Uncertainty quantification as presented in training courses for safeguards inspectors

Elisa Bonner[1], Thomas Burr[3], Thomas Krieger[4], Klaus Martin[1], Claude Norman[1], Peter Santi[2]

[1]IAEA/SGIM/Nuclear Fuel Cycle Information Analysis;
[2]IAEA/SGCP/Safeguards Training
[3]Los Alamos National Laboratory;
[4]Forschungszentrum Jülich GmbH

Summary

For safeguards evaluators to provide credible assurance that States are honoring their safeguards obligations, quantitative conclusions regarding non-diversion from States' nuclear material flows and inventories are needed. The statistical analysis used to reach these conclusions requires that each measurement method undergo uncertainty quantification (UQ). Training for safeguards inspectors includes measurement error models that must account for variation within and between groups, where a group is defined to be a calibration or inspection period. A typical model for multiplicative errors for the inspector (I) *is* $I_{jk} = \mu_{jk}(1 + S_{Ij} + R_{Ijk})$ *with* $S_{Ij} \sim N(0, \delta_{SI}^2)$ *and* $R_{Iij} \sim N(0, \delta_{RI}^2)$ *where* I_{jk} *is the inspector's measured value of item* k *in group* j, μ_{jk} *is the true value of item* k *from group* j, R_{Ijk} *is a random error of item* k *from group* j, S_{Ij} *is a short-term systematic error in group* j. *The notation* $S_{Ij} \sim N(0, \delta_{SI}^2)$ *means that values of* S_{Ij} *are assumed to have a normal distribution with mean* (0) *and variance* δ_{SI}^2. *This paper describes three main inspector UQ-related training topics. Topic one is analysis of variance to estimate the relative standard deviations (RSDs)* δ_{SI}^2 *and* δ_{RI} *(and the corresponding RSDs for the operator). Topic two is an example involving the uranium neutron coincidence collar (UNCL) to illustrate the need for inspector UQ training to include an understanding of the most important factors that impact the RSDs, which in turn also affect the rejection limits for comparing operator declarations to inspector measurements. For the UNCL method, it is important for inspectors to understand the fuel assembly design and IAEA neutron coincidence counting (INCC) software input requirements. Incorrect INCC declaration input is thought to be among the largest contributors to the observed UNCL uncertainty (as quantified by the RSDs). In response to needs arising from IAEA measurement evaluations, improved UQ methods have recently been developed, and the new methods described in topics one and two are beginning to be presented in training for safeguards inspectors, as will be described. Topic three is to use the estimated RSDs to evaluate material balances and to plan inspector sample sizes based on estimated material loss detection probabilities.*

Keywords: Nuclear Safeguards, Statistical Methodologies, Approximate Bayesian Computation (ABC), Data Analytics, Non-Destructive Assay (NDA)

1. Introduction

Inspector measurements are a cornerstone of IAEA safeguards, so it is important for inspectors to have a basic understanding of UQ. For example, suppose the operator's declared nuclear material (NM) mass for an item is 1.1 kg, and the inspector's measurement is 0.95 kg. Whether the 0.15 kg difference is a cause for concern depends on the uncertainty in the 1.1 kg and the 0.95 kg values. Effective UQ is critical in order to make quantitative safeguards conclusions based on inspector verification measurements.

This paper describes three UQ topics presented in training courses for safeguards inspectors. For background, the Guide to the Expression of Uncertainty in Measurement (GUM) provides guidance on the expression of measurement uncertainty [1]. UQ can be approached by comparing multiple measurements of the same item (top-down) or by assessing each step in the measurement procedure (bottom-up). The GUM briefly addresses top-down methods, but is most known for a bottom-up option using the measurement equation

$$Y = f(X_1, X_2, \ldots, X_p) \quad (1)$$

where Y is the estimate of the measurand, and $X_1, X_2, \ldots, X_p$ are inputs. The inputs can be measurement or adjustment factors, and can be regarded as having a joint probability distribution that can include covariances among the inputs. For example, some of the inputs can be estimated calibration parameters, others can be measured values, and others can be adjustment factors. The GUM applies uncertainty propagation to $X_1, X_2, \ldots, X_p$ and the function $f(\,)$ in Eq. (1) to estimate the uncertainty in the assay, defined as the standard deviation of Y. Informally, standard deviation quantifies measurement variability, defined as the square root of the variance, which is the average squared distance of the Y value from the mean of the

Y values. Safeguards metrology routinely partitions total variability into variability around the mean and variability around the true value into "random" and "systematic" components, respectively, and the GUM [1] combines random and systematic into a total variability as explained in inspector UQ training.

For verification purposes, paired (Operator (O), Inspector (I)) data are collected from inspections performed during site visits that occur once or a few times per year, and then for top-down UQ, several years of paired (O, I) data are analysed. An effective measurement error model must account for variation within and between groups, where a group is an inspection period. A typical top-down multiplicative error model used for the (I) (and similarly for O) is

$$I_{jk} = \mu_{jk}(1 + S_{Ij} + R_{Ijk}) \ , k = 1, \dots, n, j = 1, \dots, g, \quad (2)$$

where I_{jk} is the inspector's measured value of item k in group j, μ_{jk} is the true but unknown value of item k from group j, $R_{Ijk} \sim IIDN(0, \delta_{RI}^2)$ (IIDN is independently and identically distributed normal) is a random error of item k from group j, $S_{Ij} \sim IIDN(0, \delta_{SI}^2)$ is a short-term systematic error in group j [1-10]. Short-term systematic error remains constant for a short term when measurement conditions or settings, i.e., parameters of calibration curves, normalizations, and/or subtracted background etc. are not altered, but vary in a random way over the long term [2]. Long-term systematic error is sometimes also referred to as bias. In applications, the normality assumption is not usually critical in the error model depicted in Eq. (2); the important aspect of the modelling assumptions are that there are systematic and random components with RSDs δ_{SI} and δ_{RI}, respectively, that both bottom-up and top-down UQ aim to estimate [1-10]. Often, Top-down estimates $\hat{\delta}_{SI}$ and $\hat{\delta}_{RI}$ are larger than bottom-up estimates, and the gap is called "dark uncertainty" [2-8]. Note that in Eq. (2) the same number of measurements n per group is assumed for simplicity of presentation.

Figure 1 plots $n = 10$ simulated values of $(o - i)/o$ for each of $g = 5$ groups with parameters $\delta_{SO} = 0.001, \delta_{RO} = 0.001, \delta_{SI} = 0.027, \delta_{RI} = 0.027$ and $\delta_{\mu} = 0.003$. The within-group means are indicated by the horizontal within-group lines.

The need for quality control within UQ approaches provides motivation for excellent ongoing communication and collaboration among inspectors to reduce and better understand error variance components, which in turn provides partial validation that safeguards is properly implemented. Periodically, bottom-up and top-down estimates $\hat{\delta}_S$ and $\hat{\delta}_R$ should be compared (here the subscript I is dropped because the discussion also applies to operator measurements). It is not surprising that bottom-up UQ tends to lead to smaller estimates of $\hat{\delta}_S$ and $\hat{\delta}_R$ than does top-down, because small sources of variation are often neglected in bottom-up UQ. However, until the gap between top-down and bottom-up estimates is acceptably small, the fielded assay system is not fully understood. For example, Fig. 2 plots the estimated probability density of the bottom-up and top-down estimates $\hat{\delta}_T$ of the total RSD $\delta_T = \sqrt{\delta_S^2 + \delta_R^2}$ for the Uranium Neutron Coincidence Collar (UNCL) [5,11] measurement using approximate Bayesian Computation (ABC) [12-16]. The UNCL is the bottom-up UQ example in Section 3 that uses thermal neutrons to measure the ^{235}U content in fresh fuel assemblies. These measurements exhibit a gap between the bottom-up and top-down estimates of δ_T [5,11]. The vertical lines are the best point estimates, lying in the middle of the distribution, and the width of the distribution

