the calibration data (amount fractions, associated uncertainties, and supporting numbers of degrees of freedom), but then we find out that the calibration data yet are, to a significant extent, inconsistent with the apparently best model for the analysis function. This inconsistency manifests itself in the residuals from the fit being appreciably larger than the reported uncertainties associated with the PSMs suggest that they should be, as the right-panels of Figs. 3 and 12 show.

Table 2 Data used to select the degree of the polynomial that will serve as analysis function for lot SN

PSM	x /(µmol/mol)	$u(x)$	r	$u(r)$	ν
A	806.16	0.25	0.8071269	0.0000550	9
B	907.04	0.26	0.9095385	0.0000335	10
C	1008.82	0.28	1.0126681	0.0000487	9
D	1108.30	0.46	1.1145633	0.000105	9

And here, the same as in an interlaboratory study or key comparison whose results are mutually inconsistent, the inconsistency can be resolved by putting into play an "extra" uncertainty component, which we rightfully call "dark uncertainty" honoring the spirit and the letter of its introduction by Thompson and Ellison [35].

Regardless of whether the provenance of the "extra" uncertainty can be easily identified or not, this uncertainty component only becomes apparent when measurements are intercompared for the purpose of fitting a curve to them to form the analysis function, or when we reanalyze historical lot standards.

The standard deviation, τ_C, of the PSM effects $\{\lambda_j\}$ in Equation (1) quantifies a source of dark uncertainty. In the present context, and for some SRM lots, for example SN, two sources of dark uncertainty can be identified, apparent in Fig. 3, which we explain next, following some preliminary remarks about historical lot standards.

SN is one of several SRM lots that NIST has developed with 1000 µmol/mol nominal amount fraction of sulfur dioxide in nitrogen. Each of these SRMs has had its own LS, with the role already described in section Data acquisition and data reduction workflow. NIST usually keeps the cylinders used as LSs, rather than sell them to customers, so that their composition can be measured at later dates, for sundry purposes, in particular for stability tests.

As of this writing, there are three such historical LSs with the same nominal amount fraction of sulfur dioxide as SN, which we denote X, Y, and Z in the right panel of Fig. 3. On the one hand, each of these historical LSs has the value, denoted x in the labels of the vertical axes in this figure, of the amount fraction of sulfur dioxide that was assigned to it when the corresponding SRM was last certified. On the other hand, ratios of instrumental indications from each of these historical LSs, relative to the LS for SN, were also determined as part of the workflow for SN. These ratios were then mapped into values of amount fraction of sulfur dioxide, $\hat{\xi}$, using the analysis function built for SN via maximum likelihood estimation of the EIV regression model specified in Equation (1).

Note that, in the right panel of Fig. 3, for PSMs C and D, the difference $x - \hat{\xi}$ is more than one standard uncertainty,

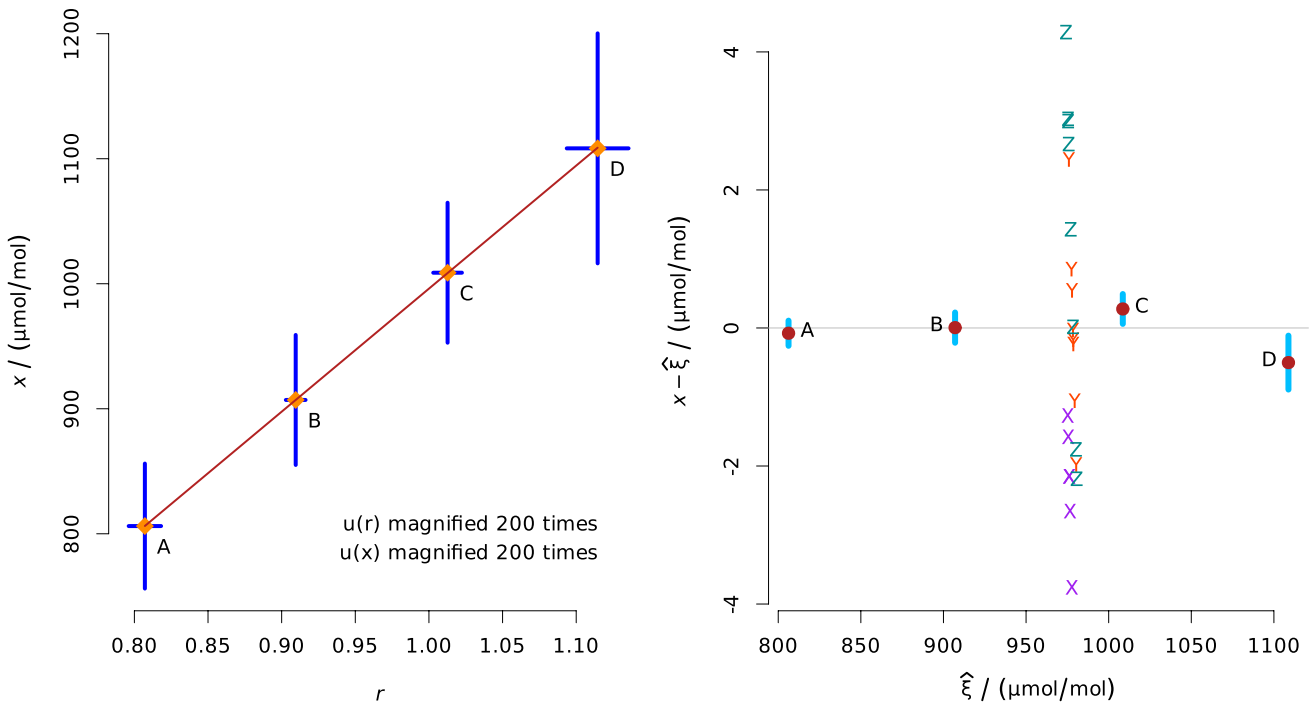

Fig. 3 LEFT PANEL: The (orange) solid diamonds represent the PSMs, which are labeled A, B, C, and D, and the blue segments represent $r \pm u(r)$ and $x \pm u(x)$, except that the uncertainties are magnified 200 times. The (red) sloping line represents the analysis function for lot SN (which is a polynomial of the first degree, hence $p = 2$). RIGHT PANEL: The solid (dark red) dots labeled A, B, C, and D represent the residual amount fractions of sulfur dioxide for the four PSMs used to build the analysis function for lot SN. The vertical (light blue) line segments represent these residuals plus or minus their associated standard uncertainties, which were evaluated using the parametric bootstrap [9]. The letters X, Y, and Z (of three different colors) represent replicated differences between the values, x, of the amount fraction of sulfur dioxide in three historical lot standards, and the estimates, $\hat{\xi}$, of their true values produced by the analysis function for lot SN (color figure online)

$u(x - \hat{\xi})$, above or below the horizontal line at 0 µmol/mol. This means that these two PSMs are (almost imperceptibly in the left panel of the figure, yet quite visibly in the right panel) "misaligned" relative to A and B. This misalignment translates into an estimate $\hat{\tau}_C = 0.21$ µmol/mol of a component of dark uncertainty. The subscript "C" serves as a reminder of the fact that this component of dark uncertainty is uncovered during calibration.

This $\hat{\tau}_C$, which was evaluated as described next, captures the dispersion of the points representing the PSMs around the EIV regression curve depicted in the left panel of the same figure, above and beyond the dispersion that the $u(x - \hat{\xi})$ intimate these points should have.

The value listed above for $\hat{\tau}_C$ was computed as $\sqrt{s^2 - g^2}$, where $s = 0.322$ µmol/mol denotes the standard deviation of the four differences $\{x_A - \hat{\xi}_A, x_B - \hat{\xi}_B, x_C - \hat{\xi}_C, x_D - \hat{\xi}_D\}$, and $g = 0.241$ µmol/mol denotes the geometric mean of the standard uncertainties associated with these differences, which were evaluated by application of the parametric bootstrap [9]. In Characterization of the analysis function, we will obtain a more reliable, model-based estimate of τ_C, which is listed in Table 4.

The ordinates of the letters X, Y, and Z, of three different colors, in the right panel of Fig. 3 represent replicated differences between the original estimates of the amount fraction of sulfur dioxide in the three available historical LSs, and the amount fractions estimated using the analysis function for SN. The abscissae of the same letters represent these amount fractions that were estimated based on the ratios relative to the LS for SN.

Both the ordinates and the abscissae of the letters depend on the measurements made of the PSMs that underlie the analysis function for the current lot of this SRM, and on the measurements made of the sets of PSMs that were used to build the analysis functions for the historical lots. The fact that, as a group, these letters are approximately centered (vertically) around the horizontal line at 0 µmol/mol, suggests that the analysis function built for SN is an approximately unbiased estimator of the amount fraction of sulfur dioxide in the historical LSs.

