ORIGINAL PAPER

Calibration and uncertainty assessment for certified reference gas mixtures

Franklin R. Guenther · Antonio Possolo

Abstract The weighted least squares method to build an analysis function described in ISO 6143, *Gas analysis—Comparison methods for determining and checking the composition of calibration gas mixtures*, is modified to take into account the typically small number of instrumental readings that are obtained for each primary standard gas mixture used in calibration. The theoretical basis for this modification is explained, and its superior performance is illustrated in a simulation study built around a concrete example, using real data. The corresponding uncertainty assessment is obtained by application of a Monte Carlo method consistent with the guidance in the Supplement 1 to the *Guide to the expression of uncertainty in measurement*, which avoids the need for two successive applications of the linearizing approximation of the conventional method for uncertainty propagation. The three main steps that NIST currently uses to certify a reference gas mixture (homogeneity study, calibration, and assignment of value and uncertainty assessment), are described and illustrated using data pertaining to an actual standard reference material.

F. R. Guenther
Material Measurement Laboratory, Analytical Chemistry Division, Gas Metrology Group, National Institute of Standards and Technology, 100 Bureau Drive, Stop 8393, Gaithersburg, MD 20899-8393, USA
e-mail: franklin.guenther@nist.gov

A. Possolo (✉)
Information Technology Laboratory, Statistical Engineering Division, National Institute of Standards and Technology, 100 Bureau Drive, Stop 8980, Gaithersburg, MD 20899-8980, USA
e-mail: antonio.possolo@nist.gov

Introduction

The National Institute of Standards and Technology (NIST) provides Standard Reference Materials (SRM) to transfer traceability to SI units to customers in a form that is useful to their processes. For analytical chemists, this usually takes the form of a sample in a suitable matrix, certified for a particular chemical species, with a specified 95% confidence limit and traceable to mass (kilogram) or to the mole.

The Gas Metrology group of NIST's Analytical Chemistry Division provides gas standard SRMs in aluminum cylinders for this purpose to customers worldwide. To assure traceability, the group uses primary standard gas mixtures (PSM) to assign composition values to these gas SRMs.

PSMs are prepared gravimetrically at NIST from pure components, then tested at NIST to assure accuracy and consistency, and validated through international bilateral and multilateral studies including "Key Comparisons" performed under the auspices of the *Comité Consultatif pour la Quantité de Matière* of the *Comité International des Poids et Mesures* [1, 5]. A high precision instrument is used at NIST to compare the potential SRM to multiple PSMs, and a certified value and uncertainty is derived from the comparison data.

ISO 6143 [10] is the internationally recognized standard used to certify gas SRMs. It describes a method to build an analysis function that, once applied to

instrumental indications obtained from gas samples, assigns values to them. The same standard also describes how to assess the uncertainty associated with these values.

This standard addresses a problem often encountered in analytical chemistry, where standards of known composition are employed to build a function, relating chemical composition c to instrumental response r, that, once applied to the instrumental response produced by a sample of unknown composition, produces an estimate of this composition.

This problem is different from the problem that conventional regression methods [6] have been conceived for, where one assumes that the values of one of the variables are known without error, and that only the values of the other are affected by measurement error. Indeed, in the situations we are concerned with, both c and r have non-negligible uncertainties. The uncertainty of the former is listed in the calibration certificate for the PSMs. The uncertainty of the latter expresses uncontrolled variations and drift in the instrumental response and in the experimental conditions. This defines the so-called "errors-in-variables" problem [8].

Generally, it is inappropriate to fit a calibration function disregarding the uncertainties of the mole fractions of measurand in the PSMs, and considering only the uncertainties of the instrumental responses. The calibration function maps instrumental responses to mole fractions, and it is the inverse of the analysis function [10, 2.4, p. 2]. If a linear relationship were so fitted by ordinary least squares then the slope of the regression line would be biased towards zero [2, 3.2.1]. The analysis function obtained by inversion of this calibration function would be similarly biased.

The method described by ISO 6143 (which is reviewed in "ISO Guide 6143—original procedure") recognizes and takes into account uncertainties in both variables, and then uses the resulting analysis function to assign values to the gas samples, and produces a complete uncertainty estimate consistently with the *Guide to the expression of uncertainty in measurement* (GUM) [11]. Since the instrumental responses and the mole fractions play symmetrical roles in this context, we use the term *calibration* generically, to denote the process whereby one determines either the analysis function, or the calibration function.

Although that standard does not specify explicitly the conditions that validate the method that it describes, examination of the method does allow one to infer what these conditions are: that the measurement errors in play all are like realized values, or outcomes, of independent Gaussian random variables whose standard deviations are known. In these circumstances, the resulting estimates are so-called maximum likelihood estimates, and therefore possess commendable statistical optimality properties ([3, 6.3], [14]).

Obviously the method of ISO 6143 can be employed in circumstances different from these, but then there is no guarantee that the estimates it produces use the data in the best manner possible. In addition, in many cases of practical importance the circumstances are different from those. "ISO Guide 6143—modified procedure" includes a motivating example to illustrate the penalty that neglecting such differences may entail.

In particular, the uncertainties associated with the instrumental indications obtained for the PSMs typically are obtained by a type A evaluation [13, 2.28] and reflect the dispersion of values of only a handful of replicates. Therefore, the corresponding uncertainty assessments are based on a small number of degrees of freedom, not on the infinitely many that would be required to render the ISO 6143 method optimal: recognizing this affects not only the values assigned to the samples, but also the associated uncertainties.

The principal contributions of this study are: (1) a modified version of the method of ISO 6143 to address situations where the uncertainty assessments are based on small numbers of degrees of freedom ("ISO Guide 6143—modified procedure"); (2) a Monte Carlo method for uncertainty analysis that is consistent with [12], and that captures all the recognized sources of uncertainty in play ("Uncertainty assessment"); and (3) the integration of (1) and (2) into the complete process that is currently used at NIST to assign values to gas SRMs ("Example: NIST SRM 1685"). "Conclusions" consolidates our conclusions and recommendations.

ISO Guide 6143—original procedure

Most calibration procedures, and ISO 6143 in particular, seek to establish a relationship between two quantities (which we assume to be scalar), ξ and η, which will be used to build a function that predicts values of one given values of the other. In practice, however, these quantities are not observable directly, and only noisy versions thereof, $x = \xi + \delta$ and $y = \eta + \epsilon$, are accessible to observation, where ϵ and δ denote non-observable measurement errors [2].

The quantity ξ may represent the mole fraction of a particular species in a gas mixture, and η may represent the "true" response that this mixture would induce in a particular instrument. In this context, when the relationship is expressed in the form $\eta = F(\xi)$, ISO 6143 calls F the *calibration function*, and when it is expressed as $\xi = G(\eta)$ this standard calls G the *analysis function*.

