Analytical and Bioanalytical Chemistry (2023) 415:1657–1673
https://doi.org/10.1007/s00216-022-04410-y

RESEARCH PAPER

Value assignment and uncertainty evaluation for anion and single-element reference solutions incorporating historical information

Brian E. Lang[1] · John L. Molloy[1] · Thomas W. Vetter[1] · Shaun P. Kotoski[1] · Antonio Possolo[2]

Received: 27 July 2022 / Revised: 16 September 2022 / Accepted: 27 October 2022 / Published online: 27 February 2023

Abstract

The National Institute of Standards and Technology, which is the national metrology institute of the USA, assigns certified values to the mass fractions of individual elements in single-element solutions, and to the mass fractions of anions in anion solutions, based on gravimetric preparations and instrumental methods of analysis. The instrumental method currently is high-performance inductively coupled plasma optical emission spectroscopy for the single-element solutions, and ion chromatography for the anion solutions. The uncertainty associated with each certified value comprises method-specific components, a component reflecting potential long-term instability that may affect the certified mass fraction during the useful lifetime of the solutions, and a component from between-method differences. Lately, the latter has been evaluated based only on the measurement results for the reference material being certified. The new procedure described in this contribution blends historical information about between-method differences for similar solutions produced previously, with the between-method difference observed when a new material is characterized. This blending procedure is justified because, with only rare exceptions, the same preparation and measurement methods have been used historically: in the course of almost 40 years for the preparation methods, and of 20 years for the instrumental methods. Also, the certified values of mass fraction, and the associated uncertainties, have been very similar, and the chemistry of the solutions also is closely comparable within each series of materials. If the new procedure will be applied to future SRM lots of single-element or anion solutions routinely, then it is expected that it will yield relative expanded uncertainties that are about 20 % smaller than the procedure for uncertainty evaluation currently in use, and that it will do so for the large majority of the solutions. However, more consequential than any reduction in uncertainty, is the improvement in the quality of the uncertainty evaluations that derives from incorporating the rich historical information about between-method differences and about the stability of the solutions over their expected lifetimes. The particular values listed for several existing SRMs are given merely as retrospective illustrations of the application of the new method, not to suggest that the certified values or their associated uncertainties should be revised.

Keywords Bayesian · Between-method differences · Certified reference material · Dark uncertainty · Gravimetry · Stability

✉ Brian E. Lang
brian.lang@nist.gov

John L. Molloy
john.molloy@nist.gov

Thomas W. Vetter
thomas.vetter@nist.gov

Shaun P. Kotoski
skotoski@terpmail.umd.edu

Antonio Possolo
antonio.possolo@nist.gov

[1] Inorganic Chemical Metrology Group, Chemical Sciences Division, Material Measurement Laboratory, National Institute of Standards and Technology, 100 Bureau Drive, Stop 8391, Gaithersburg, MD 20899-8391, USA

[2] Statistical Engineering Division, Information Technology Laboratory, National Institute of Standards and Technology, 100 Bureau Drive, Stop 8980, Gaithersburg, MD 20899-8980, USA

Introduction

The certification of Standard Reference Materials (SRMs) by the National Institute of Standards and Technology (NIST) involves careful and accurate assessment of the measurand and its associated uncertainty for value assignment. NIST's portfolio of 3100 series SRMs (single-element solutions) and 3180 series SRMs (anion solutions) supports the calibration of instrumental and classical methods of analytical chemistry.

These solutions, which are intended to be used as primary calibration standards, provide a clear and relatively short traceability chain linking measurements made in science, medicine, commerce, industry, and agriculture to the SI units of mass and amount of substance. In each of them, the certified value for the mass fraction of a single element or ion is obtained following the requirements described by Beauchamp *et al.* [1].

Of the 74 SRMs within the scope of these two programs, 67 of them are single-element solutions covering a majority of the elements in the periodic table, including but not limited to the transition metals, alkali metals, and alkali earths. Most of the single-element solutions are made from high-purity elements dissolved in acid and diluted to a final nominal mass fraction of 10 mg/g. These solutions are generally stored either in borosilicate glass ampoules or in high-density polyethylene bottles that are sealed in aluminized polyester bags to promote the long-term stability of the analyte in solution.

The 3180 series of SRMs currently comprises seven reference materials with certified mass fractions of non-metal anions in solution (referred to as anion solutions) that can include both mono-atomic and poly-atomic ions: bromide, chloride, fluoride, iodide, nitrate, phosphate, and sulfate. The anion solutions are prepared from high-purity salts dissolved in water to a nominal mass fraction of 1 mg/g.

The current procedure for certifying SRMs mandates that the mass fraction of the measurand in each of the 3100 or 3180 series SRMs be measured using two different methods applied independently of one another [1], although both methods might not be used to assign the certified value.

For the single-element solutions, the methods are gravimetric preparation from carefully assayed source materials, and high-performance inductively coupled plasma optical emission spectroscopy (HP-ICP-OES) [2–5]. For the anion solutions, the methods are gravimetric preparation from carefully assayed source materials and ion chromatography (IC), the latter following a high-performance calibration protocol similar to the HP-ICP-OES method [6][1].

The uncertainties associated with the values measured using each method (IC, HP-ICP-OES, and gravimetric preparation) are evaluated in accordance with the *Guide to the Expression of Uncertainty in Measurement* (GUM) [8]. When prepared from a carefully assayed high-purity material, the final mass fractions of the solutions are determined by combining the results from the gravimetric preparation with the results obtained using an instrumental method, either IC or HP-ICP-OES.

Presently, value assignment and uncertainty evaluation for 3100 and 3180 series SRMs accounts for between-method uncertainty using the *NIST Consensus Builder* [9], in particular the DerSimonian-Laird procedure [10] as described by Koepke *et al.* [11]. This procedure captures and propagates *dark uncertainty* [12], which is the component of uncertainty manifest in the between-method difference in excess of what the uncertainty budgets of the individual measurement methods recognize.

This manner of combining gravimetric and instrumental results suffers from two kinds of shortcomings: those that pertain specifically to the DerSimonian-Laird procedure [13–15], and those that derive from evaluating the between-method uncertainty from a single difference (between the gravimetric and instrumental determinations). The main purpose of this contribution is to improve the latter, while also improving the former.

Weber *et al.* [16] have recommended recently that, for meta-analyses including only two studies, a Bayesian approach "using a weakly informative prior for the heterogeneity may help." Since the value assignment and uncertainty evaluation for our single-element and anion solutions in fact is a form of meta-analysis, and in many cases there is significant heterogeneity (that is, the absolute value of the between-method difference significantly exceeds what should be expected in light of the method-specific uncertainties), such recommendation applies here.

The new procedure described in this contribution exploits the historical information from pairs of measurements (gravimetric and HP-ICP-OES) for 3100 series SRMs produced over the course of the past 14 years, and from pairs of measurements (gravimetric and IC) of the 3180 series SRMs produced over the course of the past 18 years, to characterize between-method differences, which will then be combined with the information about the between-method difference provided by the measurement results for a new SRM.

[1] The IUPAC *Orange Book* [7, 10.3.4] recommends that the term OES be abandoned and that AES (denoting *atomic emission spectroscopy/spectrometry*) be used instead, while noting that both OES and AES have been advocated in IUPAC documents. This contribution concerns historical information from the analytical methods we have been using for the certification of the solutions being discussed. Considering that users of these solutions are most familiar with the usage of OES in many relevant publications, including those just cited, to avoid confusion we will continue to refer to AES methods using the traditional designation "OES."

The improvements we describe in this contribution are these:

(i) To capture and use the knowledge about between-method differences that has been acquired during the long history of development of SRMs in series 3100 and 3180, and to leverage it to enhance the reliability of the evaluation of the component of uncertainty attributable to between-method differences;
(ii) To estimate the measurand (that is, to assign a value to the SRM), and to evaluate the associated uncertainty, taking into account method-specific uncertainties, between-method differences, and potential long-term instability, in an integrated and internally coherent way that also overcomes the aforementioned shortcomings of the DerSimonian-Laird procedure.

We emphasize three important facts:

(a) The proposed new method is applied here to existing SRMs only for purposes of illustration, and to assess the impact that the new method has on the evaluation of the uncertainty resulting from blending the gravimetric and instrumental determinations — in the case of a new SRM, the blended value would become the certified value.
(b) Neither assigned values nor uncertainties stated in currently valid certificates will change in consequence of this exercise. Therefore, the only authoritative value assignments to the existing SRMs in series 3100 and 3180, and associated uncertainty evaluations, continue to be those listed in their respective certificates as published by NIST, until they are replaced by new lots of the same SRM or reach the end of their periods of validity.
(c) This proposed new procedure is applied retrospectively to existing SRM lots to demonstrate how the use of historical information enhances the quality of the evaluation of the uncertainty component attributable to differences between analytical methods. Such enhancement results from the use of historical data about between-method differences, and it does reduce the overall expanded uncertainty for the large majority of the SRM lots of both single-element and anion solutions.

The scope of this work includes only the uncertainty contributions from the gravimetric preparation (“Analytical methods-gravimetric preparation’’), instrumental analysis (“Analytical methods-instrumental measurement’’), long-term instability (“Long-term instability’’), and between-method differences (“Model specification’’ and “Model fitting’’). A reassessment of contributions from other, less influential sources of uncertainty, such as short-term (shipping) instability of the solutions, may be necessary as their relative contribution increases when the relative contribution from method differences decreases.

“Historical overview’’ provides an overview of the history of the development of the SRMs in series 3100 and 3180. “Materials and methods’’ reviews the analytical methods employed for the measurements that contribute to the assignment of value to these SRMs: gravimetric preparations for both series, and different instrumental methods for the single-element solutions and for the anion solutions. In “Long-term instability’’ we describe how we evaluate the component of uncertainty attributed to potential long-term instability of the solutions.

“Example of application’’ motivates the new procedure, introduces the underlying statistical model, and describes an example of application of the new procedure to an existing SRM, comparing the results with their counterparts in the corresponding certificate.

