Anal Bioanal Chem (2012) 403:537–548
DOI 10.1007/s00216-012-5847-4

ORIGINAL PAPER

A Bayesian approach to the evaluation of comparisons of individually value-assigned reference materials

Blaza Toman · David L. Duewer · Hugo Gasca Aragon · Franklin R. Guenther · George C. Rhoderick

Received: 13 January 2012 / Revised: 3 February 2012 / Accepted: 7 February 2012 / Published online: 3 March 2012

Abstract Several recent international comparison studies used a relatively novel experimental design to evaluate the measurement capabilities of participating organizations. These studies compared the values assigned by each participant to one or more qualitatively similar materials with measurements made on all of the materials by one laboratory under repeatability conditions. A statistical model was then established relating the values to the repeatability measurements; the extent of agreement between the assigned value(s) and the consensus model reflected the participants' measurement capabilities. Since each participant used their own supplies, equipment, and methods to produce and value-assign their material(s), the agreement between the assigned value(s) and the model was a fairer reflection of their intrinsic capabilities than provided by studies that directly compared time- and material-constrained measurements on unknown samples prepared elsewhere. A new statistical procedure is presented for the analysis of such data. The procedure incorporates several novel concepts, most importantly a leave-one-out strategy for the estimation of the consensus value of the measurand, model fitting via Bayesian posterior probabilities, and posterior coverage probability calculation for the assigned 95% uncertainty intervals. The benefits of the new procedure are illustrated using data from the CCQM-K54 comparison of eight cylinders of *n*-hexane in methane.

Keywords Bayesian analysis · Degrees of equivalence · Generalized distance regression · Leave-one-out analysis · Posterior coverage probability

Electronic supplementary material The online version of this article (doi:10.1007/s00216-012-5847-4) contains supplementary material, which is available to authorized users.

B. Toman (✉)
Statistical Engineering Division,
National Institute of Science and Technology,
Gaithersburg, MD 20899-8980, USA
e-mail: blaza.toman@nist.gov

D. L. Duewer · H. G. Aragon · F. R. Guenther · G. C. Rhoderick
Analytical Chemistry Division,
National Institute of Science and Technology,
Gaithersburg, MD 20899-8390, USA

H. G. Aragon
Department of Mathematics and Statistics,
University of Massachusetts,
Amherst, MA 01003-9305, USA

Introduction

The Gas Analysis Working Group (GAWG) of the Consultative Committee for Amount of Substance–Metrology in Chemistry (CCQM) has recently conducted several international comparison studies to evaluate the capabilities of its member national metrology institutes (NMIs) for preparing and value-assigning gas mixtures [1–4]. Complete descriptions of these and many other between-NMI comparisons are publically accessible [5]. In the referenced studies, each NMI prepared one or more primary standard gas mixtures (PSMs) at pre-determined target compositions. Each NMI shipped their PSM cylinder(s) to a coordinating laboratory where the relative composition of all cylinders from all participating NMIs was measured under repeatability conditions. NMI capabilities were assessed through the comparison of the assigned values for the PSMs, consisting of both a value and its associated uncertainty, with the measurement data using a regression technology that respected the variability in both sets of data.

Note that "value assigned" is a generic term for any material that has been assigned a value and an uncertainty

on that value; these include but are not limited to certified reference materials and some proficiency test materials. PSMs are gas mixtures value-assigned using a primary reference measurement procedure [6]; such highest metrological-order materials have the shortest practical traceability chain and are typically only used within an NMI or for peer-to-peer comparisons. The CCQM comparisons are intended to evaluate higher-order capabilities and do not address whether study materials are suitable as reference materials for field measurement procedures. The GAWG's PSM comparisons are much more representative of higher-order capabilities as actually used in delivering services than studies where the individual NMIs make time- and material-constrained measurements on one or more materials prepared.

The 2006 CCQM-K54 study evaluated eight *n*-hexane-in-methane mixtures having mole fraction compositions ranging from 120 to 200 μmol/mol *n*-hexane. The data analysis published in the study's Final Report [4] determined that the assigned values for half of the mixtures were somewhat unsatisfactory. A technical root-cause was established for the discrepancy in only one of these mixtures. The relatively small number of PSMs examined in the study, the excellent precision of the repeatability measurements, and the presence of qualitatively different types of discrepancies make this a nearly ideal exemplar for exploring the fresh statistical challenges presented by this study design.

This article presents a re-analysis of the CCQM-K54 data with several new features that further illuminate the results. The following sections describe the CCQM-K54 study, review the original analysis, detail our analysis, and present the advantages of our approach. The new analysis approach is fully portable to other studies of this kind.

The CCQM-K54 experiment

Eight NMIs participated in CCQM-K54, each NMI producing and value-assigning a single cylinder of *n*-hexane mixed in methane. The target composition for each PSM was assigned by the GAWG when the design for the study was finalized. Participants were instructed to include only gravimetric preparation and purity assessment components of uncertainty their uncertainty budgets. Table 1 presents this information, where x_i is the assigned *n*-hexane value in micromoles per mole reported by the i^{th} NMI and $u(x_i)$ is the standard uncertainty associated with the assigned value.

After value-assigning their PSM, the NMIs shipped the cylinder to the coordinating laboratory. This laboratory evaluated all of the PSMs under repeatability conditions, making five independent measurements of each mixture per day on three different days using a well-characterized gas chromatographic measurement process. The measurement design included appropriate controls and atmospheric pressure measurements to identify and, if necessary, correct for within- and between-day instrumental drift. Table 2 reports the daily means and standard deviations for the pressure-adjusted indications, $\overline{y}_{ij}$ and s_{ij}, and the mean of the means and standard deviation of the means, $\overline{y}_i$ and s_i, over the three measurement campaigns. These summary estimates of the instrumental response are reported in arbitrary units.

Table 1 PSM compositions with values in micromoles per mole

Cylinder	Target value	Assigned value	
		x_i	$u(x_i)$
PSM-1	120	119.65	0.28
PSM-2	120	119.97	0.12
PSM-3	140	140.09	0.15
PSM-4	140	140.70	0.30
PSM-5	160	160.52	0.13
PSM-6	180	180.997	0.278
PSM-7	180	181.17	0.12
PSM-8	200	199.02	0.15

The original analysis

The study data was analyzed using procedures described in ISO (2001) [7]. These procedures relate instrumental indications obtained from gas samples (y) to given chemical compositions (x) using generalized distance regression (GDR), in this case fitting a straight line model

$$y = a_0 + a_1 x \tag{1}$$

Ordinary linear regression is not used because both the x and the y values have associated uncertainties.

