EPJ Nuclear Sci. Technol. **2**, 36 (2016)

DOI: 10.1051/epjn/2016026

EPJ Nuclear Sciences & Technologies

Available online at:
http://www.epj-n.org

REGULAR ARTICLE OPEN ACCESS

The impact of metrology study sample size on uncertainty in IAEA safeguards calculations

Tom Burr*, Thomas Krieger, Claude Norman, and Ke Zhao

SGIM/Nuclear Fuel Cycle Information Analysis, International Atomic Energy Agency, Vienna International Centre, PO Box 100, 1400 Vienna, Austria

Received: 4 January 2016 / Accepted: 23 June 2016

Abstract. Quantitative conclusions by the International Atomic Energy Agency (IAEA) regarding States' nuclear material inventories and flows are provided in the form of material balance evaluations (MBEs). MBEs use facility estimates of the material unaccounted for together with verification data to monitor for possible nuclear material diversion. Verification data consist of paired measurements (usually operators' declarations and inspectors' verification results) that are analysed one-item-at-a-time to detect significant differences. Also, to check for patterns, an overall difference of the operator-inspector values using a "D (difference) statistic" is used. The estimated DP and false alarm probability (FAP) depend on the assumed measurement error model and its random and systematic error variances, which are estimated using data from previous inspections (which are used for metrology studies to characterize measurement error variance components). Therefore, the sample sizes in both the previous and current inspections will impact the estimated DP and FAP, as is illustrated by simulated numerical examples. The examples include application of a new expression for the variance of the D statistic assuming the measurement error model is multiplicative and new application of both random and systematic error variances in one-item-at-a-time testing.

1 Introduction, background, and implications

Nuclear material accounting (NMA) is a component of nuclear safeguards, which are designed to deter and detect illicit diversion of nuclear material (NM) from the peaceful fuel cycle for weapons purposes. NMA consists of periodically comparing measured NM inputs to measured NM outputs, and adjusting for measured changes in inventory. Avenhaus and Canty [1] describe quantitative diversion detection options for NMA data, which can be regarded as time series of residuals. For example, NMA at large throughput facilities closes the material balance (MB) approximately every 10 to 30 days around an entire material balance area, which typically consists of multiple process stages [2,3].

The MB is defined as $\mathrm{MB} = I_{\mathrm{begin}} + T_{\mathrm{in}} - T_{\mathrm{out}} - I_{\mathrm{end}}$, where T_{in} is transfers in, T_{out} is transfers out, I_{begin} is beginning inventory, and I_{end} is ending inventory. The measurement error standard deviation of the MB is denoted σ_{MB}. Because many measurements enter the MB calculation, the central limit theorem, and facility experience imply that MB sequences should be approximately Gaussian.

To monitor for possible data falsification by the operator that could mask diversion, paired (operator, inspector) verification measurements are assessed by using one-item-at-a-time testing to detect significant differences, and also by using an overall difference of the operator-inspector values (the "D (difference) statistic") to detect overall trends. These paired data are declarations usually based on measurements by the operator, often using DA, and measurements by the inspector, often using NDA. The D statistic is commonly defined as $D = N\sum_{j=1}^{n}(O_j - I_j)/n$, applied to paired (O_j,I_j) where j indexes the sample items, O_j is the operator declaration, I_j is the inspector measurement, n is the verification sample size, and N is the total number of items in the stratum. Both the D statistic and the one-item-at-a-time tests rely on estimates of operator and inspector measurement uncertainties that are based on empirical uncertainty quantification (UQ). The empirical UQ uses paired (O_j,I_j) data from previous inspection periods in metrology studies to characterize measurement error variance components, as we explain below. Our focus is a sensitivity analysis of the impact of the uncertainty in the measurement error variance components (that are estimated using the prior verification (O_j,I_j) data) on sample size calculations in IAEA verifications. Such an assessment depends on the

* e-mail: t.burr@iaea.org

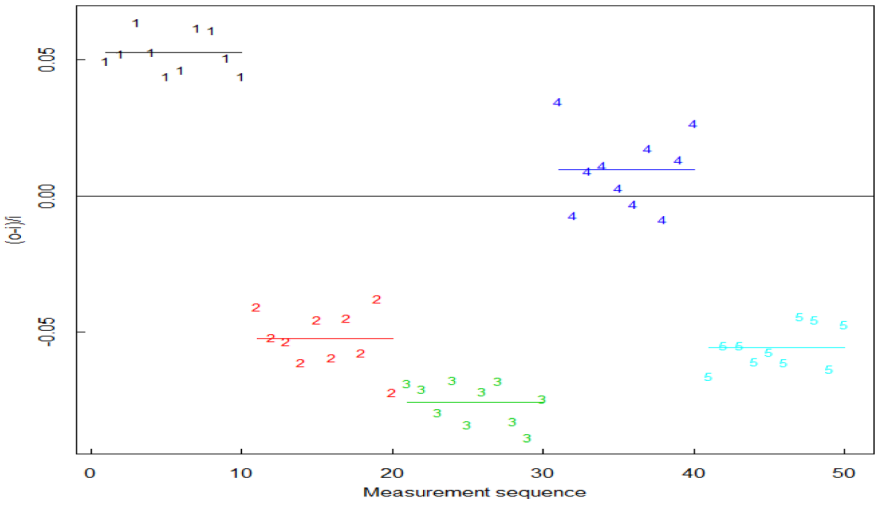

Fig. 1. Example simulated verification measurement data. The relative difference $\tilde{d}=(o-i)/o$ is plotted for each of 10 paired (o,i) measurements in each of 5 groups, for a total of 50 relative differences. The mean relative difference within each group (inspection period) is indicated by a horizontal line through the respective group means of the paired differences.

assumed measurement error model and associated uncertainty components, so it is important to perform effective UQ.