“Model specification’’ specifies the statistical model in detail, and “Model fitting’’ explains how the model is fitted to the gravimetric and instrumental measurement results for a new SRM, and how the resulting output is used for value assignment and uncertainty evaluation.

“Results’’ summarizes the results of applying the new procedure retrospectively to existing SRMs whose uncertainty component attributable to long-term instability has been evaluated separately from the uncertainty evaluations for the gravimetric and instrumental determinations of the measurand, and compares these results with the certified values and associated uncertainties for the same SRMs.

“Discussion’’ discusses the results, highlighting some SRMs that stand out in one way or another in the graphical summaries presented in “Results.’’ “Conclusion’’ summarizes our conclusions. The supplementary material associated with this article lists relevant computer codes and details practical matters about their use.

Historical overview

During the early stages of the 3100 series SRM program, in the 1980s and 1990s, the value was assigned to the measurand based on the gravimetric preparation alone. Another method was used to confirm this value but did not contribute to the certified value that was listed on the certificate of analysis. As the ability to perform highly accurate and precise measurements using instrumental methods increased, a second method started being used to assign the certified value, reducing the risk that some unknown bias might remain undetected.

The combination of results from two different methods of analysis is an instance of consensus building or meta-analysis [11, 17]. However, since two measurement methods typically yield different results, usually there will be a component of uncertainty attributable to between-method differences, which has to be evaluated and propagated to the final results.

This component of uncertainty was recognized soon after a second method started being used to assign a certified mass fraction to the 3100 and 3180 series SRMs. The approach employed to combine the results from two different methods and to incorporate the between-method uncertainty into the expanded uncertainty has evolved over time. Possolo *et al.* [18] describe the procedure that has been used most recently for the single-element solutions.

The earliest attempts at determining the uncertainty component attributable to between-method differences were rudimentary and generally conservative, aiming mostly to capture the impact of any potential bias. The methods that have been in use more recently to quantify the between-method uncertainty are more rigorous but are still amenable to improvement, which will produce more realistic uncertainty evaluations and hopefully will decrease the uncertainty of measurements made using the resulting SRMs as calibrants.

Frequently, the results from the gravimetric preparation and from the instrumental method are in very good agreement, and the corresponding estimate of dark uncertainty turns out to be 0 mg/g. Since the same two measurement methods are used for value assignment throughout the 3100 series and 3180 series SRMs, the typical between-method uncertainty ought not to be overridden by such exceptional, possibly fortuitous mutual agreement resting on a single difference. For this reason, the new procedure entertains the possibility of there being some significant dark uncertainty that remains undetected even in such cases of good mutual agreement between the measurement methods.

Since around 2006, the evaluation of the between-method uncertainty has been based on comparisons between corresponding gravimetric and HP-ICP-OES measurement results for 60 single-element solutions whose elemental mass fractions have been certified individually. This evaluation involves forming the ratios between corresponding gravimetric and HP-ICP-OES measured values, computing the relative standard deviation of these ratios, dividing it by $\sqrt{2}$, and using the result as estimate of the between-method uncertainty, which is then folded into the calculation of the combined standard uncertainty for both the gravimetric preparation and for the HP-ICP-OES determination.

This traditional method of incorporating the between-method uncertainty into the combined uncertainty has two significant shortcomings:

(i) First, splitting the between-method uncertainty evenly between the gravimetric preparation and HP-ICP-OES, and then treating the two pieces as if they were method-specific, suggests that they could somehow counterbalance each other and neutralize the effect, when one knows that this is not how between-method differences impact the results;

(ii) Second, the between-method uncertainty for the anion solutions, which are prepared gravimetrically and measured instrumentally using IC, was assumed to be the same as the method uncertainty for the single-element solutions, whose instrumental determination is made using HP-ICP-OES. This assumption was made because there was not sufficient data to evaluate the between-method uncertainty specific to the differences between the gravimetric and IC determinations for the anion solutions, and this approach was thought to provide the best estimate available.

The new procedure described in "Model specification'' and "Model fitting'' overcomes both shortcomings. However, it pools the historical information available for the single-element solutions and for the anion solutions because there are many lots of the former and few of the latter, and they appear to convey mutually consistent historical information about between-method differences. As additional lots of the anion SRMs are produced, if they will convey historical information different from the information conveyed by the single-element solutions, then the same method described here will continue to apply, except that two different repositories of historical information will be used, one for the single-element solutions, another for the anion solutions.

When the proposed new procedure is used to characterize a new SRM, as illustrated in "Example of application'' for a specific case, it blends historical information about between-method differences with the actually observed difference in the case under consideration, weighing one and the other optimally by application of Bayes's rule [19].

Materials and methods

Analytical methods — gravimetric preparation

The NIST 3100 single-element solution SRMs are generally prepared from high-purity metals assayed for purity to establish an initial link to the SI. These assays are usually performed indirectly, with the elements expected to be present measured except the element being assayed, and the overall purity being assigned by a mass balance approach [20, 21]. While counter-intuitive, this approach leads to a much lower assay uncertainty than measuring the high-purity element directly [22], a critical aspect when the assay uncertainty will dictate the precision of any measurements down the traceability chain.

For some 3100 single-element solution SRMs, it is more practical to use a high-purity salt or oxide for the gravimetric preparation, reserving the high-purity metal for calibration solutions measured with HP-ICP-OES. Unless the high-purity salt is another SRM, only one method (HP-ICP-OES) is used to assign a certified value and uncertainty, although any method biases or additional uncertainty attributable to possible instability over time would also be incorporated into the measurement result. The SRM 3180 series anion solutions are all prepared from high-purity sodium or potassium salts of the anion of interest. The anion salts are assayed for mass fraction of the anion and the associated uncertainty is evaluated.

The 3100 series SRMs are prepared by first carefully determining the mass of an appropriate quantity of the source material, then dissolving the source material in a small amount of solvent, typically concentrated acid. Dissolution is followed by dilution in a large carboy to the target element mass fraction of 10 mg/g analyte with an acidic aqueous solution fraction to promote stability of the solution over the shelf life of the material. For the 3180 series anion solution SRMs, salts are dissolved in a small amount of high-purity water and then diluted to the target anion mass fraction of 1 mg/g with high-purity water.

For both 3100 and 3180 series SRMs, solutions are homogenized by mixing and then sealed into borosilicate glass ampoules or portioned into high-density polyethylene bottles. Through careful measurement of the masses of both the primary materials and the resulting solutions, a gravimetric value can be determined, and its associated uncertainty can be evaluated. The uncertainty of the gravimetric preparation is combined with the uncertainty of the purity of the source material, to evaluate the expanded uncertainty of the gravimetric method [18].

As the stock of each of these SRMs is sold out or reaches the end of its period of validity, a new solution of the same element or anion is prepared and labeled with a new lot number for the same SRM. Thus, multiple lots of most of these SRMs have been prepared over time.

Analytical methods — instrumental measurement

HP-ICP-OES — procedure

Prior to analysis of the new lots of SRM solutions, a set of calibration standards are prepared using a high-purity material of the element of interest. The source materials for these calibration standards are assayed for purity and, if practical, are obtained from a source material containing the element or anion that is independent from the source material used to produce the respective SRM. The calibration solutions are prepared similarly to the preparation of the SRMs, but on a smaller scale and in multiple batches. Batches are validated for mass fraction against each other and against older lots to ensure continuity and stability, and to confirm that the new calibrants are fit for purpose.

Analysis of the SRM solutions by HP-ICP-OES follows an experimental protocol explicitly designed to give the most accurate and precise results possible. The details of the experimental design can be found elsewhere [2–5]. Briefly, the experimental protocol is as follows:

(a) Working solutions are prepared from calibrants and the new lot of SRM using an exact matching protocol [5]. These working solutions have nearly identical amounts of analyte, internal standard, acid, and water, so that any nonlinearity of the response of the HP-ICP-OES instrument is inconsequential. This is possible only because the new, “unknown” lot of SRM in fact is relatively well known, having already been carefully prepared gravimetrically.

(b) The SRM solutions and calibration solutions are analyzed in a randomized order, and the intensities for the analyte and internal standard are extracted. The ratios of the signal of the analyte to the internal standard for each sample are calculated and corrected for instrumental drift using the method described by Salit and Turk [2].

(c) The analyte mass fraction in the new lot of SRM is determined using the known masses of analyte in the calibration solutions.

(d) Uncertainty of the determined HP-ICP-OES mass fraction is derived from the variability of the values measured for the analytical samples, variability associated with the instrument’s sensitivity coefficient, HP-ICP-OES method uncertainty, and the uncertainty of the calibration solutions. In most cases the relative uncertainty achieved is small because the critical measurements in the process are shifted from the HP-ICP-OES to the analytical balance. Therefore most, but not all, uncertainties present other than the between-method uncertainty are related to the ability to prepare solutions carefully using gravimetric methods.

IC — procedure

Analysis of new lots of SRM solutions by IC uses the same general protocols and calculations as HP-ICP-OES, as described above. However, there are significant differences between the two methods and the related solution preparations. The anion SRMs solutions and the dilutions for IC analysis are all prepared in water rather than in acid. For IC measurements, all the ions are separated before detection, thus limiting the matrix effects, which can present challenges for some HP-ICP-OES measurements. There are also differences between IC and HP-ICP-OES detectors and their respective sensitivities. For this reason, the method uncertainty for IC may need to be determined independently of HP-ICP-OES as more measurement results are accumulated.

Long-term instability

As the solutions age, the mass fraction of the element or anion in solution may remain invariant, or it may change, either continuously over time (with or without a well-defined trend), or abruptly. In "Example of application'' and "Model specification'' we model the result of such change as a random drawing from a probability distribution that has mean 0 mg/g and standard deviation set equal to the standard uncertainty that quantifies the potential long-term instability.

Using a probability distribution to characterize the impact of potential long-term instability means that the true value of the mass fraction in solution, at any particular time during the period of validity of the certified value, is unpredictable, and the best one can do is to characterize its typical size, in the form of such standard deviation.