The root mean square (RMS) of the ordinates of the colored letters, $\hat{\tau}_H = 2.18$ µmol/mol, is an evaluation of a component of dark uncertainty different from $\hat{\tau}_C$ introduced above. τ_H quantifies the component of dark uncertainty reflecting what we call *historical uncertainty* (hence its

subscript "H"), an uncertainty that could not be perceived by examining the results for SN alone, but that becomes apparent when independent, original estimates of the amount fractions of the historical LSs are compared with the corresponding estimates produced for them by SN's analysis function. Therefore, the value of such τ_{H}, the aforementioned 2.18 µmol/mol, is an "external" evaluation of a component of uncertainty that should be recognized when evaluating the uncertainty surrounding the estimates of the measurand in SN using SN's analysis function.

This τ_{H} expresses not only lack of reproducibility when comparing gas mixtures nominally identical to the new mixture, but it also quantifies a variety of uncertainty contributions beyond mere lack of reproducibility: it can include, for example, differences between the sets of primary standards or the instrumentation used for calibration, and it can also include differences related to "aging" of historical lot standards, i.e. long-term instability of the gas mixtures over extended periods of time.

These two components of dark uncertainty, τ_{C} and τ_{H}, are combined in quadrature to produce $\tau = (\tau_{\mathrm{C}}^2 + \tau_{\mathrm{H}}^2)^{1/2}$, whose value will be assigned to the median of τ's prior probability distribution in the Bayesian procedure described in subsection Characterization of the analysis function, to fit the model in Equation (1) to the calibration data, thence to build the analysis function for the lot, and to evaluate the uncertainty associated with it.

This prior distribution for τ is the vehicle that we use to inject the knowledge about historical between-lot variability, similarly to how Lang et al [21] recognize and propagate historical information about between-method differences in the development of NIST SRMs that are single-element solutions or anion solutions.

Selection of degree for analysis function polynomial

Even though the construction of the analysis function, and the corresponding uncertainty evaluation (in particular recognizing the historical uncertainty), will be done using a Bayesian procedure described in subsection Characterization of the analysis function, for the sake of expediency, we select the degree of the analysis function polynomial using classical (non-Bayesian) statistical methods.

The first step taken to build the analysis function is to select n PSMs of similar composition as the lot being certified, whose range of amount fractions of the analyte, $x_1, \ldots, x_n$, brackets the nominal amount fraction of this lot. The standard uncertainties associated with the amount fractions of the analyte in the PSMs, $u(x_1), \ldots, u(x_n)$, evaluated during the gravimetry, are usually assumed to be based on very large (practically infinite) numbers of degrees of freedom, hence are treated as known constants.

For each of these PSMs, multiple ratios of instrumental indications, relative to the LS selected for the lot undergoing certification, are obtained. For the purpose of selecting the degree of the polynomial to be used for the analysis function, the replicates of the ratios for each PSM are summarized by their average and by the standard error of this average, which is the Type A evaluation of the standard uncertainty of an average of independent, identically distributed replicates.

The selection of the polynomial is thus based on n quintuplets $(x_1, u(x_1), r_1, u(r_1), \nu_1), \ldots, (x_n, u(x_n), r_n, u(r_n), \nu_n)$, where r_j is the average of m_j ratios for PSM j, $u(r_j)$ is the corresponding standard uncertainty, and $\nu_j = m_j - 1$ is the number of degrees of freedom that $u(r_j)$ is based on, for $j = 1, \ldots, n$. The data used to select the degree of polynomial for the analysis function of lot SN are listed in Table 2.

The analysis function G is a polynomial of degree $p-1$, for some integer $p \geqslant 2$, of the form $G(r, \boldsymbol{\beta}) = \beta_1 + \beta_2 r + \ldots \beta_p r^{p-1}$, where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ is the coefficients of the polynomial used as analysis function. The maximum likelihood estimates of $\boldsymbol{\beta}$, of the true values of the ratios, $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_n)$, and of τ_{C}, can be obtained by maximizing a function L with respect to $\boldsymbol{\beta}$, $\boldsymbol{\rho}$, and τ_{C}, such that

$$L(\boldsymbol{\beta}, \boldsymbol{\rho}, \tau_{\mathrm{C}}) = \sum_{j=1}^{n} \log \phi\left(\frac{x_j - G(\rho_j, \boldsymbol{\beta})}{\sqrt{\tau_{\mathrm{C}}^2 + u^2(x_j)}}\right) + \sum_{j=1}^{n} \log \psi_{\nu_j}\left(\frac{r_j - \rho_j}{u(r_j)}\right) - \frac{n}{2} \log(\tau_{\mathrm{C}}^2 + u^2(x_j)), \tag{2}$$

where ϕ denotes the probability density function of the Gaussian distribution with mean 0 and standard deviation 1, and ψ_ν denotes the probability density function of Student's t distribution with ν degrees of freedom. A Gaussian probability density is used for x_j because $u(x_j)$ is assumed to be based on a very large number of degrees of freedom, while a rescaled and shifted Student's t distribution is used for r_j because $u(r_j)$ is based on only a fairly small number of degrees of freedom, the aforementioned ν_j, for $j = 1, \ldots, n$.

Up to an additive constant that involves none of $\boldsymbol{\beta}$, $\boldsymbol{\rho}$, or τ_{C}, the function L is the logarithm of the likelihood function under the assumptions stated after Equation (1), that the $\{\lambda_j\}$, $\{\delta_j\}$, and $\{\epsilon_{i,j}\}$ all have Gaussian distributions. The values of the parameters that maximize L are denoted $\widehat{\boldsymbol{\beta}}$, $\widehat{\boldsymbol{\rho}}$, and $\widehat{\tau}_{\mathrm{C}}$. In these circumstances, and according with Equation (1), we also have $\widehat{\xi}_j = G(\widehat{\rho}_j, \widehat{\boldsymbol{\beta}})$, for $j = 1, \ldots, n$.

The candidate models for G are polynomials of degrees $1, \ldots, n-2$ (because a polynomial of degree $n-1$, which has n coefficients, would fit the n points that represent the calibration data exactly). These candidate models are fitted in sequence and evaluated using conventional model selection criteria: Akaike's Information Criterion,

$\text{AIC} = 2p - 2L(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\rho}}, \hat{\tau}_C)$, and the Bayesian Information Criterion, $\text{BIC} = p\log(n) - 2L(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\rho}}, \hat{\tau}_C)$ [3]. The smaller the values of these criteria the more adequate the model is for the calibration data, as illustrated in Table 3.

It is conceivable that leaving out one or more of the lower degree terms of a polynomial of degree $p-1$ will produce a better model than when all the terms are included in the model. The same criteria, AIC and BIC, can be used to compare such polynomials with those that include all the terms.

In addition to model selection criteria, the initial selection of the polynomial to represent G involves the examination of plots of residuals against fitted values to compare the magnitudes of the resulting residuals and any patterns in the relationship between residuals and fitted values. Figure 4 shows these plots for the two candidate polynomials intended to serve as analysis function for lot SN.

Characterization of the analysis function

Up until recently, NIST has been using the procedures that Guenther and Possolo [12] introduced to estimate the analysis function G and to evaluate its associated uncertainty. Next, we describe the updated version of this procedure, to fit a polynomial (whose degree will have been selected previously, as described in subsection Selection of degree for analysis function polynomial), using a Bayesian approach that is best suited to incorporate relevant historical information into the uncertainty evaluation, via a suitably tuned prior probability distribution for the dark uncertainty, τ, that comprises the contributions from both τ_C and τ_H.

Possolo and Meija [26, pp. 204–215] provide a brief, general introduction to Bayesian models in metrology, which Cecelski et al [7] introduced in the context of EIV regression in gas metrology. It should be noted that, other than the prior distribution for τ, the prior distributions used in this contribution generally do not necessarily reflect *bona fide* preexisting information about the unknown quantities (like the $\{\beta_j\}$ or the $\{\rho_j\}$), other than in the trivial sense that, based on long experience with these materials, we expect that the $\{\beta_j\}$ should be close to their maximum likelihood estimates, and the $\{\rho_j\}$ should be close to the corresponding, observed average ratios $\{r_j\}$. Instead, the prior distributions adopted for parameters other than τ serve merely as regularization prescriptions for what can be regarded as an elaborate optimization procedure that explores the whole parameter space thoroughly. The use we make of the specific data obtained for SN and PA is merely to locate the priors in the scale of amount fractions, consistent with common practice in applications of Bayesian methods to measurement science [23].