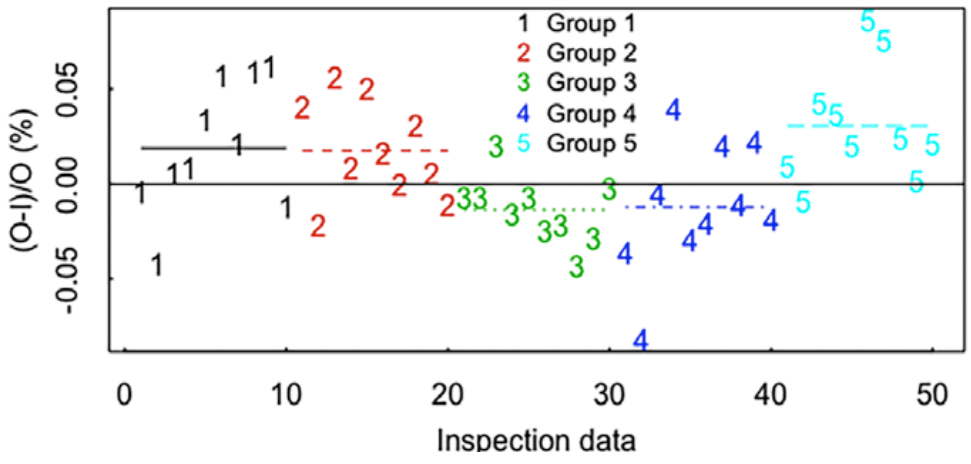

Figure 1: Ten simulated values of $(o - i)/o$ for each of g=5 groups.

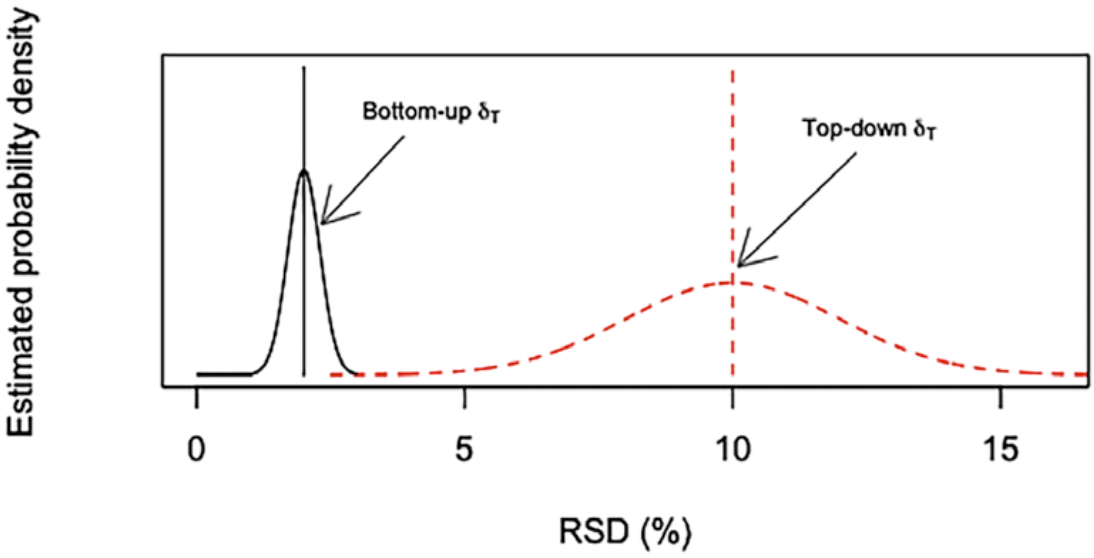

Figure 2: Uranium Neutron Collar example with a gap between top-down and bottom-up RSD estimates.

characterizes how well δ_T is estimated. Figure 2 indicates that the bottom-up estimate of δ_T is optimistically low compared to the top-down estimate because the two distributions have very little overlap (Section 3). Data sets having repeated measurements of the same item use top-down UQ to separately estimate variance arising from pure random effects from variance arising from item-specific effects (Section 2).

Inspectors need to understand both bottom-up and top-down UQ. Bottom-up UQ provides guidance regarding best measurement practice and protocol to understand and possibly reduce measurement uncertainty. This paper describes three main inspector UQ-related training topics. Topic one in Section 2 is analysis of variance to provide top-down estimate of the relative standard deviations (RSDs) δ_{SI} and δ_{RI} and (and the corresponding RSDs for the operator). Topic two in Section 3 is a bottom-up UQ example involving the UNCL to illustrate the need for inspector UQ training to include an understanding of the most important factors that impact the RSDs, which in turn also affect the rejection limits for comparing operator declarations to inspector measurements (Section 4.2). For the UNCL method, it is important for inspectors to understand the fuel assembly design and INCC (neutron coincidence counting software) input requirements. Incorrect INCC declaration input is thought to be among the largest contributors to the observed UNCL uncertainty (as quantified by the RSDs). Topic three in Section 4 is to use the estimated RSDs δ_S and δ_R for both the operator and inspector to evaluate material balances (MB) and to plan inspector sample sizes based on estimated material loss detection probabilities. Partitioning into random and systematic components has important implications for sample planning and MB evaluation.

2. Uncertainty Quantification (UQ) – an empirical (top-down) approach (ANOVA)

To be conservative (with regard to reducing false alarms), the IAEA's data evaluation group (nuclear fuel cycle information analysis) uses top-down-based UQ (rather than bottom-up UQ) to estimate the inspector's random and short-term systematic uncertainty components (RSDs) [2]. As explained in Section 4, the estimated RSDs are used in calculations to achieve target detection probabilities (DP) [17,18], and to perform error variance propagation in order to estimate the standard deviation of each material balance (MB) [1-12,19,20]. The published international target values (ITVs) for δ_R and δ_S are updated approximately every 10 years and the next updates are scheduled to be issued in early 2022 [2].

The basis of the top-down approach to UQ is an analysis of variance (ANOVA) with random effects based on operator-inspector relative differences. Such paired data arise when the operator and the inspector measure the same object once without measurement repetition. One goal is to estimate $\delta_R^2 = \delta_{RO}^2 + \delta_{RI}^2$ and $\delta_S^2 = \delta_{SO}^2 + \delta_{SI}^2$ for the relative differences. Another goal is to partition the total variance into four components, $\delta_{SO}^2, \delta_{RO}^2, \delta_{SI}^2$ and δ_{RI}^2.

Figure 1 illustrated an example in which paired operator (0) measurements (typically using Destructive Assay (DA)) and inspector (I) measurements (typically using Non-Destructive Assay (NDA)) from five previous inspection periods are used to estimate δ_R and δ_S for both the operator and inspector, and then to set alarm thresholds to detect possible data falsification in period six. The within-period variance is regarded as the random error variance, which includes the effects of item-specific bias. The between-period variance includes both random error variance (divided

by the number of measurements per period) [1-12] and short-term systematic effects such as instrument recalibration. Therefore, in Eq. (2), the errors R_{Ojk} and R_{Ijk} include "item-specific" bias because in verification data used for metrology, the measured items are not true replicates, that is, the relevant physical properties of the item being measured may vary randomly among items.

2.1 Estimating the variances of the relative differences $D = (o - i)/o$

For a balanced dataset with n paired differences in each of g groups $(N = ng)$ and under the assumption that no data are falsified by the operator, Eq. (2) yields for the relative differences $D_{jk}, j = 1, \dots, g, k = 1, \dots, n$,

$$D_{jk} = \frac{O_{jk} - I_{jk}}{O_{jk}} \approx \frac{O_{jk} - I_{jk}}{\mathbb{E}(O_{jk})} = S_j + R_j\,, \qquad (3)$$

where $S_j = S_{Oj} + S_{Ij}$ and $R_j = R_{Oj} + R_{Ij}$ and $\mathbb{E}(O_{jk})$ denotes the expected value of O_{jk}. Therefore, for the $n = 10$ sets of $g = 5$ groups values of $d = (o - i)/o$ as in Figure 1, standard ANOVA [21] can be applied to estimate δ_S^2 and δ_R^2. The validity of the approximation in Eq. (3) is shown in [10].