In both cases, the goal is to select a functional form for the function of interest, and then to estimate any free parameters in it based on m pairs of values $(x_1, y_1), \ldots, (x_m, y_m)$. This "errors-in-variables" problem, of fitting a functional form (linear, polynomial, exponential, or other) to pairs of values both of which are affected by measurement error, has a long history: [27] provides a concise review, and [2] discusses the main issues for linear and non-linear errors-in-variables models.

The case we are primarily interested in concerns the construction and use of an analysis function that, given an instrumental response, will produce an estimate of the mole fraction of some target molecular species in a gas mixture. We suppose that the functional form of this analysis function has been determined already (an issue discussed in "Example: NIST SRM 1685"), and we write it as G_β, where β denotes a vector of p scalar parameters that determine the function specifically: for example, if the analysis function should be a polynomial of the second degree, then $p = 3$.

ISO 6143 [10, 5.1] suggests a procedure, involving 11 steps, to determine the analysis function, to use it to estimate compositions of gas mixtures, and to assess the corresponding uncertainty. In particular (step H), the standard notes that the calculation of the parameters of the analysis function should take into account the uncertainties $u(x_1), \ldots, u(x_m)$ of the compositions $x_1, \ldots, x_m$ of the reference gas mixtures, as well as the uncertainties $u(y_1), \ldots, u(y_m)$ of the measured responses $y_1, \ldots, y_m$, and suggests that one should find the values of $\eta_1, \ldots, \eta_m$ and β that minimize

$$S(\beta, \eta_1, \ldots, \eta_m) = \sum_{i=1}^{m} \left[\frac{(x_i - G_\beta(\eta_i))^2}{u^2(x_i)} + \frac{(y_i - \eta_i)^2}{u^2(y_i)} \right].$$

(1)

The statistical model that makes such procedure optimal has the following ingredients: (1) $x_i = \xi_i + \delta_i$, $y_i = \eta_i + \epsilon_i$, and $\xi_i = G_\beta(\eta_i)$ for $i = 1, \ldots, m$; (2) $\delta_1, \ldots, \delta_m$, $\epsilon_1, \ldots, \epsilon_m$ are like realized values of independent, Gaussian random variables, all with mean 0; (3) $u(x_i)$ and $u(y_i)$ are the standard deviations of the random variables that x_i and y_i are realized values of, and they are based on infinitely many degrees of freedom (that is, they are known fully accurately).

In these circumstances, and disregarding an additive constant, the right-hand side of Eq. 1 equals -1 times the logarithm of the likelihood function for the data $\{(x_i, y_i)\}$. In fact, $-S(\beta, \xi_1, \ldots, \xi_m)$ is the sum of the logarithms of $2m$ Gaussian probability densities: m of them pertaining to the $\{y_i\}$, and the other m to the $\{x_i\}$. The

values $\widehat{\beta}, \widehat{\eta}_1, \ldots,$ and $\widehat{\eta}_m$ that minimize $S(\beta, \eta_1, \ldots, \eta_m)$ are the so-called maximum likelihood estimates, which generally possess desirable optimality properties [14].

Annex C of [10] suggests that the computer program B_LEAST, developed by Germany's Federal Institute for Materials Research and Testing (BAM), can be used to find these estimates. XGENLINE [26] and ODRPACK95 [31] serve the same purpose: the latter's Fortran 95 source code is publicly available. We have developed R [24] code for the same purpose, which implements the modified procedure described in "ISO Guide 6143—modified procedure", using Nelder–Mead optimization [18].

ISO Guide 6143—modified procedure

The reason why the criterion $S(\beta, \eta_1, \ldots, \eta_m)$ in Eq. 1 needs to be modified when the uncertainties $\{u(x_i)\}$ and $\{u(y_i)\}$ are based on small numbers of degrees of freedom is this: if Z is a Gaussian random variable with mean ζ and standard deviation σ, and this standard deviation is not known exactly but is estimated by $u(Z)$ on a small number k of degrees of freedom, then the ratio $(Z - \zeta)/S$ has a Student's t probability distribution with k degrees of freedom, rather than a Gaussian distribution [3, 5.3.2].

This implies that, when the $\{u(x_i)\}$ and the $\{u(y_i)\}$ are based on small numbers of degrees of freedom, minimizing $S(\beta, \eta_1, \ldots, \eta_m)$ as defined in Eq. 1 no longer corresponds to maximum likelihood estimation . Even though the Student's t distribution and the Gaussian distribution have similarities (for example, both have bell-shaped probability densities), and for large numbers of degrees of freedom they become essentially indistinguishable, for small numbers of degrees of freedom their differences are marked and consequential.

When $u(y_i)$ and $u(x_i)$ are based on small numbers ν_i and μ_i of degrees of freedom, $(y_i - \eta_i)/u(y_i)$ and $(x_i - G_\beta(\eta_i))/u(x_i)$ are like realized values of independent Student's t_{ν_i} and t_{μ_i} random variables for $i = 1, \ldots, m$. In these circumstances, the modified criterion $S^*(\beta, \eta_1, \ldots, \eta_m)$ defined in Eq. 2, and up to an additive constant, equals -1 times the sum of the logarithms of $2m$ Student's t probability densities.

$$S^*(\beta, \eta_1, \ldots, \eta_m)$$
$$= \sum_{i=1}^{m} \left[(\mu_i + 1) \log \left(1 + \frac{(x_i - G_\beta(\eta_i))^2}{\mu_i u^2(x_i)} \right) \right.$$
$$\left. + (\nu_i + 1) \log \left(1 + \frac{(y_i - \eta_i)^2}{\nu_i u^2(y_i)} \right) \right].$$

(2)

Since $\log(1+z) \approx z$ for small z, when v_i is large the second term inside the square braces is approximately equal to $\big((v_i + 1)/v_i\big)(y_i - \eta_i)^2/u^2(y_i) \approx (y_i - \eta_i))^2/u^2(y_i)$. A similar approximation holds for the first term.

That is, if the numbers of degrees of freedom of the $\{u(x_i)\}$ and the $\{u(y_i)\}$ all are large, then $S^*(\beta, \eta_1, \ldots, \eta_m)$ agrees with $S(\beta, \eta_1, \ldots, \eta_m)$. And when some of the numbers of degrees of freedom are small, the modified criterion S^* will take this into account by dampening the effective weight of the deviation, either $x_i - G_\beta(\eta_i)$ or $y_i - \eta_i$, relative to what S does.

Performance

To illustrate the relative performance of the original and modified optimization criteria discussed above, we will use data from a study of the uncertainty in the preparation of PSMs involving a single laboratory comparing standards of the same nominal composition prepared independently by nine laboratories [17].

These standards were compared based on residual deviations from a straight line fitted to the data, relating gravimetric mole fractions (of carbon monoxide in nitrogen) to ratios of instrumental indications obtained for a control cylinder and for the PSMs. This line was fitted according to the weighted least squares criterion of Eq. 1 [10, A.2].