The Bayesian version of the model in Equation (1), whose implementation in the probabilistic programming language Stan [4] is listed in Fig. 5, is specified next, where items (M1)–(M3) describe the prior distributions, and items (M5)–(M6) describe the terms that define the likelihood function.

The same as in all Bayesian models, prior distributions are chosen for all the parameters in the model, which are the $\{\beta_j\}$ (the elements of the ordered set $\boldsymbol{\beta}$ in Equation (2)), the $\{\rho_j\}$ (the elements of the ordered set $\boldsymbol{\rho}$ in the same equation), and τ. (Since the $\{\xi_j\}$ are functions of the $\{\beta_j\}$

Table 3 Both the AIC and BIC model selection criteria suggest $p = 2$ for lot SN; hence, a polynomial of the first degree is chosen for the analysis function

p	AIC	BIC
2	2.12	0.895
3	2.85	1.011

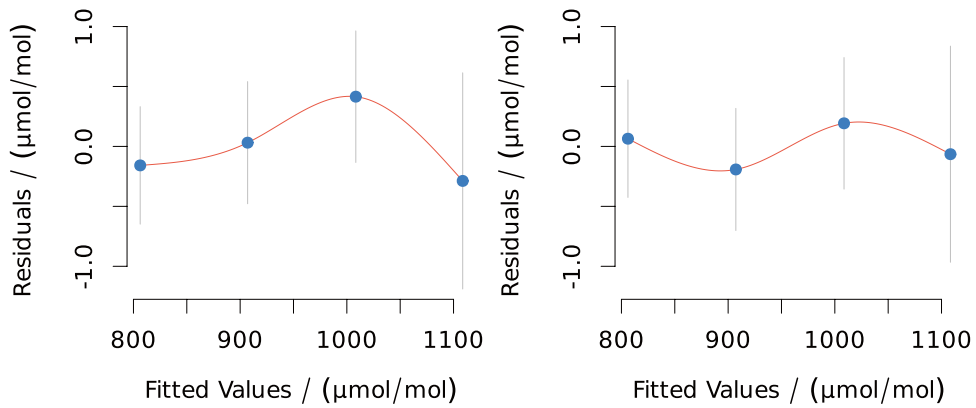

Fig. 4 Residuals plotted against fitted values (blue dots), for the polynomials of first and second degree, corresponding to $p = 2, 3$ in Table 3, that were fitted to the calibration data for lot SN, during model selection. The red curves, which are interpolating splines, serve only as visualization aids. The vertical (gray) line segments represent plus or minus $U_{95\,\%}(x_j)$ to enable comparing the sizes of the residuals with the expanded uncertainties associated with the amount fractions in the PSMs. Even though the residuals corresponding to $p = 3$ have absolute values slightly smaller than those that correspond to $p = 2$, the model corresponding to the latter is favored by both the AIC and BIC model selection criteria listed in Table 3 (color figure online)

and of the $\{\rho_j\}$, their prior distributions are determined by the prior distributions of the $\{\beta_j\}$ and of the $\{\rho_j\}$.)

(M1) $\{\beta_j\}$, which are the true values of the coefficients of the polynomial used to represent the analysis function, are assumed to be independent *a priori* and to have Gaussian prior distributions with means equal to their EIV regression maximum likelihood estimates, and with standard deviations equal to three times the standard errors of the same estimates.

(M2) $\{\rho_j\}$, which are the true values of the ratios of instrumental indications for the PSMs used for calibration, are assumed to be independent *a priori* and to have Gaussian prior distributions with means equal to the averages of the corresponding batches of replicates of the ratios, and with standard deviations equal to the standard errors of these averages.

(M3) The prior distribution chosen for the dark uncertainty τ is half-Cauchy, with median that depends on whether there is, or there is not, relevant historical information as described in subsection Two sources of dark uncertainty. This is how this prior median is chosen in these two cases:

- w/out HISTORICAL INFORMATION: If $\{\widehat{\xi}_j\}$ denote the maximum likelihood estimates of the true amount fractions of the analyte in the PSMs, then the prior median of τ is τ_C, which is the square root of the dif-

```
data
{
  int n;           // Number of standards (PSMs) for analysis function
  vector [n] r;    // r[j] = Ave. of replicates of ratio for standard j
  vector [n] ur;   // ur[j] = Std. unc. associated with ratio for standard j
  vector [n] nur; // nur[j] = Number of degrees of freedom supporting ur[j]
  vector [n] x;    // x[j] = Amount fraction in standard j
  vector [n] ux;   // ux[j] = Std. unc. associated with x[j]
  int p;           // Analysis function is polynomial of degree p-1
  vector [p] betaPriorMean;    // Prior mean for beta (EIV regression coeffs.)
  vector [p] betaPriorStdDev; // Prior std. dev. for beta
  vector [n] rhoPriorMean;     // Prior mean for rho (true ratios)
  vector [n] rhoPriorStdDev;   // Prior std. dev. of rho
  // Prior median for the dark uncertainty surrounding the amount fractions
  real tauxPriorMedian;
}
transformed data
{
  vector [n] sr; // sr[j] = Student's t scale corresponding to ur[j]
  for (j in 1:n) {sr[j] = ur[j]/sqrt(nur[j]/(nur[j]-2));};
}
parameters
{
  vector[p] beta; // EIV regression coefficients
  vector [n] rho; // rho[j] = True value of ratio for standard j
  real <lower=0> taux; // Dark uncertainty (comprising tauC and tauH)
}
transformed parameters
{
  vector [n] xi; // True values of amount fractions
  for (j in 1:n)
    {
      xi[j] = beta[1];
      for (jp in 2:p) { xi[j] = xi[j] + beta[jp]*pow(rho[j],jp-1); };
    }
}
model
{
  beta ~ normal(betaPriorMean, betaPriorStdDev); // Prior for beta
  rho ~ normal(rhoPriorMean, rhoPriorStdDev);    // Prior for rho
  taux ~ cauchy(0, tauxPriorMedian);             // Prior for taux
  // Gaussian likelihood for amount fractions
  x ~ normal(xi, sqrt(square(ux) + square(taux)));
  // Student's t likelihood for ratios
  r ~ student_t(nur, rho, sr);
}
```

Fig. 5 Stan code that implements the Bayesian version of the model for EIV regression, characterized in items (M1)–(M6), that is used to build the analysis function

ference between the variance of the $\{x_j - \widehat{\xi}_j\}$ and the square of the geometric average of the $\{u(x_j)\}$ when this difference is positive, or a value comparable to machine precision otherwise.

- w/HISTORICAL INFORMATION: The prior information about the historical dispersion of values for the lot standard is expressed using a gently informative (in the sense described by Meija et al [23]) half-Cauchy distribution for the historical component of the dark uncertainty, τ_{H}, whose median was determined based on historical data, as follows: Suppose that there are n_{H} historical LSs, whose original estimates of the amount fraction of the analyte are $x_{\mathrm{H},1}, \ldots, x_{\mathrm{H},n_{\mathrm{H}}}$. Ratios relative to the LS for the SRM under development have also been determined for each of these historical LSs, which we denote as $\{r_{\mathrm{H},i,j}\}$ for each historical LS $j = 1, \ldots, n_{\mathrm{H}}$. (The numbers of such ratios can be different for different historical LSs.) For each such ratio i of each historical LS j, we compute the estimate of the measurand, $\widehat{\xi}_{\mathrm{H},i,j}$, that corresponds to the maximum likelihood estimates of the coefficients of the analysis function, and then form the differences $D_{\mathrm{H},i,j} = x_{\mathrm{H},j} - \widehat{\xi}_{\mathrm{H},i,j}$. (Note that the first term on the right-hand side is the same for all the ratios for historical LS j.) τ_{H} is the root mean square of all the $\{D_{\mathrm{H},i,j}\}$, and the prior median of τ is $\sqrt{\tau_{\mathrm{H}}^2 + \tau_{\mathrm{C}}^2}$.

(M4) Given $\boldsymbol{\beta}$ and $\boldsymbol{\rho}$, the $\{\xi_j\}$ are Gaussian random variables with means $\{G(\rho_j, \boldsymbol{\beta})\}$ and standard deviation τ.

(M5) Conditionally upon the $\{\xi_j\}$, $\{x_j\}$ are assumed to be outcomes of Gaussian random variables with means $\{\xi_j\}$ and standard deviations $\{u(x_j)\}$ (these standard deviations are assumed to be known).