The original data analysis first combined the 3-day averages $\overline{y}_{i1}$, $\overline{y}_{i2}$, and $\overline{y}_{i3}$ for each mixture to produce a single average $\overline{y}_i$. Since the within-day variation of the instrument response was relatively small compared with the between-day, and the between-day variation did not appear to be related to the response magnitude, the standard uncertainty for the measurement response for all mixtures was estimated by pooling the eight between-day standard deviations:

$$u(\overline{y}_i) = \sqrt{\sum s_i^2/8} = 2.26\,\text{a.u.}$$

The GDR procedure estimated $(\widehat{x}_i, \widehat{y}_i)$ for each $(x_i, \overline{y}_i)$ that minimizes the criterion

$$\sum \left[\left(\frac{\widehat{x}_i - x_i}{u(x_i)} \right)^2 + \left(\frac{\widehat{y}_i - \overline{y}_i}{u(\overline{y}_i)} \right)^2 \right].$$

Table 2 Summary statistics for the pressure-adjusted instrumental indications with values in arbitrary units

	Day 1		Day 2		Day 3		Combined	
Cylinder	$\overline{y}_{i1}$	s_{i1}	$\overline{y}_{i2}$	s_{i2}	$\overline{y}_{i3}$	s_{i3}	$\overline{y}_i$	s_i
PSM-1	1,156.80	0.79	1,165.53	0.50	1,161.81	0.67	1,161.38	2.53
PSM-2	1,152.61	0.93	1,161.22	0.76	1,157.18	1.17	1,157.00	2.49
PSM-3	1,352.81	1.21	1,361.55	0.72	1,358.65	0.93	1,357.67	2.57
PSM-4	1,359.09	0.62	1,368.99	0.98	1,367.64	0.84	1,365.24	3.10
PSM-5	1,561.25	0.35	1,566.04	1.30	1,566.04	0.87	1,564.44	1.60
PSM-6	1,758.92	1.97	1,763.80	0.87	1,764.20	0.66	1,762.31	1.70
PSM-7	1,729.73	0.43	1,735.47	0.93	1,736.47	0.68	1,733.89	2.10
PSM-8	1,933.68	1.21	1,934.83	0.76	1,929.89	1.50	1,932.80	1.49

Table 3 lists the GDR estimates for the two sets of values, $\widehat{x}_i$ and $\widehat{y}_i$, their absolute residuals $\Delta x_i = |x_i - \widehat{x}_i|$ and $\Delta y_i = |\overline{y}_i - \widehat{y}_i|$, and the uncertainty-scaled residuals, $\Delta x_i/u(x_i)$ and $\Delta y_i/u(y_i)$.

As is pointed out in Guenther and Possolo [8], the $\widehat{x}_i$ and $\widehat{y}_i$ so derived are the maximum likelihood estimates [9] of θ_i and μ_i, for a statistical model in which x_i and $\overline{y}_i$ are the observed values of Gaussian random variables X_i and Y_i with $E(X_i)=\theta_i$ and $E(Y_i)=\mu_i=a_0+a_1\theta_i$. The notation $E(.)$ represents the expectation, or in other words, the average of a random variable. Maximum likelihood estimates are widely used in classical statistics because they possess good optimality properties.

Numerous assumptions are applied in performing the GDR analysis, the most important being that all of the indication means $\overline{y}_i$ are related to the chemical compositions x_i according to a linear relationship. The fit of this model needs to be evaluated before it is used to make judgments about the correctness of the specifications accompanying each mixture. The original analysis followed [7], which recommends the model be validated by checking that the residuals satisfy the requirement that for all i

$$|x_i - \widehat{x}_i| \leq k \cdot u(x_i) \text{ and } |\overline{y}_i - \widehat{y}_i| \leq k \cdot u(y). \quad (2)$$

Since the uncertainties on both the assigned values and the repeatability measurements are asserted to be associated with a "large" number of degrees of freedom, a coverage factor of k=2 was used for these tests. The interpretation is that, if this criterion is not satisfied for a particular mixture, then either the model is inappropriate or that the assigned value and/or the repeatability measurements for that mixture are suspect.

The results for PSM-5, PSM-6, and PSM-7 do not satisfy the validation criteria. A root-cause for the disparity in PSM-7's result was established by examination of the repeatability measurement chromatograms, revealing the presence of hexane isomers. The analysis then proceeded by removing PSM-7 from the GDR calculation, re-estimating $u(\overline{y}_i) = 2.28$a.u. by pooling the s_i over just the seven sets of measurements included in the regression and re-computing the estimates and residuals. Table 4 lists the results; Fig. 1 provides an overview of the data and details the repeatability measurement residuals relative to the consensus GDR solution for this reduced-dataset. The new estimates for the seven sets of results included in the GDR analysis satisfy Eq. (2).

For completeness, we consider the statistical properties of this validation procedure. Equation (2) is a classical two-sided hypothesis test of $H_{01} : \theta_i = \widehat{x}_i$ versus $H_{A1} : \theta_i \neq \widehat{x}_i$ and of $H_{02} : \mu_i = \widehat{y}_i$ versus $H_{A2} : \mu_i \neq \widehat{y}_i$ for each mixture, done independently at approximately a 0.05 confidence level. This means that, for each test, there is a 5% chance that the criterion would not be met when the null hypothesis is in

Table 3 GDR results for the original analysis

Cylinder	$\widehat{x}_i \mu$mol/mol	$\Delta x_i \mu$mol/mol	$\frac{\Delta x_i}{u(x_i)}$	$\widehat{y}_i$ a.u.	Δy_ia.u	$\frac{\Delta y_i}{u(y_i)}$
PSM-1	119.819	0.169	0.60	1,160.25	1.13	0.50
PSM-2	119.868	0.102	0.85	1,160.73	3.73	1.65
PSM-3	140.129	0.039	0.26	1,356.76	0.91	0.40
PSM-4	140.890	0.190	0.63	1,364.13	1.11	0.49
PSM-5	160.759	0.239	1.84	1,556.37	8.07	3.57
PSM-6	181.610	0.610	2.19	1,758.13	4.18	1.85
PSM-7	180.767	0.403	3.36	1,749.97	16.08	7.12
PSM-8	199.208	0.188	1.25	1,928.40	4.40	1.95

Table 4 Results of the original analysis with PSM-7 excluded from regression

Cylinder	$\widehat{x}_i$ μmol/mol	Δx_i μmol/mol	$\frac{\Delta x_i}{u(x_i)}$	$\widehat{y}_i$ a.u.	Δy_i a.u	$\frac{\Delta y_i}{u(y_i)}$
PSM-1	119.851	0.201	0.72	1,160.02	1.36	0.59
PSM-2	119.880	0.090	0.75	1,160.31	3.30	1.45
PSM-3	140.062	0.028	0.18	1,358.32	0.65	0.29
PSM-4	140.742	0.042	0.14	1,364.99	0.25	0.11
PSM-5	160.643	0.123	0.99	1,560.25	4.19	1.84
PSM-6	181.140	0.140	0.50	1,761.34	0.96	0.42
PSM-8	198.901	0.119	0.79	1,935.61	2.81	1.23

fact true simply due to random variation. Assuming that these 16 tests were truly independent, there would be a 56% (see [10, section 2.1]) chance that at least one of the tests fails by chance. However, the tests of Eq. (2) are not really independent, as the μ_i are functions of θ_i, a_0, and a_1. It is thus difficult to say what the overall probability of a chance rejection of the set of tests really is. However, unless all of the tests are completely correlated, the probability must be greater than 0.05. Hence, there is a probability higher than 5% that a false outlier will be incorrectly identified. The usual procedure to assure that a set of tests do not falsely reject too many times is called the Bonferroni inequality [10, section 2.1]. It divides the target confidence level by the number of tests and uses an expansion factor that corresponds to that adjusted confidence level. For the original analysis of results for all eight participants, this would mean applying Eq. (2) with k=2.95, identifying only PSM-5 and PSM-7 as potential outliers.