By and large, materials in both series 3100 and 3180 do remain stable during the periods of validity stated in their certificates, which are quite long. For example, the certified value for SRM 3118a (Lot No. 200511) Gadolinium (Gd) Standard Solution, whose certificate was issued in 2021, has an initial period of validity that ends in 2032; and the certified value for SRM 3185 (Lot No. 170309) Nitrate Anion (NO_3^-) Standard Solution, whose certificate was issued in 2017, has an initial period of validity that ends in 2029.

Any processes at work that may induce changes in the mass fraction in solution are very slow, therefore difficult to characterize. Rather than study these processes in detail, NIST's choice is to produce a new lot of the material as soon as there is a mere suspicion that the mass fraction may be changing.

Some of these solutions are delivered in glass ampoules, others in polyethylene bottles. The mechanism whereby ampouled materials may become unstable is not well known, but long-term stability of bottled materials has been studied and the mechanism for their change is understood better. For bottled materials, the main cause of change for the mass fraction in solution is evaporative loss through transpiration. The rate at which solutions lose solvent depends on the acid type, acid concentration, and identity of the dissolved element.

The use of aluminized polyester bags, which are heat sealed around the bottle during production, makes the transpiration across the different bottled materials more uniform and predictable. Furthermore, losses are negligible while the bottle remains in the sealed bag. However, after the bag has been opened the bottled solutions may lose approximately 0.2 % of their mass per year.

The uncertainty component attributable to long-term instability was evaluated using the procedure outlined by Linsinger *et al.* [23], from stability data collected during the period 2004–2008 for the 3100 series SRMs, and during 2002–2020 for the 3180 series SRMs. This procedure uses a linear regression of the mass fraction on age, to describe the change in mass fraction values over time for bottled or ampouled SRMs. The slope of this regression is used to determine the "shelf life" of the solutions. Based on this shelf life the expected change in mass fraction is quantified as a standard deviation, and propagated to the uncertainty associated with the certified value.

Table 1 Measurement results and evaluation of the uncertainty component attributable to potential long-term instability for SRM 3161a (Lot No. 140917) Tin (Sn) Standard Solution, where w denotes the mass fraction, $u(w)$ denotes the corresponding standard uncertainty, $U_{95\,\%}(w)/(mg/g)$ denotes the expanded uncertainty for 95 % coverage, and DF denotes the number of degrees of freedom that $u(w)$ is based on

	w/ (mg/g)	$u(w)$ / (mg/g)	DF
GRAVIMETRY	10.0001	0.00031	279
ICP-OES	10.0217	0.00264	11
INSTABILITY		0.00582	55
DARK UNCERTAINTY		0.01520	
	w/ (mg/g)	$U_{95\,\%}(w)$/ (mg/g)	
CERTIFIED VALUE	10.011	0.025	
NEW APPROACH	10.011	0.022	

Note that this approach is conservative because it translates the expected change in mass fraction over the entire lifetime of the material into an uncertainty component that applies at all times until the expiration date. An alternative would be a variable uncertainty component that would increase over the lifetime of the material based on the time elapsed since the certified value was assigned. While this would result in lower uncertainties during most of the lifetime of the material, it would stand as an obstacle to the use of the material in practice.

Example of application

Table 1 lists the results from the gravimetric and instrumental determinations of the mass fraction of tin in SRM 3161a (Lot No. 140917) Tin (Sn) Standard Solution, the certified value, and associated expanded uncertainty, and their counterparts produced by the new approach that is described in detail in "Model specification'' and "Model fitting.'' All calculations described subsequently were done using all the digits available in the corresponding digital records of analysis, not only the digits listed in this table.

Interestingly, for this SRM at least, dark uncertainty makes the largest contribution to the combined uncertainty associated with the final result, and potential long-term instability makes the second largest. The results summarized in "Results'' indicate that leveraging the historical information about between-method differences will reduce the contribution that between-method differences make to the uncertainty associated with the certified value of many SRMs.

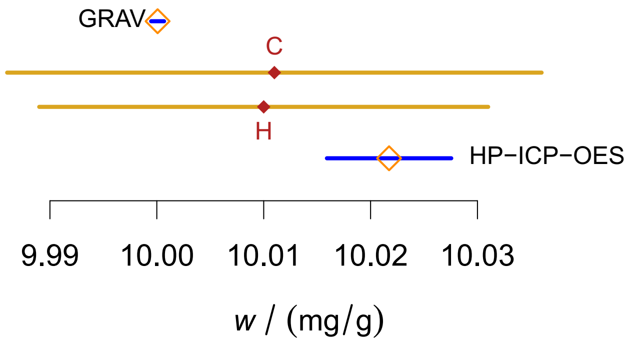

Fig. 1 The diamond outlines indicate the values measured gravimetrically (GRAV) and instrumentally (HP-ICP-OES) for SRM 3161a (Lot 140917) Tin (Sn) Standard Solution, and the solid diamonds indicate the certified value (C) and the corresponding value produced using the new procedure that uses historical information about between-method differences (H). The horizontal line segments represent 95 % coverage intervals centered at the measured values

Figure 1 suggests that the gravimetric and instrumental results are mutually inconsistent, which is confirmed both by Cochran's Q test (p-value smaller than 3.5e-16) [24], and by Welch's t test (p-value smaller than 5.5e-6) [25].

The value assignment for these SRMs is currently being done using a random effects model that is able to detect mutual inconsistency between the gravimetric and instrumental results, and takes the corresponding "excess" dispersion into account [18], employing the DerSimonian-Laird procedure [10]. The corresponding statistical model represents the measured values as

$$w_G = \omega + \lambda_G + \varepsilon_G \quad \text{and} \quad w_I = \omega + \lambda_I + \varepsilon_I, \tag{1}$$

where ω denotes the true value of the measurand, the subscripts G and I refer to the gravimetric and instrumental determinations, λ_G and λ_I denote method effects, and ε_G and ε_I denote measurement errors.

The statistical model involves the additional assumptions that λ_G and λ_I are like two independent drawings from a Gaussian distribution with mean 0 and standard deviation τ, and ε_G and ε_I are like two independent drawings from Gaussian distributions both with mean 0 but possibly different standard deviations, σ_G and σ_I. The assumption is also made that the effects attributable to differences between analytical methods, and the measurement errors, are, for all practical purposes, statistically mutually independent. The standard deviation τ quantifies what is often called *dark uncertainty* [12, 26].

Traditionally, the uncertainty contribution related to potential long-term instability is incorporated after the data reduction that produces an estimate, $\widehat{\omega}$, of the true value of the measurand, and an evaluation of the associated uncertainty. For example, using the DerSimonian-Laird procedure, as implemented in the *NIST Consensus Builder* [9, 11], to fit the model in Eq. (1) to the measurement results listed in the upper part of Table 1, yields $\widehat{\omega} = 10.011$ mg/g, $u(\widehat{\omega}) = 0.011$ mg/g, and $\widehat{\tau} = 0.0152$ mg/g.

The main shortcoming of this approach is that the estimate of dark uncertainty, $\widehat{\tau}$, is based on a single difference between measured values obtained using the two analytical methods used for certification. Even though the *NIST Consensus Builder* includes provisions that take this shortcoming into account, the fact remains that dark uncertainty is evaluated based on a single degree of freedom.

Furthermore, the DerSimonian-Laird procedure is more likely erroneously to conclude that $\tau = 0$ mg/g than the Bayesian procedure, which, differently from classical treatments like DerSimonian-Laird's and variance component estimation based on the analysis of variance [27], expresses the knowledge of dark uncertainty in the form of a probability distribution over its conceivable values. The Bayesian procedure thus characterizes the state of knowledge about dark uncertainty thoroughly and comprehensively.

A more reliable estimate of τ can be produced if one exploits the treasure trove of historical information that is available in the collection of paired gravimetric and instrumental measurements obtained for SRMs developed in the past, and combines it with the information that the difference $w_I - w_G$ provides for a newly developed SRM. "Model specification" and "Model fitting" explain how this is accomplished using a Bayesian estimation procedure.

The following relationship, discussed in detail in "Model specification," captures that historical information and expresses the size of the dark uncertainty relative to the gravimetric value as a function of the absolute value of the difference between the values measured gravimetrically and instrumentally relative to the gravimetric value:

$$\ln\left(\frac{\tau}{w_G}\right) = \alpha + \beta \ln\left(\frac{|\, w_I - w_G \,|}{w_G}\right), \tag{2}$$

where ln denotes the natural logarithm (base e). The intercept α and the slope β are estimated using historical data for all SRMs (both single-element and anion solutions) whose gravimetric and instrumental determinations yield mutually inconsistent results, hence a positive estimate of τ. One can then regard the value of τ produced by Eq. (2) as an *a priori* estimate of τ that can subsequently be updated considering the actual difference $w_I - w_G$ observed for a new SRM.

The current estimates of those intercept and slope are $\widehat{\alpha} = -0.06846$ and $\widehat{\beta} = 1.05309$, as detailed in "Model specification." Therefore, the *a priori* estimate of dark uncertainty for the data in Table 1 is (solving Eq. (2) for τ)

$$\tau_M = (10.0001\ \text{mg/g}) \times \exp\left\{(-0.06846) + 1.05309 \times \ln\left(\frac{|10.0217 - 10.0001|}{10.0001}\right)\right\} = 0.0146\ \text{mg/g}.$$

Equation (2) is similar to the renowned Horwitz equation, for example as reviewed by Horwitz and Albert [28] and by Meija [29, 30]. Applied to the certified value listed in Table 1, and following the suggestion from [28, Page 1100] about converting an estimate of between-laboratory reproducibility into an estimate of within-laboratory variability, the Horwitz equation yields 10.011 mg/g $\times 2(10.011/1000)^{-0.15}/2 = 0.2$ mg/g, which is more than 10 times larger than the corresponding value, 0.0146 mg/g, computed for τ_M above.

This discrepancy questions neither the general usefulness of the Horwitz equation, nor the relevance of Eq. (2) for the specific application to the classes of SRMs under consideration here. Such discrepancy can be attributed primarily to the very close comparability of the solutions under consideration, and to the tightly uniform control over their production, which has been carefully maintained throughout the history of these SRMs, which by now involve a large proportion of the elements in the periodic table, and most anions of greatest practical interest.

