Hindawi
Science and Technology of Nuclear Installations
Volume 2017, Article ID 2679243, 16 pages
https://doi.org/10.1155/2017/2679243

Research Article

Comprehensive Uncertainty Quantification in Nuclear Safeguards

E. Bonner,[1] T. Burr,[1] T. Krieger,[2] K. Martin,[1] and C. Norman[1]

[1]*SGIM/Nuclear Fuel Cycle Information Analysis, International Atomic Energy Agency, Vienna, Austria*
[2]*International Safeguards, Institute of Energy and Climate Research, Forschungszentrum Jülich GmbH, Jülich, Germany*

Correspondence should be addressed to T. Burr; t.burr@iaea.org

Received 12 March 2017; Accepted 13 June 2017; Published 12 September 2017

Academic Editor: Oleg Melikhov

Nuclear safeguards aim to confirm that nuclear materials and activities are used for peaceful purposes. To ensure that States are honoring their safeguards obligations, quantitative conclusions regarding nuclear material inventories and transfers are needed. Statistical analyses used to support these conclusions require uncertainty quantification (UQ), usually by estimating the relative standard deviation (RSD) in random and systematic errors associated with each measurement method. This paper has two main components. First, it reviews why UQ is needed in nuclear safeguards and examines recent efforts to improve both top-down (empirical) UQ and bottom-up (first-principles) UQ for calibration data. Second, simulation is used to evaluate the impact of uncertainty in measurement error RSDs on estimated nuclear material loss detection probabilities in sequences of measured material balances.

1. Introduction

Nuclear material accounting (NMA) provides a quantitative basis to detect nuclear material loss or diversion at declared nuclear facilities. NMA involves periodically measuring facility input transfers T_{in}, output transfers T_{out}, and physical inventory I to compute a material balance (MB) defined for balance period j as $\text{MB}_j = (I_{j-1} + T_{\text{in},j} - T_{\text{out},j}) - I_j$, where $(I_{j-1} + T_{\text{in},j} - T_{\text{out},j})$ is the book inventory. In NMA, one MB or a collection of MBs are tested for the presence of any statistically significant large differences and/or for trends, while allowing for random and systematic errors in variance propagation to estimate the measurement error standard deviation of MB_j, σ_{MB_j}. Similarly, in verification activities done by an inspector, paired operator and inspector data are tested for any large differences and/or for trends [1, 2]. Therefore, both material balance evaluation and verification activities require statistical analyses, which require UQ.

In metrology for nuclear safeguards, the term "uncertainty" characterizes the dispersion of estimates of a quantity known as the measurand, which is typically the amount of NM (such as U or Pu) in an item. To measure the amount of NM, both destructive analysis (DA, a sample of item material is analyzed by mass spectrometry in an analytical chemistry laboratory) and nondestructive assay (NDA, an item is assayed by using a neutron or gamma detector) are used. NDA uses calibration and modeling to infer NM mass on the basis of radiation particles, such as neutrons and gammas emitted by the item and registered by detectors. For any measurement technique, one can use a first-principles physics-based or "bottom-up" approach to UQ by considering each key step and assumption of the particular method. Alternatively, one can take an empirical or "top-down" approach to UQ, for example, by comparing assay results on the same or similar items by multiple laboratories and/or calibration periods.

A well-known guide for bottom-up UQ is the Guide to the Expression of Uncertainty in Measurement (GUM, [3]). The GUM also briefly mentions top-down UQ in the context of applying analysis of variance (ANOVA, [4]) to data from interlaboratory studies. Although the GUM is useful, it is being revised because it has known limitations [5–7]. For example, the GUM provides little technical guidance regarding calibration as a type of bottom-up UQ or regarding

top-down UQ [5–8]. The GUM also mixes Bayesian with non-Bayesian concepts. In a Bayesian approach, all quantities, including the true measurand value, are regarded as random. In a non-Bayesian (frequentist) approach, some quantities are regarded as random and other quantities, such as the true value of the measurand, are regarded as unknown constants. This paper uses both Bayesian and non-Bayesian concepts but specifies when each is in effect. For example, in the Bayesian approach to top-down UQ in Section 3, the true RSD values are regarded as being random.

In NDA safeguards applications, the facility operator declares the NM mass of each item. Then, some of those items are randomly selected for NDA verification measurement by inspectors. This is a challenging NDA application because often the detector is brought to the facility where ambient conditions can vary over time and because the items to be assayed are often heterogeneous in some way and/or are different from the items that were used to calibrate/validate and assess uncertainty in the NDA method. Because of such challenges, "dark uncertainty" [9] can be large, as is evident whenever bottom-up UQ predicts smaller measurement error RSDs than are observed in top-down UQ [1]. The RSD of an assay method is often defined as the reproducibility standard deviation as estimated in an interlaboratory comparison. As shown in Section 3, comparing NDA verification measurements to the operator's DA measurements can be regarded as a special case of an interlaboratory evaluation [10–12].

For top-down UQ applied to NM measurements of the same item by both the operator (often using DA) and the inspector (often using NDA), this paper describes an existing and a new approach to separately estimate operator and inspector systematic and random error variance components. Systematic and random error components must be separated because their modes of propagation are different (Section 4). Currently, random error variance estimates (from paired data) are based on Grubbs' estimator or variations of Grubbs' estimator, which was originally developed by Grubbs to estimate random error variance separately for each of the two methods applied to each of several items, without repetition of measurement by either method [13, 14]. In Section 3, Grubbs' estimator, constrained versions of Grubbs' estimator, and a Bayesian alternative [7] are described; the Bayesian option easily allows for parameter constraints and prior information regarding the relative magnitudes of variance components to be exploited to improve top-down UQ.

This paper is organized as follows. Section 2 provides a background on bottom-up UQ for NDA, describes a gamma-based NDA example and a neutron-based NDA example, and illustrates why simulation is necessary for improved UQ for calibration data. Section 3 reviews currently used top-down UQ and describes a new Bayesian option [7] that applies approximate Bayesian computation. Section 4 provides a new simulation study assessing the sensitivity of estimated NM loss detection probabilities to estimation errors in measurement error RSDs. Section 5 concludes with a summary.

2. Bottom-Up UQ

For bottom-up UQ, the GUM [3] assumes that the measured value can be expressed using a measurand equation that relates input quantities (data collected during the measurement process and relevant fundamental nuclear data such as attenuation corrections) to the output (the final measurement value). The GUM's main technical tool is a first-order Taylor approximation applied to the measurand equation

$$Y = f\left(X_1, X_2, \ldots, X_N\right), \tag{1}$$

which relates input quantities $X_1, X_2, \ldots, X_N$ (regarded as random) to the measurand Y (also regarded as random). The input quantities can include estimates of other measurands or of calibration parameters, so (1) is quite general. The variance of each X and σ_i^2 and any covariances, $\sigma_i\sigma_j\rho_{i,j}$, between pairs of X's are then propagated using the Taylor approximation to obtain $\sigma_Y^2 \approx \sum_{i=1}^{N}(\partial f/\partial x_i)^2\sigma_i^2 + 2\sum_{i=1}^{N-1}\sum_{j=i+1}^{N}(\partial f/\partial x_i)(\partial f/\partial x_j)\sigma_i\sigma_j\rho_{i,j}$ (or using simulation if the Taylor approximation is not sufficiently accurate) to estimate the variance in Y, σ_Y^2.

According to (1), the estimated value Y of the measurand is a random variable, regardless of whether the left side of (1) is expressed as Y (as in a typical Bayesian approach) or as $\widehat{\mu}_Y$ (as in a typical non-Bayesian setting). The hat notation is a frequentist convention for denoting an estimator, so $\widehat{\mu}_Y$ is an estimate of μ_Y, and μ_Y (which is also denoted as y_T, where "T" denotes the true value) denotes the unknown true value of the measurand.

2.1. Calibration UQ as an Example of Bottom-Up UQ. Typically, calibration is performed using reference materials having nominal measurand values (known to within a relatively small uncertainty), and then, in the case of linear calibration, (1) can be reexpressed as $\widehat{\mu}_Y = \widehat{\beta}_O + \widehat{\beta}_1 X$, where $\widehat{\mu}_Y$ is the estimated measurand value, $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are parameters estimated from calibration data, X is the net count rate (usually the net gamma or net neutron count rate in NDA; see Section 2.3), and the three inputs in mapping to (1) are $X_1 = \widehat{\beta}_0$, $X_2 = \widehat{\beta}_1$, and $X_3 = X$. The estimates $\widehat{\beta}_0$ and $\widehat{\beta}_1$ will vary in predictable ways (Sections 2.2 and 2.3) across repeats of the calibration.

The convention in statistical literature to reverse the roles of X and Y from that in GUM's equation (1) will be followed here, so X denotes the quantity to be inferred (the measurand value) and Y denotes the detected net radiation count rate. Then, in the case of reverse regression (see below), (1) can be expressed as

$$X = g\left(Y_1, Y_2, \ldots, Y_N\right) = \widehat{\alpha}_0 + \widehat{\alpha}_1 Y, \tag{2}$$

identifying $Y_1 = \widehat{\beta}_0$, $Y_2 = \widehat{\beta}_1$, and $Y_3 = Y$. Following calibration on data consisting of $n\,(x_i, y_i)$ pairs (lowercase denotes observed value of a random variable), the three "input quantities" $Y_1 = \widehat{\beta}_0$, $Y_2 = \widehat{\beta}_1$, and $Y_3 = Y$ have variances and covariances that can be estimated. However, in most applications of calibration in NDA, accurate estimation of these variances and covariances requires simulation because

analytical approximations as described in Section 2.2 have been shown to be inadequate (see Section 2.3).

Expressing (2) as $X = g(Y_1, Y_2, \ldots, Y_N) = \widehat{\alpha}_0 + \widehat{\alpha}_1 Y$ indicates how the estimate X is computed and how to assign systematic and random error variances to X. For example, and to introduce notation used in top-down UQ (Section 3), one could express the estimate as $X = \mu_X + S + R$, where μ_X denotes the true value of the measurand, S denotes systematic error due to estimation error in the fitted slope and intercept and/or due to correlations among the inputs, and R denotes random error. If there are no correlations among the inputs but only estimation error in the fitted slope and intercept during calibration, then expressions for the variances of S and R, denoted as σ_S^2 and σ_R^2, respectively, can be given (Sections 2.3 and 3), which allow comparison between bottom-up UQ and top-down UQ. The GUM does not discuss calibration in much detail; instead, the GUM applies propagation of variance to the steps modeled in (1), which sometimes leads to a defensible estimate of the combined variances of S and R. The GUM does not attempt to separately estimate the variances of S and R, but such separation is needed in some applications, such as assigning an uncertainty to a sum of measurand estimates ([15] and Section 4).

2.2. Extension of Standard Regression Results to Calibration. One way to express that the net count rate depends on the true measurand value is

$$Y = \beta_0 + \beta_1 X_{\text{True}} + R_Y, \tag{3}$$

which is a typical model used in regression when there is negligible error in X. If errors in predictors cannot be ignored, (3) should be modified; however, one can still regress measured Y on measured X, so, in effect, (3) can be reexpressed as $Y = \widetilde{\beta}_0 + \widetilde{\beta}_1 X + \widetilde{R}_Y$, where the tildes denote that parameter values and the random error are different from those in (3).

In inverse calibration, (3) is used, and one inverts the fitted model using $\widehat{\beta}_0$ and $\widehat{\beta}_1$ to use future measured y_{test} to predict enrichment using

$$\widehat{x}_{\text{test}} = \frac{y_{\text{test}} - \widehat{\beta}_0}{\widehat{\beta}_1}, \tag{4}$$

which is regression followed by inversion. An alternative model to (3) is reverse calibration:

$$X = \alpha_0 + \alpha_1 Y_{\text{True}} + R_X. \tag{5}$$

In reverse calibration, (5) expresses the measurand X as a function of the true net count rate y_T, but in practice one must regress X on $Y = Y_{\text{true}} + R_Y$, where R_Y is a random error. As an aside, this paper does not consider models with systematic errors such as $X = X_{\text{True}} + R_X + S_X$ or $Y = Y_{\text{true}} + R_Y + S_Y$. Cheng and Van Ness [16] point out that any additive systematic errors in X or Y could be absorbed into β_0 and α_0, respectively; however, any systematic errors in the X values used for calibration would remain a part of the total uncertainty.