A clarification is also necessary of what it means when a PSM satisfies Eq. (2). For such a mixture, it can only be said that the *data did not provide any proof* that the model does not fit well or that the measurements and/or the mixture are outliers. There is no probability estimate associated with this "absence of proof" conclusion.

After elimination of PSM-7 on technical grounds and validating the revised model, the original analysis proceeded to evaluate the degrees of equivalence, D_i, for the eight PSMs. The D_i were estimated as

$$D_i = x_i - \widehat{x}_i$$

The $\widehat{x}_i$ is the GDR-estimate for the seven PSMs included in the GDR analysis. However, for PSM-7,

$$\widehat{x}_i = \frac{\overline{y}_i - \widehat{a}_0}{\widehat{a}_1}$$

where $\widehat{a}_0$ and $\widehat{a}_1$ are the GDR estimates of intercept and slope.

For the PSMs included in the GDR analysis, the uncertainty in the degrees of equivalence, $u(D_i)$, was estimated as

$$u(D_i) = \sqrt{u^2(x_i) - u^2(\widehat{x}_i)}. \quad (3)$$

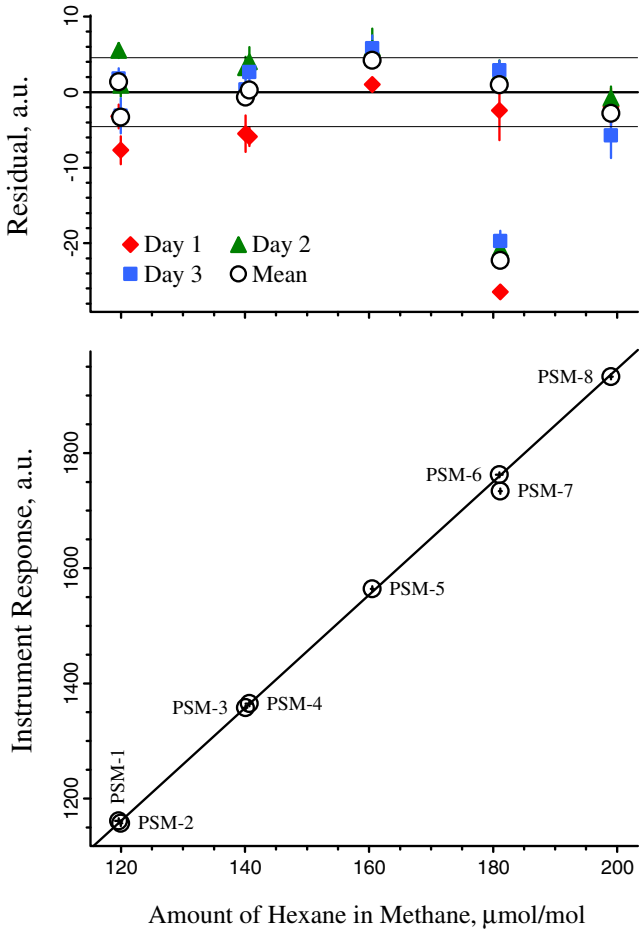

Fig. 1 Observed data and residuals from the consensus regression model. The *lower* panel displays the mean of the repeatability measurements as a function of the assigned values for the eight PSMs. The *open circles* enclose 2*u* "*error bar*" intervals along both axes; these intervals are barely visible at this graphical scale. The *line* represents the generalized distance regression solution when PSM–7 is excluded. The *upper* panel plots the repeatability measurement residuals from this consensus model for the three sets of within-day measurements and for the grand mean, with 2*u* bars on the within-day means. The *thick horizontal line* again represents the GDR solution; the *thin horizontal* lines represent the 95% level of confidence uncertainty interval (±4.56 μmol/mole) assigned to the repeatability measurement process

For PSM-7, the uncertainty was estimated as

$$u(D_i) = \sqrt{u^2(x_i) + u^2(\widehat{x}_i)}.$$

The standard uncertainties of all the $\widehat{x}_i$, $u(\widehat{x}_i)$ were estimated from first-order Taylor's series propagation of uncertainty using the GDR-estimated uncertainties and covariance of the regression parameters. Table 5 lists the resulting estimates.

Based on these results, using the criterion that

$$\frac{|D_i|}{u(D_i)} \leq 2 \tag{4}$$

the CCQM-K54 Final Report concluded that the measurements for the mixtures produced by PSM-7, PSM-2, PSM-5, and PSM-8 were inconsistent with the consensus GDR model, but that the other four mixtures were consistent with it [4].

It is again useful to review the above reasoning from a statistical viewpoint. Equation (4) represents a classical two-sided hypothesis test of the null hypothesis that the PSM contents as defined by the NMI and as estimated by the repeatability measurements are equal, again performed independently at approximately a 0.05 level of confidence. The mixtures for which $|D_i|/u(D_i) > 2$ (i.e., PSM-7, PSM-2, PSM-5, and PSM-8) are mixtures for which the null hypothesis was rejected. This procedure is at best only approximate because $u(D_i)$ as defined by Eq. 3 may be incomplete (see [11, section 1.2.3]). In any case, the same implications hold here as in the use of Eq. (2). That is, although there is approximately a 5% chance individually for each test that a discrepant result could be obtained by chance even though the actual mixture content was correctly specified, the chance for at least one of eight PSMs to be misidentified is higher than 5%.

The remaining mixtures are those for which the null hypothesis was not rejected, and for these, one may again only say that the data obtained by the experiment *did not prove* that the mixture content as defined by the NMI and as estimated by the repeatability measurements *are not equal*.

In the following section, an alternative method of analysis is proposed, one that in our view is better able to extract all of the available information from the experiment and thus better evaluate the NMI's gravimetry and purity verification capabilities.