This paper is organized as follows. Section 2 describes measurement error models and error variance estimation using Grubbs' estimation [4–6]. Section 3 describes statistical tests based on the D statistic and one-verification-item-at-a-time testing. Section 4 gives simulation results that describe inference quality as a function of two sample sizes. The first sample size n_1 is the metrology study sample size (from previous inspection periods) used to estimate measurement error variances using Grubbs' (or similar) estimation methods. The second sample size n_2 is the number of verification items from a population of size N. Section 5 is a discussion, summary, and implications.

2 Measurement error models

The measurement error model must account for variation within and between groups, where a group is, for example, a calibration or inspection period. The measurement error model used for safeguards sets the stage for applying an analysis of variance (ANOVA) with random effects [4,6–9]. If the errors tend to scale with the true value, then a typical model for multiplicative errors is

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

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, σ_μ^2 is the "item variance", defined here as $\sigma_\mu^2 = \left(\sum_{i=1}^{N}(\mu_i - \overline{\mu})^2\right)/(N-1)$, $R_{Iij} \sim N(0, \delta_{RI}^2)$ is a random error of item j from group i, and $S_{Ii} \sim N(0, \delta_{SI}^2)$ is a short-term systematic error in group i. Note that the variance of I_{ij} is given by $V(I_{ij}) = \mu_{ij}^2(\delta_{SI}^2 + \delta_{RI}^2) + \sigma_\mu^2(\delta_{SI}^2 + \delta_{RI}^2)$. The term σ_μ^2 is the called "product variability" by Grubbs [6]. Neither R_{Iij} nor S_{Ii} are observable from data. However, for various types of observed data, we can estimate the variances δ_{RI}^2 and δ_{SI}^2. The same error model is typically also used for the operator, but with $R_O \sim N(0, \delta_{RO}^2)$ and $S_O \sim N(0, \delta_{SO}^2)$. We use capital letters such as I and O to denote random variables and corresponding lower case letters i and o to denote the corresponding observed values.

Figure 1 plots simulated example verification measurement data. The relative difference $\tilde{d}=(o-i)/o$ is plotted for each of 10 paired (o,i) measurements in each of 5 groups (inspection periods), for a total of 50 relative differences. As shown in Figure 1, typically, the between-group variation is noticeable compared to the within-group variation, although the between-group variation is amplified to a quite large value for better illustration in Figure 1; we used $\delta_{RO}=0.005$, $\delta_{SO}=0.001$, $\delta_{RI}=0.01$, $\delta_{SI}=0.03$, and the value $\delta_{SI}=0.03$ is quite large. Figure 2a is the same type of plot as Figure 1, but is for real data (four operator and inspector measurements on drums of UO_2 powder from each of three inspection periods). Figure 2b plots inspector versus operator data for each of the three inspection periods; a linear fit is also plotted.

2.1 Grubbs' estimator for paired (operator, inspector) data

Grubbs introduced a variance estimator for paired data under the assumption that the measurement error model was additive. We have developed new versions of the Grubbs' estimator to accommodate multiplicative error models and/or prior information regarding the relative sizes of the true variances [4,5]. Grubbs' estimator was developed for the situation in which more than one measurement method is applied to multiple test items, but there is no replication of measurements by any of the methods. This is the typical situation in paired (O,I) data.

Grubbs' estimator for an additive error model can be extended to apply to the multiplicative model equation (1) as follows. First, equation (1) for the inspector data (the

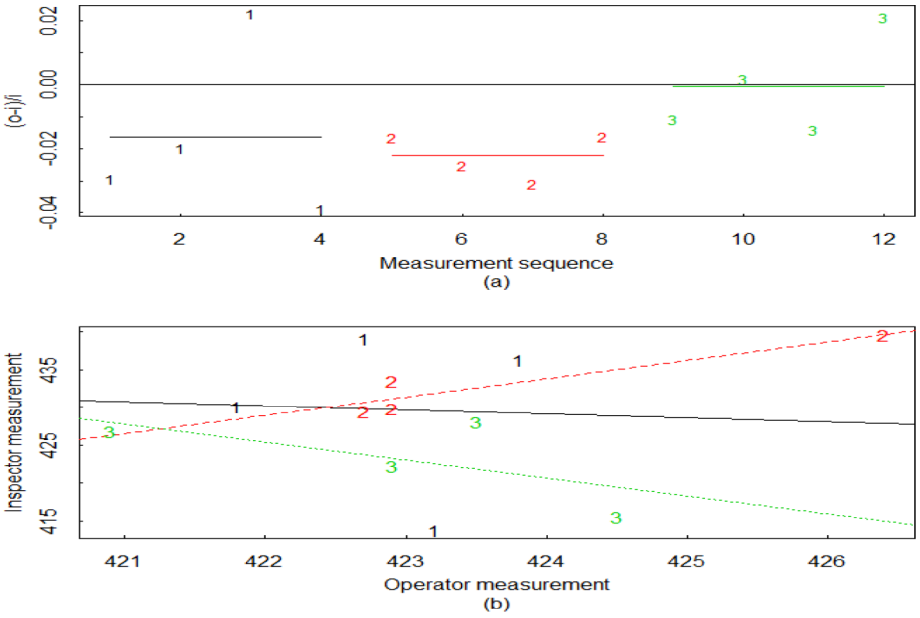

Fig. 2. Example real verification measurement data. (a) Four paired (O,I) measurements in three inspection periods; (b) inspector vs. operator measurement by group, with linear fits in each group.