NIST has used the Horwitz curve previously and for a similar purpose [31]: to update certificates of old NIST reference materials, mostly metals, whose stocks have not been exhausted yet, that were developed prior to the adoption of the current practices for uncertainty evaluation. However, in that application the Horwitz equation was tuned in light of empirical data to make it more accurate and specifically relevant to a particular class of materials of much narrower scope than had been used originally to develop said curve [32].

The historical estimate of τ from the foregoing Eq. (2) will be updated using the measurement results for the SRM of current interest, by application of the procedure described in "Model fitting,'' which also propagates the updated estimate of τ to the evaluation of the uncertainty associated with the estimate of the mass fraction of the measurand in the solution.

If the gravimetric and instrumental results appear to be mutually consistent, hence $\hat{\tau} = 0$ mg/g, the *a priori* estimate of τ is set equal to $\tau_M/10$. This choice is introduced in item (*b*) of the model specification ("Model specification''), and it is discussed under *Practical Matters* in the supplementary material.

The model for the measured values in Eq. (1) is incomplete because it does not recognize the potential, long-term instability of the solution. To include the effect of instability, we add yet another effect, κ, to the model:

$$w_G = \omega + \kappa + \lambda_G + \varepsilon_G, \quad \text{and} \quad w_I = \omega + \kappa + \lambda_I + \varepsilon_I. \quad (3)$$

The non-observable quantity κ represents the contribution from long-term instability, so that one can interpret $\omega + \kappa$ as the value of the mass fraction in solution at any future time between when the SRM was first certified, and when it reaches the end of its period of validity.

Since, in general, there is no compelling reason to expect κ to be positive or negative, we model it as outcome of a non-observable random variable with mean 0 mg/g and with standard deviation u_S based on ν_S degrees of freedom. The mean being 0 mg/g reflects our inability to state *a priori* whether long-term instability will cause the mass fraction to increase or decrease over time, while the uncertainty component u_S quantifies our estimate of the magnitude of the possible change, regardless of the direction, up or down, of this change. Using a random variable, κ, to model long-term instability does not imply that long-term instability is "chancy" in the same sense that the number of pips shown by a casino die is chancy, when it comes to rest after being rolled: it means simply that we do not know and cannot predict the value of any such change in advance.

The mass fraction of the element or anion in solution may change continuously over time, with or without a well-defined trend, or it may change abruptly: modeling it as a random drawing from a probability distribution simply means that its value at any particular moment in time is unpredictable, and the best one can do is characterize its typical size by specifying its standard deviation, u_S.

We model κ as a non-observable Student's *t* random variable centered at 0 mg/g rescaled to have standard deviation equal to the reported standard uncertainty u_S associated with long-term instability, which is based on the specified number ν_S of degrees of freedom. Since neither the gravimetric nor the instrumental results provide any information about the value of κ, its *a priori* mean value of 0 mg/g is also the mean value that it shall have at the end of the model fitting process described in "Model fitting.''

The probability distribution assigned to κ merely describes one's lack of knowledge about the future: about how much the amount fraction of the solute may change from when it was first certified, up until it reaches the end of its period of validity. Even though the measurement results obtained during certification provide no information about the value of κ, its presence in the model affords a seamless propagation of the uncertainty component u_S.

The last two rows of Table 1 show that, for this particular lot of this SRM 3161a Tin (Sn) Standard Solution, both the certified value and its associated expanded uncertainty are very close to their counterparts obtained using the approach that exploits historical information. Such close agreement will prevail for most other SRMs, but the new procedure that exploits historical information will often yield smaller

Table 2 List of symbols used in the definition of the statistical model and in the description of how it is fitted to the measurement data. Note that κ, λ_G, λ_I, ε_G, and ε_I all can be positive or negative, and have unit mg/g. The magnitudes of τ and τ_{REML} are either zero or positive, and both have unit mg/g

MODEL PARAMETERS	
ω	True value of the measurand
κ	Effect of potential long-term instability, with expected value 0 mg/g
λ_G	Method effect that quantifies the extent to which the gravimetric determination tends to be high or low
λ_I	Method effect that quantifies the extent to which the instrumental determination tends to be high or low
τ	Dark uncertainty (standard deviation of λ_G and λ_I)
τ_M	Prior mean value of τ defined in Eq. (4)
ε_G	Measurement error incurred in the gravimetric preparation
σ_G	Standard deviation of ε_G
ε_I	Measurement error incurred in the instrumental determination
σ_I	Standard deviation of ε_I
MEASUREMENT DATA	
w_G	Measured mass fraction from the gravimetric preparation
$u(w_G)$	Standard uncertainty associated with w_G
ν_G	Number of degrees of freedom supporting $u(w_G)$
w_I	Measured mass fraction from the instrumental method
$u(w_I)$	Standard uncertainty associated with w_I
ν_I	Number of degrees of freedom supporting $u(w_I)$
$u(\kappa)$	Standard uncertainty associated with κ
ν_S	Number of degrees of freedom supporting $u(\kappa)$
MEASUREMENT RESULTS	
τ_{REML}	Restricted Maximum Likelihood estimate of dark uncertainty, τ, based on $(w_G, u(w_G), \nu_G)$ and $(w_I, u(w_I), \nu_I)$ only
w_{HIST}	Mass fraction estimated by the new approach
$U_{95\,\%}(w_{HIST})$	Expanded uncertainty associated with w_{HIST}
w_{CERT}	Certified value of the mass fraction
$U_{95\,\%}(w_{CERT})$	Expanded uncertainty associated with w_{CERT}

uncertainties than those listed in current certificates, as will become clear in "Results.''

Model specification

Table 2 summarizes the symbols used for quantities that appear repeatedly throughout this section and the next, and explains their roles succinctly, supporting the description of the statistical model underlying the procedure already illustrated in "Example of application,'' and the explanation, detailed in "Model fitting,'' of how this model is fitted to the measurement data.

The new approach is based on the relationship introduced in Eq. (2) and depicted in Fig. 2, between $\ln(\tau_{REML}/w_G)$ and $\ln(|w_I - w_G|/w_G)$, for those SRMs whose preliminary estimate of the dark uncertainty, τ_{REML}, is positive.

These preliminary estimates of dark uncertainty were derived from the comparison of gravimetric and instrumental measurement results in the context of the model in Eq. (1) fitted by restricted maximum likelihood (REML) estimation [27, 33], using R function `metagen` defined in package `meta` [34, 35]. This method of estimation has been found to perform well across a wide range of situations encountered in interlaboratory studies and meta-analyses [15, 36].

Only 38 lots of the current single-element solution SRMs have been selected for this exercise because these 38 have independent evaluations of method uncertainty, which the new procedure propagates coherently and simultaneously with the propagation of the other recognized uncertainty components. Of these 38, only 25 (blue circles in Fig. 2) are used to calibrate the relationship in Eq. (4) because only for these is there a positive estimate of dark uncertainty, τ_{REML}. For only for 4 (red dots in Fig. 2) of the 11 lots of anion solution SRMs is $\tau_{REML} > 0$.

This relationship can be summarized quite accurately by the (green) sloping line also depicted in the same figure. The line was fitted to all the points represented by (blue) circles or (red) dots in Fig. 2 using the robust regression procedure implemented in R function `lmrob` of package `robustbase` [37]: its intercept is $\alpha = -0.069$ with standard uncertainty 0.215 (hence α does not differ

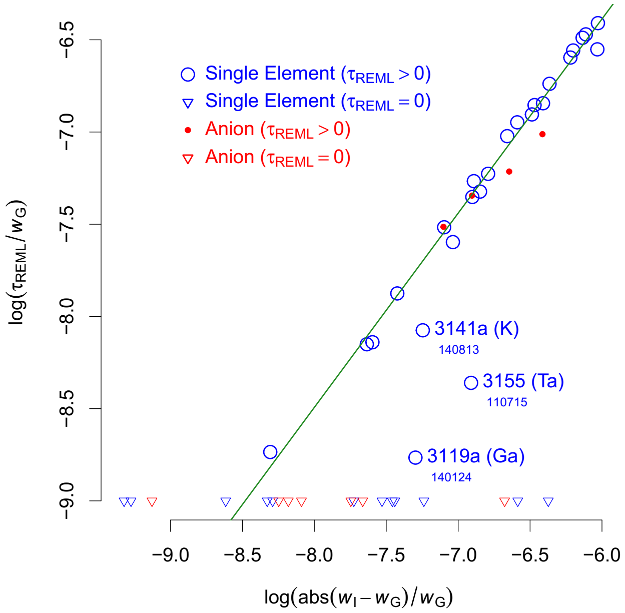

Fig. 2 The large, blue open circles pertain to 25 lots of the single-element solution SRMs for which the REML estimate of dark uncertainty, τ_{REML}, is positive, and similarly for the 4 small red dots, which pertain to individual lots of anion solution SRMs. Taken together, they were used to calibrate the relationship in Eq. (4) that captures the historical information about between-method differences. The small triangles at the bottom indicate the values of $\ln(|\, w_I - w_G \,| / w_G)$ for those SRM lots (13 for single-element solutions in blue and 7 for anion solutions in red) with $\tau_{REML} = 0$. The labels in larger font size designate SRMs, and the labels in smaller font size are lot numbers

significantly from zero), and the slope is $\beta = 1.053$ with standard uncertainty 0.031.

The robust regression line down-weights automatically the three points that are labeled with the corresponding SRM number in Fig. 2 because they are clearly off the clear trend that the regression line aims to capture. The three labeled points all pertain to single-element solutions.

There are only 4 lots of the anion solutions with $\tau_{REML} > 0$ (indicated by red dots in the same figure), and the corresponding points are fairly well aligned with the bulk of the points (blue circles) that correspond to individual lots of the single-element solutions. Thus, for now, the same relationship will be used to express the historical information about between-method differences both for single-element solutions and for anion solutions. However, since different instrumental methods are used for the single-element solutions and for the anion solutions, it is conceivable that, as the collection of anion solutions will keep growing, two different relationships may become warranted to capture the historical information separately for the two series of SRMs.