From standard ANOVA, it is well known that unbiased estimators of δ_R^2 and δ_S^2 are given by

$$\hat{\delta}_R^2 = \frac{1}{ng - g}\sum_{j=1}^{g}\sum_{k=1}^{n}(d_{jk} - \bar{\bar{d}})^2 \quad \text{and} \quad \hat{\delta}_S^2 = \frac{\sum_{j=1}^{g}(\bar{d}_j - \bar{\bar{d}})^2}{(g-1)} - \frac{\hat{\delta}_R^2}{n}, \qquad (4)$$

where $\bar{\bar{d}} = \sum_{j=1}^{g}\sum_{k=1}^{n} d_{jk}/(ng)$ is the overall unweighted average and $\bar{d}_j = \sum_{k=1}^{n} d_{jk}/n$ is the average measurement for item k. These formulas assume the same sample size (number of measurements is n) per group for simplicity of presentation here. Because the actual sample sizes n_i often vary across groups, weighted averages are actually used [21-24].

In standard one-way random effects ANOVA [10], if the relative error variances are not constant (which would mean that that the assumptions $R_{jk} \sim IIDN(0, \delta_R^2)$ and/or $S_j \sim IIDN(0, \delta_S^2)$ regarding the error variances in Eq. (2) are not correct), then it can be shown that $\sum_{j=1}^{g}\sum_{k=1}^{n}(d_{jk} - \bar{d}_j)^2/(n(g-1))$ is an unbiased estimate of the average relative variance $\sum_{j=1}^{g}\delta_{Rj}^2/g$ and $\hat{\delta}_S^2 = \sum_{j=1}^{g}(\bar{d}_j - \bar{\bar{d}})^2/(g-1) - \hat{\delta}_R^2/n$ is an unbiased estimate of the average relative variance $\sum_{j=1}^{g}\delta_{Sj}^2/g$ [21,22]. Note from Eq. (4) that it is possible that the estimate $\hat{\delta}_S^2 < 0$, in which case $\hat{\delta}_S^2$ is set to 0 (in a non-Bayesian framework, as presented here).

This same standard random-effects ANOVA just explained can also be applied to data sets for which there are repeated measurements on the same item in another common top-down approach to UQ [9]. In this case, the groups are not inspection periods, but are items, and the between-group variance is the variance of item-specific biases [9].

Typically, it is assumed that short-term systematic errors change across inspection periods from the groups used in the ANOVA. However, it may appear that the short-term systematic errors change at other times and thus the groups are unknown. The impact of unknown groups on the estimates of the variances of random and systematic errors in ANOVA is discussed in [24].

2.2 Grubbs estimator for paired (operator, inspector) data to estimate $\delta_{SO}, \delta_{RO}, \delta_{SI}$ and δ_{RI}

One-way ANOVA based on paired data allows us to estimate the measurement error variances of operators and inspectors. ANOVA requires the data to fall in groups, so that within-group and between-groups variances can be defined and estimated. In this example, the groups are the inspection periods. The basis of a Grubbs-based estimator [3-6,23,24] as applied to data assumed to be generated according to Eq. (2) in order to estimate $\delta_\mu^2, \delta_{SI}^2$ and δ_{RI}^2 ($\delta_{SO}^2, \delta_{RO}^2$ can be estimated accordingly) is that the covariance between operator and inspector measurements equals σ_μ^2, so can be estimated using (with $\bar{O}_j$ used to estimate the average true value μ)

$$\hat{\delta}_\mu^2 = \frac{1}{(n-1)g}\sum_{j=1}^{g}\frac{1}{\bar{O}_j{}^2}\sum_{k=1}^{n}(O_{jk} - \bar{O}_j)(I_{jk} - \bar{I}_j)\,. \qquad (5)$$

Note that the variance of the true values is estimated using the covariance between the operator and inspector measurements, and this provides a "teaching moment" in that the assumed error model in Eq. (2) implies zero covariance between operator and inspector measurements unless there is variability in the true values.

The measurement error model $I_{jk} = \mu_{jk}(1 + S_{Ij} + R_{Ijk})$ in Eq. (2) is the random variable μ_{jk} multiplied by the composite random variable $(1 + S_{Ij} + R_{Ijk})$. Therefore, a class exercise is to show that the variance of I_{jk} conditional on the value of S_{Ij} is given by the random variable $\mu^2(\delta_\mu^2\delta_{RI}^2 + \delta_\mu^2(1 + S_{Ij})^2 + \delta_{RI}^2)$, which has an expected value over inspection periods of

$\mu^2(\delta_\mu^2\delta_{RI}^2 + \delta_\mu^2(1+\delta_{SI}^2)+\delta_{RI}^2)$. Therefore, the expected between-group and within-group sums of squares involve both δ_{SI}^2 and δ_{RI}^2. Provided that δ_{SI}, δ_{RI}, and δ_μ are each less than approximately 0.15 (typically true for most safeguards measurements), the approximation $\mu^2(\delta_\mu^2\delta_{RI}^2 + \delta_\mu^2(1+\delta_{SI}^2)+\delta_{RI}^2) \approx \mu^2(\delta_\mu^2+\delta_{RI}^2)$ is adequate and will be used here. Then, the sample covariance between operator and inspector measurements can be subtracted from the sample variance of the inspector measurements to estimate δ_{RI}^2 (and similarly for estimating δ_{RO}^2. That is, within a single inspection period (group j), a reasonable estimate of δ_{RI}^2 is

$$\hat{\delta}_{RI,j}^2 = \frac{1}{\bar{O}_j^2(n-1)}\left(\sum_{k=1}^{n}(I_{jk}-\bar{I}_j)^2 - \sum_{k=1}^{n}(O_{jk}-\bar{O}_j)(I_{jk}-\bar{I}_j)\right).$$

The final estimate of the inspector's random error relative variance is then the average over groups,

$$\hat{\delta}_{RI}^2 = \frac{1}{g}\sum_{j=1}^{g}\hat{\delta}_{RI,j}^2. \quad (6)$$

The variance $\mathbb{V}(I_{jk})$ of I_{jk} is given by $\mathbb{V}(I_{jk}) = \mu^2(\delta_\mu^2(1+\delta_{SI}^2+\delta_{RI}^2)+\delta_{SI}^2+\delta_{RI}^2) \approx \mu^2(\delta_\mu^2+\delta_{RI}^2+\delta_{SI}^2)$, so the variance of the between group means (equal sample size with n observations per group) is $\sigma_{\text{Between},I}^2 \approx \mu^2(\delta_\mu^2/n + \delta_{RI}^2/n + \delta_{SI}^2)$, which yields by Eqs. (5) and (6)

$$\hat{\delta}_{SI}^2 = \frac{1}{\bar{O}^2}\frac{\sum_{j=1}^{g}(\bar{I}_j-\bar{I})^2}{g-1} - \frac{\hat{\delta}_\mu^2}{n} - \frac{\hat{\delta}_{RI}^2}{n}. \quad (7)$$

There is no guarantee that $\hat{\delta}_\mu^2, \hat{\delta}_{SI}^2$, or $\hat{\delta}_{RI}^2$ are non-negative, but the corresponding true quantities are non-negative $\delta_\mu^2 \geq 0, \delta_{SI}^2 \geq 0$ and $\delta_{RI}^2 \geq 0$, so constrained versions of the Grubbs' and ANOVA-based estimators are available; see [12-16] for Bayesian-type constraints and [23-24] for non-Bayesian constraints.