The weights used in this least squares calculation are the reciprocals of the squared uncertainties of those amount fractions and ratios. As explained in "ISO Guide 6143—original procedure", this procedure is optimal only when all uncertainties involved are based on infinitely many degrees of freedom. In that study, however, each PSM was measured six times, followed by six measurements of the control cylinder, to produce six ratios, hence the type A evaluation of the uncertainty of the average ratio for each PSM has only five degrees of freedom.

To compare the performance of the original ISO 6143 procedure with its modified counterpart defined in Eq. 2, we have carried out a simulation study using the data listed in [17, Table 2], with the added assumption that the uncertainties for the ratios of standard to control cylinder and for the gravimetric mole fractions, all are based on 5 degrees of freedom: this assumption is realistic for the ratios, but is hypothetical for the gravimetry, and it is introduced solely for the purpose of this simulation study.

The simulation study was done using "true" values for the $\{\xi_i\}$ (mole fractions), and for the intercept β_1 and slope β_2 of the relationship between mole fractions and instrumental ratios, set equal to the estimates that the modified procedure yielded when applied to the data aforementioned. The "true" values of the instrumental ratios were then computed as $\eta_i = \beta_1 + \beta_2 \xi_i$, and the "true" uncertainties were chosen as $\sigma_i = u(x_i)$ and $\tau_i = u(y_i)$ for $i = 1, \ldots, m$, and $m = 9$.

The following steps were repeated for $j = 1, \ldots, J$ with $J = 25{,}000$:

– Simulate $u_j(x_i)$ and $u_j(y_i)$ as realized values of $\sigma_i \sqrt{V_{i,j}/\mu_i}$ and of $\tau_i \sqrt{W_{i,j}/v_i}$, for $i = 1, \ldots, m$, where $V_{i,j}$ and $W_{i,j}$ denote independent chi-squared random variables with μ_i and v_i degrees of freedom, all mutually independent.
– Simulate $x_{i,j}$ and $y_{i,j}$ as realized values of Gaussian random variables with means ξ_i and η_i, and standard deviations σ_i and τ_i, all mutually independent, and also independent of the $\{V_{i,j}\}$ and the $\{W_{i,j}\}$.
– Minimize the criterion S defined in Eq. 1 with respect to β, η_1, $\ldots$, η_m, using $(x_{1,j}, y_{1,j}), \ldots, (x_{m,j}, y_{m,j})$ just described as data, to obtain estimates $\widehat{\beta}_{1j}$ and $\widehat{\beta}_{2j}$ of the slope and intercept of the analysis function.
– Do likewise for S^* defined in Eq. 2, to obtain $\widehat{\beta}^*_{1j}$ and $\widehat{\beta}^*_{2j}$.

We summarized the relative performance by the ratio of the mean squared error of the ISO 6143 procedure to the mean squared error of the modified procedure, separately for the intercept and for the slope of the analysis function: $\sum_{j=1}^{J}(\widehat{\beta}_{\ell j} - \beta_{\ell j})^2 / \sum_{j=1}^{J}(\widehat{\beta}^*_{\ell j} - \beta_{\ell j})^2$ for $\ell = 1, 2$. This ratio turned out to be 1.2 for both the intercept and the slope, indicating the superiority of the modified procedure.

One way of interpreting this result is to say that, to achieve performance comparable to the modified procedure, the procedure in ISO 6143 requires 20% more replicates for the measurements of each PSM (both for mole fraction and for instrumental response) than the modified procedure. This relative performance pertains to this particular situation and assumptions (including method chosen for the numerical optimization), and likely will vary from case to case.

Uncertainty assessment

Suppose that y_0 denotes the ratio of instrumental responses for a cylinder containing a gas mixture whose exact composition is unknown but is expected to lie within the range over which the estimated analysis function $G_{\widehat{\beta}}$ is applicable.

Suppose also that this analysis function resulted from steps A through H in ISO 6143, and that $\widehat{\beta}$, together with $\widehat{\eta}_1, \ldots, \widehat{\eta}_m$, minimize the criterion $S_x^*(\beta, \eta_1, \ldots, \eta_m)$ defined in Eq. 2.

The problem we now turn to consists of assessing the uncertainty $u(x_0)$ associated with the value of the measurand $x_0 = G_{\widehat{\beta}}(y_0)$ for that cylinder.

The approach in [10, Step K] follows the treatment proposed by the *Guide to the Expression of Uncertainty in Measurement* (GUM) [11]. This regards y_0 and the components of the vector of coefficients $\widehat{\beta}$ as input quantities, and x_0 as output quantity. And $\widehat{\beta}$ and the $\{\widehat{\eta}_i\}$ in turn are output quantities of a set of other measurement equations, whose input quantities are the ratios of instrumental indications and the chemical compositions of the PSMs that the analysis function is based on.

Implementing the GUM's approach in this case involves two successive first-order Taylor approximations. The second one (in the order in which they must be applied) is for the analysis function, and is used to propagate the uncertainty components captured in $u(x_0)$ and in the covariance matrix of $\widehat{\beta}$. The first one is for the function that maps the ratios of instrumental indications and the compositions of the same PSMs into $\widehat{\beta}, \widehat{\eta}_i, \ldots,$ and $\widehat{\eta}_m$. The evaluation of this (non-linear) function involves the minimization of the criterion S from Eq. 1.

However, this approach of [10, Step K] fails to capture the component of uncertainty that derives from the small number of replicates that typically are used to estimate the uncertainty of the instrumental responses. In addition, it is unclear how well the approximation implicit in the GUM's conventional formula works in this case, given the non-linearities of the participating measurement functions.

These several complications notwithstanding, the GUM's starting point, which is the measurement equation, continues to apply: the output quantity of primary interest, x_0, is a function of the input quantities mentioned above, and also of the uncertainty assessments associated with the chemical compositions and with the ratios of instrumental indications for the PSMs.

An alternative approach to the uncertainty analysis avoids reliance on the approximations of [11, G.3–G.4], and also easily addresses the issue of small numbers of degrees of freedom just alluded to. This alternative is a suitable version of the Monte Carlo method that the GUM Supplement 1 [12] describes, and is analogous to the parametric statistical bootstrap [7].

This alternative approach involves applying stochastic perturbations to the values of the participating quantities repeatedly, and computing the corresponding estimates of β and the resulting "replicates" of x_0. We describe the process in the context of the situation that Eq. 2 applies to, and where it is likely to make a difference: when the uncertainties in play all are based on small numbers of degrees of freedom.