Both inverse and reverse calibrations involve ratios of random variables, which can be problematic [7, 17]. In inverse calibration, the solution in (4) involves division by the random variable $\widehat{\beta}_1$, which has a normal distribution under typical modeling assumptions. Williams [18] notes that $\widehat{x}_{\text{test}}$ in (4) has infinite variance even if the expected value of $\widehat{\beta}_1$ is nonzero, due to division by a normal random variable [19], and hence has infinite mean squared error, while the reverse estimator has finite variance and mean squared error. In reverse calibration, the least squares solution $\widehat{x}_T = \widehat{\alpha} y = \{Y_{\text{cal}}^T X_{\text{cal}} / Y_{\text{cal}}^T Y_{\text{cal}}\} y$ also involves division of random variables (X_{cal} is the vector or matrix of X values used in calibration and Y_{cal} is the vector of Y values in calibration). Experience suggests that one can develop adequate approximations for the ratio of random variables when the ratio is almost certain to be far from infinity or zero [19]. Ignoring errors in predictors, [20] uses the following common approximation for the variance of the ratio of random variables U and V:

$$\begin{aligned} &\text{var}\left(\frac{U}{V}\right) \\ &\quad \approx \frac{(EU)^2}{(EV)^2}\left\{\frac{\text{var}\,(U)}{(EU)^2} + \frac{\text{var}\,(V)}{(EV)^2} - 2\frac{\text{cov}\,(U,V)}{(EU)\,(EV)}\right\}, \end{aligned} \tag{6}$$

where E denotes expectation to derive the approximation (for inverse calibration) for variance due to uncertainty in the estimated calibration coefficients $\widehat{\beta}_0$ and $\widehat{\beta}_1$ and in the test measurement Y_{test}:

$$\text{var}\left(\widehat{X}\right) \approx \frac{1}{\beta_1^2}\left\{\frac{\sigma_{RY}^2}{n} + \frac{\left(x_{\text{test}} - \overline{x}\right)^2 \sigma_{RY}^2}{\sum_{i=1}^n \widetilde{x}_i^2}\right\}, \tag{7}$$

where $\widetilde{x}_i = x_i - \overline{x}$ and $\overline{x}$ is the mean of the x values in the calibration data. To apply (7), β_1^2 and σ_{RY}^2 are estimated from the calibration data (assuming (3) in forward calibration or the alternate version of (3), $Y = \widetilde{\beta}_0 + \widetilde{\beta}_1 X + \widetilde{R}_Y$). Equation (7) is almost the same as the corresponding well-known result for regression; the only differences are the swapping of the roles for x and y and the appearance of β_1^2 in the denominator. For reverse regression, [20] derives $\text{var}(\widehat{x}) \approx ((\sum_{i=1}^n \widetilde{x}_i^2)/(n-2))\{1/n + (y_{\text{test}} - \overline{y})^2 / \sum_{i=1}^n \widetilde{y}_i^2\}$, where $\widetilde{y}_i = y_i - \overline{y}$ and $\widetilde{\text{x}}_i = x_i - \overline{x}$. Reference [20] also showed the long-term bias $B_{\text{inverse}} \approx (x - \overline{x})\sigma_{RY}^2 / \beta_1^2 S_{xx}$ for inverse calibration and $B_{\text{reverse}} \approx -(x - \overline{x})/(1 + \beta_1^2 S_{xx}/(n-1)\sigma_{RY}^2)$ for reverse calibration, where $S_{xx} = \sum_{i=1}^n (x_i - \overline{x})^2$. Notice that B_{inverse} decreases as n increases (because S_{xx} increase as n increases), but B_{reverse} does not decrease as n increases; however, recall that, in NDA applications, n is small, usually 3 to 10.

A common summary performance measure of an estimator combines squared bias and repeatability variance defined as RMSE = (repeatability variance) + (bias)^2; that is, $\text{RMSE} = \sqrt{E\{\widehat{X} - X_{\text{true}}\}^2} = \sqrt{E\{\widehat{X} - E\widehat{X}\}^2 + \{E\widehat{X} - X_{\text{true}}\}^2}$, where E denotes the expected value (i.e., the first moment of the underlying probability distribution) [7]. Some technical details arise regarding the best model fitting approach if the predictor Y is measured with nonnegligible error. In addition, there is controversy regarding the relative merits of inverse

and reverse calibration [7, 17, 21, 22]. Simulation can be used to choose between inverse and reverse calibration, because simulation provides accurate UQ (such as RMSE estimation) for both options. In simulations for NDA calibration, errors in the standard reference materials' nominal values (X's) are usually small compared to errors in the instrument responses Y's, which are possibly adjusted by using adjustment factors that have uncertainty (see Section 2.3).

2.3. Summary of Recent NDA Examples. Recent publications have used simulation to assess the adequacy of (7) in the context of NM measurements by gamma detection [7, 17] and neutron detection [1, 7, 23, 24].

2.3.1. Enrichment Meter Principle (EMP). The EMP aims to infer the fraction of ^{235}U in U (enrichment, defined as atom percent of ^{235}U in an item) by measuring the count rate of the strongest-intensity direct (full-energy) gamma from decay of ^{235}U, which is emitted at 185.7 keV [7, 25, 26]. The EMP makes three key assumptions: (1) the detector field of view into each item is the same as that in the calibration items, (2) the item is homogeneous with respect to both ^{235}U enrichment and chemical composition, and (3) the container attenuation of gamma-rays is the same as or similar to that in the calibration items, so empirical correction factors have modest impact and are reasonably effective. If these three assumptions are approximately met, the enrichment of ^{235}U in the U is directly proportional to the count rate of the 185.7 keV gamma-rays emitted from the item. It has been shown empirically that, under good measurement conditions, the EMP can have a random error RSD of less than 0.5% and long-term bias of less than 1% relative to the true value, depending on the specific implementation of the EMP. Implementation details include features such as the detector resolution, stability, and extent of corrections needed to adjust items to calibration conditions. However, in some EMP applications, the random error RSD can be larger than bottom-up UQ predicts and larger than the 0.5% goal. For example, assay of the ^{235}U mass in a stratum of UO_2 drums suggests that there is a larger-than-anticipated random RSD [17].

2.3.2. Uranium Neutron Coincidence Collar (UNCL). The UNCL uses an active neutron source to induce fission in ^{235}U in fresh fuel assemblies [27]. Neutrons from fission are emitted in short bursts of time and so exhibit non-Poisson bursts in detected count rates. Neutron coincidence counting is used to measure the "doubles" neutron coincident rate Y, which can be used to estimate the linear density of ^{235}U in a fuel assembly (g-^{235}U/cm) using calibration parameters, a_1 and a_2. The coincident rate Y is the observed rate of observing two neutrons in very short time gates, each of approximately 10^{-6} sec, and is attributable to fission events. The equation commonly used to convert the measured doubles rate Y to an estimate of X (grams ^{235}U per cm) is $X = kY/(a_1 - a_2kY)$, where a_1 and a_2 are calibration parameters, and $k = k_0k_1k_2k_3k_4k_5$ is a product of correction factors that adjust Y to item-, detector-, and source-specific conditions in the calibration [27]. Therefore, $X = kY/(a_1 - a_2kY)$ is a special case of GUM's equation (1) (with X and Y reversed), where the two calibration parameters a_1 and a_2 and the 6 correction factors k_0, k_1, k_2, k_3, k_4, and k_5 are among X's in (1).

Reference [23] showed that calibration is most effective (leading to smallest RMSE in $\widehat{X}$) if there is no adjustment for errors in the predictor kY and that errors in k_0, k_1, k_2, k_3, k_4, and k_5, in $k = k_0k_1k_2k_3k_4k_5$, should be included in synthetic calibration data. Note that, by working with $1/X$ and $1/Y$, one can convert $X = kY/(a_1 - a_2kY)$ to one that is linear in the transformed predictor $1/Y$.

2.3.3. Main Results for Sections 2.3.1 and 2.3.2. The main results for Sections 2.3.1 and 2.3.2 can be summarized in four main points as follows.

(1) If possible, both classical (see (2)) and reverse (see (3)) regression methods should be compared; however, reverse regression tends to do either as well as or better than classical regression. Analytical approximations such as (7) have been shown not to be sufficiently adequate in some settings, so simulation is recommended to compare classical and reverse regression and to estimate variance components in $X = \mu_X + S + R$ (Section 3).

(2) Error sources that are expected to be present in test measurements, such as container thickness measurements, can be simulated in synthetic calibration data. Such error sources often lead to item-specific biases (Burr et al., 2016).

(3) If reverse regression is used, then there is no need to adjust for errors in the predictors Y in (3). If inverse regression is used, then it is better to adjust for errors in predictors.

(4) Figure 1 plots (a) the observed and predicted bias and (b) the observed and predicted RMSE in a generic NDA example involving either gamma or neutron counting. It is not well known that calibration applications lead to bias, and [7, 17] showed that the bias cannot be easily removed, because measurement errors obscure the true measurand value and hence the true bias. Note in Figure 1(a) that the observed bias (in simulated data) is not in close agreement with the predicted bias, which is obtained from the expressions in Section 2.2. Therefore, long-term bias should be estimated using simulation rather than relying on the approximate expressions $B_{\text{inverse}} \approx (x-\overline{x})\sigma_{RY}^2/\beta_1^2 S_{xx}$ for inverse calibration and $B_{\text{reverse}} \approx -(x-\overline{x})/(1+\beta_1^2 S_{xx}/(n-1)\sigma_{RY}^2)$ for reverse calibration, where $S_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2$. Similarly, Figure 1(b) illustrates that the observed RMSE is not well predicted by the expressions in Section 2.2, so, again, simulation is needed for adequate estimation of the RMSE. Note that the smallest RMSE is for reverse regression. Burr et al. [7, 17] show that reverse regression tends to have smaller RMSE than inverse regression but that if inverse regression is used, then methods to adjust for errors in predictors should be used.

Figure 1 summarizes the results of 10^5 simulations of $Y = \beta_0 + \beta_1 X_{\text{True}} + R_Y = 1 + 0.1X_{\text{True}} + R_Y$ with $\delta_X = 0.01$ and $\delta_Y = 0.15$, using 5 (x, y) calibration pairs (with x scaled to lie in $(0, 1)$, at 0, 0.25, 0.5, 0.75, and 1) and 10 testing pairs as shown. All simulations in this paper are done in R [25].

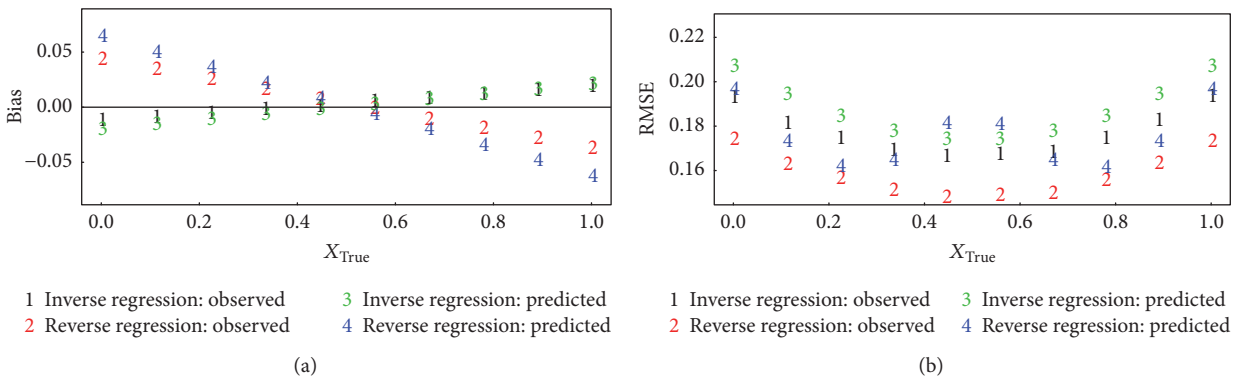


FIGURE 1: Simulation results for both inverse and reverse regression for (a) the observed and predicted bias versus X_{True} and (b) the observed and predicted RMSE versus X_{True}.