The original Grubbs' estimate [25] is for additive error models. The ABC framework [12-16] makes Grubbs' type estimation straightforward (constrained according to the prior distribution for each parameter) for either additive or multiplicative models. Using Eqs. (5), (6) and (7), the five summary statistics used in this application of ABC for n *(O,I)* pairs in each of g groups are $\hat{\delta}_\mu^2, \hat{\delta}_{RI}^2, \hat{\delta}_{RO}^2, \hat{\delta}_{SI}^2, \hat{\delta}_{SO}^2$.

As an example, Figure 3 plots the ABC-based posterior probability density function (pdf) for δ_{RI} using the ABC threshold $\varepsilon = 0.001, 0.02, 0.1$. The ABC-based [12-16] estimate of the pdf for δ_{RI} is computed as follows. First, ABC simulates synthetic data from Eq. (2) using many (105 in this example) candidate RSD values from a wide prior pdf, Second, ABC accepts all those candidate RSD values into the posterior pdf for which the corresponding five summary statistics above are close (within some small tolerance denoted ε) to those in the test data computed in this case as simulated data from Eq. (2). Regarding whether ABC is well calibrated, for $\varepsilon = 0.005, 0.01$ or 0.1, the predicted root mean squared error (RMSE) is 0.027 and the observed average standard deviation is 0.028, and the actual posterior probability coverages are 0.99, 0.96, and 0.91. The nominal interval coverages are 0.99, 0.95, and 0.90, so ABC is well calibrated. However, using $\varepsilon = 0.001$ leads in this example to a few unreasonably large accepted trial values of δ_{RI}, which shifts the mean upward from the true value of 0.027. Fortunately, the calibration check comparing nominal to actual coverages detects that $\varepsilon = 0.001$ leads to poor calibration, and so it is to be avoided. References [5,12,14-16] describe calibration checks for ABC in safeguards measurement applications, anticipating that the user will experiment to find an effective value of ε such that ABC is well calibrated. All analyses presented here are done in R [24].

Inspector training currently presents the original Grubbs' estimate [23] above in Eqs. (5)-(7) in a non-Bayesian framework. This paper emphasizes the need for all Safeguards professionals, including inspectors, to have a basic understanding of bottom-up and top-down UQ. The fact that dark uncertainty typically exists is most easily illustrated using a plot such as Fig. 2, where uncertainty in the bottom-up and top-down RSD estimates is also provided so that apparent gaps can be assessed for statistical significance. This section used simulated data from Eq. (2) to illustrate ABC as an effective option to provide a pdf for the top-down estimate of δ_{RI}. A Bayesian framework [3,4,54,7,8,12] naturally provides the uncertainty in the RSD estimates and is beginning to be presented as an option in training courses. Note that the non-negativity constraint on true RSDs forces truncation of negative estimates whether a Bayesian or non-Bayesian approach is used. Note also that while top-down UQ might lead to more realistic RSD estimates, there can be large uncertainty (wide posterior pdf) in top-down RSD estimates. Also, bottom-up RSD estimates are continually being improved, which leads to better understanding of the measurement process. It should be emphasized that the IAEA has unique opportunities to assess the quality of top-down and bottom-up UQ because inspector measurements are typically made using NDA while operator measurements are typically made using DA. As a consequence, dark uncertainty such as item-specific bias that is different for NDA than for DA of the same item is exposed when comparing operator declarations to inspector verification measurements of the same items. Therefore, topics one and two describe recent improvements to UQ methods, and the new methods are beginning to be presented in inspector training, which is

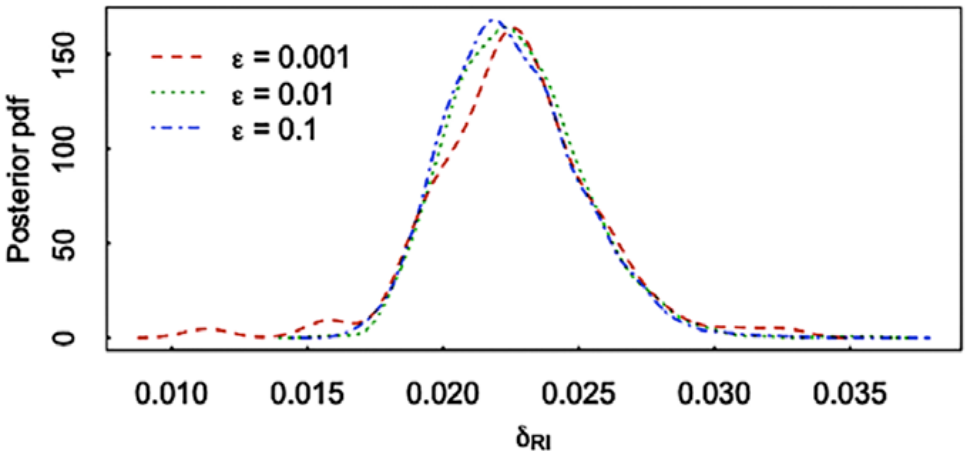

Figure 3: The posterior pdf for Grubbs'-type estimation of δ_{RI}. The posterior mean is 0.023 and the true value of δ_{RI} is 0.027 as was used to simulate the data in Fig. 1.

helping to overcome communication challenges involving the expression and quantification of uncertainty.

One of the main data sources used to estimate the IAEA's ITVs, is paired operator-inspector data such as that in Fig. 1 and analysed here without repeated measurements of items [2-5,21]. Recall that item-specific bias is often evident from applying Grubbs' estimation to paired data such as that used to generate the $d = (o - i)/o$ values in Fig. 1. The evidence for item-specific bias is that the estimated δ_{RI} is larger than predicted from bottom-up UQ (on the basis of non-overlapping pdfs such as in Fig. 2, but for δ_{RI}). As a result of such evidence, bottom-up UQ has only recently begun to consider sources of item-specific bias, such as departures from calibration items [2-5] or modelling assumptions [5,11,12]. Item-specific bias is random across items, so the effective random error RSD is

$$\delta_{R_{\text{effective}}} = \sqrt{\delta^2_{R_{repeat}} + \delta^2_{item-specific\ bias}} \quad .$$

The reported ITV values include item-specific bias effects if they are based on such paired data, and so δ_{RI} is actually $\delta_{R_{\text{effective}}}$.

3. A bottom-up UQ example: Uranium Neutron Coincidence Collar (UNCL)

The UNCL uses an active neutron source to induce fission in ^{235}U in fresh fuel assemblies [3,5,8]. 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 coincidence rate Y, which can be used to estimate the linear density of ^{235}U in a fuel assembly (grams ^{235}U per cm) using calibration parameters a_1 and a_2. The 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

$$X = \frac{kY}{a_1 - a_2 kY}, \tag{8}$$

where a_1 and a_2 are to be estimated, and $k = k_1k_2k_3k_4k_5$ is a product of correction factors that adjust Y to item-, detector-, and source-specific conditions in the calibration [5,11]. Therefore, Eq. (8) is a special case of GUM's Eq. (1), but with X and Y reversed here compared to that used by the GUM in Eq. (1) because conventionally in calibration, X is the measurand and Y is the measurement data, such as the neutron count rate. In Eq. (8), the net doubles rate Y, the two calibration parameters a_1 and a_2, and the correction factors in $k = k_1k_2k_3k_4k_5$ are among the X 's in Eq. (1).

Reference [8] showed that calibration is most effective (leading to smallest RMSE in the estimate of X, denoted $\hat{X}$) if there is no adjustment for errors in the predictor kY, and that errors $k_1k_2k_3k_4k_5$ should be included in synthetic calibration data. Note that by working with $1/X$ and $1/Y$, one can convert Eq. (8) to one that is linear in the transformed predictor $1/Y$.