Table 5 Degrees of equivalence

Cylinder	$\widehat{x}_i \mu$mol/mol	$u(\widehat{x}_i) \mu$mol/mol	D_i	$u(D_i)$	$\frac{D_i}{u(D_i)}$
PSM-1	119.85	0.21	−0.20	0.19	−1.1
PSM-2	119.88	0.11	0.09	0.04	2.3
PSM-3	140.06	0.13	0.03	0.07	0.4
PSM-4	140.74	0.20	−0.04	0.23	−0.2
PSM-5	160.64	0.11	−0.12	0.05	−2.4
PSM-6	181.14	0.20	−0.14	0.19	−0.7
PSM-7	178.34	0.28	2.83	0.30	9.4
PSM-8	198.90	0.14	0.12	0.05	2.4

Re-analysis based on the Bayesian approach

In our view, the classical statistical approach adopted by the original analysis is not able to extract all of the available information from the study. Classical statistical methods rely on probability distributions of data conditional on parameters, here, for example, that the X_i and Y_i follow Gaussian distributions with means θ_i and μ_i. The probabilities associated with them, like the 0.05 probability of false rejection of a null hypothesis, do not provide direct measures of how likely the null hypothesis is. Classical hypothesis tests are not capable of answering the questions that we would really like to have answered, for example: "Given that Eq. (2) was not satisfied, what is the probability that the mixture is *truly* an outlier?" Similarly, "Given that Eq. (2) was satisfied, what is the probability that the mixture is properly specified?" Computation of these probabilities requires the inversion from a distribution of data conditional on parameters to one for the parameters conditional on the data. This is provided by Bayes Theorem, and for this reason, we adopt the Bayesian approach [12] which is better suited to this analysis and at the same time is fully consistent with the ISO Guide to the Expression of Uncertainty in Measurement [13] and its supplement 1 [14]. We believe that the advantage of using this approach is fully demonstrated by this example.

In addition to the choice of statistical paradigm, there are a number of features of the original analysis method that we view as problematic and in need of improvement. Perhaps the most important drawback of the method applied [4] is that it uses the information about the mixture being verified to first fit the GDR model and then to verify this same information. When intrinsically discrepant data are present (as was the case above with PSM-7), this results in incorrect assessment. With this observation in mind, we take inspiration from the procedure described in ISO Guide 6143 [7] for building an analysis function to be used to value-assign gas certified reference materials (CRMs). The Guide 6143 procedure uses PSMs to build the function and then uses the measured instrumental indication y of the CRM to estimate its corresponding x. In the same way, we may proceed here. For each PSM, in turn, use the others to estimate the analysis function and then apply it to the instrument indication $\bar{y}_i$ to estimate its $\widehat{x}_i$. This "leave-one-out" approach is what was actually used in the evaluation of the degrees of equivalence

for PSM-7 because it was identified as a true outlier caused by incomplete purity assessment of the *n*-hexane used in its preparation. Here, we suggest that all of the materials be treated equally.

Another potential problem with the original analysis is its reliance on the Gaussian distribution. Specifically, this determines the expansion factors k in the two criteria which determine the fit of the model. In Guenther and Possolo [8], it was suggested that, when the uncertainties of the x_i and the y_i are based on a small number of observations and thus have small degrees of freedom, it is more appropriate to use a Student's t distribution rather than a Gaussian. The maximum likelihood estimates are then obtained by optimization of the slightly different criterion given in Guenther and Possolo [8]. Here, a similar procedure is followed to account for the uncertainty in the y_i. Since the $u(x_i)$ are asserted to be expandable to approximate 95% coverage intervals using k=2, they are all assumed to be associated with a "large" number of degrees of freedom and thus to be well characterized as Gaussian distributions.

Since the original analysis pooled the repeatability measurements before performing GDR, it did not allow for the estimation of any uncertainty component due to incomplete elimination of between-day differences in the measurement process. The confounding of uncertainty components may lead to over-estimating the $u(y_i)$. For this reason, our reanalysis does not pool the data.

The statistical model

We use a statistical model where the five measurements of PSM i on day j are observed values y_{ijk} of Gaussian random variables

$$Y_{ijk}|\alpha_{ij},\sigma_{ij}^2 \sim N\left(\alpha_{ij},\sigma_{ij}^2\right), i=1,\ldots,8, j=1,\ldots,3, k=1,\ldots,5. \quad (5)$$

The notation $Y|\alpha$ represents conditioning, that is, the probability distribution of the random variable Y given a specific value for the random variable α. This is necessary in the Bayesian framework because parameters such as means and variances have distributions which represent our state of knowledge about them.

To account for potential differences between measurements made on different days, we assume that the day means α_{i1}, α_{i2}, and α_{i3} are related to each other because these measurements are made on the presumptively unchanging contents of the same cylinder. However, the measurements may also reflect possible "day" effects that the atmospheric pressure measurements did not completely adjust for. This can be modeled as

$$\alpha_{ij}|\mu_i,\tau^2 \sim N\left(\mu_i,\tau^2\right), i=1,\ldots,8, j=1,\ldots,3. \quad (6)$$

This makes the "day" effect a random variable with variance τ^2. Pooling across mixtures is expressed in this model by having the same τ^2 for all eight mixtures, implying that the "day" effect is due to the same cause for measurements of all PSMs. This pooling is sensible as all measurements were made under the same conditions and is far less severe than the pooling done in the original analysis where all of the σ_{ij}^2 are assumed to be equal.

Assuming the same linear relationship between the analyte content and the measured values assumed in Eq. (1) we again have

$$\mu_i = a_0 + a_1\,\theta_i. \quad (7)$$

Combining Eqs. (5) to (7) it follows that

$$\overline{Y}_{ij}|a_0,a_1,\theta_i,\tau^2,\sigma_{ij}^2 \sim N\left(a_0+a_1\theta_i,\tau^2+\frac{\sigma_{ij}^2}{5}\right). \quad (8)$$

Note that the repeatability measurement data in Table 2 are the day averages, $\overline{y}_{ij}$, and day standard deviations, $s_{ij}=\sqrt{\sum_{k=1}^{5}\left(y_{ijk}-\overline{y}_{ij}\right)^2/4}$. For each mixture and day, the variance σ_{ij}^2 can be estimated via these s_{ij} since random variables S_{ij}^2 follow a Gamma distribution with degrees of freedom 2 and $\frac{1}{2\sigma_{ij}^2}$:

$$S_{ij}^2|\sigma_{ij}^2 \sim Gamma\left(2,\frac{1}{2\sigma_{ij}^2}\right). \quad (9)$$

This is the same condition, in a slightly different form, as was used in Guenther and Possolo [8].

Finally, we assume that, for each mixture, the x_i are observed values of Gaussian random variables

$$X_i|\theta_i,u^2(x_i) \sim N\left(\theta_i,u^2(x_i)\right). \quad (10)$$

So far, with the exception of details dealing with the uncertainty of the indications, this statistical model is essentially the same as what was used by the original analysis. But, unlike the classical, the Bayesian model requires additional components in the form of the so-called prior distributions [12] for parameters a_0, a_1, τ, and the θ_i and σ_{ij}. These probability distributions represent our prior knowledge of these parameters, that is, our knowledge before the experiment was performed. If no such prior knowledge exists, as is true here, they are assigned so-called noninformative prior distributions. One possible choice is to assign in each case a uniform density on a wide interval. As this choice is not the only prior distribution that can be reasonably made, it is important to consider possible alternatives and see if the analysis results are insensitive to the

choice of non-informative prior. This is the case for this particular data set.