Given the values of the gravimetric and instrumental determinations made for a new SRM, w_G and w_I, the aforementioned robust regression line can then be used to compute an *a priori* estimate of method uncertainty based on the historical record, as follows:

$$\tau_M = w_G \exp\left\{ \alpha + \beta \ln\left(\frac{|\, w_I - w_G \,|}{w_G} \right) \right\}, \tag{4}$$

where, the same as in Eq. (2), ln denotes the natural logarithm (base *e*).

The value of τ_M produced by Eq. (4) will be used as the mean of the prior distribution for the method uncertainty in the Bayesian data reduction described in "Model specification,'' which updates this initial assessment, τ_M, taking into account the actual gravimetric and instrumental determinations made for the new SRM being certified, and their associated uncertainties.

The model in Eq. (3) becomes a *hierarchical* (or *multilevel*) Bayesian model [38] because we envision its being assembled incrementally: starting by fixing ω, then adding the method-specific effects λ_G and λ_I, next the method-specific measurement errors ε_G and ε_I, and finally the future expression of the degradation of the solution, κ.

Consistently with the Bayesian paradigm, all the unknown quantities in the model — so-called *parameters* (including ω, κ, τ) — are treated as non-observable random variables, whose (prior) probability distributions need to be specified, reflecting knowledge about their values before obtaining the gravimetric and instrumental measurement results for a new SRM.

The measurement results, $(w_G, u(w_G))$ and $(w_I, u(w_I))$, are modeled as observed values of random variables whose probability distributions also have to be specified and depend on the values of those parameters.

We are not interested in the measurement errors ε_G and ε_I themselves, only in their standard deviations, σ_G and σ_I, which impact the uncertainty of the value assigned to the SRM. The standard uncertainties $u(w_G)$ and $u(w_I)$ are estimates of σ_G and σ_I, which are parameters in the model.

Also, rather than estimating the method effects, λ_G and λ_I, we focus on $\mu_G = \omega + \lambda_G$ and $\mu_I = \omega + \lambda_I$, which are the conditional expectations of w_G and w_I given ω, λ_G, and λ_I. The method effects capture possible biases inherent to the gravimetric and instrumental methods, while the goal is to estimate ω unencumbered by any such biases. The modeling assumption made for λ_G or λ_I ensures that both w_G and w_I have ω as their common (unconditional) expected value.

The Bayesian specification of the model in Eq. (3) involves the following choices:

(a) ω has a Gaussian prior distribution with mean equal to the REML estimate of ω based on the gravimetric and instrumental results and computed using the aforementioned R function `metagen`, and standard deviation that is twice the standard deviation of the REML estimate of ω.

(b) τ has a prior distribution that is either gamma with mean τ_M and coefficient of variation equal to 0.3 when $\tau_{REML} > 0$, or exponential with median $\tau_M/10$ when $\tau_{REML} = 0$.

The choice of 0.3 reflects the fact that, for most of the points supporting the trend line in Fig. 2, the actual value of τ_{REML} lies within 60 % of the corresponding value of τ_M. This choice and the choice of prior distribution for τ when $\tau_{REML} = 0$ are discussed in section *Practical Matters* of the supplementary material.

(c) $\mu_G = \omega + \lambda_G$ and $\mu_I = \omega + \lambda_I$ both have Gaussian distributions with mean ω and standard deviation τ.

(d) κ has a prior Student's t distribution with ν_S degrees of freedom, rescaled to have standard deviation u_S.

Since the measurement results provide no information about κ, both its prior and posterior means are 0 mg/g (up to the fluctuations resulting from the vagaries of Monte Carlo Markov Chain sampling described in "Model fitting''), and the sole consequence of including this term in the model is the effective propagation of the uncertainty associated with potential long-term instability of the solution.

(e) σ_G and σ_I have prior gamma distributions with means $u(w_G)$ and $u(w_I)$ and coefficients of variation both equal to 0.1.

The choice of 0.1 expresses the belief that the method-specific uncertainty evaluations are not grossly incorrect in the sense that σ_G and σ_I should be within 20 % of the corresponding evaluations of u_G and u_I. Other choices may also be reasonable, depending on how the mass fraction is measured for a particular solution and on how accurate the corresponding uncertainty budgets are believed to be.

(f) $u(w_G)$ has a gamma distribution with shape $\nu_G/2$ and rate $\nu_G/(2\sigma_G^2)$, and similarly for $u(w_I)$.

This part of the model expresses the established fact that, under the Gaussian model, sample variances have rescaled chi-squared sampling distributions.

(g) w_G has a Gaussian distribution with mean μ_G and standard deviation σ_G, and similarly for w_I.

Model fitting

The immediate inputs for value assignment to a new SRM are these two triplets: $(w_G, u(w_G), \nu_G)$ and $(w_I, u(w_I), \nu_I)$. The Bayesian model specified in Eq. (3) and in "Model specification'' is fitted to these data to obtain an estimate of the true value of the measurand, and an evaluation of the associated uncertainty.

The uncertainty component attributable to differences between analytical methods (gravimetric and instrumental) expresses both historical information about such differences, which is encapsulated in the prior distribution for τ specified in item (*b*) of "Model specification,'' and the difference $w_I - w_G$ actually observed for this new SRM. These two pieces of information are merged using Bayes's rule [19] during the process of fitting the model to the new measurement results.

The model fitting process explores the set of possible values for the parameters to gather information about where their true values are likely to be. The estimates of the quantities of interest are then derived based on the resulting "map" of probabilities for the location of the true value of $\theta = (\omega, \kappa, \tau, \mu_G, \mu_I, \sigma_G, \sigma_I)$.

Rather than applying Bayes's rule directly, which would be impracticable for a model as complex as described in "Model specification,'' we employ an established procedure that produces a large sample drawn from the conditional probability distribution of θ given the measurement results for the new SRM. This procedure is called *Markov Chain Monte Carlo* (MCMC) sampling [39].

The (multivariate) probability distribution of θ that MCMC samples is the so-called *posterior distribution* of θ, which results from updating the prior information specified in *(a)–(g)* of "Model specification'' with the fresh information provided by the measurement results obtained for the certification of the new SRM. MCMC is implemented in R [35] and Stan [40] codes listed in the supplementary material associated with this article.

The output of primary interest issuing from the MCMC procedure is a large sample of values of $\omega + \kappa$ that reflect all the prior information about all the parameters in the model, as well as the fresh measurement results. The effect of potential long-term instability, κ, is present in the model as a virtual carrier of the uncertainty associated with such instability, whose expected value is zero both *a priori* and *a posteriori*, but whose standard deviation is both recognized and suitably propagated.

The estimate of the measurand is the average of the MCMC sample of values of ω, which is the same as the average of the MCMC sample values of $\omega + \kappa$ because κ has (both prior and posterior) mean 0 mg/g. The associated standard uncertainty is the standard deviation of the sample of values of $\omega + \kappa$ (this standard deviation typically is larger than the standard deviation of the sample of values of ω because it includes the contribution from potential long-term instability).

A 95 % credible interval for ω, the true value of the measurand, can be constructed in many different ways: any interval that encompasses 95 % of the sample of values of $\omega + \kappa$ qualifies. Typically, one chooses the interval that is centered at the estimate of the measurand, so that half of the interval's length can be used as the associated expanded uncertainty for 95 % coverage. This "direct" construction of an expanded uncertainty was employed in "Example of application'' for SRM 3161a (Lot No. 140917) Tin (Sn) Standard Solution.

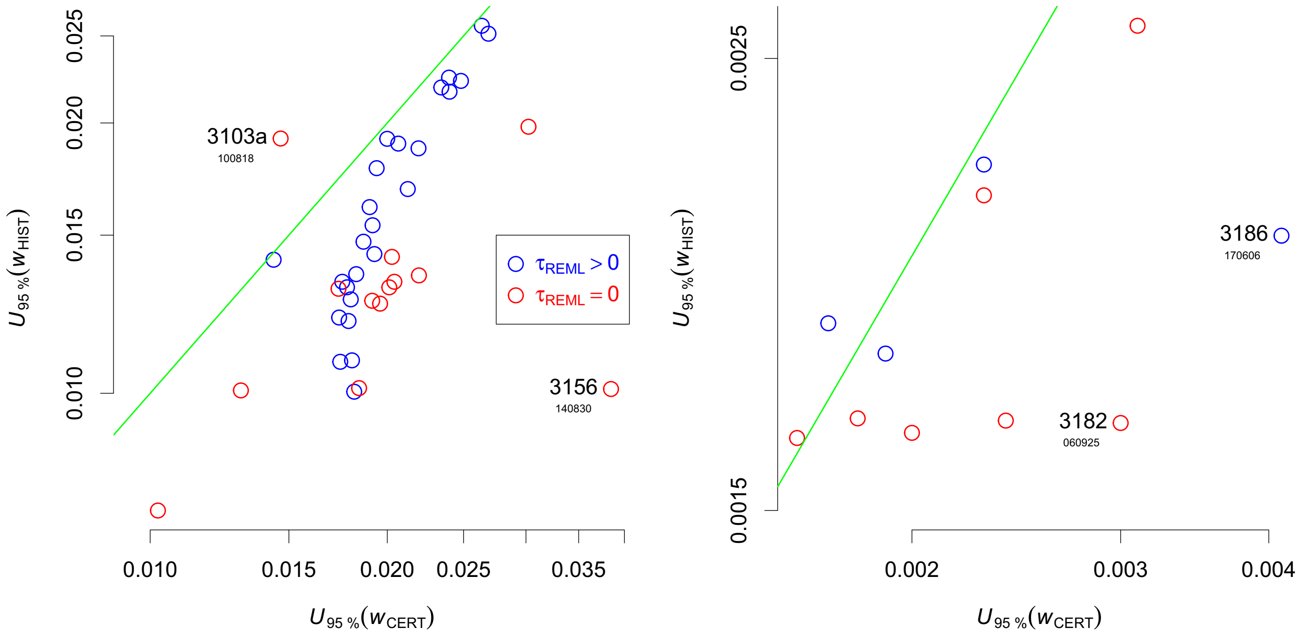

Fig. 3 Expanded uncertainties produced by the new approach, $U_{95\,\%}(w_{\mathrm{HIST}})$, versus the expanded uncertainties associated with the certified values, $U_{95\,\%}(w_{\mathrm{CERT}})$, separately for the single-element solutions (left panel) and for the anion solutions (right panel). The (green) lines have slope 1 and intercept 0. For the large majority of the single-element solutions, the new approach tends to produce smaller expanded uncertainties than are listed in the corresponding certificates. The labels designate specific SRMs, with the lot number given below them in a smaller font size

If the posterior distribution of $\omega + \kappa$ is markedly skewed (that is, asymmetrical), then one may opt for the shortest interval that encompasses 95 % of the sample of values of $\omega + \kappa$. In such case (because the distribution that expresses the uncertainty surrounding ω is asymmetrical), the concept of expanded uncertainty is not particularly meaningful. However, in many such cases, $\widehat{\omega} \pm 2u(\widehat{\omega})$ is an approximate credible interval with 95 % confidence [41].