operator data is analysed in the same way) implies that the within-group mean squared error (MSE), $\sum_{j=1}^{n}(I_j-\overline{I})^2/(n-1)$, has expectation $\sigma_\mu^2\delta_{SI}^2+(\sigma_\mu^2+\overline{\mu}^2)\delta_{RI}^2+\sigma_\mu^2$, where $\overline{\mu}$ is the average value of μ_{ij} (assuming that each group has the same number of paired observations n). Second, the between-group MSE, $(\sum_{j=1}^{n}n(\overline{I}_j-\overline{I})^2)/(g-1)$, has expectation $(\sigma_\mu^2+n\overline{\mu}^2)\delta_{SI}^2+(\sigma_\mu^2+\overline{\mu}^2)\delta_{RI}^2+\sigma_\mu^2$. Therefore, both δ_{SI}^2 and δ_{RI}^2 are involved in both the within- and between-groups MSEs, which implies that one must solve a system of two equations and two unknowns to estimate δ_{SI}^2 and δ_{RI}^2 [4,5]. By contrast, if the error model is additive, only σ_{RI}^2 is involved in the within-group MSE, while both σ_{RI}^2 and σ_{SI}^2 are involved in the between-group MSE. The term σ_μ^2 in both equations is estimated as in the additive error model, by using the fact that the covariance between operator and inspector measurements equals σ_μ^2 [4,5]. However, σ_μ^2 will be estimated with non-negligible estimation error in many cases. For example, see Figure 2b where the fitted lines in periods 1 and 3 have negative slope, which implies that the estimate of σ_μ^2 is negative in periods 1 and 3 (but the true value of σ_μ^2 cannot be negative in this situation). We note that in the limit as σ_μ^2 approaches zero, the expression for the within-group MSE reduces to that in the additive model case (and similarly for the between-group MSE).

3 Applying uncertainty estimates: the *D* statistic and one-at-a-time-verification measurements

This paper considers two possible IAEA verification tests. First, the overall D test for a pattern is based on the average difference, $D=N\sum_{j=1}^{n}(O_j-I_j)/n$. Second, the one-at-a-time test compares the operator to the corresponding inspector measurement for each item and a relative difference is computed, defined as $d_j=(o_j-i_j)/o_j$. If $d_j>3\delta$, where $\delta=\sqrt{\delta_o^2+\delta_i^2}$, where $\delta_O^2=\delta_{OR}^2+\delta_{OS}^2$ and $\delta_I^2=\delta_{IR}^2+\delta_{IS}^2$ (or some other alarm threshold close to the value of 3 that corresponds to a small false alarm probability), then the jth item selected for verification leads to an alarm. Note that the correct normalization used to define the relative difference is actually $d_j=(o_j-i_j)/\mu_j$, which has standard deviation exactly δ. But μ_j is not known in practice, so a reasonable approximation is to use $d_j=(o_j-i_j)/o_j$, because the operator measurement o_j is typically more accurate and precise than the inspectors's NDA measurement i_j. Provided $\sqrt{\delta_{OR}^2+\delta_{OS}^2}\le 0.20$ (approximately), one can assume that $d_j=(o_j-i_j)/o_j$ is an adequate approximation to $d_j=(o_j-i_j)/\mu_j$ [10]. Although IAEA experience suggests that $\sqrt{\delta_{IR}^2+\delta_{IS}^2}$ sometimes exceeds 0.20, usually $\sqrt{\delta_{OR}^2+\delta_{OS}^2}\le 0.20$ [8].

3.1 The *D* statistic to test for a trend in the individual differences $d_j=o_j-i_j$

For an additive error model, $I_{ij}=\mu_{ij}+S_{Ii}+R_{Iij}$, it is known [11] that the variance of the D statistic is given by $\sigma_D^2=N^2((\sigma_R^2/n)+\sigma_S^2)$, where $\sigma_R^2=\sigma_{RO}^2+\sigma_{RI}^2$ and $\sigma_S^2=\sigma_{SO}^2+\sigma_{SI}^2$ are the absolute (not relative) variances. If one were sampling from a finite population without measurement error to estimate a population mean, then $\sigma_D^2=N^2(\sigma^2/n)((N-n)/N)$, where $f=(N-n)/N$ is the finite population correction factor, and σ^2 is a quasi-variance term (the "item variance" as defined previously in a slightly different context), defined here as

$\sigma^2 = (\sum_{i=1}^{N}(d_i - \overline{d})^2/(N-1))$. Notice that without any measurement error, if $n = N$ then $f = 0$, so $\sigma_D^2 = 0$, which is quite different from $\sigma_D^2 = N^2((\sigma_R^2/n) + \sigma_S^2)$. Figure 1 can be used to explain why $\sigma_D^2 = N^2((\sigma_R^2/n) + \sigma_S^2)$ when there are both random and systematic measurement errors. And, the fact that $\sigma_D^2 = N^2(\sigma^2/n)f = 0$ when $n = N$ and there are no measurement errors is also easily explainable.