3. Top-Down UQ to Estimate Variance Components

In facilities under international safeguards agreements, inspectors measure randomly selected items to monitor for possible data falsification by the operator that could mask NM diversion [1, 28]. These paired (O, I) data (O denotes operator measurement; I denotes inspector measurement) are assessed using one-item-at-a-time testing to detect significant differences and also by using an average of the operator-inspector values to detect trends in the context of material balance evaluation [1]. Conclusions from such an assessment depend on the assumed measurement error model and associated random and systematic uncertainty components, so it is important to perform effective UQ [1, 7, 17, 28].

The paired (O, I) data are collected during relatively short (one week) inspections that occur once or a few times per year, and then several years of paired (O, I) inspection data are included in top-down UQ. The measurement error model must account for variation within and between groups, where, in this context, a group is an inspection period. A typical top-down model used for additive errors for the inspector (I) (and similarly for the operator O) is

$$I_{ij} = \mu_{ij} + S_{Ii} + R_{Iij}, \tag{8}$$

where I_{ij} is the inspector's measured value of item j (from 1 to n) in group i (from 1 to g), μ_{ij} is the true but unknown value of item j from group i, $R_{Iij} \sim N(0, \sigma^2_{RI})$ is a random error on item j from group i, and $S_{Ii} \sim N(0, \sigma^2_{SI})$ is a short-term systematic error in group i [28]. The error variance components σ^2_{SI} and σ^2_{RI} can be estimated using a specialized version of random-effects one-way ANOVA described in Section 3.1. NDA measurements often have larger uncertainty at larger true values, which implies a multiplicative rather than an additive error model. However, provided that the individual RSDs are fairly small, resulting in a total RSD of approximately 10% or less, a multiplicative error model such as $I_{ij} = \mu_{ij}(1 + S_{Ii} + R_{Iij})$ can be analyzed in the same manner as an additive error model, by analyzing on the log scale [1, 2]. Therefore, for brevity, only an additive error model such as in (8) is presented here. Bonner et al. [1] provide new expressions for a multiplicative error model that should be used if the total RSD is approximately 10% or larger. Note that one could write (8) in more cluttered notation as $Y_{I_{ij}} = \mu_{ij} + S_{Ii} + R_{Iij}$. That is, I_{ij} is the inspector's measured value of item j, which is obtained using various inputs, denoted with X's on the right side of (1). And one could also consider other error models, such as error models that allow for nonconstant absolute or relative random and/or systematic SD [28, 29].

The GUM [3] briefly describes ANOVA in the context of top-down UQ using measurement results from multiple laboratories and/or assay methods to measure the same measurand; however, the GUM is mostly concerned with bottom-up UQ. The GUM does not explicitly present any measurement error models such as (8) but only considers the model for the measurand, (1). However, the GUM implicitly endorses the notion of a measurement error model (or "observation equation," [14]) such as (8) in its top-down UQ. Note that if total measurement error is partitioned into random and systematic components, then the variance of a sum of n measured NM values (which is often needed in safeguards assessments; see Section 4) is $n^2\sigma^2_{SI} + n\sigma^2_{RI}$ for an additive model such as (8).

To illustrate, Figure 2 plots $d = O - I$ data simulated from (8) with $n = 10$, $g = 5$, $\sigma_{RO} = 1$, $\sigma_{SO} = 0.1$, $\sigma_{RI} = 3$, $\sigma_{SI} = 1$, $\mu_{\text{True}} = 100$, and the standard deviation of the true values, $\sigma_\mu = 1$. The horizontal lines depict the five group means. Equation (8) implies that there is a need to partition error variance of d into "between" (B) and "within" (W) groups, as in

$$\sum_{i=1}^{g}\sum_{j=1}^{n}\left(d_{ij} - \bar{d}\right)^2 = \sum_{i=1}^{g}\sum_{j=1}^{n}\left(d_{ij} - \bar{d}_i\right)^2 + n\sum_{i=1}^{g}\left(\bar{d}_i - \bar{d}\right)^2 = \text{SSW} + \text{SSB}. \tag{9}$$

3.1. Grubbs' Estimator as an Example of Top-Down UQ to Estimate σ^2_{RO}, σ^2_{SO}, σ^2_{RI}, σ^2_{SI}. Standard random-effects ANOVA [4] requires repeated measurements on some items in order to estimate σ^2_{RI} and then σ^2_{SI} in data assumed to be produced by a model such as (8). However, for most (O, I) data, repeated measurements of the same item are not available, so

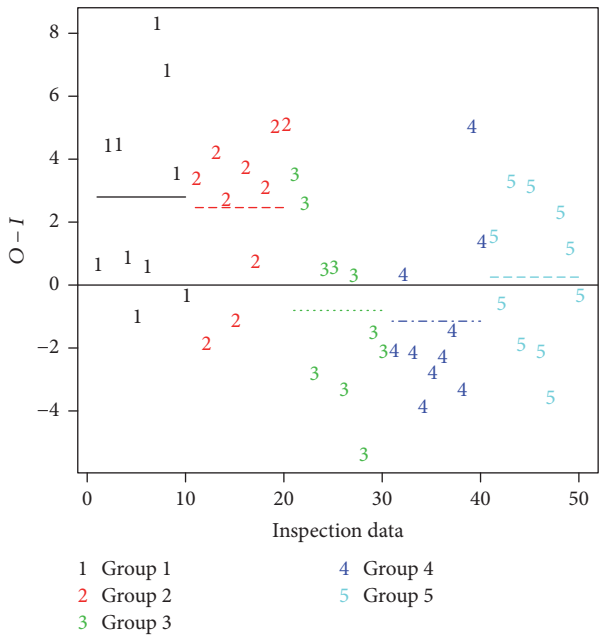


FIGURE 2: Values of $d = O-I$ in data simulated from (7) with $n = 10$, $g = 5$, $\sigma_{RO} = 1$, $\sigma_{SO} = 0.1$, $\sigma_{RI} = 3$, $\sigma_{SI} = 1$, $\mu_{\text{True}} = 100$, and the SD of the true values $\sigma_\mu = 1$.

this section describes Grubbs' estimator. Grubbs' estimator was developed for situations in which more than one measurement method is applied to each of multiple test items (which may contain different material amounts), but there is no replication of measurements by any of the methods. Grubbs' estimator will be described for additive measurement error models; a new version of Grubbs' estimator for multiplicative error models is described in [1]. Note that the variance σ^2_{RI} of the random error variance component R_{Iij} includes the effects of "item-specific" bias (see Section 2, [7, 28]), which could not be estimated if available data were only from repeated measurements of the same or very similar items. Note also that Grubbs' estimator does not consider the possibility of falsification by the operator, so it is intended to be applied to paired (O, I) data that has no evidence of falsification.

The basis of Grubbs' estimator within one group to estimate σ^2_{RO} and σ^2_{RI} is that the covariance between operator and inspector measurements equals the variance of the true item masses, σ^2_μ, while the variance of I, σ^2_I, conditional on the value of S is given by $\sigma^2_I = \sigma^2_\mu + \sigma^2_{RI}$. Therefore, the sample covariance within a single inspection period between operator and inspector measurements can be subtracted from the sample variance of the inspector measurements to estimate σ^2_{RI} (and similarly for estimating σ^2_{RO}). That is, within one inspection period (group) (lowercase $i(o)$ denotes the observed values of $I(O)$), Grubbs' estimator is given by

$$\widehat{\sigma}^2_{RI} = \frac{1}{n-1}\left\{\sum_{j=1}^{n}\left(i_j - \bar{i}\right)^2 - \sum_{j=1}^{n}\left(o_j - \bar{o}\right)\left(i_j - \bar{i}\right)\right\}. \quad (10)$$

The estimates from (10) from each of the g groups are averaged to get the final estimate of the inspector's random error variance, and similarly, the estimate of σ^2_μ is the average of the sample covariances $\widehat{\sigma}^2_\mu = \sum_{j=1}^{n}(o_j - \bar{o})(i_j - \bar{i})/(n-1)$ computed within each group.

To estimate σ^2_{SI} in (7), a minor extension of standard random-effects ANOVA to account for σ^2_μ shows that $E\{\sum_{j=1}^{g} n(\bar{I}_j - \bar{\bar{I}})^2/(g-1)\} = \sigma^2_{RI} + \sigma^2_\mu + n\sigma^2_{SI}$, so a method-of-moments-based estimate of σ^2_{SI} is $\widehat{\sigma}^2_{SI} = (\sum_{j=1}^{g}(\bar{i}_j - \bar{\bar{i}})^2)/(g-1) - (\widehat{\sigma}^2_{RI} + \widehat{\sigma}^2_\mu)/n$. There is no guarantee that $\widehat{\sigma}^2_{RI}$ or $\widehat{\sigma}^2_{SI}$ are nonnegative, but the corresponding true quantities are nonnegative (i.e., $\sigma^2_{RI} \geq 0$ and $\sigma^2_{SI} \geq 0$), so constrained versions of Grubbs-based and ANOVA-based estimators can be used (Section 3.2, [7, 30]).

Grubbs showed (and simulation in R [25] also verified) that his estimator for σ^2_{RI} has variance $\sigma^2_{\widehat{\sigma}^2_{RI}} = 2\sigma^4_{R_I}/(n-1) + (1/(n-1))(\sigma^2_{R_O}\sigma^2_{R_I} + \sigma^2_{R_O}\sigma^2_\mu + \sigma^2_{R_I}\sigma^2_\mu)$, which is relatively large, particularly if σ^2_μ is comparable in magnitude to σ^2_{RI} [7] and/or n is small, so [13] proposed an option to mitigate the impact of σ^2_μ on $\sigma^2_{\widehat{\sigma}^2_{RI}}$. The method in [13] relied on knowing the value, or approximate value of $\sigma^2_{RI}/\sigma^2_\mu$, and studied the sensitivity to misspecifying the ratio $\sigma^2_{RI}/\sigma^2_\mu$. The Bayesian option in Section 3.2 specifies a probability distribution for $\sigma^2_{RI}/\sigma^2_\mu$ prior to observing the (O, I) data, as an example of enforcing nonnegativity constraints and including prior information, such as information from bottom-up UQ regarding σ^2_μ, σ^2_{RI}, and/or σ^2_{SI}, and similarly for the operator or for variance ratios such as $\sigma^2_{RI}/\sigma^2_{RO}$.

A measurement error model that can be used in top-down UQ for the type of data in Figure 2 was given in (8). In the case of inverse regression as a type of bottom-up UQ, using (5), one can modify the top-down error model of (8) to

$$I_{ij} = \mu_{ij} + S_{1Ii} + S_{2Ii} + R_{Iij}, \quad (11)$$

where $R \sim N(0, \sigma^2_R)$ (σ^2_R is random error variance that includes the effects of errors in X and Y); $S_1 = \widehat{\alpha}_0 - \alpha_0$ (an additive subcomponent of systematic error; see below), and $S_2 = (\widehat{\alpha}_1 - \alpha_1)(Y_{\text{Test}} - \overline{y})$ (a multiplicative subcomponent of systematic error).