Results

Figure 3 compares the expanded uncertainties produced by the new procedure with their counterparts that qualify the certified values, separately for the single-element solutions and for the anion solutions. The SRM lots depicted (38 for single-element solutions and 11 for anion solutions) are those for which there are separate evaluations of the uncertainty component attributable to potential long-term instability.

For the majority of the lots of both the single-element and anion solutions, the new approach produces smaller expanded uncertainties, $U_{95\,\%}(w_{\mathrm{HIST}})$, than those that are listed in the corresponding certificates, $U_{95\,\%}(w_{\mathrm{CERT}})$.

For SRM 3103a (As) the new procedure increases the expanded uncertainty appreciably, relative to the expanded uncertainty associated with the certified value: this is one of the SRMs for which the REML procedure suggested that there is no between-method difference. For two of the anion solutions, SRM 3182 (Lot 060925) Chloride Anion (Cl^-) Standard Solution and SRM 3186 (Lot 170606) Phosphate Anion (PO_4^{3-}) Standard Solution, $U_{95\,\%}(w_{\mathrm{CERT}})$ appreciably exceeds $U_{95\,\%}(w_{\mathrm{HIST}})$. The labeled SRMs are discussed in detail in "Noteworthy SRMs."

Figure 4 depicts both the expanded uncertainties (those listed in the certificates of these SRMs and the corresponding ones produced by the new approach) and the differences between the estimates of mass fraction produced by the new approach and the corresponding certified values. The results for the single-element solutions and the anion solutions are displayed in the left panel and right panel, respectively.

Discussion

Table 3 lists the SRMs in series 3100 and 3180 with largest absolute difference between $U_{95\,\%}(w_{\mathrm{CERT}})$ and $U_{95\,\%}(w_{\mathrm{HIST}})$. These differences are depicted graphically in Fig. 4. The table also lists the corresponding values of the ratio

$$R = \frac{U_{95\,\%}(w_{\mathrm{HIST}})/w_{\mathrm{HIST}}}{U_{95\,\%}(w_{\mathrm{CERT}})/w_{\mathrm{CERT}}} \tag{5}$$

between the relative historical and certified expanded uncertainties. For example, $R = 0.6$ for SRM 3101a Aluminum (Al) Standard Solution Lot No. 140903: this means that the proposed procedure for uncertainty evaluation would have

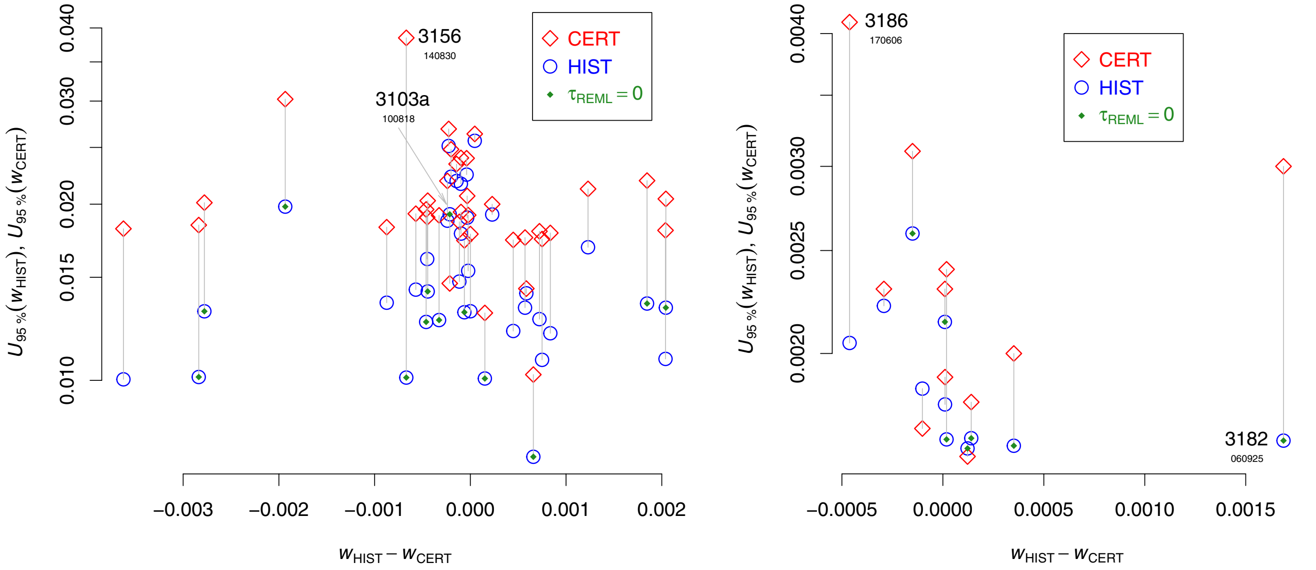

Fig. 4 Both the expanded uncertainties produced by the new approach and the corresponding ones listed in the certificates of these SRMs, $U_{95\,\%}(w_{\mathrm{HIST}})$, $U_{95\,\%}(w_{\mathrm{CERT}})$, versus the differences between the estimates of mass fraction produced by the new approach and the corresponding certified values, $w_{\mathrm{HIST}} - w_{\mathrm{CERT}}$. The single-element solutions are shown in the left panel and the anion solutions are shown in the right panel. Each pair of symbols linked by a vertical (gray) line segment pertain to the same SRM. The blue circles represent results of the new approach (HIST), and the red diamonds represent results listed in the certificates (CERT). The horizontal coordinate is the difference between the value assigned by the new procedure, w_{HIST}, and the certified value, w_{CERT}: all of these differences are negligibly small. The vertical coordinates of the circle and corresponding diamond are the expanded uncertainties for 95 % coverage, $U_{95\,\%}(w_{\mathrm{HIST}})$ and $U_{95\,\%}(w_{\mathrm{CERT}})$. Both expanded uncertainties include the contribution from long-term instability, but evaluated and propagated differently for HIST and for CERT. The tiny (green) diamonds inside some circles indicate those SRMs for which the REML estimate of dark uncertainty is 0 mg/g. The labels in larger font size designate SRMs, and the labels in smaller font size are lot numbers

Table 3 SRM lots with largest absolute difference between the relative expanded uncertainties$U_{95\,\%}(w_{\mathrm{CERT}})/w_{\mathrm{CERT}}andU_{95\,\%}(w_{\mathrm{HIST}})/w_{\mathrm{HIST}}$. R is the ratio of the relative historical expanded uncertainty to the relative certified expanded uncertainty.$R <$ lindicates that the proposed procedure, using historical information, produces a smaller relative expanded uncertainty than the procedure used during certification. Three of the four noteworthy SRMs discussed in "Noteworthy SRMs'' are on this list

SRM	LOT	ELEMENT OR ANION	w_{CERT} / (mg/g)	$U_{95\,\%}(w_{\mathrm{CERT}})$	w_{HIST}	$U_{95\,\%}(w_{\mathrm{HIST}})$	R
3101a	140903	Al	10.011	0.018	10.009	0.010	0.6
3106	180815	Bi	10.002	0.018	9.999	0.010	0.6
3111a	130228	Cs	10.006	0.020	10.008	0.013	0.7
3119a	140124	Ga	9.996	0.018	9.998	0.011	0.6
3129a	100714	Li	9.969	0.030	9.967	0.020	0.7
3134	130418	Mo	9.999	0.022	10.001	0.013	0.6
3156	140830	Te	10.005	0.038	10.004	0.010	0.3
3180	110530	I^-	1.0006	0.0024	1.0006	0.0017	0.7
3182	060925	Cl^-	0.9987	0.0030	1.0004	0.0017	0.6
3186	170606	PO_4^{3-}	1.0005	0.0041	1.0000	0.0020	0.5

produced a relative expanded uncertainty two times smaller than the relative expanded uncertainty corresponding to the measurement result listed in the corresponding certificate.

Figure 3 shows that for some SRM lots the proposed procedure would have reduced the reported uncertainty, while for others it would have increased it. This retrospective study reveals that, overall, the new procedure would have reduced the relative expanded uncertainty for 97 % of the lots of the single-element solution SRMs, and for 82 % of the lots of the anion solution SRMs. Also, the median of the ratios of expanded uncertainties corresponding to the proposed procedure and to the current procedure, across all lots of all SRMs, for the single-element solutions and for the anion solutions, was 0.8.

We carry out these comparisons retrospectively only for purposes of illustration of the new method, not to suggest that the certified values or their associated uncertainties should be revised. Next, "Noteworthy SRMs'' provides details about the SRMs whose representative points are labeled in Figs. 3 and 4.

Noteworthy SRMs

The following examples show that the new procedure being proposed neither replaces knowledge about the chemical nature of these systems, nor does it dispense with the need for exercising professional judgment in all cases about how best to estimate the measurand and to characterize measurement uncertainty. Instead, the new procedure improves the quality of the uncertainty evaluations by drawing on the wealth of accumulated, relevant historical information and by blending it with the information provided by the measurement results specific to each SRM.