For a multiplicative error model (our focus), it can be shown [11] that

$$\sigma_D^2 = \frac{N}{n}\delta_R^2\sum_{j=1}^{N}\mu_j^2 + \text{Total}^2\delta_S^2 + \frac{N-n}{n}N\sigma_\mu^2\delta_S^2, \qquad (2)$$

where $\text{Total} = \sum_{j=1}^{N}\mu_j = N\overline{\mu}$ and $\sigma_\mu^2 = (\sum_{i=1}^{N}(\mu_i - \overline{\mu})^2)/(N-1)$, and so to calculate σ_D^2 in equation (2), one needs to know or assume values for σ_μ^2 (the item variance) and the average of the true values, $\overline{\mu}$. In equation (2), the first two terms are analogous to $N^2((\sigma_R^2/n) + \sigma_S^2)$ in the additive error model case. The third term involves σ_μ^2 and decreases to 0 when $n = N$. Again, in the limit as σ_μ^2 approaches zero, equation (2) reduces to that for the additive model case; and regardless whether σ_μ^2 is large or near zero, the effect of δ_S^2 cannot be reduced by taking more measurements (increasing n in Eq. (2)).

In general, the multiplicative error model gives different results than an additive error model because variation in the true values, σ_μ^2, contributes to σ_D^2 in a multiplicative model, but not in an additive model. For example, let $\sigma_R^2 = \overline{\mu}^2\delta_R^2$ and $\sigma_S^2 = \overline{\mu}^2\delta_S^2$, so that the average variance in the multiplicative model is the same as the variance in the additive model for both random and systematic errors. Assume $\delta_R = 0.10$, $\delta_S = 0.02$, $\overline{\mu} = 100$ (arbitrary units), and $\sigma_\mu^2 = 2500$ (50% relative standard deviation in the true values). Then the additive model has $\sigma_D = 270.8$ and the corresponding multiplicative model with the same average absolute variance has $\sigma_D = 310.2$, a 15% increase. The fact that $\text{var}(\mu)$ contributes to σ_D^2 in a multiplicative model has an implication for sample size calculations such as those we describe in Section 4. Provided the magnitude of $S_{Iij} + R_{Iij}$ is approximately 0.2 or less (equivalently, the relative standard deviation of $S_{Iij} + R_{Iij}$ should be approximately 8% or less), one can convert equation (1) to an additive model by taking logarithms, using the approximation $\log(1+x) \approx x$ for $|x| \leq 0.20$. However, there are many situations for which the log transform will not be sufficiently accurate, so this paper describes a recently developed option to accommodate multiplicative models rather than using approximations based on the logarithm transform [4,5].

The overall D test for a pattern is based on the average difference, $D = N\sum_{j=1}^{n}(O_j - I_j)/n$. The D-statistic test is based on equation (2), where $\delta_R^2 = \delta_{OR}^2 + \delta_{IR}^2$ is the random error variance and $\delta_S^2 = \delta_{OS}^2 + \delta_{IS}^2$ is the systematic error variance of $\tilde{d} = (o-i)/\mu \approx (o-i)/o$, and σ_μ^2 is the absolute variance of the true (unknown) values. If the observed D value exceeds $3\sigma_D$ (or some similar multiple of σ_D to achieve a lot false alarm probability) then the D test alarms.

The test that alarms if $D \geq 3\sigma_D$ is actually testing whether $D \geq 3\hat{\sigma}_D$, where $\hat{\sigma}_D$ denotes an estimate of σ_D; this leads to two sample size evaluations. The first sample size n_1 involves metrology data collected in previous inspection samples used to estimate $\delta_R^2 = \delta_{OR}^2 + \delta_{IR}^2$, $\delta_S^2 = \delta_{OS}^2 + \delta_{IS}^2$, and σ_μ^2 needed in equation (2). The second sample size n_2 is the number of operator's declared measurements randomly selected for verification by the inspector. The sample size n_1 consists of two sample sizes: the number of groups g (inspection periods) used to estimate δ_S^2 and the total number of items over all groups, $n_1 = gn$ in the case (the only case we consider in examples in Sect. 4) that each group has n paired measurements.

3.2 One-at-a-time sample verification tests

The IAEA has historically used zero-defect sampling, which means that the only acceptable (passing) sample is one for which no defects are found. Therefore, the non-detection probability is the probability that no defects are found in a sample of size n when one or more true defective items are in the population of size N. For one-item-at-a-time testing, the non-detection probability is given by

$$\text{Prob(discover 0 defects in sample of size n)} = \sum_{i=Max(0,n+r-N)}^{Min(n,r)} A_i \times B_i, \qquad (3)$$

where the term A_i is the probability that the selected sample contains i truly defective items, which is given by the hypergeometric distribution with parameters on i, n, N, r, where i is the number of defects in the sample, n is the sample size, N is the population size, and r is the number of defective items in the population. More specifically,

$$A_i = \binom{r}{i}\binom{N-r}{n-i}/\binom{N}{n},$$

the above equation is the probability of choosing i defective items from r defective items in a population of size N in a sample of size n, which is the well-known hypergeometric distribution. The term B_i is the probability that none of the i truly defective items is inferred to be defective based on the individual d tests. The value of B_i depends on the metrology and the alarm threshold. Assuming a multiplicative error model for the inspector measurement (and similarly for the operator), implies that, for an alarm threshold of $k = 3$, for $\tilde{D}_j = ((O_j - I_j)/O_j) \approx ((O_j - I_j)/\mu_j)$ we have to calculate $B_i = P(\tilde{D}_1 \leq 3\delta, \tilde{D}_2 \leq 3\delta, \ldots, \tilde{D}_i \leq 3\delta)$, where $\delta = \sqrt{\delta_R^2 + \delta_S^2}$, which is given by the multivariate normal integral

$$B_i = \frac{1}{(2\pi)^{i/2}|\Sigma_i|^{1/2}}\int_{-\infty}^{3\delta}\cdots\int_{-\infty}^{3\delta}\exp\left\{\frac{-(z-\lambda)^T\Sigma_i^{-1}(z-\lambda)}{2}\right\} dz_1 dz_2 \ldots dz_i,$$

where each of the components of λ are equal to $1\,\text{SQ}/r$ (SQ is a significant quantity; for example, $1\,\text{SQ} = 8\,\text{kg}$ for Pu, and r was defined above as the number of defective items in