Assume that inverse regression is performed using mean-centered data ($\widetilde{x} = x - \overline{x}_{\text{train}}$ and $\widetilde{y} = y - \overline{y}_{\text{train}}$), so $\text{cov}(\widehat{\alpha}_0, \widehat{\alpha}_1) = 0$, which simplifies interpretation. The equation $S_1 = \widehat{\alpha}_0 - \alpha_0$ is then interpreted to mean that S_1 has mean zero and variance σ^2_R/n. Note that one cannot use paired (O, I) data to estimate σ^2_{S1} and σ^2_{S2}, because the effects of σ^2_{S1} and σ^2_{S2} are confounded. However, simulation such as that used to produce Figure 1 provides a way to perform bottom-up prediction of top-down SD (or RSD) that can be compared to the SD (or RSD) observed in top-down evaluations. If the bottom-up predicted SD agrees well with that observed in top-down (O, I) data, then bottom-up UQ via simulation (because analytical approximations such as those plotted in Figure 1 in the calibration context have been shown to not be

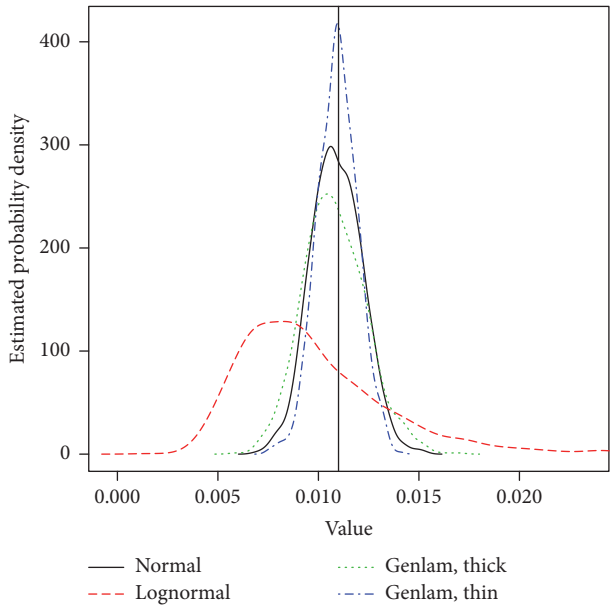


FIGURE 3: The estimated probability density of Grubbs' estimator for σ_{RO} when the random errors have either a normal, lognormal, or generalized lambda (thin tail or thick tail) distribution. The true value of σ_{RO} is 0.0111, indicated by the vertical line.

sufficiently accurate) can separately estimate σ^2_{S1} and σ^2_{S2}, or, if fitting a zero-intercept model, (8) could be modified to a multiplicative error model, $I_{ij} = \mu_{ij}(1 + S_{Ii} + R_{Iij})$, where $S_I = (\widehat{\alpha}_1 - \alpha_1)(Y_{\text{Test}} - \overline{y})$, again with the variance of S_I estimated by simulation [7, 17, 31].

3.2. New Bayesian Approach to Grubbs-Type Estimation. Recall that the variance of Grubbs' estimator can be large, so [7, 31] review alternatives to Grubbs' estimator based on constrained optimization, such as Jaech's [31] constrained maximum likelihood estimator (CMLE), which assumes that the random and systematic measurement errors are normally distributed. Also, although the impact of σ^2_μ can be relatively large on Grubbs' estimator, versions of Grubbs' estimators that are constrained so that $\widehat{\sigma}^2_{R_I} + \widehat{\sigma}^2_{R_O} = \widehat{\sigma}^2_d$, where $\widehat{\sigma}^2_d$ is the sample variance of the differences $d = O - I$, have exhibited lower RMSE than Grubbs' estimator in limited testing to date. Any estimator of $\sigma^2_{R_O}, \sigma^2_{R_I}, \sigma^2_{S_O}$, and $\sigma^2_{S_I}$ must be accompanied by its respective uncertainty. The uncertainty in CMLE or constrained least squares estimators [32] can be approximated using approximate analytical results and asymptotic results or by resampling methods such as the bootstrap. The quality of such approximations is not yet known; so a Bayesian alternative is presented here which does not rely on such approximations or the bootstrap for assessing uncertainty in $\widehat{\sigma}^2_{R_O}, \widehat{\sigma}^2_{R_I}, \widehat{\sigma}^2_{S_O}$, or $\widehat{\sigma}^2_{S_I}$.

Another reason to consider the Bayesian alternative is that Grubbs' estimator exhibits dependence on the underlying measurement error distribution. Figure 3 plots the estimated probability density for Grubbs' estimator for $\widehat{\sigma}_{RO}$, when the underlying random error distribution is either the normal, lognormal, or generalized lambda with thin or thick tails, for the relatively large sample size of 5 groups and 10 measurements per group. The probability density for $\widehat{\sigma}_{RO}$ was estimated using a kernel-density estimator in R [25]. There is a relatively large uncertainty in $\widehat{\sigma}_{RO}$ (Section 4 evaluates one impact of uncertainty in estimated RSDs), with an RSD in $\widehat{\sigma}_{RO}$ of 11%, 8%, 13%, and 37%, respectively, for the four distributions in Figure 3, and an RSD in $\widehat{\sigma}_{SO}$ of approximately 50% for all four distributions; also, for 2 groups and 5 measurements per group, the RSD in $\widehat{\sigma}_{RO}$ is approximately 50% for all four distributions. One implication of Figure 3 is that uncertainties in $\widehat{\sigma}^2_{R_O}, \widehat{\sigma}^2_{R_I}, \widehat{\sigma}^2_{S_O}$, or $\widehat{\sigma}^2_{S_I}$ depend on the error distributions. An advantage of the Bayesian option is that the width of the Bayesian posterior adjusts to accommodate nonnormal underlying distributions [7].

One option to improve Grubbs' estimator is to impose constraints, such as $\sigma_{RO} \le \sigma_{RI}$, or, more flexibly, by assigning a prior probability to the ratio σ_{RO}/σ_{RI} which puts most of the probability on values less than 1. The uncertainty in constrained estimators is simple to estimate in a Bayesian approach and is often difficult to estimate in a non-Bayesian framework. Within an inspection group for fixed S_O and S_I, the paired (O, I) data from (8) has a bivariate distribution with covariance matrix $\Sigma_W = \begin{pmatrix} \sigma^2_\mu + \sigma^2_{R_O} & \sigma^2_\mu \\ \sigma^2_\mu & \sigma^2_\mu + \sigma^2_{R_I} \end{pmatrix}$ and [33] provided a Bayesian approach to estimate σ^2_μ, $\sigma^2_{R_O}$, and $\sigma^2_{R_I}$, assuming a bivariate normal likelihood, without imposing constraints on any of the variance ratios. In principle, one could extend the Bayesian approach in [33] to allow for a nonnormal likelihood and/or to allow for constraints on any of the variance ratios. In practice, such extensions are technically difficult and rarely attempted.

In any Bayesian approach, prior information regarding the sizes or relative sizes of σ^2_μ, $\sigma^2_{R_O}$, and $\sigma^2_{R_I}$ must be provided. If the prior is "conjugate" for the likelihood, then the posterior is in the same likelihood family as the prior, in which case analytical methods are available to compute posterior prediction intervals for quantities of interest, so that a wide variety of priors and likelihoods can be accommodated; modern Bayesian methods do not rely on conjugate priors but use numerical methods to obtain samples of σ^2_μ, $\sigma^2_{R_O}$, and $\sigma^2_{R_I}$ from their approximate posterior distributions [34]. For numerical methods such as Markov Chain Monte Carlo, the user must specify a prior distribution for σ^2_μ, $\sigma^2_{R_O}$, and $\sigma^2_{R_I}$ and a likelihood (which need not be normal). The Bayesian approach presented next is approximate Bayesian computation (ABC), which does not require a known likelihood for the data and can accommodate constraints on variances and/or ratios of variances by choice of the prior distributions.

The "output" of any Bayesian analysis is the posterior distribution and so the output of ABC is an estimate of the posterior distributions of σ^2_μ, $\sigma^2_{R_O}$, and $\sigma^2_{R_I}$. No matter what type of Bayesian approach is used, a well-calibrated Bayesian approach satisfies several requirements. The requirement of interest here is that, in repeated applications of ABC, approximately 95% of the middle 95% of the posterior distribution

for each of σ_μ^2, $\sigma_{R_O}^2$, and $\sigma_{R_I}^2$ should contain the respective true values.

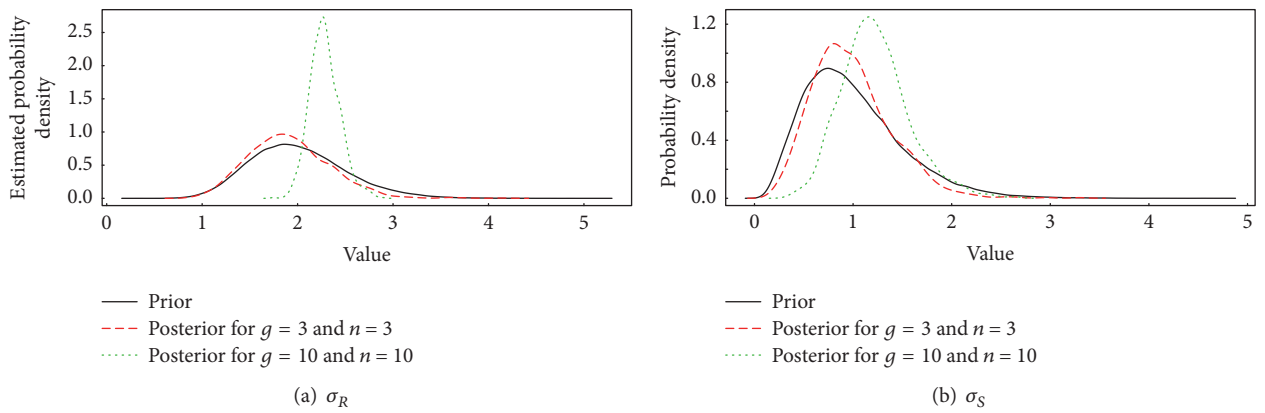
In ABC, one assumes that a model has input parameters θ and outputs data $y_M = y(\theta)$ (M for "model") and that there is corresponding observed real data y_{obs}. Here, the model is (8), which specifies how to generate synthetic O and I data and does require a likelihood; however, the true likelihood used to generate the data need not be known to the user. Synthetic data is generated from the model for many trial values of θ, and trial θ values are accepted as contributing to the estimated posterior distribution for $\theta|y_{\text{obs}}$ if the distance $D(y\text{obs}, y(\theta))$ between $y$obs and $y_M = y(\theta)$ is reasonably small. Alternatively, for most applications, it is necessary to reduce the dimension of y_{obs} to a small set of summary statistics S and instead accept trial values of θ if $D(S(y\text{obs}), S(y(\theta))) < T$, where T is a user-chosen threshold. Here, y_{obs} is the paired (O, I) data in each inspection group, and $S(y_{\text{obs}})$ includes within- and between-groups sums of squares and within-group covariance between O and I. Specifically, recall that the estimator of σ_{RI}^2 in (8) is $\hat{\sigma}_{RI}^2 = (1/(n-1))\{\sum_{j=1}^{n}(i_j - \bar{i})^2 - \sum_{j=1}^{n}(o_j - \bar{o})(i_j - \bar{i})\}$ and that a method-of-moments-based estimate of σ_{SI}^2 is $\hat{\sigma}_{SI}^2 = (\sum_{j=1}^{g}(\bar{i}_j - \bar{\bar{i}})^2)/(g-1) - (\hat{\sigma}_{RI}^2 + \hat{\sigma}_\mu^2)/n$. And σ_μ^2 can be estimated using $\hat{\sigma}_\mu^2 = \sum_{j=1}^{n}(o_j - \bar{o})(i_j - \bar{i})$. The quantities $\hat{\sigma}_{RI}^2$, $\hat{\sigma}_{SI}^2$, $\hat{\sigma}_{RO}^2$, $\hat{\sigma}_{SO}^2$, and $\hat{\sigma}_\mu^2$ are therefore good choices for summary statistics to be used for ABC. Because trial values of θ are accepted if $D(S(y_{\text{obs}}), S(y(\theta))) < T$, an approximation error to the posterior distribution arises which several ABC options attempt to mitigate. Such options involve weighting the accepted θ values by the actual distance $D(S(y_{\text{obs}}), S(y(\theta)))$ (`abctools` in `R` [25]). As an aside, if the error model is multiplicative rather than additive, expressions given in [1] can be used as summary statistics.