Future preparations of these and other SRMs may require the development of customized versions of the general approach described in the previous sections to fold in the unique information about them while still using the accumulated historical information about the between-method differences, to improve the quality of the uncertainty evaluations for them.

SRM 3103a (Lot No. 100818) Arsenic

SRM 3103a (Lot No. 100818) Arsenic (As) Standard Solution is markedly different from all the other single-element solutions in the collection under consideration. Arsenic is rather unique in that its valence state in solution can make quite a difference during measurement using ICP-OES.

Yu *et al.* [42] demonstrated that solutions used during development of this SRM contained a mixture of trivalent and pentavalent arsenic, and that ICP-OES was approximately 8 % more sensitive to pentavalent than to trivalent arsenic. Narukawa *et al.* [43] confirmed the qualitative difference, even if they estimated the differential relative sensitivity to be only 4 %, and Narukawa *et al.* [44] studied this phenomenon further.

For this reason, when preparing arsenic solutions for certification, an oxidizing agent is added to ensure that all the arsenic ions present are in their higher valence state, which makes the measurement of these solutions markedly different from how the solutions of the other elements are measured.

SRM 3156 (Lot No. 140830) Tellurium

SRM 3156 (Lot No. 140830) Tellurium (Te) Standard Solution stands out because $U_{95\,\%}(w_{\mathrm{CERT}})$ is almost four times larger than $U_{95\,\%}(w_{\mathrm{HIST}})$. For this element, too, $\tau_{\mathrm{REML}} = 0$ mg/g. However, during certification, the measured mass fraction of tellurium in solution varied appreciably from day to day while it was being measured repeatedly by HP-ICP-OES.

This day-to-day variability was recognized and expressed by first obtaining a consensus value and associated uncertainty based only on the replicated instrumental determinations. This consensus value was then combined with the gravimetric result to produce a final consensus value for assignment to the material.

The end result was that $u(w_{\mathrm{I}})$ was about 50 times larger than $u(w_{\mathrm{G}})$ for SRM 3156. Fewer than 5 % of the SRMs in this series have a ratio $u(w_{\mathrm{I}})/u(w_{\mathrm{G}})$ this large or larger. However, since our model does not exploit the historical relationship between $u(w_{\mathrm{I}})$ and $u(w_{\mathrm{G}})$, the peculiarity just mentioned does not explain the large difference between $U_{95\,\%}(w_{\mathrm{HIST}})$ and $U_{95\,\%}(w_{\mathrm{CERT}})$.

This difference may be attributable to the fact that our model uses the reported standard uncertainties differently from how they were used during certification, to weight the gravimetric and instrumental determinations as they are combined for value assignment. Since the difference between $u(w_{\mathrm{I}})$ and $u(w_{\mathrm{G}})$ is so pronounced in this case, the impact of the different weighting schemes upon the evaluation of the uncertainty associated with the assigned value may be large enough to explain why $U_{95\,\%}(w_{\mathrm{CERT}})$ should be almost four times as large as $U_{95\,\%}(w_{\mathrm{HIST}})$.

SRM 3182 (Lot No. 060925) Chloride

The certificate of SRM 3182 (Lot No. 060925) Chloride Anion (Cl^-) Standard Solution lists a value for the expanded uncertainty that is 1.8 times larger than its counterpart, $U_{95\,\%}(w_{\mathrm{HIST}})$, produced by the new approach. This is noteworthy but not surprising because both the certified value and the associated $U_{95\,\%}(w_{\mathrm{CERT}})$ are based on measurement results different from those that we have used in the calculation of both w_{HIST} and $U_{95\,\%}(w_{\mathrm{HIST}})$.

This material was originally certified in 2007, its certification period was extended in 2011, and then expired in 2017, when it was replaced by a new lot. The original certification used X-ray fluorescence (XRF) as instrumental method of analysis, rather than IC. The instrumental result we have used to obtain w_{HIST} and $U_{95\,\%}(w_{\mathrm{HIST}})$ is not that XRF result: it is the first measurement of this lot that was made using IC, during a stability study supporting the 2011 extension of the certification period.

The original gravimetric and XRF results, which were markedly mutually inconsistent, were combined in 2006 using a statistical procedure proposed by Levenson *et al.* [45], which came to be known by its acronym, BOB ("Type B On Bias"). NIST deprecated this procedure in 2009, and replaced it with procedures based on conventional random effects models, either classical [46, 47] or Bayesian [48].

SRM 3186 (Lot No. 170606) Phosphate

The certificate of SRM 3186 (Lot No. 170606) Phosphate Anion (PO_4^{3-}) Standard Solution lists an expanded uncertainty that is 2 times larger than its counterpart from the new procedure, which makes this material stand out in the right panel of Figs. 3 and 4.

The certified value and the associated uncertainty for this material include a second IC measurement using an independent phosphate source material to account for a potential bias in the phosphate assay of the source material. The inclusion of this second IC measurement makes for a relative uncertainty larger than for the other anion solution SRMs.

Since the new procedure would have required a modification to be able to take the additional measurement into account, it neglected the second IC measurement, which impacted the estimate of the measurand substantially (apparent in the right panel of Fig. 4) and yielded a smaller expanded uncertainty than is associated with the certified values (Figs. 3 and 4).

Conclusions

A new procedure was presented for evaluating the uncertainty component attributable to the between-method differences between the gravimetric preparation and the instrumental method of analysis, HP-ICP-OES for the single-element solutions and IC for the anion solutions. This procedure blends historical information about such differences with the difference observed when measuring a new solution.

The procedure was illustrated using the methods of gravimetric preparation and instrumental measurement of series 3100 and series 3180 SRMs. Since the new procedure exploits the rich history of production of these materials, the resulting uncertainty evaluations are believed to be more reliable than those obtained neglecting such history.

The retrospective evaluation of the new procedure suggests that it is likely to yield appreciable reductions in expanded uncertainty for future lots of both the single-element solution SRMs and of the anion solution SRMs. In fact, the new procedure reduced the relative expanded uncertainty for 97 % of the single-element solution lots and for 82 % of the anion solution lots, by comparison with the original procedure.

As already pointed out toward the end of "Discussion'' the median of the ratios of expanded uncertainties corresponding to the proposed procedure and to the current procedure, across all lots of all SRMs, for the single-element solutions and for the anion solutions, was 0.8. This means that the new procedure tends to yield expanded uncertainties that are about 20 % smaller than the current procedure. And what enables it to do so is the treasure trove of information accumulated in the long history of development of these reference materials at NIST.

The new procedure that we have described and illustrated in applications, improves the quality of the uncertainty evaluations and enhances their value as links in the traceability chain for measurements made by users of two flagship series of SRMs, because it draws on the wealth of relevant historical information accumulated in the course of decades of development and measurement of these materials [49], and blends it with the information provided by the measurement results specific to each SRM.

As the collections of SRM lots in these two series continue to grow, the value added by the blending procedure will continue to increase accordingly, provided the measurement methods used remain stable and are well-characterized.

We have illustrated the application of the new method retrospectively, using data for each SRM lot as if it were a new lot, building on the historical information provided by all the other lots of the same or other SRMs. However, we did it only for purposes of illustration, not to suggest that the certified values or their associated uncertainties should be revised.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00216-022-04410-y.

Acknowledgements The authors are grateful to their NIST colleagues who kindly provided detailed reviews of a draft, and offered useful comments and suggestions for improvement: Michael Epstein, Adam Pintar, and Michael Winchester. The authors are also highly appreciative of the suggestions for improvement offered by two anonymous reviewers, in particular for pointing out the similarity between Eq. (2) and the Horwitz equation [28].

Declarations

The authors declare no competing interests. The research reported herein did not involve human or animal subjects, or any biological materials, as objects of research. This research was conducted as part of the authors' duties as employees of the National Institute of Standards and Technology, an agency of the federal government of the United States of America, under the U.S. Department of Commerce.

References

1. Beauchamp CR, Camara JE, Carney J, Choquette SJ, Cole KD, DeRose PC, Duewer DL, Epstein MS, Kline MC, Lippa KA,

Lucon E, Phinney KW, Possolo A, Sharpless KE, Sieber JR, Toman B, Winchester MR, Windover D. Metrological tools for the reference materials and reference instruments of the NIST Materials Measurement Laboratory. In: NIST Special Publication 260-136 (2020 Edition). National Institute of Standards and Technology. 2020. https://doi.org/10.6028/NIST.SP.260-136-2020