1. Choose a suitably large integer K (typically in the range 1,000–5,000)
2. For $k = 1, \ldots, K$

   a. Simulate $v_{1,k}, \ldots, v_{m,k}$ as realized values of m independent chi-squared random variables with $\mu_1, \ldots, \mu_m$ degrees of freedom, respectively, and compute perturbed versions of the standard uncertainties of $x_1, \ldots, x_m$ as $u_k(x_1) = u(x_1)\sqrt{\mu_1/v_{1,k}}, \ldots, u_k(x_m) = u(x_m)\sqrt{\mu_m/v_{m,k}}$.

   b. Simulate $x_{1,k}, \ldots, x_{m,k}$ as realized values of m independent Gaussian random variables with means $x_1, \ldots, x_m$, and standard deviations $u(x_1), \ldots, u(x_m)$, respectively.

   c. Simulate $w_{1,k}, \ldots, w_{m,k}$ as realized values of m independent chi-squared random variables with $v_1, \ldots, v_m$ degrees of freedom, respectively, and compute perturbed versions of the standard uncertainties of $y_1, \ldots, y_m$ as $u_k(y_1) = u(y_1)\sqrt{v_1/w_{1,k}}, \ldots, u_k(y_m) = u(x_m)\sqrt{v_m/w_{m,k}}$.

   d. Simulate $y_{1,k}, \ldots, y_{m,k}$ as realized values of m independent Gaussian random variables with means $y_1, \ldots, y_m$, and standard deviations $u(y_1), \ldots, u(y_m)$, respectively.

   e. Minimize the criterion S^* of Eq. 2, for example using R function `optim` and Nelder-Mead's method, with respect to $\beta, \eta_1, \ldots, \eta_m$, to obtain estimates β_k^* and $\eta_1^*, \ldots, \eta_m^*$.

   f. Compute $x_{k,0}^* = G_{\beta_k^*}(y_0)$.
   Since $u(y_0)$ typically will be greater than 0 and will also be based on a small number of degrees of freedom, the uncertainty associated with y_0, too, has to be propagated. In general, this involves perturbing y_0 similarly to how we have perturbed other quantities above, which will then produce as many replicates of $x_{k,0}^*$ as there will have been perturbed values of y_0. In the example of "Example: NIST SRM 1685", there are replicates of y_0 to begin with (which are ratios of instrumental readings for many gas cylinders filled from a single batch); hence, in the context of this example, no additional perturbations need to be synthesized by simulation.

3. Produce uncertainty assessments for x_0 and for $\widehat{\beta}$.

   a. A histogram or a kernel density estimate [25] built from $\{x_{1,0}^*, \ldots, x_{K,0}^*\}$ provides useful insight into the probability distribution that the $\{x_{k,0}^*\}$ are a sample from, which fully characterizes the dispersion of values that can reasonably be attributed to x_0 [11, 2.2.3]. (Similarly for $\{\beta_1^*, \ldots, \beta_K^*\}$.)

   b. The standard deviation of $x_{1,0}^*, \ldots, x_{K,0}^*$ is an assessment of $u(x_0)$, and the covariance matrix of $\beta_1^*, \ldots, \beta_K^*$ characterizes the uncertainty of the estimate $\widehat{\beta}$ of β.

   c. If the replicates $\{x_{1,0}^*, \ldots, x_{K,0}^*\}$ are ordered from smallest to largest and $0 < \gamma < 1$ is the desired coverage probability, then the range of the middlemost γK of these replicates defines a coverage interval for x_0 with probability γ.

Certifying a reference gas mixture

The method that NIST's Gas Metrology group uses currently to assign values and uncertainty assessments to certified gas mixtures involves statistical data analysis and uncertainty analysis that are consistent with [10] while extending it as described in "ISO Guide 6143—modified procedure", to take into proper account the typically small numbers of replicates of instrumental responses obtained in the certification process. This process comprises the following three main steps.

1. **Homogeneity study** of the candidate lot of cylinders containing samples of the target gas mixture, including determining whether there are materially and statistically significant sources of heterogeneity that can be mitigated by partitioning the lot into acceptably homogeneous sub-lots (which may then become separate SRMs). Also including a quantitative assessment of the contributions that identifiable experimental factors make to the lot's heterogeneity.

2. **Calibration** that develops an *analysis function* to map ratios of indications from the analytical instrument, obtained for primary standard mixtures and for the lot standard, to amount-of-substance fractions of the measurand. The lot standard typically is a cylinder of larger volume than the cylinders that are the units of the reference material available for purchase, and it is filled from the same batch of the gas mixture that the sample cylinders are filled from. The lot standard is used primarily in the calibration step, but it may also be re-used subsequently, in other studies, in the role of PSM.

3. **Assignment** of a reference value based on the values of the measurand that the analysis function assigns to the individual cylinders in the lot, and assessment of its uncertainty.

The homogeneity study involves classifying the experimental factors that may be causes of heterogeneity into two groups: (1) a first group comprising those factors reflecting intrinsic properties of the gas cylinders and such that, if significant, can meaningfully be used to partition the lot into acceptably homogeneous sub-lots; (2) a second group comprising all the other factors. Based on this classification, a model is fitted to the data that gauges the significance of the factors in the first group, and that estimates the sizes of the uncertainty contributions from the second group.

Calibration is performed consistently with [10], in particular with its clauses 4.f, 5.1–5.2, and A.2, and with the corresponding modifications introduced in "ISO Guide 6143—modified procedure".

The certified value assigned to the SRM is the arithmetic average of the mole fractions that the analysis function assigns to the cylinders in the lot, possibly after removal of apparently outlying values: it is qualified by an assessment of its associated uncertainty. The component of this uncertainty that captures the effects of heterogeneity of the lot and of the construction of the analysis function, obtains by application of the Monte Carlo process described in "Uncertainty assessment". This uncertainty component can then be combined with others (for example, lot instability) that will have been assessed separately, using the techniques described in the GUM [11].

Homogeneity study

The examination of graphical displays that depict the dispersion of the ratios of instrumental indications according to the levels of several factors that have been identified as potential sources of heterogeneity is an indispensable part of the homogeneity study. Of these displays, boxplots [28] of subsets of the data partitioned according to values of the relevant factors, singly or in combination, are particularly useful, as illustrated in "Example: NIST SRM 1685".

The experimental factors (including batch the sample cylinder was filled from, sample cylinder, block of measurements between consecutive measurements of the lot standard, day when measurements were made, port of the manifold used to connect a sample cylinder to the instrument, etc.) that are potential sources of

heterogeneity should first be classified into one of the following two types:

Type I: Factors that relate to some intrinsic property of the cylinders (for example, batch the cylinders were filled from), and such that the corresponding heterogeneity, if statistically and materially significant, can be mitigated by partitioning the lot into acceptably homogeneous sub-lots;

Type II: Factors that relate to other properties (for example, number of the port of the manifold used to connect a sample cylinder to the instrument) whose uncertainty components may either lead to rejection of the lot (if excessively large), or that remain unmitigated and are folded into the combined standard uncertainty of the lot's reference value.