To summarize, ABC consists of three steps: (1) sample parameter values from their prior distribution $p_{\text{prior}}(\theta)$; (2) for each simulated value of θ in (1), simulate data from (8); (3) accept a fraction of the sampled prior values in (1) by checking whether the summary statistics computed from the data in (2) satisfy $D(S(y_{\text{obs}}), S(y(\theta))) < T$. If desired, aiming to improve the approximation to the posterior, adjust the accepted θ values on the basis of the actual $D(y_{\text{obs}}, \theta)$ value. ABC requires the user to make three choices: the summary statistics, the threshold T, and the measure of distance d. Reference [35] introduced a method to choose summary statistics that use the estimated posterior means of the parameters based on pilot simulation runs. Reference [36] used an estimate of the change in posterior $p_{\text{posterior}}(\theta)$ when candidate summary statistics are added to the current set of summary statistics. Reference [37] illustrated a method to evaluate whether a candidate set of summary statistics leads to a well-calibrated posterior in the same sense that is used in this paper; that is, nominal posterior probability intervals should have approximately the same actual coverage probability.

To illustrate application of ABC to top-down UQ, simulations using (8) were performed. Recall that an additive model is a reasonable approximation to a multiplicative model if the error variances are small and the data is analyzed on a log scale or if there are effects in addition to calibration which impact the random and systematic errors, such as random changes in background count rates which are not adjusted for. Simulations from a multiplicative error model were also investigated, with good results, such as those described next for the additive model; the summary statistics for ABC to use Grubbs-type estimators for a multiplicative model are given in [1, 7].

The simulations were performed in `R` using three steps. In the first step, ABC requires a training collection of parameter values and corresponding summary statistics for each of many simulations. So, in each of 10^5 simulations, the values for $\sigma_{RO}, \sigma_{SO}, \sigma_{RI}, \sigma_{SI}, \sigma_\mu$ were sampled from their respective prior probability densities. In the second step, (8) was used to generate I and O data. In the third step, the expressions for $\hat{\sigma}_{RO}^2$, $\hat{\sigma}_{SO}^2$, $\hat{\sigma}_{RI}^2$, $\hat{\sigma}_{SI}^2$, and $\hat{\sigma}_\mu^2$ given above were used as summary statistics, resulting in a parameter matrix and corresponding summary statistics matrix, each having 10^5 rows and 5 columns. Then, in a separate set of 10^5 simulations, the same first and second steps were repeated, and for each simulated set of parameters and summary statistics, the parameter and summary statistics matrices from the third step in training were used in the `abc` function in the `abctools` package, using an acceptance fraction of 0.01 (meaning that 1% of the trial values for the true parameters were accepted), to produce an approximate posterior for each of the five parameters. This posterior can be used to assess the actual coverage probability, for example, of an interval that contains 95% of the posterior probability.

It was found [7] that the actual coverages for σ_{RO}, σ_{SO}, σ_{RI}, σ_{SI} were essentially the same (to within simulation uncertainty) as the nominal coverages, at 90%, 95%, and 99% probabilities, for a normal distribution and all of the nonnormal distributions investigated (uniform, gamma, lognormal, beta, t, and generalized lambda with thick or thin tails), each having the same $\sigma_{RO}, \sigma_{SO}, \sigma_{RI}, \sigma_{SI}$ values. In addition, when a normal distribution was used to train ABC and any of the evaluated nonnormal distributions were used to test ABC, very nearly the same actual coverages (to within approximately ± 0.01) were obtained. As another check of robustness, one prior distribution was used for training ABC and a prior that was wider by a factor of 1.3 was used in testing. In that case, the actual coverages dropped from approximately 0.95 to approximately 0.90, so this implementation is less robust to using a wider prior in testing than in training. Also, the RMSE of the ABC estimator was compared to each of several non-Bayesian constrained estimators that are currently being evaluated. Not surprisingly, provided that the prior used in training was approximately the same as that used in testing, the ABC estimator had lowest RMSE. The intent here is not to make any general RMSE performance claims; instead, these results provide a second indication that the ABC implementation is well calibrated, in the sense that if the assumed prior equals the true prior then the RMSE of the ABC estimator is highly competitive with that of the non-Bayesian estimator and in the sense that the width of the posterior accurately describes uncertainty. Finally, a well-calibrated Bayesian analysis allows one to evaluate the effect



(a) σ_R (b) σ_S

FIGURE 4: Prior and posterior probability densities for data from realization of (8) for σ_{RI} and σ_{SI} for $g = 3$ and $n = 3$ and for $g = 10$ and $n = 10$. The true values are $\sigma_{RI} = 2.3$ (% relative) and $\sigma_{SI} = 1.2$ (% relative).

of increasing sample size on uncertainty in the estimated values of $\sigma_{RO}, \sigma_{SO}, \sigma_{RI}, \sigma_{SI}$.

To illustrate ABC output, Figure 4 plots the prior density and the estimated posterior density for σ_{RI} and σ_{SI} for $g = 3$ and $n = 3$ and for $g = 10$ and $n = 10$. Because the posterior densities are well calibrated, they can be used to reliably assess whether top-down estimates of σ_{RI} and σ_{SI} are in agreement to within their respective uncertainties of the corresponding bottom-up estimates of σ_{RI} and σ_{SI} from Section 2. The priors used in Figure 4 are wide, with no relative information regarding variance ratios assumed, and the parameter σ_μ is assigned a prior with a mean value of 0.10, with a relative standard deviation in σ_μ of 10%. And Figure 4 shows that, in this case, having only $g = 3$ and $n = 3$ per group does not lead to a narrow posterior, but, with $g = 10$ and $n = 10$, the posterior is fairly narrow.

4. Application of RSD Estimates: Statistical Testing of Materials Accounting Data

4.1. Sequential Statistical Testing of MB Sequences. Recall that NMA evaluates one or more MBs, where the MB is defined for balance period j as $\mathrm{MB}_j = (I_{j-1} + T_{\mathrm{in},j} - T_{\mathrm{out},j}) - I_j$ [38, 39]. Typically, many measurements are combined to estimate the terms $T_{\mathrm{in}}, I_{\mathrm{begin}}, T_{\mathrm{out}}$, and I_{end} in the MB; therefore, the central limit effect and years of experience suggest that MBs in most facilities will be approximately normally distributed with the mean equal to the true NM loss λ_j and standard deviation σ_j, which is expressed as $X_j \sim N(\lambda_j, \sigma_j)$, where X denotes the MB and the notation σ_j is a shortened version of $\sigma_{\mathrm{MB},j}$. A sequence of n MBs will be assumed to have approximately a multivariate normal distribution [38–43], $(X_1, X_2, \ldots, X_n) \sim MVN(\boldsymbol{\lambda}, \Sigma)$, where the n-by-n covariance matrix is

$$\Sigma = \begin{pmatrix} \sigma_1^2 \sigma_{12}^2 \cdots \sigma_{1n}^2 \\ \sigma_{21}^2 \sigma_2^2 \cdots \sigma_{2n}^2 \\ \vdots \\ \sigma_{n1}^2 \sigma_{n2}^2 \cdots \sigma_n^2 \end{pmatrix}. \quad (12)$$

The magnitude of σ_j determines the amount of NM loss, $\lambda = \sum_{k=1}^{n} \lambda_k$, which leads to high detection probability (DP). For example, suppose the facility tests for NM loss only, not for NM gain, and assume that $X_j \sim N(\lambda_j, \sigma_j)$ is an adequate model. Then, if a false alarm probability (FAP) of $\alpha = 0.05$ is desired, the alarm threshold is $1.65\sigma_j$. In the case of testing for loss only, it follows that the loss detection probability $1 - \beta$ for $\lambda = 3.3\sigma$ and $1 - \beta > 0.95$ if $\lambda > 3.3\sigma_j$, where β is the nondetection (false negative) probability. The factor 3.3 arises from symmetry of the Gaussian distribution, requiring $\alpha = \beta = 0.05$, and the fact that $1.65 = 3.3/2$ is the 0.95 quantile of the $N(0, 1)$ distribution. One common goal is for DP $= 1 - \beta$ to be at least 0.95 if $\lambda \geq 1$ SQ (significant quantity, which is 8 kg for Pu), which is accomplished if $\sigma_j \leq$ SQ/3.3. If $\sigma_j >$ SQ/3.3, this can be mitigated by reducing measurement errors to achieve $\sigma_j \leq$ SQ/3.3 (if feasible) and/or by closing the balances more frequently, so there is less nuclear material transferred per balance period, which reduces σ_j [38, 39]. The DP of other safeguards measures such as enhanced containment and surveillance with smart cameras and/or remote radiation detection is difficult to quantify and is outside the scope of this paper. This section concludes with four remarks.

Remark 1. Large throughput facilities try to make σ_j as small as reasonably possible and often try to keep σ_j small as a percent of throughput but cannot achieve $\sigma_j \leq$ SQ/3.3. For example, suppose that there is a measurement error relative standard deviation of 0.5% of throughput. And suppose that the FAP/DP goals are $\alpha = 0.05$ and $1 - \beta = 0.95$ and annual throughput is 100 SQ. Then, $\sigma_j = 0.5$ SQ $=$ SQ/2 $>$ SQ/3.3, so protracted diversion of 1 SQ over one year will not have a high DP. Therefore, one reason for frequent MB accounting is that abrupt diversion over hours or days is very likely to be detected [40]. As a complementary approach that is beyond the scope here, process monitoring [39] methods have the potential to detect off-normal facility operation that could misdirect NM to undeclared locations.

Remark 2. In the 1980s, some believed that a plant reporting a MB every 30 days would have higher DP than that same plant

reporting a MB every year. However, [44] showed that, for optimal (from the diverter's viewpoint) protracted diversion with the per-period loss being proportional to the row sums of the covariance matrix Σ of the MB sequence, annual MB testing has higher DP than monthly MB testing, and so, for such protracted diversion, "less frequent balance closure is better." However, [44] conceded that NRTA has shorter detection times and higher DP against abrupt diversion. Publications [45–50] soon followed involving joint sequential tests: one tuned to detect protracted diversion and one more tuned for abrupt diversion. Such joint Page's tests can be tuned to have high DP against abrupt loss while still having reasonably high DPs against protracted loss. Two types of combined Page's tests are included in the simulation study in Section 4.3.

Remark 3. The assumption $(X_1, X_2, \ldots, X_n) \sim MVN(\boldsymbol{\lambda}, \Sigma)$ implies that Σ is known without estimation error. In practice, Σ is estimated using variance propagation applied to $X_j = \mathrm{MB}_j = (I_{j-1} + T_{\mathrm{in},j} - T_{\mathrm{out},j}) - I_j$ and there will be estimation error in the estimate $\widehat{\Sigma}$ [40].

The simulation study in Section 4.3 includes a sensitivity study to assess the impact of estimation error $\widehat{\Sigma}$ on the estimated false alarm probabilities (FAPs) and detection probabilities (DPs).

Remark 4. This section focuses on operator MB sequences. In international safeguards, the inspector randomly selects items to verify NM declarations made by the operator. The difference statistic, $D = N\sum_{j=1}^{n}((o_j - i_j)/n)$, defined as the average difference in the sample of size n (extrapolated to the population of size N) between operator declarations (almost always based on operator measurements) and inspector measurements can be used as a test statistic, or the D statistic [2] can be used to compute the inspector MB statistic (or sequence). Inspector MB = MB − D, which could be analyzed using sequential statistical methods such as those in Section 4.3. Alternatively, one can test individually each of the n paired differences $d_j = o_j - i_j$ and the overall D statistic, and if none are found to be statistically significant, then the IAEA could rely on operator MB evaluation as in Section 4.3.