Application of Bayes Theorem [12] results in a posterior distribution of the quantities of interest: the distribution of $\theta_i|\overline{y}_{i1},\overline{y}_{i2},\overline{y}_{i3},s_{i1},s_{i2},s_{i3},x_i$ or, in words, the distribution of the parameters conditional on the observed data. This is the inversion of probability that was previously discussed and in most cases needs to be achieved numerically as closed-form solutions are available only in the simplest cases. The most common numerical approach is Markov Chain Monte Carlo (MCMC), often applied using freeware systems such as WinBUGS and OpenBUGS [15]. However accomplished, the MCMC process produces "draws" (representative values) from the posterior distributions of the parameters. Using a large number of such draws, the means, standard deviations, and probability intervals for the parameters can be estimated. Once the model is fully defined (and debugged), obtaining reliable estimates using this computationally intensive procedure typically requires only a few wall-clock minutes on a contemporary personal computer.

The analysis

The steps of the proposed analysis are described in this section. Before proceeding with the actual calculation of the estimates, as it was in the original analysis, it is necessary to examine how well the linear model fits the data, that is, to identify any outlying observations. In the original analysis, this was done using Eq. (2). In a Bayesian analysis, the usual method of checking model fit is to compute posterior predictive probabilities (Bayesian posterior p values) $P\left(\overline{Y}_i > \overline{y}_{iobs}\right)$ [16], where $\overline{y}_{iobs}$ are the observed overall means for each mixture obtained from Table 2. (See [17] for an application of posterior predictive probabilities in a more usual interlaboratory study design). These Bayesian p values measure how likely it is to obtain the value $\overline{y}_{iobs}$ given our model and all of the observed data.

For the CCQM-K54 data, the probabilities are computed using the posterior predictive distribution of $\overline{Y}_i$, that is, the likelihood function

$$\overline{Y}_i|a_0,a_1,\theta_i,\tau^2,\sigma_{ij}^2 \sim N\left(a_0+a_1\theta_i,\tau^2+\frac{1}{15}\sum_{j=1}^{3}\sigma_{ij}^2\right)$$

integrated over the posterior distributions of the parameters a_0, a_1, τ, θ_i, and σ_{ij}. This calculation is also achieved using MCMC. Fitting the model to all eight observation pairs using the WinBUGS program listed in the Electronic supplementary material produces the Bayesian p values given in Table 6. The 0 value for PSM-7 identifies it as a definite outlier while the values for PSM-5 and PSM-6 are too low for a good fit. In this particular application, we benefit from the knowledge that the assigned value for PSM-7 was technically flawed. Observing that the target n-hexane content of PSM-6 was the same as that for PSM-7, it is quite plausible that including the miss-assigned value for PSM-7 strongly distorts the result for PSM-6. As in the original analysis, refitting the model without PSM-7 produces better results. The Bayesian p values given in the third column of Table 6 are all larger than 0.05.

Table 6 Bayesian posterior p values

Cylinder	All cylinders	PSM-7 excluded
PSM–1	0.34	0.30
PSM–2	0.12	0.13
PSM–3	0.44	0.39
PSM–4	0.32	0.46
PSM–5	0.003	0.06
PSM–6	0.04	0.39
PSM–7	0	–
PSM–8	0.15	0.14

The rest of the analysis is now done without the PSM-7 data. The following steps are performed for each cylinder in turn:

1. For a particular cylinder i', apply the model given in Eqs. (5) to (10) to the cylinders 1,2,…,$i'-1$,$i'+1$,…,8 using their measurements y_{ijk} and x_i to produce posterior densities of a_0, a_1, and τ.
2. Use the measurements for cylinder i', $y_{i'jk}$, $j=1,\ldots,3$, $k=1,\ldots,5$, to apply the model given in Eqs. (5) to (9) with a non-informative (uniform) prior for the quantity $\theta_{i'}$, and the posterior densities of the a_0, a_1, and τ. Implement via MCMC to obtain draws from the posterior distribution of $\theta_{i'}$.
3. The mean of these draws is an estimate of the true amount of measurand, and the standard deviation is an

Table 7 Re-analysis results

	Arbitrary units		µmol/mol		µmol/mol		
Cylinder	$\widehat{y}_i$	$u(\widehat{y}_i)$	$\widehat{x}_i$	$u(\widehat{x}_i)$	D_i	$u(D_i)$	$(D_{i,0.025}, D_{i,0.975})$
PSM-1	1,161.4	2.81	120.07	0.35	−0.42	0.45	−1.32, 0.48
PSM-2	1,157.1	2.74	119.23	0.36	0.74	0.38	−0.01, 1.49
PSM-3	1,357.8	2.92	139.97	0.33	0.11	0.36	−0.59, 0.84
PSM-4	1,365.2	2.93	140.74	0.31	−0.05	0.43	−0.89, 0.79
PSM-5	1,564.3	2.67	161.17	0.31	−0.65	0.34	−1.31, −0.01
PSM-6	1,762.5	2.92	181.33	0.36	−0.34	0.45	−1.24, 0.55
PSM-7	1,733.9	2.74	178.31	0.33	2.86	0.35	2.17, 3.55
PSM-8	1,933.0	2.66	198.11	0.41	0.91	0.44	0.01, 1.72

estimate of the standard uncertainty. These are labeled as $\widehat{x}_i$and $u(\widehat{x}_i)$ in Table 7. The endpoints of the central 95% of the draws, $D_{i,0.025}$ and $D_{i,0.975}$, estimate the 95% coverage interval.

4. The posterior means and standard deviations of the μ_i can also be obtained and are labeled $\widehat{y}_i$ and $u(\widehat{y}_i)$.

The computations were done using the WinBUGS program listed in the Electronic supplementary material and produced the results given in Table 7.