Type I factors are treated as *fixed effects*, and type II factors as *random effects*, in a Gaussian mixed effects model [22]. Such model is fitted to the ratios of the instrumental readings for the cylinders in the lot by the method of restricted maximum likelihood estimation. Model adequacy is assessed by examining QQ-plots [4] of the residuals and of the estimates of the random effects. Indications of potentially serious violations of the model assumptions will require further study and appropriate actions to resolve them.

The QQ-plots just mentioned, and possibly other diagnostics (one of these is illustrated below in relation with the identification of outliers), should also be used to identify individual cylinders that appear to be outliers and that one may like to consider removing from the lot, to reduce the risk of sample cylinders being sold that may appear to be non-compliant with the certification.

If one or more of the type I factors is statistically significant, and the corresponding effect is judged materially significant by the chemist, then the lot is partitioned into sub-lots according to the value (or combination of values if more than one such factor is detected) of the significant factor(s), and the sub-lots treated as separate SRMs.

The lot will be accepted if the chemist determines that none of the uncertainty components associated with type II factors individually is excessively large and if all of them together (when combined in root sum of squares) are acceptably small. The contribution that unmitigated heterogeneity makes to the uncertainty of the reference value is incorporated in the subsequent data reduction step that is concerned with uncertainty assessment.

Calibration

Calibration involves measurements made of m (typically around 6) PSMs, and the mole fractions of measurand in them as determined in previous studies. The corresponding data are m pairs of values (r_1, c_1), ..., (r_m, c_m), where r_i denotes the average of n_i (typically around 10) replicates of ratios of instrumental indications obtained for PSM i and simultaneously also for the lot standard, and c_i denotes the amount fraction of the measurand in the same PSM, for $i = 1, \ldots, m$.

ISO 6143 [10, p. 6] recommends that one should determine the analysis function G directly: this function maps values of ratios $\{r_i\}$ to mole fractions $\{c_i\}$ of the measurand. The construction of the analysis function G should be done taking into account the fact that both those ratios and these mole fractions have uncertainties that are not negligible, that generally they have comparable magnitudes, and at least some of them are based on small numbers of degrees of freedom.

The uncertainty $u(c_i)$ of the mole fraction of measurand in PSM i usually results from some prior study, and may be just read off the corresponding certificate. ISO 6143 [10, Step G, p. 7] suggests that $u(r_i)$ should be the standard deviation of the average of no fewer then ten independent replicates of the ratio: in other words, that this standard uncertainty be set equal to the standard deviation of the replicates $r_{i,1}, \ldots, r_{i,n_i}$ divided by $\sqrt{n_i}$. However, it is conceivable that, in addition to this, it may have to include contributions from other sources of uncertainty, in which case one will combine them in root sum of squares, in accordance with the GUM, and will pool degrees of freedom in some reasonable manner (for example, as described in [11, G4]).

Before the analysis function G can be built, a functional form needs to be chosen for it: in many cases G will be a polynomial of low degree, but it may also be a spline or a local regression function [15]. The choice of specific functional form for the analysis function G is guided by diagnostic plots of the residuals corresponding to candidate models, for example as illustrated in Fig. 4 and further discussed in "Example: NIST RSM 1685". Formal model selection criteria, for example *Akaike's Information Criterion* and the *Bayesian Information Criterion* [9], also provide valuable indications about which, among several possible models, may be best.

Once a particular functional form has been selected for G, the function still depends on a vector parameter β with p components (for example, the three coefficients of a polynomial of the second degree), whose "best" value will be determined based on the

calibration data. For NIST's gas SRMs, the uncertainties $\{u(r_i)\}$, of the ratios of the instrumental indications, typically are based on small numbers of degrees of freedom $\{v_i = n_i - 1\}$, and the modified criterion introduced in "ISO Guide 6143—modified procedure" should be used.

The uncertainties $\{u(c_i)\}$, of the mole fractions of the PSMs, typically are assumed known with certainty. This can be accommodated by setting the corresponding numbers of degrees of freedom $\{\mu_i\}$ to suitably large numbers (say, 100), or a specialized version of the criterion S^* of Eq. 2 should be used, like this:

$$S_r^*(\beta, \rho_1, \ldots, \rho_m) = \sum_{i=1}^{m} \left[\frac{(c_i - G_\beta(\rho_i))^2}{2u^2(c_i)} + \frac{v_i + 1}{2} \right.$$

$$\left. \times \log \left(1 + \frac{(r_i - \rho_i)^2}{v_i u^2(r_i)} \right) \right]. \quad (3)$$

Reference value and uncertainty assessment

Application of the procedure explained in "Uncertainty assessment ", with r_i in the role that y_i plays there, and c_i in the role of x_i, produces K replicates of the analysis function, $G_{\beta_1^*}, \ldots, G_{\beta_K^*}$.

Now, suppose that $b_1, \ldots, b_L$ denote the ratios (relative to the lot standard) of the instrumental indications for the L cylinders in the lot. The corresponding mole fractions of measurand in them, as produced by application of the analysis function, are $f_1 = G_{\hat{\beta}}(b_1), \ldots, f_L = G_{\hat{\beta}}(b_L)$. The reference mole fraction value is $\hat{f}$, the average of $f_1, \ldots, f_L$, computed possibly disregarding any cylinders that may have been set aside owing to their ratios appearing to be outliers.

For each $k = 1, \ldots, K$ and $l = 1, \ldots, L$, $f_{l,k}^* = G_{\beta_k^*}(b_l)$ is the kth bootstrap replicate of the mole fraction of the measurand in cylinder l. The uncertainty component that comprises the contributions from lot heterogeneity and calibration is $u_{\text{HC}}(\hat{f})$, the standard deviation of the KL bootstrap replicates $\{f_{l,k}^*\}$.

Typically, $u_{\text{HC}}(\hat{f})$ will be substantially larger than the standard deviation of $f_1, \ldots, f_L$: while the latter arguably captures unmitigated heterogeneity, the former also captures the calibration uncertainty.

Example: NIST SRM 1685

NIST SRM 1685b [19] is a gas mixture of nitric oxide in nitrogen with nominal mole fraction 250 μmol/mol. The certified value was assigned, and its uncertainty

was assessed, following the procedure described in "Certifying a reference gas mixture". The results listed in this Section are provided only for purposes of illustration of this procedure: the values that should be used in relation with this SRM are listed in the certificate that accompanies each cylinder sold by NIST.

Six PSMs were used for calibration, whose mole fractions of nitric oxide ranged from 201.95 to 274.51 μmol/mol. The ratio between each PSM's chemiluminescent intensity and the lot standard's was measured 15 times: the averages of these ratios, the standard deviations of these averages, and the mole fractions and corresponding uncertainties for the PSMs, comprised the data that the analysis function was derived from.