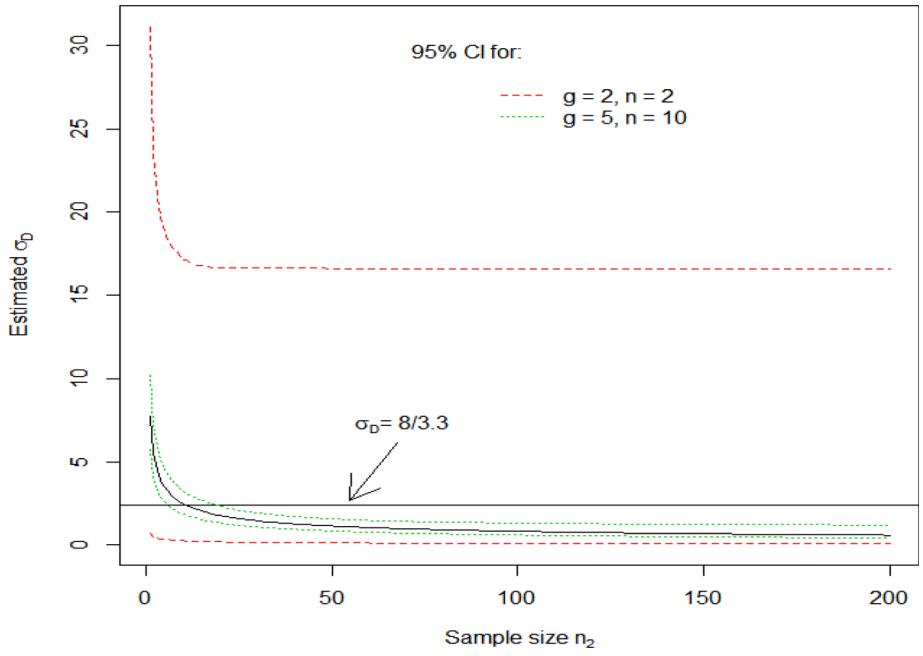

Fig. 3. The estimate of σ_D versus sample size n_2 for two values of n_1 (case A: $g=2$, $n=2$ so $n_1=4$, or case B: $g=5$, $n=10$ so $n_1=50$).

the population). The term $\sum_i$ in the B_i calculation involved in the multivariate normal integral is a square matrix with i rows and columns with values $(\delta_R^2+\delta_S^2)$ on the diagonal and values δ_S^2 on the off-diagonals.

4 Simulation study

The left hand side of equations (2) and (3) can be considered a "measurand" in the language used in the guide to expressing uncertainty in measurement [12]. Although the error propagation in the GUM is typically applied in a "bottom-up" uncertainty evaluation of a measurement method, it can also be applied to any other output quantity y (such as $y=\sigma_D$ or $y=$DP) expressed as a known function $y=f(x_1, x_2, \ldots, x_p)$ of inputs $x_1, x_2, \ldots, x_p$ (inputs such as $\delta_R^2=\delta_{OR}^2+\delta_{IR}^2$, $\delta_S^2=\delta_{OS}^2+\delta_{IS}^2$, and σ_μ^2). The GUM recommends linear approximations ("delta method") or Monte Carlo simulations to propagate uncertainties in the inputs to predict uncertainties in the output. Here we use Monte Carlo simulations to evaluate the uncertainties in the inputs $\delta_R^2=\delta_{OR}^2+\delta_{IR}^2$, $\delta_S^2=\delta_{OS}^2+\delta_{IS}^2$, and σ_μ^2 and also to evaluate the uncertainty in $y=\sigma_D$ or $y=$DP as a function of the uncertainties in the inputs. Notice that equation (2) is linear in δ_R^2 and δ_S^2, so the delta method to approximate the uncertainty in $y=\sigma_D$ would be exact; however, there is non-zero covariance (a negative covariance) between $\hat{\delta}_R^2$ and $\hat{\delta}_S^2$ that would need to be taken into account in the delta method.

We used the statistical programming language R [13] to perform simulations for example true values of $\delta_{OR}^2, \delta_{OS}^2, \delta_{IR}^2, \delta_{IS}^2, \sigma_\mu^2, \overline{\mu}, N$, and the amount of diverted nuclear material. For each of 10^5 or more simulation runs, normal errors were generated assuming the multiplicative error model (1) for both random and systematic errors (see Sect. 4.2 for examples with non-normal errors). The new version of the Grubbs' estimator for multiplicative errors was applied to produce the estimates $\hat{\delta}_{OR}^2$, $\hat{\delta}_{IR}^2$, $\hat{\delta}_{OS}^2$, $\hat{\delta}_{IS}^2$, and $\hat{\sigma}_\mu^2$, which were then used to estimate $y=\sigma_D$ in equation (2) and $y=$DP in equation (3). Because there is large uncertainty in the estimates $\hat{\delta}_{OR}^2$, $\hat{\delta}_{IR}^2$, $\hat{\delta}_{OS}^2$, $\hat{\delta}_{IS}^2$ unless σ_μ^2 is nearly 0, we also present results for a modified Grubbs' estimator applied to the relative differences $\tilde{D}_j=(O_j-I_j)/O_j$ that estimates the aggregated variances $\delta_R^2=\delta_{OR}^2+\delta_{IR}^2$ and $\delta_S^2=\delta_{OS}^2+\delta_{IS}^2$, and also estimates σ_μ^2. Results are described in Sections 4.1 and 4.2.