Several recent UNCL measurements have exhibited a gap between the bottom-up and top-down total RSD estimate, $\delta_T = \sqrt{\delta_S^2 + \delta_R^2}$ [3,5,8]. Recall that Figure 2 is an example of the estimated pdf for δ_T using approximate Bayesian computation (ABC, see Section 2) for both the top-down

and bottom-up RSD estimates [3-5,7]. The bottom-up estimate shown in Fig. 2 was presented in [3,8] using ABC applied to simulated data from Eq. (8), where X is the ^{235}U content, and the item-specific adjustment factor $k = k_1k_2k_3k_4k_5$ is a product of correction factors that adjust the measured neutron doubles rate Y to item-, detector-, and source-specific conditions in the calibration. The correction factors are currently being examined more closely using modelling [3,5,8,11] and [5] indicates the need for better inspector training on measurement protocol, including the importance for inspectors to understand the fuel assembly design and the INCC software input because these requirements impact some of the factors in $k = k_1k_2k_3k_4k_5$. Incorrect INCC declaration input is thought to [5] be among the largest contributors to the observed UNCL uncertainty (as quantified by the RSDs), and consistent source positioning is also a non-negligible contributor to item-specific effects on $k = k_1k_2k_3k_4k_5$ [3,5,8,11].

Another potential contributor to the observed UNCL uncertainty (and more generally the uncertainty on any NDA measurement) is the purely random uncertainty associated with the measurement. A large factor that drives random error variance is measurement time. Given that the inspector determines UNCL measurement duration, an important aspect in training inspectors on UQ is to promote an understanding of how uncertainty quoted by INCC software (or any analysis software used for NDA techniques) relates to the total RSD δ_T that is calculated based on a top-down analysis using operator-inspector paired data. In the case of the UNCL, the uncertainties quoted by INCC are based on a bottom-up approach that takes into account estimated uncertainties in the calibration constants, uncertainties that are known for the various correction factors, as well as the random uncertainties associated with the doubles coincidence counting rate. To ensure that the quality of the current measurement is consistent with the historical δ_T that was determined for this particular set of UNCL measurements, the inspector is encouraged to compare the quoted uncertainty from the INCC code to the top-down historical δ_T for this measurement. If the bottom-up uncertainty estimation produced by the analysis software is similar in size to, or smaller than, the historical RSD, the quality of the current measurement is presumed to be consistent with the historical data used in modelling the top-down historical RSD for this measurement technique. On the other hand, if the bottom-up uncertainty estimation is significantly larger than the top-down historical RSD, it is presumed that the random uncertainty of the current measurement is unacceptably high which requires an increase in the measurement time to reduce the bottom-up uncertainty estimation. Training inspectors how to perform a quality control check on each measurement using such a simple guideline helps reduce the chances that the historical δ_T will become large over time due to the inclusion of poor-quality data.

4. Two main applications for RSD estimates of $\delta_{SI}, \delta_{RI}, \delta_{SO}$ and δ_{RO}

Within safeguards, there are two main applications for the estimated values of $\delta_{SI}, \delta_{RI}, \delta_{SO}$ and δ_{RO}. The first application is material balance evaluation where $\delta_{SI}, \delta_{RI}, \delta_{SO}$ and δ_{RO} are used in variance propagation to estimate the standard deviation of the material balance. The second application is designing inspection plans to have a desired detection probability.

4.1 Material balance evaluation

The MB sequence is fundamental to material accounting [19,20]. For example, in a sequence of 12 monthly MBs over a one-year analysis period, a key task is to classify the period as having no loss or having non-zero loss. Nuclear material accountancy (NMA) at a facility that processes nuclear material requires measuring facility input transfers T_{in}, output transfers T_{out}, and inventory I to compute a material balance defined for balance period j as $MB_j = (I_{j-1} + T_{in,j} - T_{out,j}) - I_j$, which equals "book inventory" minus "physical inventory," where $(I_{j-1} + T_{in,j} - T_{out,j})$ is the book inventory.

Typically, many measurements are combined to estimate the terms T_{in}, T_{out}, I_{begin} and I_{end} in the MB; therefore, the central limit effect and years of experience suggests that MBs will be approximately normally distributed with mean equal to the true NM loss μ and standard deviation σ_{MB}, which is expressed as $MB \sim N(\mu, \sigma_{\text{MB}})$, where X denotes the MB [19,20]. Therefore, a sequence of n MBs are assumed to have approximately a multivariate normal distribution, $MB_1, MB_2, \ldots, MB_n \sim \text{MVN}(\mu, \Sigma)$, where the $n \times n$ covariance matrix is

$$\Sigma = \begin{pmatrix} \sigma_{11}^2 & \sigma_{12}^2 & \cdots & \sigma_{1n}^2 \\ \sigma_{21}^2 & \sigma_{22}^2 & \cdots & \sigma_{2n}^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1}^2 & \sigma_{n2}^2 & \cdots & \sigma_{nn}^2 \end{pmatrix},$$

with variances on the diagonal and covariances on the off-diagonals.

One common goal is for the loss DP to be at least 0.95 if $\mu \geq 1$ SQ (significant quantity, which is 8 kg Pu or Uranium-233 or 25 kg of uranium-235 in HEU), which is accomplished if $\sigma_{MB} \leq SQ/3.61$ (Figure 4). The factor 3.61=1.96+1.65 where the alarm rule $|MB| < 1.96\,\sigma_{MB}$ is used in two-sided testing (testing for either loss or gain of NM) for approximately a 0.05 false alarm probability. Therefore, if a loss of $1.96\,\sigma_{MB}$ occurs, then the DP is only 0.50. But if a loss of (1.96+1.65) σ_{MB}, then the DP is 0.95

actual irradiation (case 2) are the same when the uncertainties in the fits are taken into account, for all fuel assemblies except fuel 19. For fuel 19, the three RADs do not overlap below 15 μs, which results in a slightly different slope of the exponential fit and therefore also in τ. The τ values for case 9 appear, when compared to those in the default case (case 1), to be biased towards lower values for 18 of the 20 fuel assemblies. For ten fuel assemblies, the τ values for case 9 are in fact lower than for those of case 1, and in ten cases they are the same when the uncertainties in the fits are taken into account. This means that for 50% of the fuel assemblies, changing the fuel geometry and irradiation conditions results in a τ value which is lower when compared to the default case, with an average difference of 1.36 μs.

4.2.3 Relating effects in RAD and τ

Figure 4 plots τ versus the RAD amplitudes to visualize their interdependence. It shows τ versus the values of the RAD amplitudes at 5 μs, the data point closest in time to when the τ fit begins (at 4 μs), for all 20 fuel assemblies.

Comparing the RAD amplitudes for the two irradiation history cases (case 1 and case 2), shows that the largest impact on the RAD amplitude is found for one specific fuel assembly (fuel 16) and that minor discrepancies can be found for a small number of fuel assemblies. It can also be seen that for a large fraction of the fuel assemblies, mainly those with a relatively high RAD amplitude, the difference in RAD amplitude for the default case and the realistic case is noticeable. It can also be seen in Figure 4 that the RAD amplitudes in case 9 are considerably lower than for case 1 (the default case).

With respect to τ, Figure 4 shows that although the RAD amplitude does not change much with irradiation history, τ is seen to vary in all directions (remain the same, decrease and increase). Inspecting the irradiation histories of the fuel assemblies for which τ increases the most (fuels 2, 15, 5) or decreases the most (fuels 3, 19, 16, 1), provides no clear explanation for how τ changes. Both groups contain fuel assemblies with similar values of initial enrichment, burnup and cooling time. With respect to irradiation history and irradiation conditions, there is also no clear difference between the fuel assemblies in the two groups. Both of them contain fuel assemblies with unusually long and unusually short irradiation periods, low as well as high neutron fluxes, and fuel assemblies that have spent long periods of time outside the reactor before being reinserted. Only one of these seven fuel assemblies (fuel 19) has experienced an irradiation history similar to that in the default case. However, considering statistical uncertainties of Monte Carlo simulations, it is not surprising that the τ values fall outside the range of uncertainty from the fits in about a third of the cases. We have also made ten repeated simulations of the same case and seen that an average of the τ predictions are indeed normally distributed around the mean.