In the Bayesian model, it is possible to quantify the knowledge about the difference (call it δ_i) between the true quantity of the mixture content (posterior knowledge of θ_i based on measurements) and the assigned value provided by the laboratory (lab-specified knowledge of θ_i in terms of x_i and $u(x_i)$) in terms of a probability distribution. We will call the expected value of δ_i the degree of equivalence D_i. This is in fact the same as in the original analysis, that is, $D_i = x_i - \widehat{x}_i$, the difference between the specified value and the posterior mean of θ_i of the mixture content. The posterior standard deviation of this distribution estimates the standard uncertainty, $u(D_i)$. Unlike the classical confidence interval on which Eq. (4) is based, the 95% uncertainty interval ($D_{i,0.025}$, $D_{i,0.975}$) is computed from the endpoints of the central 95% of

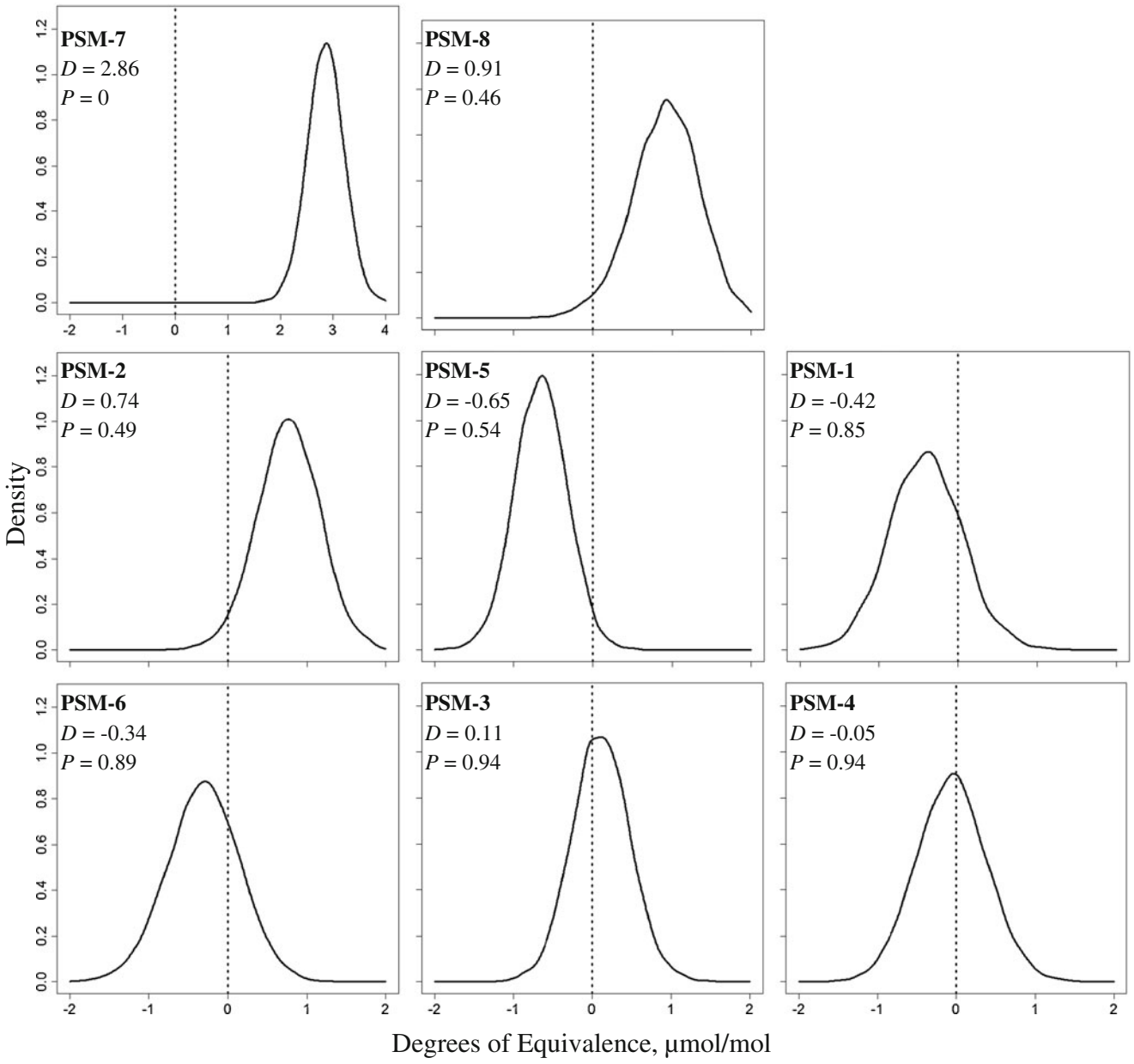

Fig. 2 Density plots of the degrees of equivalence. The curve in each panel represents a Bayesian estimate of the probability density function for the δ_i for the PSM relative to the consensus reference function. The degree of equivalence D_i is also given. The *vertical lines* mark the ideal equivalence value of zero

posterior distribution and has the property that $P\left(D_{i,0.025} < \delta_i < D_{i,0.975}\right) = 0.95$.

We first note that it is quite likely that the true quantity of the mixture content is in fact what was specified by the NMI when the uncertainty interval includes the value 0. This conclusion can be drawn about all of the mixtures except for PSM-5, PSM-7, and PSM-8. The 95% interval for PSM-7 is far from 0; those for PSM-5 and PSM-8 are marginally so.

The advantage of the Bayesian approach over the classical is that we can say much more than this about the mixtures. Figure 2 shows the probability densities of δ_i for all of the PSMs, one panel per PSM in order of improving $|D_i|$. By observing the position of 0 under the curve, it becomes clear that agreement between what is specified and what is measured is much more likely for some of the mixtures than for others. For example, it is much more likely for PSM-3 than for PSM-2 although the shapes of the density functions are rather similar.

To better quantify this observation, we can compute the probability that δ_i lies in a given interval, thus giving a direct