4.1 The *D* statistic to test for a trend in the individual differences $d_j=(o_j-i_j)/o_j$

Figure 3 plots 95% CIs for σ_D versus sample size n_2 using the modified Grubbs' estimator applied to the relative differences $\tilde{D}_j=(O_j-I_j)/O_j$ for the parameter values $\delta_{RO}=0.01$, $\delta_{SO}=0.001$, $\delta_{RI}=0.05$, $\delta_{SI}=0.005$, $\overline{\mu}=1$, $\sigma_\mu=0.01$, $N=200$ for case A (*defined here and throughout as* $n_1=4$ with $g=2$, $n=2$) and for case B (*defined here and throughout as* $n_1=50$ with $g=5$, $n=10$) . We used 10^5 simulations of the measurement process to estimate the quantiles of the distribution of $y=\sigma_D$. We confirmed by repeating the sets of 10^5 simulations that simulation error due to using a finite number of simulations is negligible. Clearly, and not surprisingly, the sample size in Case A leads to CI length that seems to be too wide for effectively quantifying the uncertainty in σ_D. The traditional Grubbs' estimator performs poorly unless σ_μ is very small, such as $\sigma_\mu=0.0001$. We use the traditional Grubbs'

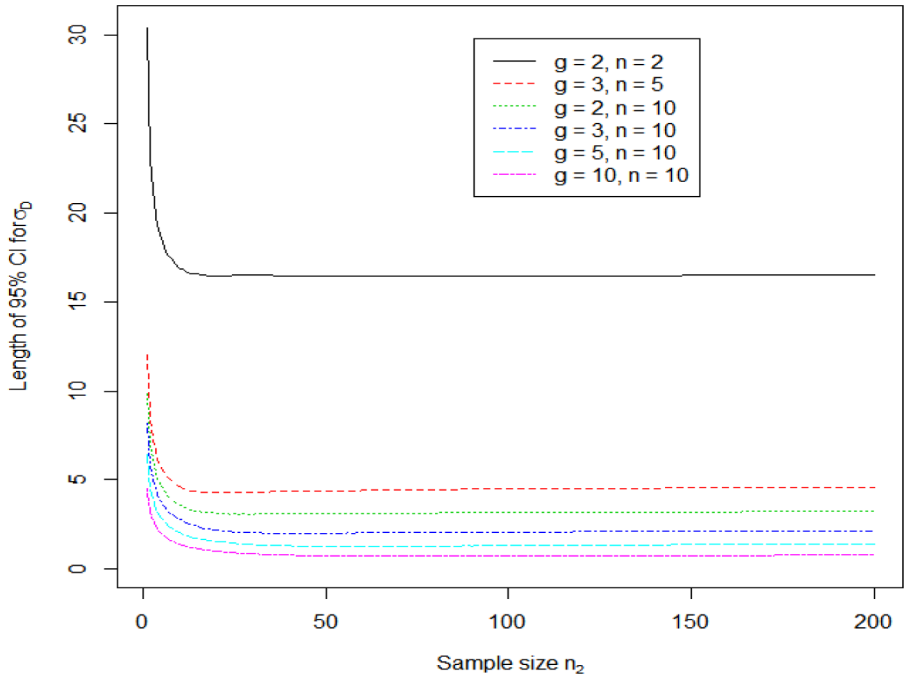

Fig. 4. Estimated lengths of 95% confidence intervals for σ_D versus sample size n_2 for six values of n_1 ($g=2$, $n=2$ so $n_1=4$, $g=3$, $n=5$ so $n_1=15$, etc.).

estimator in Section 4.2. The modified estimator that estimates the aggregated variances performs well for any value of σ_μ.

Figure 4 is similar to Figure 3, except Figure 4 plots the length of 95% CIs for 6 possible values of n_1 (see the figure legend). Again, the case A sample size is probably too small for effective estimation of σ_D. In this example, the smallest length CI is for $g=5$ and $n=100$, but $n=100$ is unrealistically large, while $g=3$ and $n=10$ or $g=5$ and $n=10$ are typically possible with reasonable resources. The length of these 95% CIs is one criterion to choose an effective sample size n_1.

Another criterion to choose an effective sample size n_1 is the root mean squared error (RMSE, defined below) in estimating the sample size n_2 needed to achieve $\sigma_D=8/3.3$ (the 3.3 is an example value that corresponds to a 95% DP to detect an 8 kg shift (1 SQ for Pu) while maintaining a 0.05 FAP when testing for material loss). In this example, the RMSE in estimating the sample size n_2 needed to achieve $\sigma_D=8/3.3$ is approximately 12.9 for case A and 8.0, 7.3, 6.8, 6.7, and 6.3, respectively, for the other values of n_1 considered in Figure 4. These RMSEs are repeatable to within ± 0.1 across sets of 10^5 simulations so the RMSE values are in the same order as the CI lengths in Figure 4. The RMSE is defined as

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{10^5} (\hat{n}_{2,i} - n_{2,true})^2}{10^5}},$$

where $\hat{n}_{2,i}$ is the estimated sample size n_2 in simulation i that is needed in order to achieve $\sigma_D=8/3.3$, and $n_{2,\text{true}}$ is the true sample size n_2 ($n_{2,\text{true}}=22$ in this example; see Fig. 3 where the true value of σ_D versus n_2 is also shown) needed to achieve $\sigma_D=8/3.3$.