4.2. Propagation of Variance to Estimate Σ. Estimating Σ is a key step required in frequent NMA. To illustrate, a simplified example model of a generic electrochemical facility with one input stream, one output stream, and one key inventory item will be used here [38]. Each measurement method is modeled here using a multiplicative measurement error model for the operator (O): $M_i = \mu_i(1 + S_i + R_i)$, with $S_i \sim N(0, \delta_S^2)$ and $R_i \sim N(0, \delta_R^2)$, where M_i is the operator's measured value of item i, μ_i is the true but unknown value of item i, R_i is a random error of item i, and S_i is a short-term systematic error for item i. Then, the diagonal terms of Σ are calculated as

$$\begin{aligned}\sigma_i^2 = {} & {T_{\mathrm{in}_i}}^2\left(\delta_{\mathrm{in},R}^2 + \delta_{\mathrm{in},S}^2\right) + {T_{\mathrm{out}_i}}^2\left(\sigma_{\mathrm{out},R}^2 + \sigma_{\mathrm{out},S}^2\right) \\ & + I_i^2\delta_{\mathrm{in}\nu,R}^2 + I_{i-1}^2\delta_{\mathrm{in}\nu,R}^2 + \left(I_i - I_{i-1}\right)^2\delta_{\mathrm{in}\nu,S}^2.\end{aligned} \tag{13}$$

The off-diagonal terms in Σ are calculated as

$$\begin{aligned}\sigma_{ij}^2 = {} & T_{\mathrm{in}_i}T_{\mathrm{in}_j}\delta_{\mathrm{in},S}^2 + T_{\mathrm{out}_i}T_{\mathrm{out}_j}\delta_{\mathrm{out},S}^2 \\ & + \left(I_iI_j + I_{i-1}I_{j-1}\right)\delta_{\mathrm{in}\nu,S}^2 \\ & - I_iI_{j-1}\left(\delta_{\mathrm{in}\nu,S}^2 + \delta_{\mathrm{in}\nu,R}^2\left[\text{if } j - i = 1\right]\right) \\ & - I_{i-1}I_j\left(\delta_{\mathrm{in}\nu,S}^2 + \delta_{\mathrm{in}\nu,R}^2\left[\text{if } i - j = 1\right]\right).\end{aligned} \tag{14}$$

In the last two terms, the random error of the inventory term is only applied if the condition is true. Reference [31] showed that the POV results for σ_i^2 and σ_{ij}^2 are obtained by appropriate interpretation of GUM's equation (1) in Section 2. For this simplified version of an example from [38], this leads to examples of 12-by-12 covariance matrices for monthly MBs over a one-year period; three example matrices Σ_1, Σ_2, and Σ_3 are listed next.

Three example covariance matrices, Σ_1, Σ_2, and Σ_3, for a generic electrochemical facility, are given [38]:

$$\begin{matrix} 1.00 & -0.48 & 0.01 & 0.01 & 0.01 \\ -0.48 & 1.00 & -0.48 & 0.01 & 0.01 \\ 0.01 & -0.48 & 1.00 & -0.48 & 0.01 \\ 0.01 & 0.01 & -0.48 & 1.00 & -0.48 \\ 0.01 & 0.01 & 0.01 & -0.48 & 1.00 \end{matrix} \tag{15}$$

Equation (15) is Σ_1, scaled to unit variance, only displaying 5 by 5 of the 12 by 12. This is the nominal case with 4 kg input per period and 40 kg inventory; $\delta_R = \delta_S = 0.01$.

$$\begin{matrix} 1.00 & 0.17 & 0.33 & 0.33 & 0.33 \\ 0.17 & 1.00 & 0.17 & 0.33 & 0.33 \\ 0.33 & 0.17 & 1.00 & 0.17 & 0.33 \\ 0.33 & 0.33 & 0.17 & 1.00 & 0.17 \\ 0.33 & 0.33 & 0.33 & 0.17 & 1.00 \end{matrix} \tag{16}$$

Equation (16) is Σ_2, scaled to unit variance. This is the case with 4 kg input per period and 4 kg inventory.

$$\begin{matrix} 1.00 & 0.58 & 0.67 & 0.67 & 0.67 \\ 0.58 & 1.00 & 0.58 & 0.67 & 0.67 \\ 0.67 & 0.58 & 1.00 & 0.58 & 0.67 \\ 0.67 & 0.67 & 0.58 & 1.00 & 0.58 \\ 0.67 & 0.67 & 0.67 & 0.58 & 1.00 \end{matrix} \tag{17}$$

Equation (17) is Σ_3, scaled to unit variance. This is the case with 4 kg input per period and 40 kg inventory and δ_S increased from 0.01 to 0.02.

4.3. Sequential Testing of Material Balance in Nuclear Material Accountancy. The assumption $(X_1, X_2, \ldots, X_n) \sim MVN(\boldsymbol{\lambda}, \Sigma)$ implies that $Y = \Sigma^{-1/2}X \sim MVN(\Sigma^{-1/2}\boldsymbol{\lambda}, I)$, where I is the identity matrix. The transform $Y = \Sigma^{-1/2}X$ is known

in safeguards as the standardized independently transformed MUF (SITMUF, where MUF is another name for the MB), which is most conveniently computed using the Cholesky decomposition [43]. There are two main advantages of applying statistical tests to Y rather than to X. First, alarm thresholds depend only on the sequence length n for Y and not on the form of the covariance matrix Σ. Because it is best to calculate thresholds using simulation, this is a logistic advantage. Second, the variance of the Y sequence decreases over time, so if a diversion occurs late in the analysis period, the DP is larger for the Y sequence than for the X sequence. Note that one cannot claim higher DP for the Y sequence than for the X sequence in general, because the true loss scenario is never known, and the DP can be larger for X than for Y for some loss scenarios, which is demonstrated in Section 4.

The value of Y_i can be calculated using $Y = \Sigma^{-1/2}X$, but, more intuitively as the residual from the X sequence, $Y_j = \{X_j - E(X_j \mid X_{j-1}, X_{j-2}, \ldots, X_1)\}/\widetilde{\sigma}_j$, where the standard deviation $\widetilde{\sigma}_j$ is given by $\widetilde{\sigma}_j = \sqrt{\sigma_{jj}^2 - f\Sigma^{-1}f^T}$, where $f = \Sigma_{j,1:(j-1)}$, the 1 to $(j-1)$ entries in the jth row of Σ.

Several reasonable statistical tests have been evaluated in [38, 41, 44–51] and are included in the simulation study in Section 4.4, including the following 13 tests:

(1) MUF test: this compares each MUF value to a threshold, which is the same as a Shewhart test in quality control (QC). The test alarms on period j if $\text{MUF}_j/\sigma_j \geq T$ for some threshold T

(2) SITMUF test: this compares each SITMUF value $Y = \Sigma^{-1/2}X$ to a threshold, which is the same as a Shewhart test in QC. The test alarms on period j if $\text{SITMUF}_j/\sigma_{\text{SITMUF},j} \geq T$ for some threshold T

(3) Page's test applied to MUF: Page's test to test for loss is a sequence of sequential probability ratio tests, defined as $P_j = \max(0, P_{j-1} + x_j/\sigma_j - k)$, where $P_0 = 0$ [52]. The test alarms on period j if $P_j > T$. The parameter k is a control parameter that is optimal for detecting a shift from zero loss to loss λ if $k = \lambda/2$. The alarm threshold T (usually denoted as h in literature on Page's test, but this paper uses T for alarm threshold) is chosen so that the FAP per analysis period (usually one year) is 0.05 or whatever FAP is specified

(4) Page's test applied to SITMUF: Page's test to test for loss is a sequence of sequential probability ratio tests, as $P_j = \max(0, P_{j-1} + y_j - k)$. The alarm threshold T is chosen so that the FAP per analysis period (usually one year) is 0.05 or whatever FAP is specified

(5) Combined Page's tests applied to MUF: the use of Page's test with a large value of k and small value of T has good DP for abrupt loss, and the use of Page's test with a small value of k and large value of T has good DP for protracted loss. Therefore, a reasonable option is to use a combination of two Page's tests, one with large k and one with small k

(6) Applying combined Page's tests to SITMUF

(7) CUMUF: at period j, $\text{CUMUF}_j = \sum_{i=1}^{j} x_i$ is the sum of all MUF values from period 1 to j

(8) GEMUF: it has been shown that if the loss vector $\boldsymbol{\lambda}$ is known, then the Neyman-Pearson test statistic is $\boldsymbol{\lambda}^T\Sigma^{-1}\mathbf{x}$, which is known as a matched filter in some literature. The GEMUF statistic substitutes $\mathbf{x}^T$ for $\boldsymbol{\lambda}^T$, so $\text{GEMUF} = \mathbf{x}^T\Sigma^{-1}\mathbf{x}$. In simulation studies, $\boldsymbol{\lambda}$ is known, so the NP test statistic is useful for calculating the largest possible DP. The GE in GEMUF is a German abbreviation of Geschätzter, which means "estimated," so GEMUF means estimated MUF, and GEMUF is the same as the Mahalanobis distance from the $\mathbf{0}$ vector and Hotelling's multivariate T statistic [51]

(9) A nonsequential version of the Neyman-Pearson test, $\boldsymbol{\lambda}^T\Sigma^{-1}\mathbf{x}$, is useful to calculate the largest possible DP for given Σ and $\boldsymbol{\lambda}$. For completeness, four other combined tests are also considered

(10) SITMUF and CUMUF

(11) Page's on SITMUF and CUMUF

(12) SITMUF, Page's on SITMUF, and GEMUF

(13) Page's on SITMUF, CUMUF, and GEMUF

This section concludes with four remarks.

Remark 1 (SITMUF transform). The SITMUF transform is recommended for two reasons. First, simulation is typically used to select alarm thresholds, and it is convenient to always work on the same scale when selecting alarm thresholds, so the fact that $Y = \Sigma^{-1/2}X \sim MVN(\Sigma^{-1/2}\boldsymbol{\lambda}, I)$ is convenient. Note that alarm thresholds could be selected on the basis of exact or approximate analytical results for some, but not all, of the tests. For example, there are approximate expressions for T and k [53]. Second, the standard deviation $\widetilde{\sigma}_j$ is given by $\widetilde{\sigma}_j = \sqrt{\sigma_j^2 - f\Sigma^{-1}f^T}$, where $f = \Sigma_{j,1:(j-1)}$, the 1 to $(j-1)$ entries in the jth row of Σ, so the standard deviation of the MUF residuals decreases in the later periods. Therefore, the independence transform is analogous to a bias adjustment, leading to smaller prediction variance in later periods, which tends to increase the DP for SITMUF compared to MUF (there are exceptions where the DP for MUF is larger than the DP for SITMUF; see Section 4.4, DP results).

Remark 2 (choosing thresholds). Thresholds can be chosen in many ways and can be assumed to be constant for each period or not. Therefore, simulation DP results in Section 4.4 are not claimed to be optimal but are example DP results.

Remark 3 (performance criteria). The main performance criterion for comparing tests is the DP. But the average time to detection and robustness to misspecifying the covariance matrix Σ are also important.

Remark 4. There are other tests in the literature, including the power one and scan tests [38, 41–43].

Table 1: DPs for Σ_1. Boldface entries have the largest DPs (excluding NP) for the respective column.