The SRM was purchased from a commercial specialty gas vendor under contract to NIST in 75 aluminum cylinders with water volume of 6 L each, and the ratio of chemiluminescent intensity of the contents of each cylinder and of the lot standard was measured eight times for 73 cylinders, and seven times for two cylinders, in a designed experiment whose factors were the lot the cylinder belongs to (the SRM was filled as two sub-lots), the cylinder ID, port of the (COGAS) manifold used to connect the cylinder to the instrument, and the subset (break-set) of uninterrupted measurements within each day, and the day within each set of measurements between consecutive measurements of the lot standard.

Figure 1 shows several boxplots that are informative about the lot's homogeneity. The fact that the boxplots in each panel are distributed into two groups at different levels suggests sub-lot heterogeneity. Since the rectangular boxes all are of comparable heights in each panel, and these heights are indicative of the dispersion of the ratios that the boxplots summarize, there is no obvious indication of heterogeneity of the dispersion (standard deviation) of values.

Table 1 summarizes the results of fitting a Gaussian mixed effects model to the ratios for the cylinders, as described in "Homogeneity study". The model was fitted to the data using function lme from the nlme package [21] for the R [24] environment for statistical computing and graphics.

The upper portion of the table shows that sub-lot is a significant source of heterogeneity, with a differential effect amounting to about 0.5% of the median ratio: since sub-lot is a type I factor, these two facts suggests that the lot should be split into two separate SRMs. The lower portion of the table shows the components of uncertainty ascribable to type II factors, considering their nested structure (where "nested" is used in the sense that it commonly has in the context of analysis
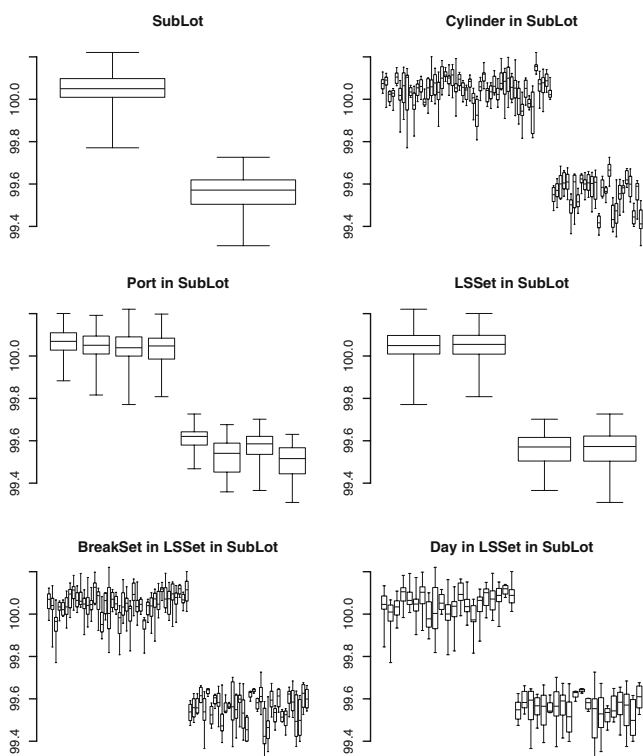
Fig. 1 Homogeneity boxplots. The vertical axis shows values of the ratios of instrumental indications as percentages of the median of all ratios $\{100 r_i / \text{median}(\{r_j\})\}$. Each boxplot summarizes the replicates of the ratios obtained for all combinations of values of the experimental factors for which there are at least five ratios. The *top* and *bottom* of each boxplot bracket the middlemost 50% of the batch of ratios it represents, with the *whiskers* extending to the minimum and the maximum values

Table 1 Summary of the results of fitting a Gaussian, linear mixed effects model to the ratios of instrumental indications

	Estimate	SE	SD	*t* value
Type I factors				
Mean	0.9986	0.000137		7227
SubLotB	−0.004875	0.000145		−34
Type II factors				
Sample ID			4.7395×10^{-4}	
BreakSet/(Day/LSSet)			3.2303×10^{-4}	
Day/LSSet			2.3223×10^{-8}	
LSSet			0	
Port			2.0843×10^{-4}	
Residual			4.8860×10^{-4}	

The factor *BreakSet* is nested within *Day*, and *Day* is nested within *LSSet*, where *nested* is used with the sense that it commonly has in the context of analysis of variance [16, p. 153]. *LSSet* ratios for cylinders that were measured between successive measurements of the lot standard, *BreakSet* ratios measured without significant interruption, *Port* the port of the manifold used to connect a sample cylinder to the instrument

of variance [16, p. 153]). The largest, for the model residuals, amounts to less than 0.05% of the median ratio.

The main conclusions are that the two sub-lots are significantly different, and that all the uncertainty contributions from unmitigated sources of heterogeneity amount to less than 0.05% of the median ratio. The QQ-plots in Fig. 2 offer no compelling reason to question the adequacy of the Gaussian mixed effects model.

Since the lot has been deemed heterogeneous and has been partitioned into two sub-lots (one with 49 cylinders and the other with 26) the reference values were assigned, and the associated uncertainties were assessed, separately for each of them.

Figure 3 shows a histogram of the ratios $\{b_l\}$ for the cylinders in one of the sub-lots, with those ratios that possibly are outliers marked by red dots: these have been diagnosed by the criterion $|b_l - \tilde{b}|/\tilde{s} > 3$, where $\tilde{b}$ and $\tilde{s}$ denote robust indications of location and scale [29, Section 5.5] for the corresponding batch of ratios. There are no such cylinders in the other sub-lot.