(M6) Conditionally upon ρ_j, the average of the replicates of the ratios for PSM j is modeled as an observed value of a Student's t distribution with ν_j degrees of freedom, rescaled to have standard deviation $u(r_j)$ and shifted to be centered at r_j, for $j = 1, \ldots, n$. This modeling choice takes into account the small number of replicates that $u(r_j)$ is based on, while circumventing the need to estimate the true standard deviation of the replicates. It is based on the fact that if r_j is an average of $\nu_j + 1$ ratios that are a sample from a Gaussian distribution whose mean is ρ_j and whose standard deviation is unknown, then $(r_j - \rho_j)/u(r_j)$ is Student's t with ν_j degrees of freedom.

The Stan code listed in Fig. 5 was fitted to the calibration data for lot SN using facilities from package `rstan` for the R environment for statistical computing and graphics [29, 33]. Four independent Markov Chain Monte Carlo (MCMC) samplers were run in parallel, all instances of the No-U-Turn-Sampler (NUTS) [15] that is available in Stan. Each sampler took 250000 warm-up steps followed by 250000 sampling steps, with the result of every 25th step having been recorded, to reduce the impact of auto-correlations. As a result, after merging the samples produced by the four samplers, we obtained a sample of size $K = 40000$ from the joint posterior distribution of the parameters.

The diagnostics for effective sample size, and for the convergence criterion RHAT, all suggest that the sampler indeed was sampling from the joint posterior distribution of the parameters after the warm-up period. RHAT is the ratio between the standard deviation of an estimate of a parameter derived from all the chains together, and the root mean square of the separate, within-chain standard deviations. If the chains have not reached their common equilibrium state, then this ratio will be greater than 1 [37]. Table 4 lists the posterior means and standard deviations of the samples of values of each of the parameters.

The Bayesian estimates of τ listed in Table 4 are appreciably different: one corresponds to the case where the information provided by the historical LSs is ignored ($\widetilde{\tau} = 0.132\,\mu\mathrm{mol/mol}$), the other to the case where it is taken into account ($\widetilde{\tau} = 0.458\,\mu\mathrm{mol/mol}$). Figure 6 depicts the prior and posterior probability densities for τ in both cases. Note that both prior probability densities (red curves) achieve their maxima at 0 µmol/mol, and that their posterior counterparts (blue curves) achieve their maxima away from 0 µmol/mol, markedly more so for the posterior when historical uncertainty is taken into account. These observations confirm previous findings: the component of dark uncertainty attributable to mutual "misalignment" of the PSMs, quantified in $\widehat{\tau}_{\mathrm{C}}$, is much smaller than the component, $\widehat{\tau}_{\mathrm{H}}$, that captures historical uncertainty, and that, taken together, they add up to a composite Bayesian estimate of dark uncertainty, $\widetilde{\tau}$, that is significantly greater than 0 µmol/mol.

The output of the MCMC procedure that samples the joint posterior distribution of all the parameters in the model, delivered $K = 40000$ sets of values of the coefficients of the analysis function, $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K$ (each of these is an ordered set with p elements), K sets of true values of the ratios for the PSMs, $\boldsymbol{\rho}_1, \ldots, \boldsymbol{\rho}_K$ (each being an ordered set with n elements), and K values of the dark uncertainty, τ. These samples are the raw materials from which uncertainty evaluations will be derived for the amount fractions of the cylinders in a lot, as will be described in section ``Measurement results for individual cylinders.

Even though the corresponding Bayesian estimates of β_1 and β_2 listed in Table 4 are slightly different, their differences are not statistically significant, and they are indistinguishable as depicted in Fig. 7. The wider (dark gray) uncertainty band depicted in the same figure fully encloses 95 % of the 40000 versions of the analysis function that corresponds to

Table 4 Posterior means and standard deviations of the parameters in the Bayesian version of the EIV regression model of Equation (1) for lot SN, either neglecting the information provided by the historical LSs (2nd and 3rd columns), or taking it into account (4th and 5th columns)

	w/out HISTORICAL INFORMATION		w/HISTORICAL INFORMATION		
	MEAN	SD	MEAN	SD	
β_1	11.95	1.40	12.06	1.92	/(μmol/mol)
β_2	984.10	1.49	983.95	2.02	/(μmol/mol)
ρ_1	0.80713	0.00004	0.80713	0.00004	
ρ_2	0.90954	0.00002	0.90954	0.00002	
ρ_3	1.01267	0.00003	1.01267	0.00003	
ρ_4	1.11455	0.00007	1.11456	0.00007	
τ	0.13	0.16	0.46	0.49	/(μmol/mol)
ξ_1	806.24	0.25	806.24	0.45	/(μmol/mol)
ξ_2	907.03	0.18	907.01	0.37	/(μmol/mol)
ξ_3	1008.52	0.22	1008.48	0.40	/(μmol/mol)
ξ_4	1108.78	0.33	1108.73	0.51	/(μmol/mol)

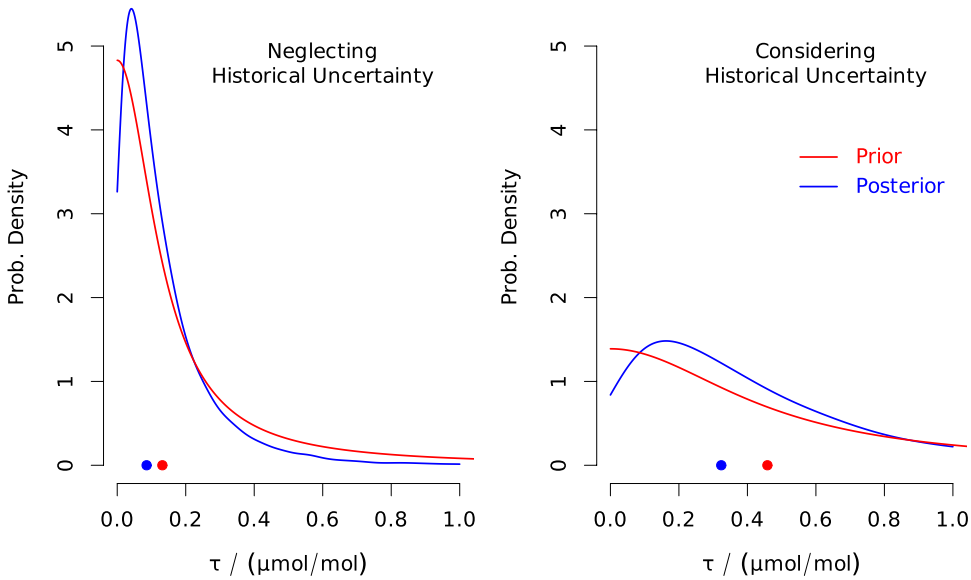

Fig. 6 Prior (red curves) and posterior (blue curves) probability densities of τ, and corresponding medians (dots) for lot SN. LEFT PANEL: Neglecting the information provided by the historical LSs. RIGHT PANEL: Considering the information provided by the historical LSs (color figure online)

the model that takes historical uncertainty into account. The narrower (light gray) band corresponds to the model that neglects historical uncertainty.

Measurement results for individual cylinders

The estimate of the amount fraction of the analyte in each cylinder and the evaluation of the associated uncertainty are derived from the results of applying the different versions of the analysis function obtained via MCMC sampling, $G_1, \ldots, G_K$, to the replicates of the ratios determined for the cylinder.

Suppose that the lot comprises L cylinders, and that $m_1, \ldots, m_L$ denote the numbers of ratios determined for the different cylinders. We take the following steps for each cylinder $l = 1, \ldots, L$:

(L1) For each replicate of the ratio $r_{l,i}$ that has been determined for cylinder l, compute the amount fraction $x_{l,i,k} = \max(0, G_k(r_{l,i}) + z_k)$, where G_k denotes a ver-

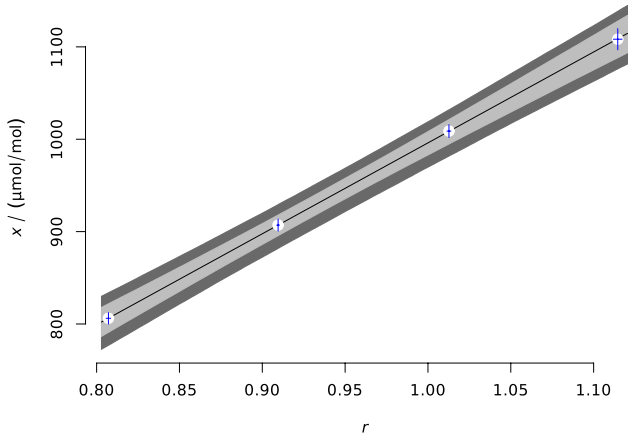

Fig. 7 Analysis function (thin, black sloping line) for lot SN and uncertainty bands, whose vertical thickness is magnified 25 times, for 95 % coverage. The (dark gray) wider band expresses contributions from the two kinds of dark uncertainty, τ_C and τ_H, which were discussed in subsection Two sources of dark uncertainty and illustrated in Fig. 3. The (light gray) narrower band does not include the contribution from historical uncertainty. The (white) dots represent the values of the averages of the replicates of the ratios and of the amount fractions for the PSMs used for calibration. The small, blue crosses centered on these dots represent measured values plus or minus one standard uncertainty, also magnified 25 times (color figure online)

sion of the analysis function, and z_k is drawn from a Gaussian distribution with mean 0 µmol/mol and standard deviation τ_k, for $i = 1, \ldots, m_l$, and for $k = 1, \ldots, K$. $\{x_{l,i,k} : i = 1, \ldots, m_c, k = 1, \ldots, K\}$ are Monte Carlo replicates of a prediction of a determination of the value of x_l that can be made using a measurement procedure comparable to the procedure used for the material's certification, taking into account the dark uncertainty (either $\widehat{\tau}_C$, when the historical uncertainty is neglected, or both $\widehat{\tau}_C$ and $\widehat{\tau}_H$, when the historical uncertainty is taken into account).