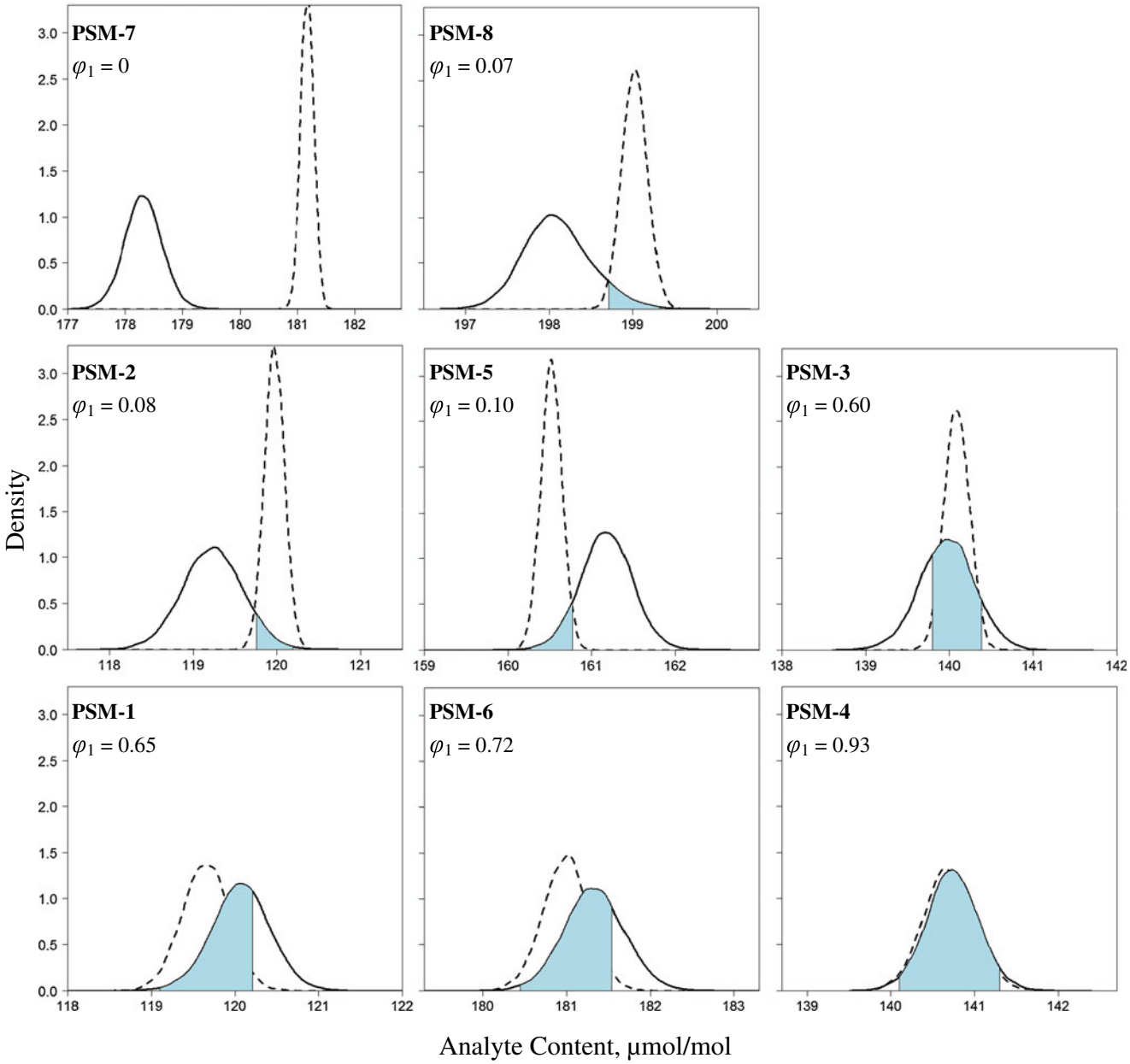

Fig. 3 Assigned and consensus probability densities for the analyte content, θ. The *dashed line* in each panel represents the Gaussian probability density function for the analyte content as assigned by the NMI. The *solid line* represents the density function for the Bayesian estimate of analyte content as calibrated through the consensus reference function. The *shaded area* represents the probability content φ_1, the fractional area of the posterior distribution of θ_i that lies within the 95% uncertainty interval that was assigned by the NMI

probability of such an event. For example, a measure of how likely it is that the assigned value of a mixture is well specified is the probability that δ_i is contained within in the interval from $-2u(D_i)$ to $2u(D_i)$: $\mathrm{P}_i = P(-2u(D_i) < \delta_i < 2u(D_i))$. The D_i and P_i values for each PSM are provided in the corresponding panel of Fig. 2.

A more intuitive method of evaluating the specification of the mixture content under the Bayesian paradigm is the following. For each mixture, we can obtain the probability content, φ_{1i}, of its certified 95% uncertainty interval computed under the posterior distribution for θ_i. If the mixture is well specified, φ_{1i} should be large since the usual understanding of the certified uncertainty interval is that it contains the true value with probability 0.95. The posterior density represents our best knowledge of θ_i after the experiment. Figure 3 compares the intervals for all of the PSMs, one panel per PSM with the panels in order of increasing φ_{1i}.

Figure 3 shows that the assigned value interval for PSM-7 is not covered and those for PSM-8, PSM-2, and PSM-5 are poorly covered. The assigned values for all of these PSMs are biased relative to the consensus model and have uncertainties that appear to be too small. These mixtures were identified by the original analysis as "not consistent with the [Key Comparison Reference Value] KCRV". The

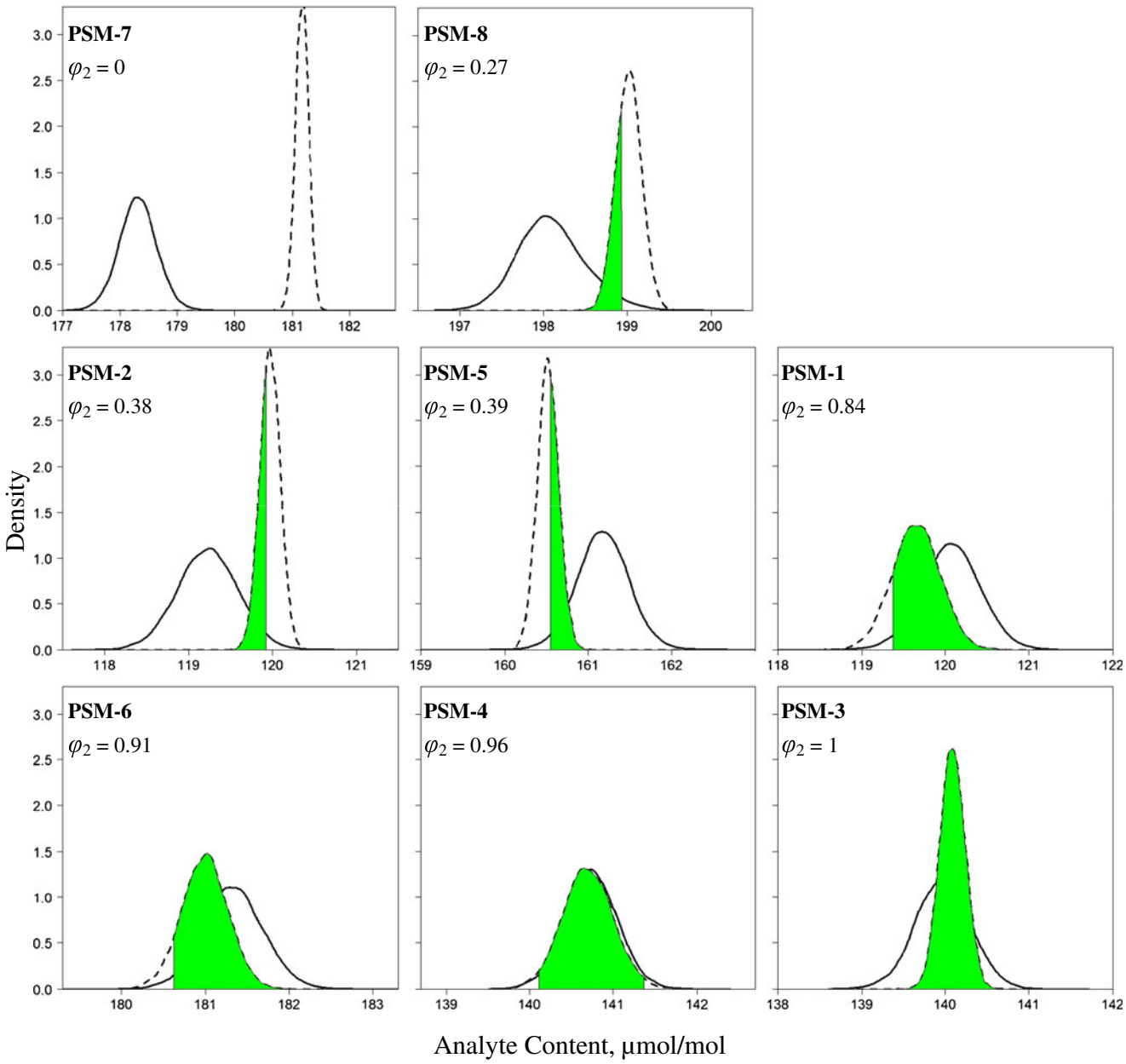

Fig. 4 Assigned and consensus probability densities for the analyte content, θ. As in Fig. 3 but with the shaded area representing the probability content φ_2, the fractional area of the 95% uncertainty interval assigned by the NMI that lies within posterior distribution of θ_i