DP	Loss 1 (3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3)/10	Loss 2 (3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0) (0, 0, 0, 0, 0, 0, 3, 0, 0, 0, 0, 0) (0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3)	Loss 3 (0, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0)	Loss 4 (0,1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1)
MUF	0.07	0.58, 0.58, 0.58	0.15	0.19
SITMUF	0.20	0.65, 0.89, 0.81	0.82	0.58
Page on MUF	0.18	0.75, 0.85, 0.74	0.76	0.58
Page on SITMUF	0.71	0.82, 0.99, 0.56	**1.0**	**0.99**
Combined Page MUF	**0.74**	**0.83, 0.99, 0.80**	**1.0**	**0.99**
Combined Page, SITMUF	**0.74**	**0.83, 0.98, 0.80**	**1.0**	**0.99**
CUMUF	0.26	0.91, 0.64, 0.18	0.50	0.65
GEMUF	0.13	0.87, 0.70, 0.11	0.57	0.48
NP	0.82	0.99, 1.0, 0.99	1.0	1.0
SITMUF and CUMUF	0.23	0.80, 0.87, 0.80	0.81	0.62
SITMUF, Page, CUMUF	0.28	0.85, 0.85, 0.75	0.76	0.58
SITMUF, Page, GEMUF	0.20	0.84, 0.84, 0.72	0.74	0.61
Page, CUMUF, GEMUF	0.23	0.85, 0.75, 0.75	0.76	0.58

Table 2: DPs for Σ_2.

DP	Loss 1 (0.3, 0.3, 0.3, 0.3, 0.3, 0.3, 0.3, 0.3, 0.3, 0.3, 0.3, 0.3)	Loss 2 (3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0) (0, 0, 0, 0, 0, 0, 3, 0, 0, 0, 0, 0) (0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3)	Loss 3 (0, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0)	Loss 4 (0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1)
MUF	0.07	0.60, 0.60, 0.60	0.18	0.18
SITMUF	0.06	0.62, 0.76, 0.80	0.14	0.14
Page on MUF	**0.13**	0.13, 0.15, 0.12	**0.23**	**0.23**
Page on SITMUF	0.09	0.24, 0.63, 0.58	0.18	0.19
Combined Page MUF	0.10	0.52, 0.76, 0.79	0.21	0.21
Combined Page SITMUF	0.10	0.52, 0.76, 0.79	0.20	0.20
CUMUF	0.07	0.81, 0.05, 0.05	0.09	0.19
GEMUF	0.06	0.79, 0.41, 0.11	0.14	0.14
NP	0.15	0.98, 0.98, 0.98	0.44	0.62
SITMUF and CUMUF	0.06	**0.79, 0.71, 0.75**	0.18	0.13
SITMUF, Page, CUMUF	0.11	0.69, 0.57, 0.63	0.16	0.17
SITMUF, Page, GEMUF	0.11	0.52, 0.65, 0.79	0.19	0.21
Page, CUMUF, GEMUF	0.11	0.69, 0.57, 0.63	0.16	0.20

4.4. Simulation Study. Example DP results are in Tables 1–3 for Σ_1, Σ_2, and Σ_3, respectively, using 10^5 simulations (so are repeatable to within ± 0.01) to choose alarm thresholds and to estimate DPs. Tables 1–3 indicate that, as expected, there is no overall best test. However, Page's test on SITMUF or a combined Page's test on SITMUF is often among the best performers. Whether Page's test on SITMUF has larger DP than Page's test on MUF depends on the covariance matrix.

To illustrate the behavior of some of the tests using Σ_1 and loss 3 in Tables 1–3 ((0, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0)), Figure 5 plots the true loss and example MUF and SITMUF values (and the average SITMUF value over all simulations) in (a) and

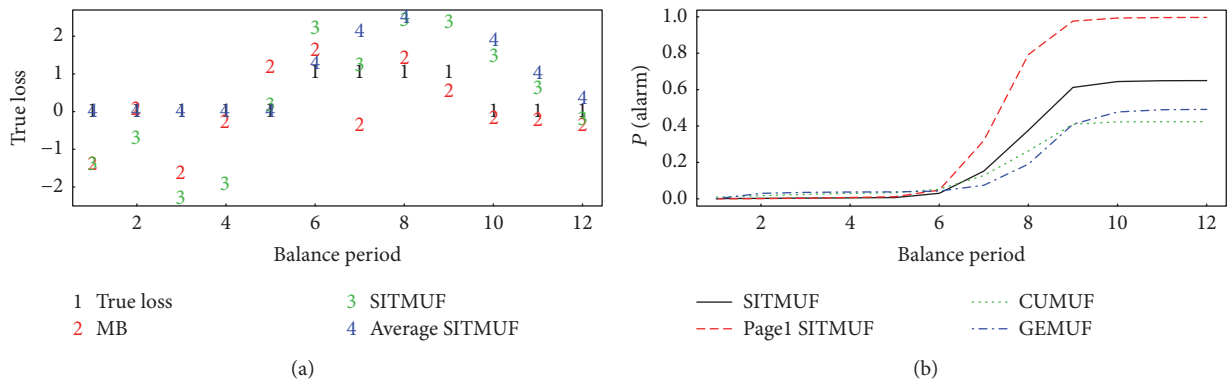


Figure 5: For $\Sigma = \Sigma_1$, (a) the true loss, MB, SITMUF, and average SITMUF over all simulations versus balance period and (b) the alarm probability versus balance period (10^5 simulations) for SITMUF, Page1 SITMUF, CUMUF, and GEMUF.

Table 3: DPs for Σ_3.

DP	Loss 1 (0.3, 0.3, 0.3, 0.3, 0.3, 0.3, 0.3, 0.3, 0.3, 0.3, 0.3, 0.3)	Loss 2 (3, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0) (0, 0, 0, 0, 0, 3, 0, 0, 0, 0, 0, 0) (0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 3)	Loss 3 (0, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0)	Loss 4 (0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1)
MUF	0.07	0.66, 0.66, 0.66	0.15	0.17
SITMUF	0.05	0.81, 0.98, 0.81	0.43	0.20
Page on MUF	**0.11**	0.10, 0.10, 0.10	0.12	0.16
Page on SITMUF	0.07	0.20, 0.93, 0.20	**0.83**	0.18
Combined Page MUF	0.08	0.52, 0.98, 0.52	0.80	**0.22**
Combined Page SITMUF	0.07	0.52, 0.98, 0.52	0.79	**0.22**
CUMUF	0.06	0.80, 0.05, 0.80	0.05	0.07
GEMUF	0.06	0.91, 0.85, 0.96	0.18	0.20
NP	0.11	1.0, 1.0, 1.0	0.97	0.83
SITMUF and CUMUF	0.06	**0.85, 0.97, 0.98**	0.35	0.17
SITMUF, Page, CUMUF	0.10	0.24, 0.82, 0.88	0.14	0.16
SITMUF, Page, GEMUF	0.10	0.24, 0.82, 0.88	0.14	0.16
Page, CUMUF, GEMUF	0.10	0.09, 0.09, 0.09	0.11	0.15

plots example DPs in (b). Figure 6 plots the residual standard deviation $\tilde{\sigma}_j = \sqrt{\sigma_{jj}^2 - f\Sigma^{-1}f^T}$ versus balance period for Σ_1.

A sensitivity analysis was also performed by simulating 30% RSD in $\widehat{\delta}_R$ and 50% RSD in $\widehat{\delta}_S$ (see Section 3.2). With these relatively large RSDs in $\widehat{\delta}_R$ and $\widehat{\delta}_S$, there is large uncertainty in the DP. For example, the 95% interval for DPs based on 10^5 simulations is {(0.09, 0.41), (0.59, 0.99), (0.34, 1.0), **(0.96, 1.0)**, **(0.96, 1.0)**, **(0.96, 1.0)**, (**0.11, 0.92**), (0.26, 0.97), (0.99, 1.0), (0, 56, 1.0), (0.35, 1.0), (0.93, 1.0), (0.31, 1.0)}, respectively, for the 13 tests for the loss in column 3 in Table 1 (0, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0). The three least-sensitive DPs are in boldface and are for Page on SITMUF, combined Page on MUF, and combined Page on SITMUF. The most sensitive DP is also in boldface and underlined and is for CUMUF.

5. Summary

Statistical analyses used to support safeguards conclusions require UQ, usually by estimating the RSD in random and systematic errors associated with each measurement method. This paper reviewed why UQ is needed in nuclear safeguards and examined recent efforts to improve both bottom-up and top-down UQ for calibration data. A key issue in bottom-up UQ using calibration with only a few calibration standards is that existing analytical approximations to estimate variance components are not sufficiently accurate, so this paper illustrated that simulation is needed. Once calibration UQ is well quantified, whenever improved bottom-up UQ predicts smaller measurement error RSDs than are observed in top-down UQ [1], this is evidence of significant unknown

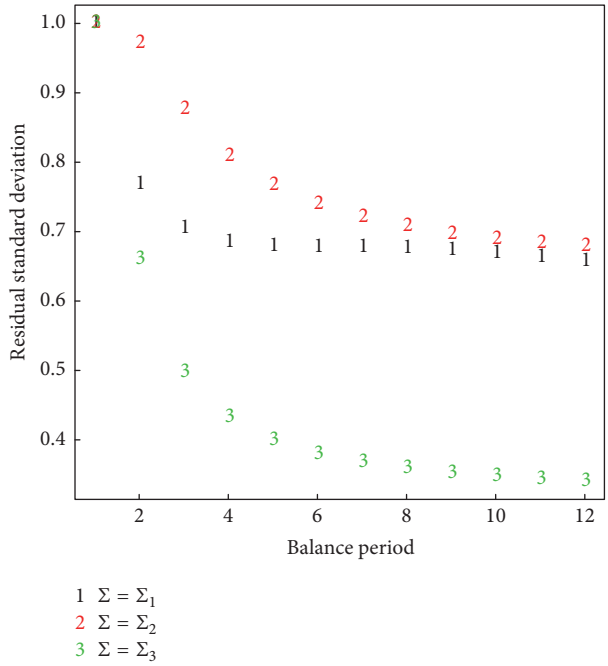


FIGURE 6: Residual standard deviation $\tilde{\sigma}_i = \sqrt{\sigma_{ii}^2 - f\Sigma^{-1}f^T}$ versus balance period for Σ_1.

NDA error sources ("dark uncertainty," [9]), which can then potentially be identified.

The RSD of an assay method is often defined as the reproducibility standard deviation as estimated in an interlaboratory comparison. Recent options for top-down UQ, such as constrained estimators or Bayesian estimators that use prior information, offer possible improvements over existing variance component estimators. Any such improvements in estimated RSDs should be accompanied by uncertainties in the RSDs, which means that uncertainty in the estimated uncertainties matters [2, 7]. The ABC approach in Section 3 appears to provide a robust estimate of the posterior distribution of the RSDs.

There are other types of bottom-up UQ used in safeguards not considered here. For example, the FRAM (fixed-energy, response function analysis with multiple efficiencies) gamma-based method [54] does not rely on calibration, and FRAM's uncertainties are impacted by physical mismatch between test items and assay assumptions, which leads to item-specific bias, and also by uncertainties in nuclear data such as half-lives. This paper also used simulation to evaluate the impact of uncertainty in measurement error RSDs on estimated nuclear material loss detection probabilities in sequences of measured material balances. Many different sequential statistical tests were evaluated. In a related context, [2] evaluated the impact of uncertainty in measurement error RSDs on estimated DPs in verification data.

Acronyms

ANOVA: Analysis of variance
ABC: Approximate Bayesian computation
CMLE: Constrained maximum likelihood estimator
CUMUF: Cumulative material unaccounted for
DA: Destructive analysis
EMP: Enrichment meter principle
FRAM: Fixed-energy, response function analysis with multiple efficiencies for gamma spectroscopy
GUM: Guide to the Expression of Uncertainty in Measurement
MB, MUF: Material balance and material unaccounted for are synonyms
NDA: Nondestructive analysis
NMA: Nuclear material accounting
RSD: Relative standard deviation
RMSE: Root-mean-square error
SQ: Significant quantity
SITMUF: Standardized independently transformed MUF
UQ: Uncertainty quantification: top-down UQ is empirical and bottom-up UQ is first-principles
UNCL: Uranium neutron coincidence collar.