(L2) Compute the estimate, $\widetilde{x}_l$, of the amount fraction of the analyte in cylinder l, and the associated standard uncertainty, $u(\widetilde{x}_l)$, by applying R functions `huberM` and `Qn`, respectively, both defined in package `robustbase` [22], to the $\{x_{l,i,k}\}$. The former delivers a robust summary of these Monte Carlo replicates of the amount fraction, in the form of an M-estimate of location [16], and the latter provides a robust estimate of the standard deviation of the replicates [30].

The amount fractions assigned to the individual cylinders of lot SN, and their associated uncertainties, which correspond to the case where one neglects the historical uncertainty, and to the case where the historical uncertainty is taken into account, are displayed side-by-side for each cylinder, in Fig. 8.

Lot homogeneity

The measured values of the amount fraction of the analyte in the individual cylinders, $\{x_c\}$, are examined from different viewpoints, also considering their associated uncertainties, $\{u(x_{c,i})\}$, to ascertain whether the lot is sufficiently homogeneous to warrant assignment of a single value to the whole lot. If it is, then all its cylinders are assigned the same value and uncertainty. If it is not, then the lot may be split into sufficiently homogeneous sub-lots, each with its own assigned value, or the different cylinders will be assigned different values and possibly different associated uncertainties.

The "dip" statistical test of unimodality [13] serves to evaluate whether there are significantly different, multiple peaks (or modes) in a probability density estimate of the

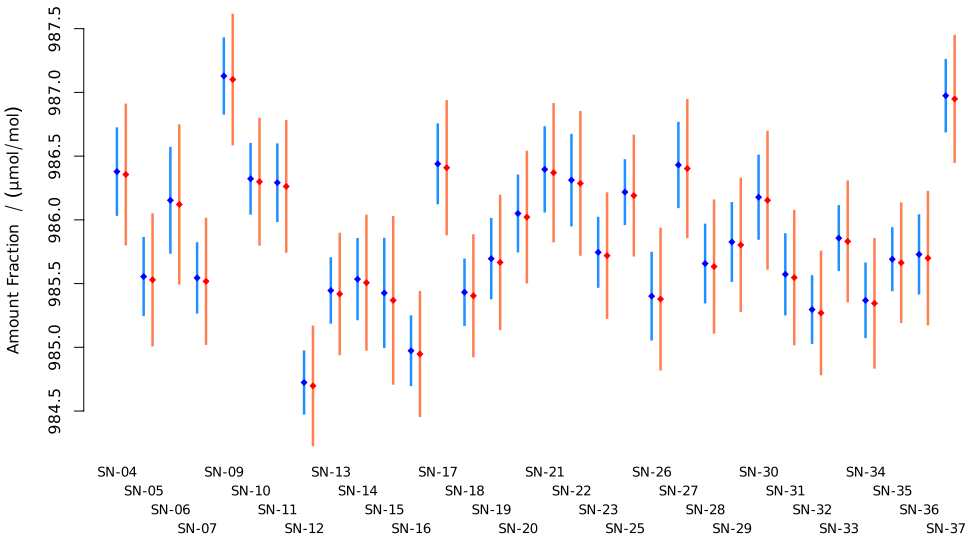

Fig. 8 Cylinder values (solid diamonds), $\{\widetilde{x}_l\}$, and associated standard uncertainties, where the vertical line segments represent $\{\widetilde{x}_l \pm u(\widetilde{x}_l)\}$, that correspond to the case where the historical uncertainty is neglected (blue, shorter segments), and to the case where it is taken into account (red, longer segments), side-by-side for each cylinder, for the 32 cylinders in lot SN (color figure online)

$\{x_c\}$. If there are, then this speaks in favor of splitting the lot. However, this test does not take the uncertainties, $\{u(x_c)\}$, into account. The "dip" test for lot SN (Fig. 9) does not reject the hypothesis of unimodality.

Cochran's Q test [8], also known as the chi-squared test of homogeneity, has been widely used to assess whether multiple, independent measurement results (measured values and associated uncertainties) for the same measurand are mutually consistent, even if its shortcomings are well-known [14]. Iyer et al [18] discussed several alternative tests for the same purpose, among them Welch's F test [40], which takes into account the numbers of degrees of freedom that the uncertainties are based on. Figure 10 presents R code that implements both Cochran's Q and Welch's F tests.

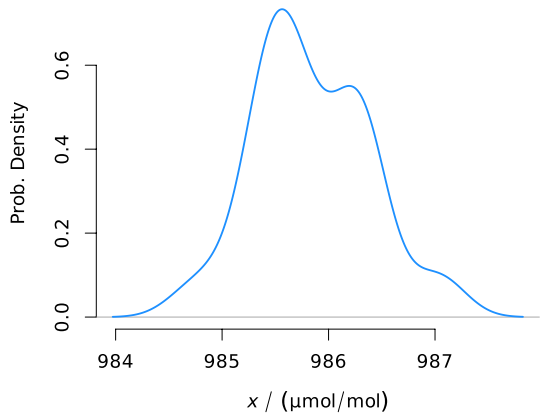

Fig. 9 The kernel estimate [32] of the probability density of the cylinder values in lot SN reveals several "bumps" (local maxima), which the "dip" statistical test [13], yielding p-value 0.2, suggests are not significant, thus not rejecting the hypothesis of unimodality

Mutual consistency or homogeneity in this context means that the measured values can be regarded as values drawn from probability distributions with the same mean, but possibly different standard deviations. Both Cochran's Q test and Welch's F test assume that these distributions are approximately Gaussian. Cochran's Q test assumes further that the uncertainty evaluations are based on very large (practically infinite) numbers of degrees of freedom.

Note that the assumption of independence made by both tests, for the measurement results for the different cylinders, is questionable because they all share the same errors that affect the analysis function. Applied to the measurement results for lot SN that are depicted in Fig. 8, neither Cochran's Q test (p-value 0.26) nor Welch's F test (p-value 0.31) reject the hypothesis of mutual consistency, thus warranting the assignment of a single value to lot SN.

Consensus value and uncertainty evaluation

If the lot is sufficiently homogeneous, particularly once the cylinder values will have been qualified with an uncertainty evaluation that includes the contribution from historical uncertainty, then the value assigned to the SRM is a consensus of the values assigned to the individual cylinders, $\{x_c\}$, computed taking into account their associated uncertainties $\{u(x_c)\}$.

The *NIST Decision Tree* (NDT) [27] is a flexible, easy-to-use, web-based application that offers suggestions about how to combine measurement results for the same measurand, obtained independently of one another. In particular, it can be applied to the measurement results for the individual

Fig. 10 R code that implements Cochran's Q test and Welch's F test of mutual consistency of measurement results. Note that in Equation (2.5) of Milliken and Johnson [24], where they describe Welsh's test, where it says "$(t-1)$" it should have said "$(t-2)$" instead

```
## Tests of mutual consistency of measurement results
## n   = Number of measurement results
## x   = Measured values
## ux  = Standards uncertainties
## nux = Numbers of degrees of freedom supporting the ux

cochranQ = function (x, ux, digits=5)
{
    w = 1/ux^2
    mu = sum(w*x)/sum(w)
    Q = sum(w*(x-mu)^2)
    Qnu = length(x)-1
    Qp = pchisq(Q, df=Qnu, lower.tail=FALSE)
    return(signif(c(Q=Q, nu=Qnu, p=Qp), digits))
}

welchF = function (x, ux, nux=NULL, digits=5)
{
    if (is.null(nux)) {nux = rep(60.4, length(x))}
    w = 1/ux^2
    mu = sum(w*x)/sum(w)
    Q = sum(w*(x-mu)^2)
    Lambda = sum((1-w/sum(w))^2/nux)
    n = length(x)
    Fc = (Q/(n-1)) / (1+2*(n-2)*Lambda/(n^2-1))
    Fcp = 1-pf(Fc, n-1, (n^2-1)/(3*Lambda))
    return(signif(c(Fc=Fc, nu1=n-1, nu2=(n^2-1)/(3*Lambda), p=Fcp), digits))
}
```

cylinders obtained as described in section Measurement results for individual cylinders.