When comparing the RADs and the τ values for the default case (case 1) and the most realistic case (case 9), it is

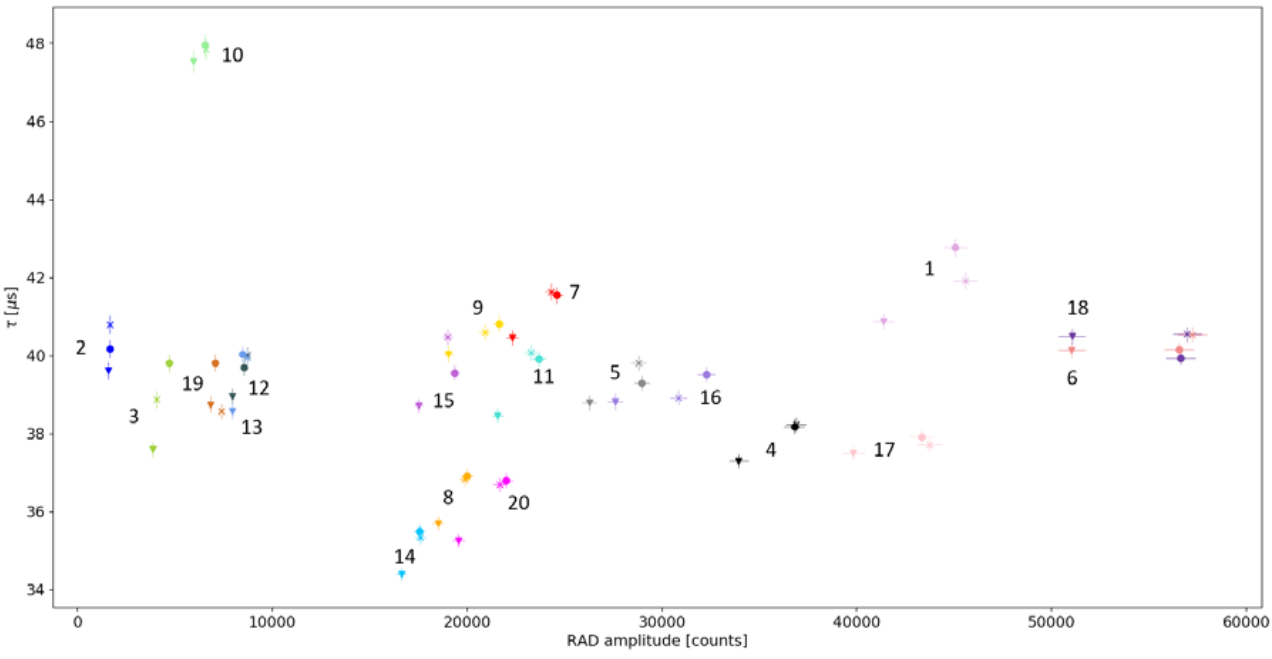

Figure 4: TValues of τ versus the RAD amplitude at 5 μs for all 20 fuel assemblies. The different markers denote the different cases: circles=case 1, x=case 2, and triangles=case 9. The different colours correspond to different fuel assemblies and the labels show the fuel ID.

(Figure 4). If $\sigma_{MB} > SQ/3.61$, this can be mitigated either by reducing the typical magnitude of measurement errors to achieve $\sigma_{MB} > SQ/3.61$ (if feasible), and/or by closing the balances more frequently so there is less nuclear material transferred per balance period, which reduces σ_{MB} .

In order to address large throughput facilities Near Real Time Accountancy (NRTA) was introduced in the mid 1990's. NRTA is frequent material balance closure, such as one per 10 to 30 days instead of only annually in conjunction with the annual physical inventory taking at the time of plant shutdown. As explained in inspector UQ training, large throughput facilities cannot typically achieve DP ≥ 0.95 for a loss of $\mu \geq 1$ SQ over a long time period such as one year. And, NRTA is not a panacea, because, as shown in [20], if a facility slowly diverts NM over, for example, one year, then a single yearly statistical test based on the annual cumulative material balance (known as CUMUF), $\sum_{i=1}^{n} X_i$ has larger DP than frequent statistical testing during the year. Of course, if the facility diverts NM abruptly, such as over one day, then NRTA will have much larger DP than a single annual statistical test. It is therefore generally accepted that NRTA is a valuable safeguards measure, despite leading to slightly smaller DP than in using annual MBs for protracted loss detection. Most safeguards studies consider a yearly analysis period, corresponding to the time of the annual scheduled physical inventory. But, if the facility diverts material, for example, $SQ/2$ in year one and $SQ/2$ in year two, then the DP is lowered compared to diverting one SQ in the analysis year. See Figure 5; however, the required diversion time would be longer than one calendar year, in this figure, lasting from period 7 to 18.

Grubbs' estimation [23] (or ABC based on Grubbs') produces the parameter estimates $\hat{\delta}_{SO}^2$, $\hat{\delta}_{RO}^2$, $\hat{\delta}_{SI}^2$ and $\hat{\delta}_{RI}^2$, which, as just explained, are needed for MB evaluation. Note that verifications rely on relative difference $(o-i)/o$ and δ_R^2 and δ_S^2 as given by Eq. (4) are the only parameters required for verification checks as described in Section 4.2. However, because MB evaluation is conducted separately for the operator, the estimates $\hat{\delta}_{SO}^2$ and $\hat{\delta}_{RO}^2$ are required.

The law of error propagation is described in the GUM [1] in the context of bottom-up UQ. The original law of error propagation by Gauss was designed for random errors only. Gauss realized after his publication that this was not always adequate. Therefore, the law of error propagation was modified to allow for measurement values to be correlated. The mode of error propagation for correlated values is a minor extension from purely independent (random) values. Specifically, formula (E.3) of JCGM 100:2008 [1] illustrates error propagation applied to the measurand equation $Y = f(X_1, X_2, \ldots, X_N)$ using the approximate result (based on a linear Taylor series approximation)

$$\sigma_Y^2 \approx \sum_{i=1}^{N} \left(\frac{\partial f}{\partial x_i}\right)^2 \sigma_i^2 + 2 \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \frac{\partial f}{\partial x_i} \frac{\partial f}{\partial x_j} \sigma_i \sigma_j \rho_{i,j} \,, \quad (9)$$

where σ_i^2 is the variance of X_i, $\rho_{ij} = v(X_i, X_j)/(\sigma_i \sigma_j)$ is the correlation coefficient of X_i and X_j, and $v(X_i, X_j)$ is the covariance of X_i and X_j. The first term on the right side of

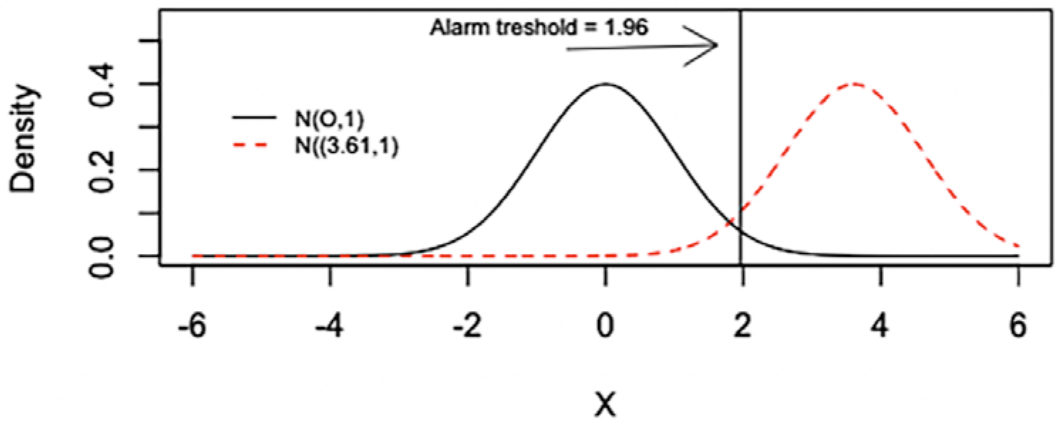

Figure 4: A $N(0,1)$ and a $N(3.61,1)$ distribution with threshold 1.96 for 2-sided testing.