The selection of a functional form for the analysis function, among all polynomials of degree no greater than 4, was driven by Fig. 4, which suggests that both quadratic ($p = 3$) and cubic ($p = 4$) polynomials may be suitable models for the calibration data. In our experience, linear and quadratic models typically are
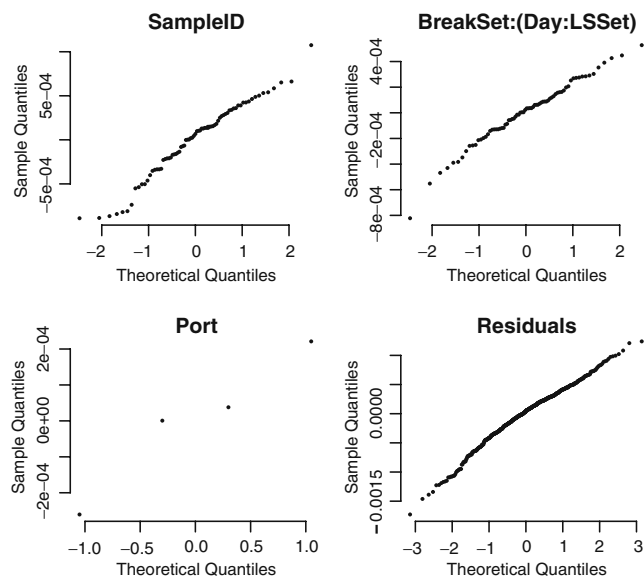
Fig. 2 QQ-plots for type II factors. The QQ-plots for Day:LSSet and LSSet are not shown because the corresponding effects are insignificant (Table 1). The *Theoretical Quantiles*, plotted relative to the horizontal axis, are for Gaussian samples of the same sizes as the corresponding sets of effects. The fair linearity of the plots suggests that the Gaussian assumption is tenable. *Port*, *BreakSet* and *LSSet* are as described in the caption of Table 1
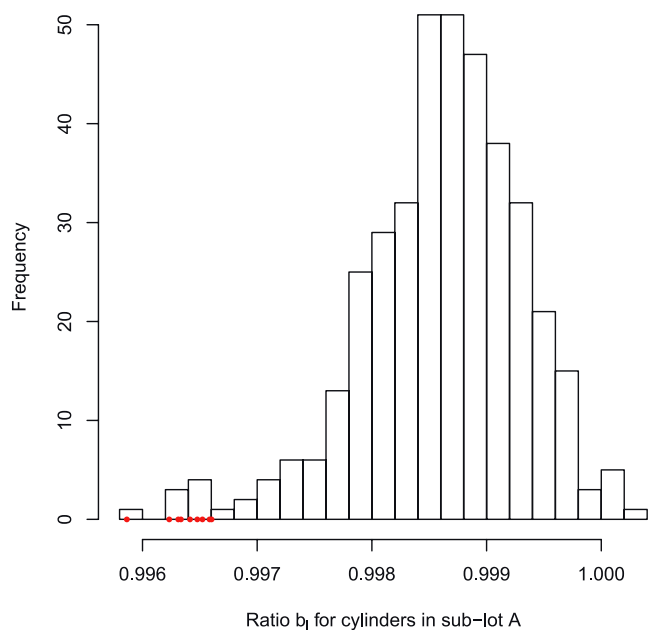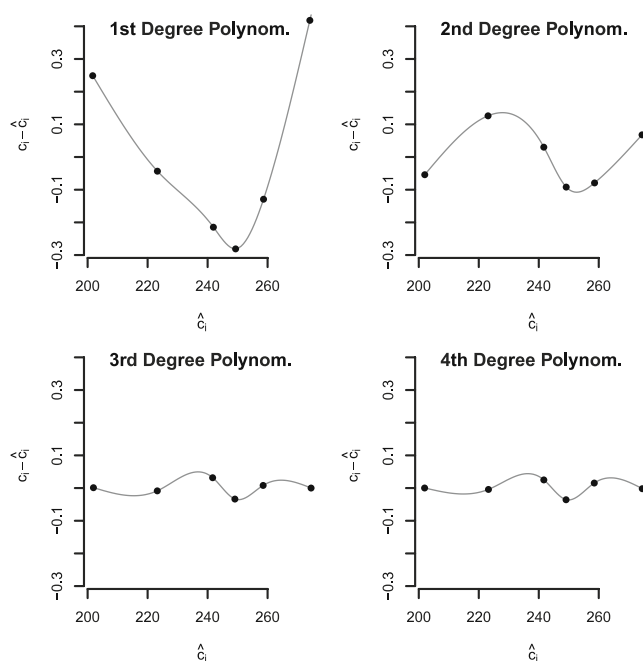
Fig. 3 Sub-lot A: ratios and outliers. Histogram of the ratios $\{b_l\}$ for the cylinders with LSSet 1 or 3, with those ratios that possibly are outliers marked by *red dots*: these have been diagnosed as satisfying the criterion $|b_l - \tilde{b}|/\tilde{s} > 3$, where $\tilde{b}$ and $\tilde{s}$ denote robust indications of location and scale [29, Section 5.5] for the corresponding batch of ratios

Fig. 4 Model selection. Each plot pertains to a candidate polynomial model for the analysis function G_β, and shows how the corresponding residuals $\{c_i - \hat{c}_i\}$ vary with the fitted values $\{\hat{c}_i\}$ where $\hat{c}_i = G_{\hat{\beta}}(r_i$ for $i = 1, \ldots, m$. The plots for polynomials of the first and second degrees display obvious structure that essentially disappears for a cubic G_β. In our experience, linear and quadratic models typically are adequate for analysis functions, but in this case we used a cubic

adequate for analysis functions, but in this case we used a cubic. To fit the model to the data we used the criterion specified in Eq. 3 because each of the $\{u(r_i)\}$ is based on 14 degrees of freedom only.

The average of the mole fractions assigned to the cylinders in the sub-lot where outlying ratios had been detected (but without considering such cylinders) was 244.23 μmol/mol, and their standard deviation was 0.15 μmol/mol (note that this is not the standard uncertainty of this sub-lot's mole fraction, whose evaluation is explained below). For the other sub-lot, the corresponding values were 243.02 and 0.2 μmol/mol.

The uncertainty analysis described in "Reference value and uncertainty assessment" was applied with $K = 9,000$. The standard deviations of the $\{f_{l,k}\}$ in each of the sub-lots were $u_{\mathrm{HC,A}}(f) = 0.295$ μmol/mol and $u_{\mathrm{HC,B}}(f) = 0.321$ μmol/mol: this summarizes the components of uncertainty (of the reference values) attributable to sub-lot heterogeneity and calibration. Combining them in root sum of squares with the uncertainty component attributable to instability of the gaseous mixture—assessed at 0.1%, hence $0.001 \times 244.23 = 0.244$ μmol/mol for sub-lot A, and $0.001 \times 243.02 = 0.243$ μmol/mol for sub-lot B, finally produces the combined standard uncertainty $\sqrt{0.295^2 + 0.244^2} \approx 0.38$ μmol/mol for sub-lot A and $\sqrt{0.321^2 + 0.243^2} \approx 0.40$ μmol/mol for sub-lot B.

On the assumption that the mole fractions in the cylinders of sub-lot A are like a sample from a Gaussian distribution (an assumption that is tenable once the "outliers" detected above are disregarded), one may recast the uncertainty assessment in the form of a *tolerance interval* [20, 7.2.6.3].

For example, the interval ranging from 243.18 to 245.28 μmol/mol is an approximate tolerance interval for the mole fractions in the cylinders of sub-lot A, which includes 99% of such fractions in these cylinders with confidence 95%. (These cylinders' actual mole fractions range from 243.79 to 244.63 μmol/mol.)

The end-points of the tolerance interval are of the form $x \pm k_{1-\alpha,\gamma} u(x)$, where x denotes the average mole fraction of the sub-lot, and $u(x)$ denotes the corresponding combined standard uncertainty. The factor $k_{1-\alpha,\gamma}$, which depends on the desired confidence level $1 - \alpha$ (95% above), and on the proportion γ (99% above) of the population of values of the measurand that the interval should include (denoted k_2 in [20, 7.2.6.3]), was computed using function K.factor in package tolerance [30] for the R [23] environment for statistical computing and graphics.