Another criterion to choose an effective size n_1 is the detection probability to detect specified loss scenarios. We consider this criterion in Section 4.3.

4.2 Uncertainty on the uncertainty on the uncertainty

The term "uncertainty" typically refers to a measurement error standard deviation, such as σ_D. Therefore, Figures 3 and 4 involve the "uncertainty of the uncertainty" as a function of n_1 (defined as $n_1=ng$, so more correctly, as a function of g and n) and n_2. Figures 5–7 illustrate the "uncertainty of the uncertainty of the uncertainty" (we commit to stopping at this level-three usage of "uncertainty"). The "uncertainty of the uncertainty" depends on the underlying measurement error probability density, which is sometimes itself uncertain. Figure 5 plots the familiar normal density and three non-normal densities (uniform, gamma, and generalized lambda, [14]). Figure 6 plots the estimated probability density (using the 10^5 realizations) of the estimated value of δ_{IR} using the traditional Grubbs' estimator for each of the four distributions (the true value of δ_{IR} is 0.05) and the five true standard deviations are the same as in Section 4.1 for generating the random variables ($\delta_{RO}=0.01$, $\delta_{SO}=0.001$, $\delta_{RI}=0.05$, $\delta_{SI}=0.005$, $\overline{\mu}=1$, $\sigma_\mu=0.01$, $N=200$). Figure 7 is similar to Figure 3 (for $g=5$, $n=10$), except it compares CIs assuming the normal distribution to CIs assuming the generalized lambda distribution. That is, Figure 7 plots the estimated CI, again for the model parameters as above, for σ_D for the normal and for the generalized lambda distributions. In this case, the CIs are wider for the generalized lambda distribution than for the normal distribution. Recall (Fig. 5) that standard deviation of the four estimated probability densities are:

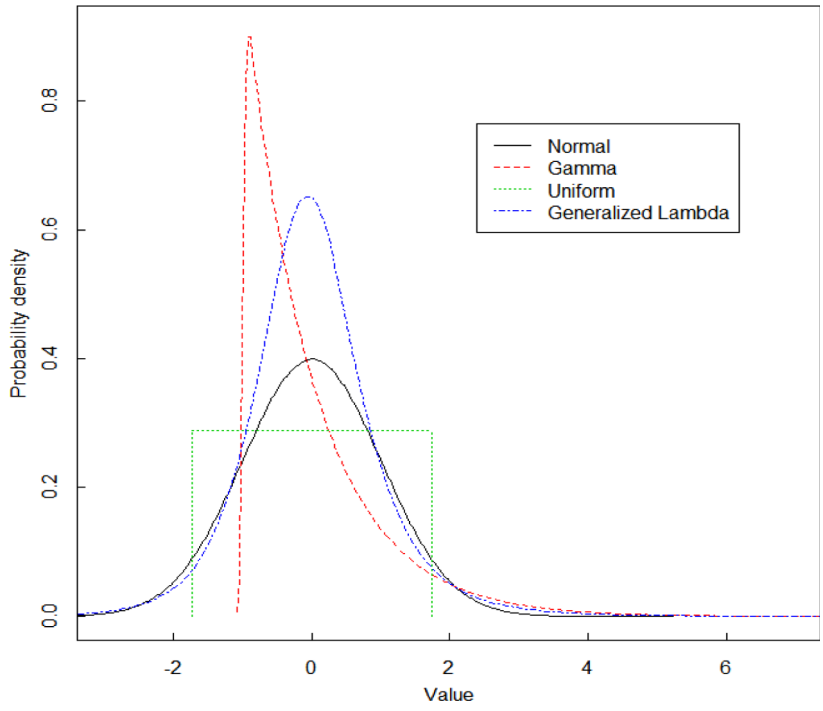

Fig. 5. Four example measurement error probability densities: normal, gamma, uniform, and generalized lambda, each with mean 0 and variance 1.

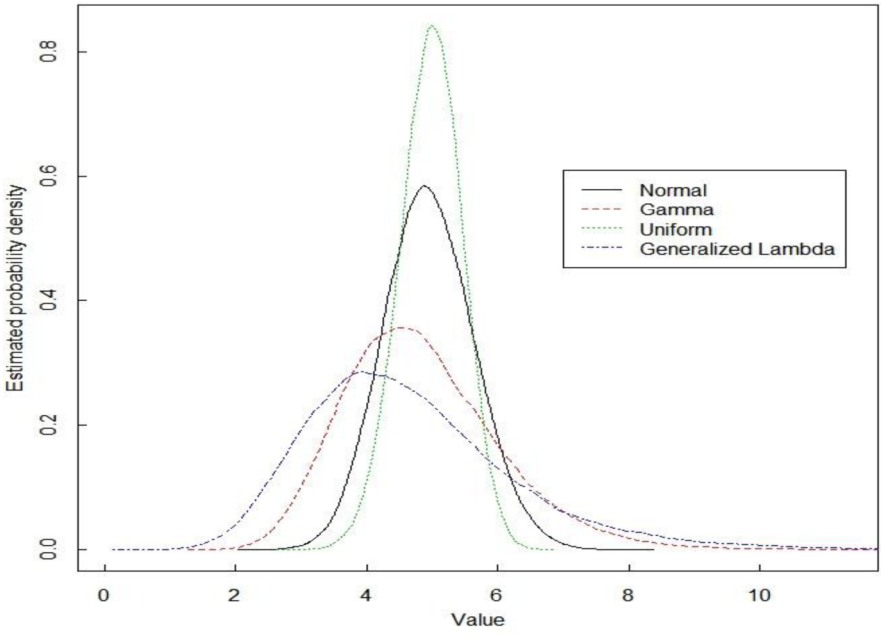

Fig. 6. The estimated probability density for $\hat{\delta}_{IR}$ in the four example measurement error probability densities (normal, gamma, uniform, and generalized lambda, each with mean 0 and variance 1) from Figure 4.