remaining four PSMs achieve coverage over 60% but only PSM–4 is close to 95%. The figure shows that PSM-6 and PSM-1 are somewhat biased relative to the consensus model, but their uncertainties are large enough to provide good probability coverage. PSM-3 has a very small bias, but because its specified uncertainty is rather small, its φ_{1i} is only 0.60. In comparison, PSM–4 achieves coverage probability of 0.93 since its specified uncertainty is close to the uncertainty obtained in the experiment.

This illustrates an important point: The posterior uncertainties are a function not only of the measurement repeatability but also of the consensus fit of the set of PSMs to the linear model. One could argue that this particular experiment resulted in posterior uncertainties that are too large. If the posterior uncertainty for PSM–3 was smaller, that is, more in line with the assigned uncertainty, then the φ_{1i} for PSM–3 would improve. However, because of their biases, the φ_{1i} of some of the other PSMs would be made worse. In fact, if one could achieve the perfect experiment, then the posterior density of each θ_i would be a point mass at the true value. In such a case, φ_{1i} would equal 1 if the point mass was within the 95% specified uncertainty interval and 0 otherwise. If the posterior densities remained centered, as they are in Fig. 3, then the solid curves would shrink to the posterior mean (corresponding to the center). The φ_{1i} of PSM–3, PSM–1, PSM–6, and PSM–4 would be equal to 1 and the others would be 0. From this, one can conclude that if a PSM has φ_{1i} of 0.5 or more then it is likely well specified.

Another way to view this data is to consider the probability content φ_{2i} of the 95% posterior probability interval under the specified probability density. Since the posterior probability interval has 0.95 probability of containing the true value, large φ_{2i} is desirable. Figure 4 compares the intervals for all of the PSMs, one panel per PSM with the panels in order of increasing φ_{2i}. Again it is instructive to consider the perfect experiment which would result in the 95% posterior probability interval being a single point. Then φ_{2i} would be the probability, computed under the specified density, that θ_i is equal to this point. If we again assume that the centers of the posterior densities remain the same, then the solid curves in Fig. 4 shrink to their center points. PSM–3 and PSM–4 would then have high values of φ_{2i} while the others would be much smaller, with the values of PSM–7, PSM–8, PSM–2, and PSM–5 being close to 0.

Summary

We examined the published analysis of CCQM-K54 critically and identified some aspects that should be more fully considered in future studies. We identified and explained how the original analysis's reliance on the frequentist theory of probability was not able to extract all of the available information from the data. We presented an alternative analysis which was able to draw stronger, more quantitative conclusions. Our analysis used several novel concepts, most importantly a Bayesian framework, a leave-one-out strategy for the estimation of $\widehat{x}_i$, model fitting via Bayesian posterior probabilities and posterior coverage probability calculation for the specified 95% uncertainty intervals. We showed the added benefit that these brought to the analysis of CCQM-K54 data. Our analysis methods are appropriate for other comparisons of individually value-assigned reference materials. With extension of the statistical model to accommodate measurements on multiple units, the methods can be made appropriate for comparisons of batch-assigned materials.

Acknowledgment We thank the GAWG and its members for pioneering the comparison of multiple reference materials value-assigned by different organizations using measurements made under repeatability conditions by one organization.

References

1. van der Veen AMH, Brinkmann FNC, Arnautovic M et al (2007) International comparison CCQM-P41 greenhouse gases. 2. Direct comparison of primary standard gas mixtures. Metrologia 44:08003
2. Wielgosz RI, Esler M, Viallon J et al (2008) International comparison CCQM-P73: nitrogen monoxide gas standards (30–70) μmol/mol. Metrologia 45:08002
3. Lee J, Lee JB, Moon DM et al (2010) Final report on international key comparison CCQM-K53: oxygen in nitrogen. Metrologia 47:08005
4. van der Veen AMH, Chander H, Ziel PR et al (2010) International comparison CCQM-K54: primary standard gas mixtures of hexane in methane. Metrologia 47:08019
5. The BIPM key comparison database, http://kcdb.bipm.org/
6. JCGM 200:2008. International vocabulary of metrology—basic and general concepts and associated terms (VIM). Joint Committee for Guides in Metrology (JCGM), Sèvres, France (2008) http://www.bipm.org/en/publications/guides/vim.html
7. ISO. ISO 6143:2001(E) Gas analysis—comparison methods for determining and checking the composition of calibration gas mixtures. International Organization for Standardization (ISO), Geneva (2001)
8. Guenther FR, Possolo A (2011) Calibration and uncertainty assessment for certified reference gas mixtures. Anal Bioanal Chem 399:489–500
9. Lehman EL (1999) Elements of large-sample theory. Springer, New York
10. Miller RG (1981) Simultaneous statistical inference, 2nd edn. Springer, New York
11. Fuller W (2006) Measurement error models. New York, NY: Wiley & Sons
12. Jeffreys H (1961) Theory of probability. New York, NY: Oxford University Press
13. JCGM 100:2008. Evaluation of measurement data—guide to the expression of uncertainty in measurement. Joint Committee for Guides in Metrology (JCGM), Sèvres, France (2008) http://www.bipm.org/en/publications/guides/gum.html
14. JCGM 101:2008. Evaluation of measurement data—supplement 1 to the "Guide to the expression of uncertainty in measurement"—

propagation of distributions using a Monte Carlo method. Joint Committee for Guides in Metrology (JCGM), Sèvres, France (2008) http://www.bipm.org/en/publications/guides/gum.html

15. Lunn DJ, Spiegelhalter D, Thomas A, Best N (2009) The BUGS project: evolution, critique and future directions (with discussion). Stat Med 28:3049–3082
16. Gelman A, Meng XL, Stern H (1996) Posterior predictive assessment of model fitness via realized discrepancies (with discussion). Stat Sin 6:733–807
17. Kacker R, Forbes A, Kessel R, Sommer K-D (2008) Bayesian posterior predictive *p*-value of statistical consistency in interlaboratory evaluations. Metrologia 45:512–523