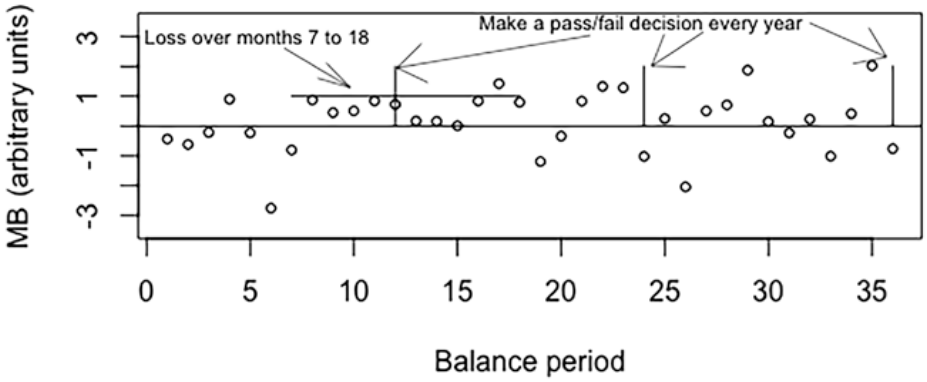

Figure 5: TMB sequences over 36 months using fixed-period (annual) decision periods.

Eq. (9) is the original law of Gauss for independent errors. The second term allows for correlated errors.

It is instructive in UQ training to illustrate how the MBE assumptions implement propagation of variance. Suppose that the total material mass declared by the operator is simply $Y = X_1 + X_2 + \cdots + X_N$ where X_i is the mass of item i. Assuming this model, then the variance of Y denoted by σ_Y^2 is given by applying Eq. (9) and accounting for the fact that the random and systematic error estimates propagate differently. Note that $\frac{\partial f}{\partial x_i} = 1$ for $f(X_1, X_2, \ldots, X_N) = X_1 + X_2 + \cdots + X_N$, and also note that the variance of an individual item is assumed to be the same for all items, that is $\sigma_i^2 = \sigma_X^2$ for all i, and the correlations $\rho_{i,j}$ are also assumed constant for each i, j. It then follows that Eq. (9) is an exact expression, and

$$\begin{aligned} \sigma_Y^2 &= \sum_{i=1}^{N} \left(\frac{\partial f}{\partial x_i}\right)^2 \sigma_i^2 + 2 \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \frac{\partial f}{\partial x_i} \frac{\partial f}{\partial x_j} \sigma_i \sigma_j \rho_{i,j} \\ &= N(\sigma_S^2 + \sigma_R^2) + N(N-1)\sigma_S^2 = N^2 \left(\sigma_S^2 + \frac{\sigma_R^2}{N}\right), \end{aligned} \tag{10}$$

because in nuclear safeguards the total error variance for measurement of one item is assumed as in Eq. (10) to be $\sigma_X^2 = \sigma_R^2 + \sigma_S^2$ and $\sigma_S^2 = \mathrm{cov}(X_i, X_j) = \rho \sigma_X^2$ is the variance of the short-term systematic measurement errors. Note that Eq. (10) is assuming negligible variation in the true values, so assuming $\mu_1 = \mu_2 = \cdots = \mu_N = \mu$, the same result is obtained by using Eq. (2),

$$Y = \sum_{i=1}^{N} \mu_i (1 + S + R_i) = \sum_{i=1}^{N} \mu_i + N\mu S + \mu \sum_{i=1}^{N} R_i,$$

which has variance $N^2(\sigma_S^2 + \sigma_R^2/N)$. Similarly, the absolute variance of $I_{11} + I_{12}$ is easily shown in a course exercise reviewing variance propagation to be $\sigma_{I_{11}+I_{12}}^2 = (\mu_{11}^2 + \mu_{11}^2)\delta_{RI}^2 + (\mu_{11} + \mu_{11})^2 \delta_{SI}^2$, which equals $2^2(\sigma_{SI}^2 + \sigma_R^2/2)$ for the case $N = 2$.

4.2 Sample size calculations for verification tests

Data such as the paired (O, I) data in Figure 1 are collected for verifying operator declarations. Recall the example from Section 1: suppose the operator's declared NM mass for an item is 1.1 kg, and the inspector's verification measurement is 0.95 kg. Whether the 0.15 kg difference is a cause for concern depends on the uncertainty (as quantified using RSDs as explained in Section 2) in the 1.1 kg and the 0.95 kg values.

Section 2.2 described how Grubbs' estimation can be applied to such (O, I) data to estimate the four RSDs $\delta_{SI}, \delta_{RI}, \delta_{SO}, \delta_{RO}$. Section 2.1 applied standard ANOVA to estimate the total RSD $\delta_T = \sqrt{\delta_{SO}^2 + \delta_{RO}^2 + \delta_{SI}^2 + \delta_{RI}^2}$ of $d = (o - i)/o$. The IAEA has historically used zero-defect sampling, which means that the only acceptable (passing) sample is one for which no defects are found according to the pass/fail test $|d| < 3\delta_T$ [17,18]. Because here we consider the case that the operator overstates the amount of NM present in a defective item by a certain amount (see below), the pass/fail test is one-sided: An alarm is raised if and only if $d > 3\delta_T$.

The non-detection probability β is the probability that no defects are found in a sample of size n when r (r is one or more) true defective items are in the population of size N. For one-item-at-a-time testing and under the assumption that only one measurement method is applied (extensions to more than one method see [18]), the non-detection probability β is given by

$$\beta = \text{Prob(discover 0 defects in a sample of size } n\text{)} = \sum_{i=Max(0,n-(N-r))}^{Min(n,r)} (A_i \times B_i) \quad (11)$$

where the selection probability term A_i is the probability $\mathbb{P}(N,r,n,i)$ that the selected sample contains i truly defective items, which is given by the hypergeometric distribution, i.e.,

$$\mathbb{P}(N,r,n,i) = \frac{\binom{r}{i}\binom{N-r}{n-i}}{\binom{N}{n}}.$$

The non-identification probability, B_i, is the probability that none of the i truly defective items is inferred to be defective based on the individual tests whether $d < 3\delta_T$. The value of B_i depends on the metrology, the defect size (defined as the amount by which the declared item nuclear material mass differs from its best accountancy value), and the alarm threshold (which is typically $3\delta_T$). Assuming a purely random error model, i.e., a multiplicative error model as in Eq. (2) for the inspector measurement (and similarly for the operator) with no systematic errors, and the case that the operator overstates the material present by the amount $M/(\bar{x}r)$ in each defective item, the non-identification probability [17,18] is

$$B_i = \left(\Phi\left(\frac{3\delta_T - \frac{M}{\bar{x}r}}{(1-\frac{M}{\bar{x}r})\,\delta_T} \right) \right)^i,$$

where δ_T is the total RSD (random plus systematic) for the one measurement method, M is the diverted amount of NM (usually M is the significant quantity), $\bar{x}$ is the average amount of NM per item, and $\Phi()$ is the cumulative normal distribution function.