When the mole fractions in the cylinders of a lot or sub-lot are significantly different from a sample from

a Gaussian distribution, one may compute a tolerance interval using a (non-parametric) statistical procedure that, contrary to the one just illustrated, does not rely on the Gaussian model. One such procedure (as implemented in function `nptol.int` of the aforementioned R package `tolerance`), applied to the values assigned to the cylinders in sub-lot A (rescaled so that their standard deviation equals the corresponding combined standard uncertainty), produces a tolerance interval ranging from 243.12 to 245.27 µmol/mol (hence wider than its parametric counterpart).

Conclusions

While ISO 6143 correctly recognizes that the uncertainties of mole fractions in PSMs used for calibration, and the uncertainties in instrumental responses for gas samples, both ought to be taken into account when building the analysis function that assigns values to these samples (cylinders), it does not take into account the fact that the uncertainties in one or both of the mole fractions and the ratios typically are based on rather small numbers of degrees of freedom.

We overcome this limitation by suggesting a modified criterion, in Eq. 2, that should be minimized when deriving the coefficients of the analysis function, and then we use data published in [17] for a numerical illustration of the superiority of the modified procedure when those numbers of degrees of freedom indeed are small. Under the particular conditions considered in this illustration, the modified procedure was shown to achieve about 20% greater efficiency (in estimation of the coefficients of the analysis function), than the conventional procedure.

We have also shown that the original ISO 6143 weighted least squares criterion is a limiting case of the modified criterion, when the numbers of the degrees of freedom become large that the participating standard uncertainties are based on. For this reason, the modified criterion can be used generally, with a suitably large number of degrees of freedom for those uncertainties that are known with virtual certainty.

The solution that we propose for this version of the errors-in-variables problem, including the Monte Carlo method that we describe to evaluate the uncertainty of the results, however, are of general applicability in analytical chemistry, and their potential value is not restricted to applications in the certification of reference gas mixtures.

The uncertainty analysis that we describe in "Uncertainty assessment" involves no linearizing ap-

proximations, and is consistent with the approach described in the GUM Supplement 1 [12]. Not surprisingly, this is computationally more expensive than the approximate method, yet is not particularly challenging for the computing resources typically available today. More importantly, it involves no compounding of analytical approximations.

In "Example: NIST SRM 1685", we used measurements made to characterize NIST SRM 1685b, to illustrate end-to-end the process described in "Certifying a reference gas mixture" that NIST's Gas Metrology group currently uses to assign a certified value, and to assess the corresponding uncertainty, for a reference gas mixture. This includes a homogeneity study, the construction of the analysis function, and the production of the measurement result, which comprises an assigned value and an assessment of its uncertainty.

References

1. BIPM (2010) The BIPM key comparison database. http://kcdb.bipm.org/
2. Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM (2006) Measurement error in nonlinear models—a modern perspective, 2nd edn. Chapman & Hall/CRC, Boca Raton
3. Casella G, Berger RL (2002) Statistical inference, 2nd edn. Duxbury, Pacific Grove
4. Chambers J, Cleveland W, Kleiner B, Tukey P (1983) Graphical methods for data analysis. Wadsworth, Belmont
5. Comité International des Poids et Mesures (CIPM) (1999) Mutual recognition of national measurement standards and of calibration and measurement certificates issued by national metrology institutes. Bureau International des Poids et Mesures (BIPM), Pavillon de Breteuil, Sèvres, France, Technical supplement revised in October 2003
6. Draper NR, Smith H (1981) Applied regression analysis, 2nd edn. Wiley, New York
7. Efron B, Tibshirani RJ (1993) An introduction to the bootstrap. Chapman & Hall, London
8. Fuller WA (1987) Measurement error models. Wiley, New York
9. Hastie T, Tibshirani R, Friedman J (2001) The elements of statistical learning: data mining, inference, and prediction. Springer, New York
10. 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)

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

12. Joint Committee for Guides in Metrology (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. International Bureau of Weights and Measures (BIPM), Sèvres, France. http://www.bipm.org/en/publications/guides/gum.html. BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP and OIML, JCGM 101:2008

13. Joint Committee for Guides in Metrology (2008) International vocabulary of metrology—basic and general concepts and associated terms (VIM). International Bureau of Weights and Measures (BIPM), Sèvres, France. http://www.bipm.org/en/publications/guides/vim.html. BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP and OIML, JCGM 200:2008

14. Lehmann EL (1999) Elements of large-sample theory. Springer, New York

15. Loader C (1999) Local regression and likelihood. Springer, New York

16. Miller RG (1986) Beyond ANOVA, basics of applied statistics. Wiley, New York

17. Milton MJT, Guenther F, Miller WR, Brown AS (2006) Validation of the gravimetric values and uncertainties of independently prepared primary standard gas mixtures. Metrologia 43:L7–L10

18. Nelder JA, Mead R (1965) A simplex algorithm for function minimization. Comput J 7:308–313

19. NIST (2008) Standard reference material 1685b: nitric oxide in nitrogen. National Institute of Standards and Technology, Gaithersburg, MD. http://www.nist.gov/ts/msd/srm/. Certificate of analysis

20. NIST/SEMATECH (2006) NIST/SEMATECH e-handbook of statistical methods. National Institute of Standards and Technology, U.S. Department of Commerce, Gaithersburg, MD. http://www.itl.nist.gov/div898/handbook/

21. Pinheiro J, Bates D, DebRoy S, Sarkar D, R Core Team (2009) nlme: linear and nonlinear mixed effects models. http:/www.r-project.org/. R package version 3.1-96

22. Pinheiro JC, Bates DM (2000) Mixed-effects models in S and S-Plus. Springer, New York

23. R Development Core Team (2009) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. http://www.R-project.org. ISBN 3-900051-07-0

24. R Development Core Team (2010) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna. http://www.R-project.org. ISBN 3-900051-07-0

25. Silverman BW (1986) Density estimation. Chapman & Hall, London

26. Smith IM (2005) XGENLINE version 8.1 stand-alone executable—software documentation. Tech. rep., National Physical Laboratory, Teddington, Middlesex, UK. Version 1.0

27. Stefanski LA (2000) Measurement error models. J Am Stat Assoc 95(452):1353–1358

28. Tukey JW (1977) Exploratory data analysis. Addison-Wesley, Reading

29. Venables WN, Ripley BD (2002) Modern applied statistics with S, 4th edn. Springer, New York. http://www.stats.ox.ac.uk/pub/MASS4. ISBN 0-387-95457-0

30. Young DS (2009) Tolerance: functions for calculating tolerance intervals. http://CRAN.R-project.org/package=tolerance. R package version 0.1.0

31. Zwolak JW, Boggs PT, Watson LT (2007) Algorithm 869: ODRPACK95: a weighted orthogonal distance regression code with bound constraints. ACM Trans Math Softw 33(4):27