To use the NDT in practice, when there is more than a handful of cylinders in a lot, the best way involves preparing a CSV file with as many rows as there are cylinders in the lot, and with four entries per row: cylinder label, measured valued x_c, associated standard uncertainty $u(x_c)$, and effective number of degrees of freedom, ν_c, that $u(x_c)$ is based on, with the headers specified in the NDT's user's manual, which is available for download from the NDT's web site. Since the previously described workflow does not provide values for the $\{\nu_c\}$, these can all be set equal to 60.4, which corresponds to the conventional *coverage factor* $k = 2$ as recommended by Taylor and Kuyatt [34, §6.5] for use at NIST.

The following steps are taken to assign a single value, x_{SRM}, to the SRM, and to evaluate the corresponding uncertainty, $u(x_{\mathrm{SRM}})$. This uncertainty is a prediction uncertainty, so that about 95 % of the individual cylinder values will differ from the value assigned to the SRM by less than the corresponding expanded uncertainty, $U_{95\,\%}(x_{\mathrm{SRM}})$:

(S1) Determine the value, x_{SRM}, to be assigned to the lot, as the consensus value computed using the *NIST Decision Tree* (or, alternatively, using the *NIST Consensus Builder* [20], or facilities implemented in R package `metafor` [39], among many others), and the associated uncertainty, $u(x_{\mathrm{SRM}})$. Since the same analysis function is used to assign values to the different cylinders, the cylinder values are correlated. However, the NDT ignores such correlations when it computes x_{SRM}.

(S2) A prediction interval for the amount fraction of the analyte in any cylinder in the lot can be built using a sample comprising a large number, K, of sets of values of the parameters in the model, drawn from the posterior predictive distribution that corresponds to the Bayesian model described in the foregoing, as follows:

(S2a) First, repeat these steps for $i = 1, \ldots, I$, where I denotes a large, positive integer of the same order of magnitude as K: (i) draw a value, r^*, uniformly at random from the set of all ratios that were determined for all cylinders in the lot; (ii) draw a value $\boldsymbol{\beta}^*$, also uniformly at random, from the MCMC sample of values of the coefficients of the polynomial generated by the Stan code in Fig. 5; (iii) similarly draw a value τ^* from the MCMC sample of values of τ; (iv) draw a residual e^* uniformly at random from the set of all differences $\{x_j - \widetilde{\xi}_{j,k}\}$; (v) simulate a drawing from a Gaussian distribution with mean $G(r^*, \boldsymbol{\beta}^*)$ and with standard deviation τ^*, and add e^* to it, to obtain x_i^*.

(S2b) Determine the endpoints of an interval centered at the consensus value for all the cylinders, x_{SRM} from (S1), that includes 95 % of the $\{x_i^*\}$ simulated in (S2a).

The *NIST Decision Tree* recommends the adaptive-weighted average (AWA, which is a modified version of the DerSimonian-Laird procedure described by Koepke et al [20]) for combining the cylinder-specific measurement results for lot SN, which are depicted in Fig. 8, into a consensus value.

We have overridden this recommendation and used the hierarchical Bayesian model with Gaussian laboratory effects and Gaussian measurement errors instead (often referred to as HGG, where H means "hierarchical", the first G indicates that the laboratory effects are Gaussian, and the second G indicates that the measurement errors are Gaussian). This modeling choice is more closely aligned with the Bayesian approach than the AWA, even though it produces results that are very close to those produced by the AWA.

The HGG procedure produces the estimate $x_{\mathrm{SN}} = 985.8(1)$ µmol/mol, and a 95 % credible interval for its true value ranges from 985.6 µmol/mol to 986.0 µmol/mol. Figure 11 depicts the measurement results for the individual cylinders in Fig. 8, as well as the consensus value for the lot, the associated uncertainty, and a 95 % prediction interval built as described above, under (S2), which ranges from 983.7 µmol/mol to 987.9 µmol/mol.

Lot PA's calibration challenge

PA is one of several SRM lots that NIST has developed with 50 µmol/mol nominal amount fraction of propane in air. Each of these SRMs has had its own LS, with the role already described in section Data acquisition and data reduction workflow. Currently, there are six such historical LSs with the same nominal amount fraction of propane as PA, which we denote U, ..., Z in the right panel of Fig. 12.

On the one hand, each of these historical LSs has the value of the amount fraction of propane that was assigned to it when the corresponding SRM was last certified. On the other hand, ratios between instrumental indications obtained during the analysis of these historical LSs, and the instrumental indications obtained for the LS of PA, were also determined as part of the workflow for PA. These ratios were then mapped into values of the amount fraction of propane, $\widehat{\xi}$, using the analysis function built for PA via maximum likelihood estimation of the parameters of the EIV regression model specified in Equation (1).

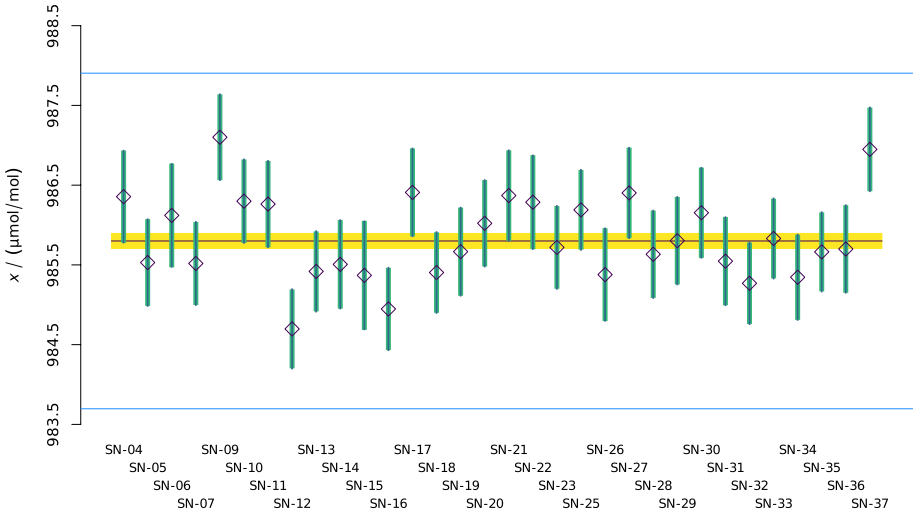

Fig. 11 Cylinder values (diamonds), $\{\widetilde{x}_c\}$, and associated standard uncertainties, for lot SN, where the thick (green) vertical line segments represent $\{\widetilde{x}_c \pm u(\widetilde{x}_c)\}$, that correspond to the case where the historical uncertainty is taken into account. The thin (dark blue) vertical line segments represent $\{\widetilde{x}_c \pm (\tau_C^2 + u^2(\widetilde{x}_c))^{1/2}\}$. The thin (dark brown), horizontal line indicates the consensus value, x_{SRM}, with the height of the (yellow) band around it representing $x_{SRM} \pm u(x_{SRM})$. The thin (light blue) horizontal lines indicate the endpoints of the corresponding, 95 % prediction interval (color figure online)