Using Eq. (11), the non-detection probability β is given by

$$\beta(N,n,r,\delta,M,\bar{x}) = \sum_{=Max(0,n-(N-r))}^{Min(r,n)} \frac{\binom{r}{i}\binom{N-r}{n-i}}{\binom{N}{n}} \left(\Phi\left(\frac{3\delta_T - \frac{M}{\bar{x}r}}{(1-\frac{M}{\bar{x}r})\,\delta_T} \right) \right)^i. \quad (12)$$

For simplicity, sample size calculations currently regard all errors as random, but the total RSD $\delta_T = \sqrt{\delta_S^2 + \delta_R^2}$ includes the effects of both systematic and random errors [17,18]. From Section 2, it is evident that δ_T must be estimated using $\hat{\delta}_T$, which has estimation error; therefore because the DP is estimated by substituting $\hat{\delta}_T$ in Eq. (12) for δ_T, the calculated DP also has estimation error and confidence bands can be constructed around plots of DP versus sample size or plots of DP versus the number of defective items for a chosen sample size n. Also, reference [10] uses concepts from tolerance interval construction to show how to control the false alarm rate in future measurements when δ_R and δ_S are estimated from a few previous inspection periods as in Figure 1.

It is to be emphasized that Eq. (11) is quite general; it allows for the non-identification probability B_i to be a user-defined probability density function, such as the familiar normal density, or any other specified density that is suggested by measurement evaluations. Most commonly, B_i as given above is assumed. Also, the requested sample size n is then based on the minimum detection probability (maximum non-detection probability) over a range of possible r values (r is the true number of defective items in the population of size N), assuming each defective item has the same defect size.

5. Summary

IAEA safeguards training courses serve many types of students, including inspectors, who need to understand bottom-up and top-down UQ. Bottom-up UQ is primarily presented in NDA training. Top-down UQ is presented by the statistical analysis team and includes Grubbs' estimation as described in Section 3. Grubbs' estimation training includes related topics such as screening for outliers, choosing the appropriate groups if by inspection period is not the appropriate grouping [24], alternatives to Grubbs' estimation to reduce variability in RSD estimates, and subsampling to make more homogeneous strata if item masses have large variability (which increases the variability in the Grubbs' estimates). Section 4 briefly described two main applications for the estimates of the four main RSDs, $\delta_{SI}, \delta_{RI}, \delta_{SO}$, and δ_{RO}.

As UQ utilizes a holistic means to assess and estimate total uncertainty, performing a proper UQ can be challenging and tedious. Effective bottom-up and top-down UQ help to further reveal the so called "dark uncertainty" which exists in-between. Without this complete assessment, all sources of uncertainty cannot be properly identified and accounted for. The aim for effective UQ is part of ongoing quality control for measurements, and collaborations regarding any gaps between bottom-up and top-down RSD estimates (with attention to uncertainty in the RSD estimates – see Fig. 2) which can lead to fruitful communication among nuclear safeguards professionals.

6. References

[1] Evaluation of measurement data — Guide to the expression of uncertainty in measurement, JCGM 100:2008

GUM 1995 with minor, https://www.bipm.org/utils/common/documents/jcgm/JCGM_100_2008_E.pdf, 2008.

[2] Zhao, K., et al., International Target Values 2010 for Measurement Uncertainties for Safeguarding Nuclear Safeguards, STR 368, IAEA, 2010.

[3] Burr, T., Croft, S., Jarman, K., Nicholson, A., Norman, C., Walsh, S., Improved Uncertainty Quantification in NonDestructive Assay for Nonproliferation, Chemometrics, 159, 164-173, 2016.

[4] Bonner, E., Burr, T., Krieger, T., Martin, K., Norman, C., Comprehensive Uncertainty Quantification in Nuclear Safeguards, Science and Technology of Nuclear Installations, 1-16, 10.1155/2017/2679243, 2017.

[5] Bonner, E., Burr, T., Guzzardo, T., Krieger, T., Norman, C., Zhao, K., Beddingfield, D., Geist, W., Laughter, M., Lee, T., Ensuring the Effectiveness of Safeguards through Comprehensive Uncertainty Quantification, Journal of Nuclear Materials Management 44(2), 53-61, 2016.

[6] Walsh, S., Burr, T., Martin, K., Discussion of the IAEA Error Approach to Producing Variance Estimates for use in Material Balance Evaluation and the International Target Values, and Comparison to Metrological Definitions of Precision, Journal of Nuclear Materials Management, 45(2), 4-14, 2017.

[7] Burr, T., Croft, S., Favalli, A, Krieger, T., Weaver, B., Bottom-up and Top-Down Uncertainty Quantification for Measurements, Chemometrics and Intelligent Laboratory Systems 209, 104224, 2021.

[8] Burr, T., Croft, S., Krieger, T., Martin, K., Norman, C. Walsh, S., Uncertainty Quantification for Radiation Measurements: Bottom-Up Error Variance Estimation Using Calibration Information, Applied Radiation and Isotopes, 108,49-57, 2015.

[9] Burr, T., Sampson, T., Vo, D., Statistical Evaluation of FRAM Gamma-ray Isotopic Analysis Data, Applied Radiation and Isotopes 62, 931-940, 2005.

[10] Burr, T., Bonner, E., Krzysztoszek, K., Norman, C., Setting Alarm Thresholds in Measurements with Systematic and Random Errors, Stats 2(2), 259-271, doi.org/10.3390/stats2020020, 2019.

[11] Croft, S., Burr, T., Favalli, A., Nicholson, A., 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 A 811. 70-75, 2016.

[12] Burr, T., Krieger, T., Norman, C., Approximate Bayesian Computation applied to Metrology for Nuclear Safeguards ESARDA Bulletin No. 57, 52-59, 2018.

[13] Carlin, B., John, B., Stern, H., Rubin, D., Bayesian Data Analysis (1st ed), Chapman and Hall, 1995.

[14] Burr, T., Skurikhin, A, Selecting Summary Statistics in Approximate Bayesian Computation for Calibrating Stochastic Models, BioMed Research International, Vol20 2013, Article ID 210646, 10 pages, 2013. doi:10.1155/2013/210646, 2013.

[15] Blum, M., Nunes, M., Prangle, D., Sisson, S., A Comparative Review of Dimension Reduction Methods in Approximate Bayesian Computation, Statistical Science 28(2), 189–208, 2013.

[16] Nunes, M. Prangle, D., abctools: an R package for Tuning Approximate Bayesian Computation Analyses, The R Journal 7(2), 189–205, 2016.

[17] Krieger, T., Burr, T., Avenhaus, R., Statistical tests for verification sampling plans, Proceedings 61st Annual INMM meeting, 2019.

[18] Burr, T., Krieger, T., Norman, C., Zhao, K, The Impact of Metrology Study Sample Size on Verification Sample Size Calculations in IAEA Safeguards, European Journal Nuclear Science and Technology, 2, 36, 2016.

[19] Picard, R., Sequential Analysis of Material Balances, Journal of Nuclear Materials Management 15(2), 38-42, 1987.

[20] Avenhaus, R., Jaech, J., On Subdividing Material Balances in Time and/or Space, Journal of Nuclear Materials Management, 10, 24-34, 1981.

[21] Miller, R., Beyond ANOVA: Basics of Applied Statistics, Chapman & Hall, 1998.

[22] Vardeman, S., Wendelberger, J., The Expected Sample Variance of Uncorrelated Random Variables with a Common Mean and Some Applications in Unbalanced Random Effects Models, Journal of Statistics Education 13, 1, 2005, available online at lib.dr.iastate.edu/cgi/viewcontent.cgi?article=1064&context=stat_las_pubs.

[23] Martin, K., Böckenhoff, A., Analysis of variance of paired data without repetition of measurement, Allgemeines Statistisches Archiv, volume 90, pages 365–384, 2006.

[24] Burr, T., Martin, K., Norman, C., Zhao, K., Analysis of Variance for Item Differences in Verification Data with Unknown Groups, Hindawi, Science and Technology of Nuclear Installations, Volume 2019, Article ID 1769149, 10 pages, https://doi.org/10.1155/2019/1769149, 2019.

[25] Grubbs, F., On Estimating Precision of Measuring Instruments and Product Variablity, Journal of the American Statistical Association, 43, 243-264, 1948.

[26] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, www.R-project.org, 2017.