2. Salit ML, Turk GC. A drift correction procedure. Analytical Chemistry. 1998;70(15):3184–90. https://doi.org/10.1021/ac980095b.
3. Salit ML, Turk GC, Lindstrom AP, Butler TA, Beck CM, Norman B. Single-element solution comparisons with a high-performance inductively coupled plasma optical emission spectrometric method. Analytical Chemistry. 2001;73(20):4821–9. https://doi.org/10.1021/ac0155097.
4. Salit ML, Turk GC. Traceability of single-element calibration solutions. Analytical Chemistry. 2005;77(7):136–41. https://doi.org/10.1021/ac053354n.
5. Winchester MR, Butler TA, Turk GC. Improving the high-performance inductively coupled plasma optical emission spectrometry methodology through exact matching. Analytical Chemistry. 2010;82(18):7675–83. https://doi.org/10.1021/ac101471a.
6. Brennan RG, Butler TA, Winchester MR. Achieving 0.2 % relative expanded uncertainty in ion chromatography analysis using a high-performance methodology. Analytical Chemistry. 2011;83(10):3801–7. https://pubs.acs.org/doi/10.1021/ac200290y
7. Inczedy J, Lengyel T, Ure AM, of Pure IU, Chemistry A. IUPAC Compendium on Analytical Nomenclature, Definitive Rules 1997, 3rd edn. Blackwell Science. 1998. IUPAC Orange Book. https://media.iupac.org/publications/analytical_compendium/
8. Joint Committee for Guides in Metrology (JCGM): Evaluation of Measurement Data — Guide to the Expression of Uncertainty in Measurement. International Bureau of Weights and Measures (BIPM), Sèvres, France. 2008. BIPM, IEC, IFCC, ILAC, ISO, IUPAC, IUPAP and OIML, JCGM 100:2008, GUM 1995 with minor corrections. https://www.bipm.org/en/publications/guides/gum.html
9. Koepke A, Lafarge T, Toman B, Possolo A. NIST Consensus Builder — User's Manual. National Institute of Standards and Technology. 2017. National Institute of Standards and Technology. https://consensus.nist.gov
10. DerSimonian R, Laird N. Meta-analysis in clinical trials. Controlled Clinical Trials. 1986;7(3):177–88. https://doi.org/10.1016/0197-2456(86)90046-2.
11. Koepke A, Lafarge T, Possolo A, Toman B. Consensus building for interlaboratory studies, key comparisons, and meta-analysis. Metrologia. 2017;54(3):34–62. https://doi.org/10.1088/1681-7575/aa6c0e.
12. Thompson M, Ellison SLR. Dark uncertainty. Accreditation and Quality Assurance. 2011;16:483–7. https://doi.org/10.1007/s00769-011-0803-0.
13. Hoaglin DC. Misunderstandings about *Q* and 'Cochran's *Q* test' in meta-analysis. Statistics in Medicine. 2016;35:485–95. https://doi.org/10.1002/sim.6632.
14. Veroniki AA, Jackson D, Viechtbauer W, Bender R, Bowden J, Knapp G, Kuss O, Higgins JPT, Langan D, Salanti G. Methods to estimate the between-study variance and its uncertainty in meta-analysis. Research Synthesis Methods. 2016;7:55–79. https://doi.org/10.1002/jrsm.1164.
15. Langan D, Higgins JPT, Jackson D, Bowden J, Veroniki AA, Kontopantelis E, Viechtbauer W, Simmonds M. A comparison of heterogeneity variance estimators in simulated random-effects meta-analyses. Research Synthesis Methods. 2019;10(1):83–98. https://doi.org/10.1002/jrsm.1316.
16. Weber F, Knapp G, Glass A, Kundt G, Ickstadt K. Interval estimation of the overall treatment effect in random-effects meta-analyses: Recommendations from a simulation study comparing frequentist, Bayesian, and bootstrap methods. Research Synthesis Methods. 2021;12(3):291–315. https://doi.org/10.1002/jrsm.1471.
17. Strawderman WE, Rukhin AL. Simultaneous estimation and reduction of nonconformity in interlaboratory studies. Journal of the Royal Statistical Society Series B (Statistical Methodology). 2010;72(2):219–34. https://doi.org/10.2307/40541584.
18. Possolo A, Bodnar O, Butler TA, Molloy JL, Winchester MR. Value assignment and uncertainty evaluation in single-element reference solutions. Metrologia. 2018;55(3):404–13. https://doi.org/10.1088/1681-7575/aabd57.
19. Possolo A, Toman B. Tutorial for metrologists on the probabilistic and statistical apparatus underlying the gum and related documents. National Institute of Standards and Technology, 2011. https://doi.org/10.13140/RG.2.1.2256.8482. https://www.itl.nist.gov/div898/possolo/TutorialWEBServer/TutorialMetrologists2011Nov09.xht
20. Kipphardt H, Matschat R, Rienitz O, Schiel D, Gernand W, Oeter D. Traceability system for elemental analysis. Accreditation and Quality Assurance. 2006;10(11):633–9. https://doi.org/10.1007/s00769-005-0084-6.
21. Westwood S, Choteau T, Daireaux A, Josephs RD, Wielgosz RI. Mass balance method for the SI value assignment of the purity of organic compounds. Analytical Chemistry. 2013;85(6):3118–26. https://doi.org/10.1021/ac303329k.
22. Vogl J, Kipphardt H, Richter S, Bremser W, Torres MRA, Manzano JVL, Buzoianu M, Hill S, Petrov P, Goenaga-Infante H, Sargent M, Fisicaro P, Labarraque G, Zhou T, Turk GC, Winchester M, Miura T, Methven B, Sturgeon R, Jährling R, Rienitz O, Mariassy M, Hankova Z, Sobina E, Krylov AI, Kustikov YA, Smirnov VV. Establishing comparability and compatibility in the purity assessment of high purity zinc as demonstrated by the CCQM-p149 intercomparison. Metrologia. 2018;55(2):211–21. https://doi.org/10.1088/1681-7575/aaa677.
23. Linsinger TPJ, Pauwels J, Lamberty A, Schimmel HG, van der Veen AMH, Siekmann L. Estimating the uncertainty of stability for matrix CRMs. Fresenius' Journal of Analytical Chemistry. 2001;370:183–8. https://doi.org/10.1007/s0021601007.
24. Cochran WG. The combination of estimates from different experiments. Biometrics. 1954;10(1):101–29. https://doi.org/10.2307/3001666.
25. Welch BL. The generalization of 'Student's' problem when several different population variances are involved. Biometrika. 1947;34:28–35. https://doi.org/10.1093/biomet/34.1-2.28.
26. Analytical Methods Committee. Dark uncertainty. Analytical Methods. 2012;4:2609–12. https://doi.org/10.1039/C2AY90034C. AMC Technical Briefs No. 53.
27. Searle SR, Casella G, McCulloch CE. Variance components. John Wiley & Sons, 2006
28. Horwitz W, Albert R. The Horwitz Ratio (HorRat): A useful index of method performance with respect to precision. Journal of AOAC International. 2006;89(4):1095–109. https://doi.org/10.1093/jaoac/89.4.1095.
29. Meija J. A chemical uncertainty principle challenge. Analytical and Bioanalytical Chemistry. 2007;387:1583–4. https://doi.org/10.1007/s00216-006-1059-0.
30. Meija J. Solution to the chemical uncertainty principle challenge. Analytical and Bioanalytical Chemistry. 2007;388:995–6. https://doi.org/10.1007/s00216-007-1312-1.
31. Sieber JR, Epstein MS, Possolo AM. A Retuned Horwitz procedure for upgrading certificates of older standard reference materials. NIST Special Publication 260-198. National Institute of Standards and Technology, Gaithersburg, MD, 2019. https://doi.org/10.6028/NIST.SP.260-198
32. Horwitz W. Evaluation of analytical methods used for regulation of foods and drugs. Analytical Chemistry. 1982;54(1):67–76. https://doi.org/10.1021/ac00238a765.

33. McCulloch CE, Searle SR, Neuhaus JM. Generalized, Linear, and Mixed Models, 2nd edn. John Wiley & Sons, 2008
34. Balduzzi S, Rücker G, Schwarzer G. How to perform a meta-analysis with R: a practical tutorial. Evidence-Based Mental Health. 2019;22:153–60. https://doi.org/10.1136/ebmental-2019-300117.
35. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, 2021. R Foundation for Statistical Computing. https://www.R-project.org/
36. Langan D, Higgins JPT, Simmonds M. Comparative performance of heterogeneity variance estimators in meta-analysis: a review of simulation studies. Research Synthesis Methods. 2017;8(2):181–98. https://doi.org/10.1002/jrsm.1198.
37. Maechler M, Rousseeuw P, Croux C, Todorov V, Ruckstuhl A, Salibian-Barrera M, Verbeke T, Koller M, Conceição ELT, di Palma MA. Robustbase: Basic Robust Statistics. 2021. R package version 0.93-9. http://CRAN.R-project.org/package=robustbase
38. Gelman A, Hill J. Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge University Press. 2007 https://doi.org/10.1017/CBO9780511790942.
39. Gelman A, Carlin JB, Stern HS, Rubin DB. Bayesian data analysis, 2nd edn. Chapman & Hall / CRC, 2003
40. Carpenter B, Gelman A, Hoffman M, Lee D, Goodrich B, Betancourt M, Brubaker M, Guo J, Li P, Riddell A. Stan: A probabilistic programming language. Journal of Statistical Software. 2017;76(1):1–32. https://doi.org/10.18637/jss.v076.i01.
41. Freedman D, Pisani R, Purves R. Statistics, 4th edn. W. W. Norton & Company, 2007
42. Yu LL, Butler TA, Turk GC. Effect of valence state on ICP-OES value assignment of SRM 3103a arsenic spectrometric solution. Analytical Chemistry. 2006;78:1651–6. https://doi.org/10.1021/ac051732i.
43. Narukawa T, Kuroiwa T, Chiba K. Mechanism of sensitivity difference between trivalent inorganic As species [As(III)] and pentavalent species [As(V)] with inductively coupled plasma spectrometry. Talanta. 2007;73:157–65. https://doi.org/10.1016/j.talanta.2007.03.021.
44. Narukawa T, Chiba K, Kuroiwa T, Inagaki K. Differences in sensitivity between As(III) and As(V) measured by inductively coupled plasma spectrometry and the factors affecting the incoherent molecular formation (IMF) effect in the plasma. Journal of Analytical Atomic Spectrometry. 2010;25:1682–7. https://doi.org/10.1039/C0JA00011F.
45. Levenson MS, Banks DL, Eberhardt KR, Gill LM, Guthrie WF, Liu HK, Vangel MG, Yen JH, Zhang NF. An approach to combining results from multiple methods motivated by the ISO GUM. Journal of Research of the National Institute of Standards and Technology. 2000;105(4):571–9. https://doi.org/10.6028/jres.105.047.
46. Vangel MG, Rukhin AL. Maximum likelihood analysis for heteroscedastic one-way random effects ANOVA in interlaboratory studies. Biometrics. 1999;55:129–36. https://doi.org/10.1111/j.0006-341X.1999.00129.x.
47. Rukhin A, Biggerstaff B, Vangel M. Restricted maximum likelihood estimation of a common mean and the Mandel-Paule algorithm. Journal of Statistical Planning and Inference. 2000;83:319–30. https://doi.org/10.1016/S0378-3758(99)00098-1.
48. Toman B. Bayesian approaches to calculating a reference value in key comparison experiments. Technometrics. 2007;49(1):81–7. https://doi.org/10.1198/004017006000000273.
49. Moody JR, Greenberg RR, Pratt KW, Rains TC. Recommended inorganic chemicals for calibration. Analytical Chemistry. 1988;60(21):1203–18. https://doi.org/10.1021/ac00172a001.