Additional Points

List of Symbols. $\mathrm{MB}_j = (I_{j-1} + T_{\mathrm{in},j} - T_{\mathrm{out},j}) - I_j$, where $(I_{j-1} + T_{\mathrm{in},j} - T_{\mathrm{out},j})$ is the book inventory. $I_{ij} = \mu_{ij} + S_{Ii} + R_{Iij}$, where I_{ij} is the inspector's measured value of item j (from 1 to n) in group i (from 1 to g), μ_{ij} is the true but unknown value of item j from group i, $R_{Iij} \sim N(0, \sigma_{RI}^2)$ is a random error on item j from group i, and $S_{Ii} \sim N(0, \sigma_{SI}^2)$ is a short-term systematic error in group i. $\mathrm{RMSE} = \sqrt{E\{\widehat{X} - X_{\mathrm{true}}\}^2} = \sqrt{E\{\widehat{X} - E\widehat{X}\}^2 + \{E\widehat{X} - X_{\mathrm{true}}\}^2}$. $Y = f(X_1, X_2, \ldots, X_N)$ is GUM's equation for the measurand Y and inputs $X_1, X_2, \ldots, X_N$.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] E. Bonner, T. Burr, T. Guzzardo et al., "Improving the effectiveness of safeguards through comprehensive uncertainty quantification," *Journal of Nuclear Materials Management*, vol. 44, no. 2, pp. 53–61, 2016.

[2] T. Burr, T. Krieger, C. Norman, and K. Zhao, "The impact of metrology study sample size on uncertainty in IAEA safeguards calculations," *EPJ Nuclear Sciences & Technologies*, vol. 2, p. 36, 2016.

[3] JCGM 104:2009, Evaluation of measurement data- an introduction to the "Guide to the Expression of Uncertainty in Measurement," 2009.

[4] R. Miller, *Beyond ANOVA: basics of applied statistics*, Chapman & Hall, 1986.

[5] W. Bich, "Revision of the 'guide to the expression of uncertainty in measurement'. Why and how," *Metrologia*, vol. 51, no. 4, pp. S155–S158, 2014.

[6] C. Elster, "Bayesian uncertainty analysis compared with the application of the GUM and its supplements," *Metrologia*, vol. 51, no. 4, pp. S159–S166, 2014.

[7] T. Burr, S. Croft, K. Jarman, A. Nicholson, C. Norman, and S. Walsh, "Improved uncertainty quantification in nondestructive assay for nonproliferation," *Chemometrics and Intelligent Laboratory Systems*, vol. 159, pp. 164–173, 2016.

[8] R. Willink, *Measurement Uncertainty and Probability*, Cambridge University Press, Cambridge, Mass, USA, 2013.

[9] M. Thompson and S. L. R. Ellison, "Dark uncertainty," *Accreditation and Quality Assurance*, vol. 16, no. 10, pp. 483–487, 2011.

[10] T. Deutler, "Grubbs-type estimators for reproducibility variances in an interlaboratory test study," *Journal of Quality Technology*, vol. 23, no. 4, pp. 324–335, 1991.

[11] ISO 21748:2010 Guidance for the use of repeatability, reproduciblity, and trueness estimates in measurement uncertianty estimation.

[12] K. Martin and A. Böckenhoff, "Analysis of short-term systematic measurement error variance for the difference of paired data without repetition of measurement," *AStA. Advances in Statistical Analysis. A Journal of the German Statistical Society*, vol. 91, no. 3, pp. 291–310, 2007.

[13] F. Lombard and C. J. Potgieter, "Another look at the Grubbs estimators," *Chemometrics and Intelligent Laboratory Systems*, vol. 110, no. 1, pp. 74–80, 2012.

[14] A. Possolo and B. Toman, "Assessment of measurement uncertainty via observation equations," *Metrologia*, vol. 44, no. 6, pp. 464–475, 2007.

[15] M. Thompson, "Scoring the sum of correlated results in analytical proficiency testing," *Analytical Methods*, vol. 2, no. 7, pp. 976-977, 2010.

[16] C.-L. Cheng and J. W. Van Ness, *Kendall's Library of Statistics 6 Statistical Regression with Measurement Error*, vol. 6, Oxford University Press Inc., New York, NY, USA, 1999.

[17] T. Burr, S. Croft, T. Krieger, K. Martin, C. Norman, and S. Walsh, "Uncertainty quantification for radiation measurements: Bottom-up error variance estimation using calibration information," *Applied Radiation and Isotopes*, vol. 108, pp. 49–57, 2016.

[18] E. J. Williams, "A Note on Regression Methods in Calibration," *Technometrics*, vol. 11, no. 1, pp. 189–192, 1969.

[19] G. Marsaglia, "Ratios of normal variables," *Journal of Statistical Software*, vol. 16, no. 4, pp. 1–10, 2006.

[20] P. A. Parker, G. G. Vining, S. R. Wilson, J. L. Szarka III, and N. G. Johnson, "The prediction properties of classical and inverse regression for the simple linear calibration problem," *Journal of Quality Technology*, vol. 42, no. 4, pp. 332–347, 2010.

[21] R. G. Krutchkoff, "Classical and inverse regression methods of calibration," *Technometrics. A Journal of Statistics for the Physical, Chemical and Engineering Sciences*, vol. 9, pp. 425–439, 1967.

[22] R. G. Krutchkoff, "Classical and Inverse Regression Methods of Calibration in Extrapolation," *Technometrics*, vol. 11, no. 3, pp. 605–608, 1969.

[23] S. Croft, T. Burr, A. Favalli, and A. Nicholson, "Analysis of calibration data for the uranium active neutron coincidence counting collar with attention to errors in the measured neutron coincidence rate," *Nuclear Instruments and Methods in Physics Research, Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 811, pp. 70–75, 2016.

[24] T. Burr, S. Croft, D. Dale, A. Favalli, B. Weaver, and B. Williams, "Emerging applications of bottom-up uncertainty quantification in nondestructive assay," *ESARDA Bulletin*, vol. 53, pp. 54–61, 2015.

[25] R: A language and environment for statistical Computing, R Development Core Team, R Foundation for Statistical Computing, Vienna, Austria, 2008.

[26] ASTM C1514, Standard test method for measurement of ^{235}U fraction using the enrichment meter principle, 2008.

[27] P. Henriksen, H. Menlove, J. Stewart, S. Qiao, T. Wenz, and G. Verrecchia, "Neutron collar calibration and evaluation for assay of LWR fuel assemblies containing burnable neutron absorbers," Tech. Rep. LA-11965-MS, 1990.

[28] T. L. Burr and G. S. Hemphill, "Multiple-component radiation-measurement error models," *Applied Radiation and Isotopes*, vol. 64, no. 3, pp. 379–385, 2006.

[29] W. Horwitz, "Evaluation of Analytical Methods Used for Regulation of Foods and Drugs," *Analytical Chemistry*, vol. 54, no. 1, pp. 67A–76A, 2012.

[30] J. Jaech, *Statistical Analysis of Measurement Errors*, Exxon Monograph Series, Wiley Sons, New York, 1985.

[31] T. Burr, K. Martin, and T. Krieger, "The analysis of measurement errors as outlined in GUM and in the IAEA statistical methodologies for safeguards: a comparison in the IAEA's statistical methodologies for safeguards, Linking Strategy, Implementation," in *Proceedings of the Symposium on International Safeguards*, 2014.

[32] J. Hartung, "Nonnegative minimum biased invariant estimation in variance component models," *The Annals of Statistics*, vol. 9, no. 2, pp. 278–292, 1981.

[33] N. Draper and I. Guttman, "Two simultaneous measurement procedures: a Bayesian approach," *Journal of the American Statistical Association*, vol. 70, pp. 43–46, 1975.

[34] S. Moussaoui, C. Carteret, D. Brie, and A. Mohammad-Djafari, "Bayesian analysis of spectral mixture data using Markov Chain Monte Carlo Methods," *Chemometrics and Intelligent Laboratory Systems*, vol. 81, no. 2, pp. 137–148, 2006.

[35] P. Fearnhead and D. Prangle, "Constructing summary statistics for approximate BAYesian computation: semi-automatic approximate BAYesian computation," *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, vol. 74, no. 3, pp. 419–474, 2012.

[36] P. Joyce and P. Marjoram, "Approximately sufficient statistics and Bayesian computation," *Statistical Applications in Genetics and Molecular Biology*, vol. 7, no. 1, Article 26, 2008.

[37] T. Burr and A. Skurikhin, "Selecting summary statistics in approximate bayesian computation for calibrating stochastic models," *BioMed Research International*, vol. 2013, Article ID 210646, 10 pages, 2013.

[38] T. Burr and M. S. Hamada, "Revisiting statistical aspects of nuclear material accounting," *Science and Technology of Nuclear Installations*, vol. 2013, Article ID 961360, 15 pages, 2013.

[39] T. Burr, M. S. Hamada, L. Ticknor, and J. Sprinkle, "Hybrid statistical testing for nuclear material accounting data and/or process monitoring data in nuclear safeguards," *Energies*, vol. 8, no. 1, pp. 501–528, 2015.

[40] T. Burr and M. S. Hamada, "Bayesian updating of material balances covariance matrices using training data," *nternational Journal of Prognostics and Health Monitoring*, vol. 5, no. 1, 13 pages, 2014.

[41] T. Speed and D. Culpin, "The role of statistics in nuclear materials accounting: issues and problems," *Journal of the Royal Statistical Society B*, vol. 149, no. 4, pp. 281–313, 1986.

[42] A. S. Goldman, J. P. Shipley, and R. R. Picard, "Statistical methods for nuclear materials safeguards: An overview," *Technometrics*, vol. 24, no. 4, pp. 267–274, 1982.

[43] R. R. Picard, "Sequential analysis of materials balances," *Nuclear materials management*, vol. 15, no. 2, pp. 38–42, 1987.

[44] R. Avenhaus and J. Jaech, "On subdividing material balances in time and/or space," *Nuclear materials management*, vol. 10, no. 3, pp. 24–33, 1981.

[45] B. Jones, "Calculation of diversion detection using the SITMUF sequence and pages test: application to evaluation of facility designs," in *Proceedings of the 7th ESARDA Symposium on Safeguards and Nuclear Material Management*, Liege, Belgium, 1985.

[46] B. Jones, "Calculation of diversion detection using the SITMUF sequence and pages test: response to abrupt and protracted diversion," in *Proceedings of the in Proceedings of the International Symposium on Nuclear Material Safeguards, IAEA-SM-293/23*, Vienna, Austria, 1986.

[47] B. Jones, "Comparison of near real time materials accountancy using SITMUF and Pages test with conventional accountancy," in *in Proceedings of the 9th ESARDA Symposium on Safeguards and Nuclear Material Management*, London, UK, 1987.

[48] B. Jones, "Near real time materials accountancy using SITMUF and a joint Pages test: dependence of response on balance frequency," in *Proceedings of the in Proceedings of the 3rd International Conference on Facility Operations-Safeguards Interface*, San Diego, Calif, USA, 1987.

[49] B. Jones, "Near real time materials accountancy using SITMUF and a joint Pages test: comparison with MUF and CUMUF tests," *ESARDA Bulletin*, vol. 15, pp. 0–26, 1988.

[50] B. Jones, "Near real time materials accountancy using SITMUF and a joint pages test: improvement of the test," *ESARDA Bulletin*, vol. 16, pp. 13–19, 1989.

[51] R. Seifert, *The GEMUF test: a new sequential test for detecting loss of material in a sequence of accounting periods, Nuclear safeguards technology*, 1986, http://www.iaea.org/inis/collection/NCLCollectionStore/_Public/18/086/18086510.pdf.

[52] E. S. Page, "Continuous inspection schemes," *Biometrika*, vol. 41, pp. 100–114, 1954.

[53] D. Brook and D. A. Evans, "An approach to the probability distribution of cusum run length," *Biometrika*, vol. 59, pp. 539–549, 1972.

[54] T. L. Burr, T. E. Sampson, and D. T. Vo, "Statistical evaluation of fram γ-ray isotopic analysis data," *Applied Radiation and Isotopes*, vol. 62, no. 6, pp. 931–940, 2005.