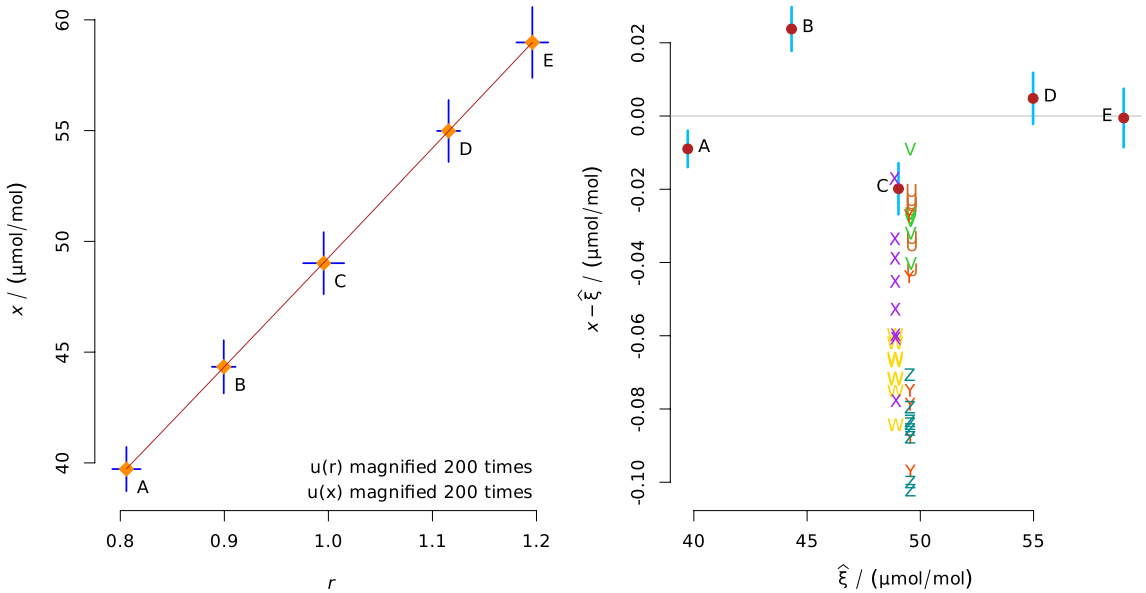

Fig. 12 LEFT PANEL: The (orange) solid diamonds represent the PSMs, which are labeled A, ..., E, the blue segments represent $r \pm u(r)$ and $x \pm u(x)$, except that the uncertainties are magnified 200 times, and the (red) sloping line represents the analysis function for lot PA (which is a quadratic polynomial). RIGHT PANEL: The solid (dark red) dots labeled A, ..., E represent the residual amount fractions of propane for the five PSMs used to build the analysis function for lot PA, and the vertical (light blue) line segments represent these residuals plus or minus the corresponding standard uncertainties of the amount fractions of propane in these PSMs. The letters U, ..., Z, of six different colors, represent replicated differences between the values, x, of the amount fraction of propane in six historical lot standards, and the estimates, $\widehat{\xi}$, of the amount fraction of propane in them produced by the analysis function for lot PA (color figure online)

Note that, for PSMs B and C in the right panel of Fig. 12, the difference $x - \widehat{\xi}$ is more than two standard uncertainties above or below the horizontal line at 0 µmol/mol. This means that these two PSMs are (almost imperceptibly on the left panel of the figure, yet statistically significantly)

"misaligned" relative to the other three. This misalignment translates into an estimate $\widehat{\tau}_C = 0.015\,\mu mol/mol$ of the component of dark uncertainty attributable to calibration, estimated as already described for lot SN in subsection Two sources of dark uncertainty.

The ordinates of the letters U, ..., Z, of six different colors, in the right panel of Fig. 12 represent replicated differences $x - \widehat{\xi}$ between the original estimates of the amount fraction of propane in historical LSs, and the amount fractions estimated using the analysis function for PA. The abscissae of those letters represent these estimates.

Here, differently from Fig. 3, all the differences $x - \widehat{\xi}$ are negative. That is, PA's analysis function appears to overestimate the amount fractions of propane in the historical LSs labeled U, ..., Z. The PSMs used to calibrate the analysis functions for all these historical lots had propane in a balance of nitrogen, while the PSMs used for PA had propane in a balance of air.

We have observed biases apparently related to the nature of the balance gas (air versus nitrogen) in mixtures similar to PA, which sometimes are positive and other times are negative, but, in either case, we recognize and propagate the corresponding uncertainty via the historical component of dark uncertainty, τ_H.

Consistent with the provisions of the GUM, reflecting "the concept that there is no inherent difference between an uncertainty component arising from a random effect and one arising from a [...] systematic effect" [19, E.1.1], the apparent bias affecting PA is taken into account together with the dispersion of the values of the historical lot standards, and the historical uncertainty is evaluated as the root mean square (RMS) of the ordinates of the letters, $\widehat{\tau}_H = 0.062\,\mu mol/mol$.

Combining the contributions from these two sources of dark uncertainty that were evaluated in the foregoing, $\widehat{\tau}_C$ and $\widehat{\tau}_H$, in root sum of squares, yields $\widehat{\tau} = 0.063\,\mu mol/mol$ for the prior median of the dark uncertainty specified in step (M3) of subsection Characterization of the analysis function. Figure 13 depicts the prior and posterior distributions of the composite dark uncertainty τ when the information provided by the historical LSs is neglected or taken into account. For this lot PA, τ is significantly greater than 0 µmol/mol both when historical uncertainty is considered and when it is disregarded, unlike their counterparts for lot SN, whose posterior distributions are depicted in Fig. 6.

Conclusions

One of the greatest challenges in certifying a reference gas mixture is ensuring that the uncertainty associated with the assigned value is realistic (*cf.* [19, E.1]) and will remain valid for the duration of the period of validity [2], which for most gas mixture SRMs developed at NIST ranges from 4 to 8 years [6, Table 4]. The procedures described in this contribution serve this purpose, being the result of refinements and improvements of the procedures for certifying reference gas mixtures, and for the evaluation of the associated uncertainty, that were originally presented by Guenther and Possolo [12] thirteen years ago.

The principal novelty of this contribution is the rigorous evaluation of historical uncertainty in the determination of the analysis function, and its propagation to the uncertainty surrounding this function. Historical uncertainty captures the historical dispersion of results for gas mixtures nominally identical to the mixture being certified, for reasons that may remain unexplained but that become apparent and can

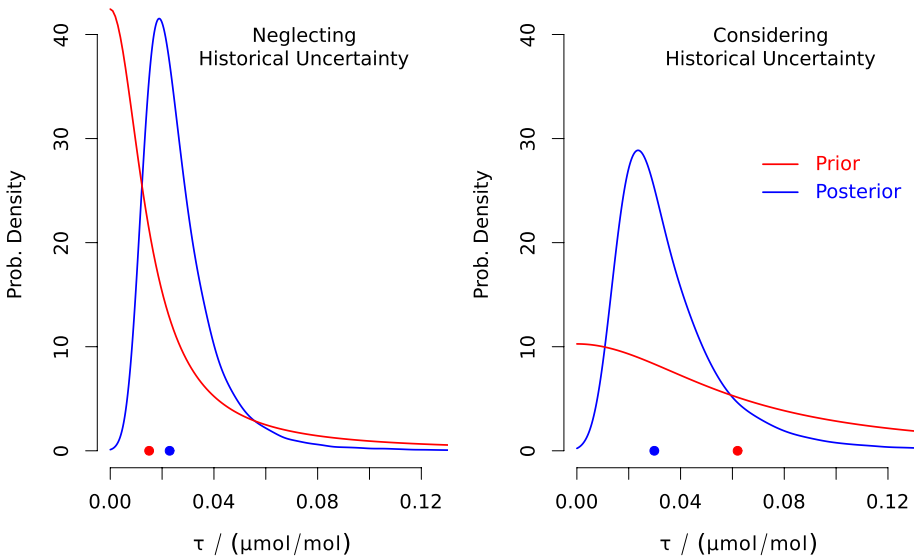

Fig. 13 Prior (red curves) and posterior (blue curves) probability densities of τ, and corresponding medians (dots) for lot PA. LEFT PANEL: Neglecting the information provided by the historical LSs. RIGHT PANEL: Considering the information provided by the historical LSs (color figure online)

be quantified when historical lot standards are remeasured against the lot standard of the SRM currently in development, and corresponding measured values are compared for these lot standards.

In general, the incorporation of such historical information produces larger, more realistic uncertainties than when such information is either not available or is neglected. This effect is the opposite of the effect that the incorporation of prior information typically induces, which is to reduce uncertainty. However, it is the desired and proper effect because, in our long experience developing these reference materials, the classical estimate of the amount fraction of a particular gas in a mixture of gases often appears to be surrounded by an uncertainty that is unrealistically small. In response, *ad hoc* "corrections" to the evaluated uncertainty have traditionally been applied, to enhance the credibility of the reported uncertainty.

Our contribution — and this indeed is its key aspect — dispenses with such *ad hoc* "corrections", and offers an honest assessment that recognizes explicitly that the conventional uncertainty evaluations can be overoptimistic, and takes into account historical information that, in most cases, suggests that the actual uncertainty should to be larger than what a classical evaluation indicates.

Other improvements over the procedure described by Guenther and Possolo [12] include:

- Recognizing that, during calibration of the analysis function, which is used for value assignment to the units of the SRM, mutual inconsistencies may be uncovered between the PSMs, which can be quantified using the concept of dark uncertainty for EIV regression that Cecelski et al [7] described in detail.
- Developing and applying a Bayesian version of the EIV regression model used by Guenther and Possolo [12] and described in ISO 6143 [17], which provides not only an estimate of the analysis function, but also the elements necessary to characterize the uncertainty that surrounds it. Not only does this improvement integrate the processes of estimation and uncertainty evaluation (for the analysis function), it also has the side-effect of reducing the computational run-time very substantially, by comparison with the Monte Carlo uncertainty evaluation that Guenther and Possolo [12] proposed originally.
- Employing state-of-the-art methods and tools, originally developed to reduce data from interlaboratory studies and meta-analyses, to combine the cylinder-specific measurement results, when these are sufficiently mutually consistent to warrant the assignment of a single value, and associated prediction value and uncertainty, to the SRM as a whole.

Acknowledgements The authors are grateful to their NIST colleagues Kimberly Harris, Michael Nelson, and Blaza Toman, who generously provided detailed reviews of a draft, and offered many suggestions for its improvement.

Declarations

Conflict of interest The authors have no conflicts of interest to declare that are relevant to the content of this article.

Ethical approval The research reported herein did not involve human or animal subjects or any biological materials, as objects of research.

Consent for publication This research was conducted as part of the authors' duties as employees of the National Institute of Standards and Technology, an agency of the federal government of the United States of America, under the U.S. Department of Commerce.

References

1. Bates D, Mächler M, Bolker B et al (2015) Fitting linear mixed-effects models using lme4. J Stat Softw 67(1):1–48. https://doi.org/10.18637/jss.v067.i01
2. Beauchamp CR, Camara JE, Carney J, et al (2021) Metrological Tools for the Reference Materials and Reference Instruments of the NIST Materials Measurement Laboratory. NIST Special Publication 260-136 (2021 Edition), National Institute of Standards and Technology, Gaithersburg, MD, doi:10.6028/NIST.SP.260-136-2021
3. Burnham K, Anderson D (2002) Model selection and multimodel inference: a practical information-theoretic approach, 2nd edn. Springer-Verlag, New York, NY
4. Carpenter B, Gelman A, Hoffman M et al (2017) Stan: a probabilistic programming language. J Stat Softw 76(1):1–32. https://doi.org/10.18637/jss.v076.i01
5. Carroll RJ, Ruppert D, Stefanski LA et al (2006) Measurement Error in Nonlinear Models – A Modern Perspective, 2nd edn. Chapman & Hall/CRC, Boca Raton, Florida
6. Cecelski CE, Harris KJ, Goodman CA, et al (2022a) Certification of NIST Gas Mixture Standard Reference Materials®. NIST Special Publication 260-222, National Institute of Standards and Technology, Gaithersburg, MD, doi:10.6028/NIST.SP.260-222
7. Cecelski CE, Toman B, Liu FH et al (2022) Errors-in-variables calibration with dark uncertainty. Metrologia 59(4):045002. https://doi.org/10.1088/1681-7575/ac711c
8. Cochran WG (1954) The combination of estimates from different experiments. Biometrics 10(1):101–129. https://doi.org/10.2307/3001666

9. Efron B, Tibshirani RJ (1993) An Introduction to the Bootstrap. Springer-Science+Business Media, Dordrecht, The Netherlands, doi:10.1201/978-0-4292-4659-3
10. Faraway JJ (2016) Extending the linear model with R, 2nd edn. Chapman & Hall/CRC, Boca Raton, Florida
11. Fuller WA (1987) Measurement Error Models. John Wiley & Sons, New York, NY
12. Guenther FR, Possolo A (2011) Calibration and uncertainty assessment for certified reference gas mixtures. Anal Bioanal Chem 399:489–500. https://doi.org/10.1007/s00216-010-4379-z
13. Hartigan JA, Hartigan PM (1985) The dip test of unimodality. Ann Stat 13(1):70–84. https://doi.org/10.1214/aos/1176346577
14. Hoaglin DC (2016) Misunderstandings about *Q* and 'Cochran's *Q* test' in meta-analysis. Stat Med 35:485–495. https://doi.org/10.1002/sim.6632
15. Hoffman MD, Gelman A (2014) The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. J Machine Learn Res 15(47):1593–1623
16. Huber PJ, Ronchetti EM (2009) Robust Stat, 2nd edn. John Wiley & Sons, Hoboken, NJ
17. ISO (2001) Gas analysis — Comparison methods for determining and checking the composition of calibration gas mixtures. International Organization for Standardization (ISO), Geneva, Switzerland, international Standard ISO 6143:2001(E)
18. Iyer HK, Wang CM, Vecchia DF (2004) Consistency tests for key comparison data. Metrologia 41(4):223–230. https://doi.org/10.1088/0026-1394/41/4/001
19. Joint Committee for Guides in Metrology (JCGM) (2008) Evaluation of Measurement Data — Guide to the Expression of Uncertainty in Measurement. International Bureau of Weights and Measures (BIPM), Sèvres, France, https://www.bipm.org/documents/20126/2071204/JCGM_100_2008_E.pdf/cb0ef43f-baa5-11cf-3f85-4dcd86f77bd6, BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP and OIML, JCGM 100:2008, GUM 1995 with minor corrections
20. Koepke A, Lafarge T, Possolo A et al (2017) Consensus building for interlaboratory studies, key comparisons, and meta-analysis. Metrologia 54(3):S34–S62. https://doi.org/10.1088/1681-7575/aa6c0e
21. Lang BE, Molloy JL, Vetter TW et al (2023) Value assignment and uncertainty evaluation for anion and single-element reference solutions incorporating historical information. Anal Bioanal Chem 415:1657–1673. https://doi.org/10.1007/s00216-022-04410-y
22. Maechler M, Rousseeuw P, Croux C, et al (2023) robustbase: Basic Robust Statistics. http://robustbase.r-forge.r-project.org/, r package version 0.99-1
23. Meija J, Bodnar O, Possolo A (2023) Ode to Bayesian Methods in Metrology. Metrologia 60:052001. https://doi.org/10.1088/1681-7575/acf66b
24. Milliken GA, Johnson DE (2009) Analysis of Messy Data, Volume 1: Designed Experiments, 2nd edn. Chapman & Hall/CRC, Boca Raton, FL
25. Pinheiro JC, Bates DM (2000) Mixed-Effects Models in S and S-Plus. Springer-Verlag, New York, NY,. https://doi.org/10.1007/b98882
26. Possolo A, Meija J (2022) Measurement Uncertainty: A Reintroduction, 2nd edn. Sistema Interamericano de Metrologia (SIM), Montevideo, Uruguay, doi:10.4224/1tqz-b038
27. Possolo A, Koepke A, Newton D et al (2021) Decision tree for key comparisons. J Res Natl Inst Standards Technol 126:126007. https://doi.org/10.6028/jres.126.007
28. R Core Team (2020) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/
29. R Core Team (2023) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/
30. Rousseeuw PJ, Croux C (1993) Alternatives to the median absolute deviation. J Am Stat Assoc 88:1273–1283
31. Searle SR, Casella G, McCulloch CE (2006) Variance Components. John Wiley & Sons, Hoboken, NJ
32. Silverman BW (1986) Density Estimation for Statistics and Data Analysis. Chapman & Hall/CRC, Boca Raton, FL
33. Stan Development Team (2023) RStan: the R interface to Stan. https://mc-stan.org/, r package version 2.32.3
34. Taylor BN, Kuyatt CE (1994) Guidelines for Evaluating and Expressing the Uncertainty of NIST Measurement Results. NIST Technical Note 1297, National Institute of Standards and Technology, Gaithersburg, MD, https://physics.nist.gov/Pubs/guidelines/TN1297/tn1297s.pdf
35. Thompson M, Ellison SLR (2011) Dark uncertainty. Accred Quality Assurance 16:483–487. https://doi.org/10.1007/s00769-011-0803-0
36. Tukey JW (1977) Exploratory Data Anal. Addison-Wesley, Reading, MA
37. Vehtari A, Gelman A, Simpson D et al (2021) Rank-Normalization, Folding, and Localization: An Improved $\widehat{R}$ for Assessing Convergence of MCMC (with Discussion). Bayesian Anal 16(2):667–718. https://doi.org/10.1214/20-BA1221
38. Viallon J, Choteau T, Flores E et al (2023) CCQM-K68.2019, nitrous oxide (N_2O) in air, ambient level, final report. Metrologia 60(1A):08011. https://doi.org/10.1088/0026-1394/60/1A/08011
39. Viechtbauer W (2010) Conducting meta-analyses in R with the metafor package. J Stat Software 36(3):1–48. https://doi.org/10.18637/jss.v036.i03
40. Welch BL (1951) On the comparison of several mean values: An alternative approach. Biometrika 38(3/4):330–336. https://doi.org/10.2307/2332579
41. Wilk MB, Gnanadesikan R (1968) Probability plotting methods for the analysis of data. Biometrika 55(1):1–17. https://doi.org/10.1093/biomet/55.1.1