0.14, 0.25, 0.10, and 0.36 for the normal, gamma, uniform, and generalized lambda, respectively. Therefore, one might expect the CI for σ_D to be shorter for the normal than for a generalized lambda distribution that has the same relative standard deviation as the corresponding normal distribution.

4.3 One-at-a-time testing

For one-at-a-time testing, Figure 8 plots 95% confidence intervals for the estimated DP versus sample size n_2 for cases A and B (see Sect. 4.1). The true parameter values used in equation (3) were $\delta_{RO} = 0.1$, $\delta_{SO} = 0.05$, $\delta_{RI} = 0.1$,

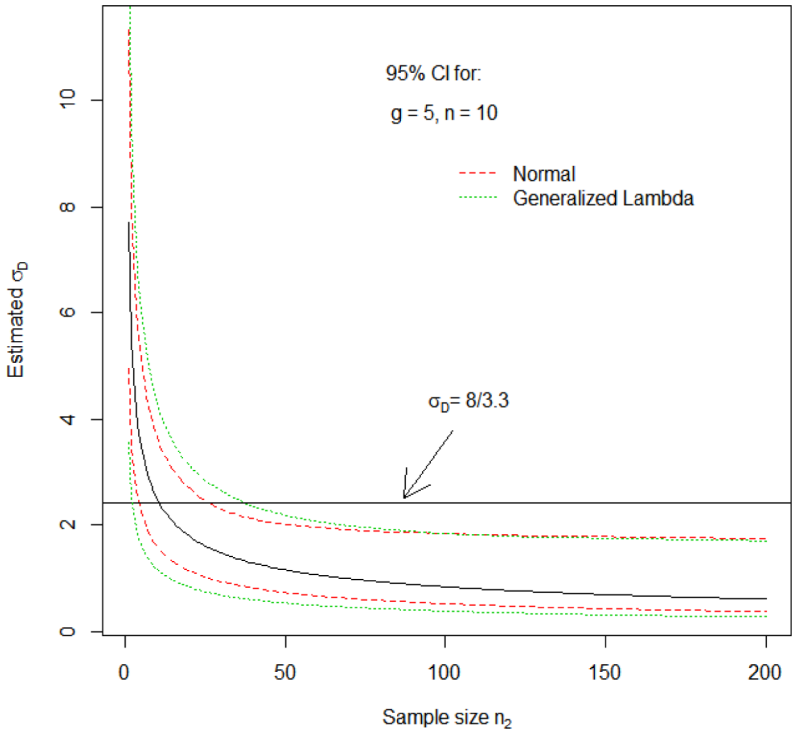

Fig. 7. 95% confidence intervals for the estimate of σ_D versus sample size n_2 for case B, assuming the measurement error distribution is either the normal or the generalized lambda distribution.

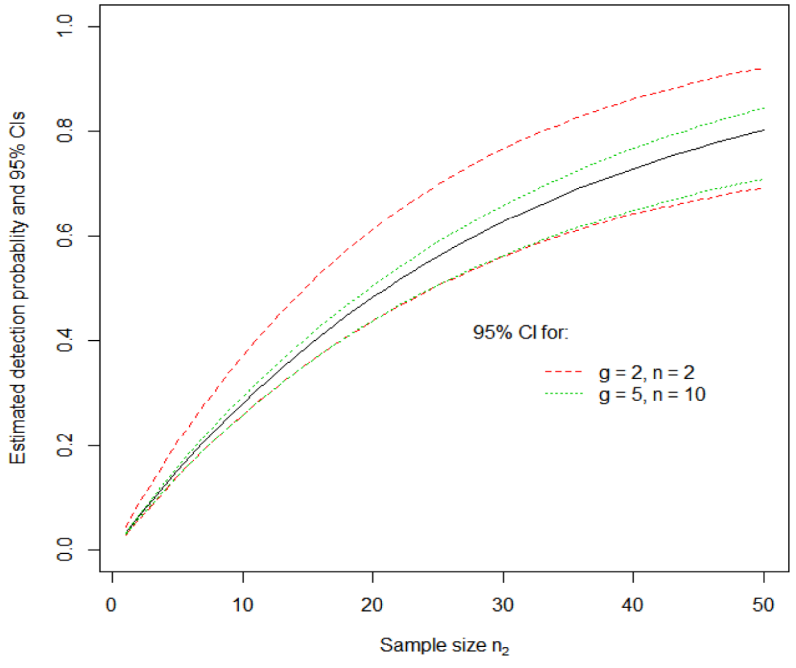

Fig. 8. Estimated detection probability and 95% confidence interval versus sample size n_2 for cases A and B. The true detection probability is plotted as the solid (black) line.

$\delta_{SI} = 0.05$, $\overline{\mu} = 15$, $\sigma_\mu = 0.01$. And, a true mean shift of 8 kg in each of 10 falsified items was used (representing data falsification by the operator to mask diversion of material). The CIs for the DP were estimated by using the observed 2.5% and 97.5% quantiles of the DP values in 10^5 simulations. As in Section 4.1, we confirmed by repeating the sets of 10^5 simulations that simulation error due to using a finite number of simulations is negligible. The very small case A sample leads to approximately the same lower 2.5% quantile as did case B; however, the upper 97.5% quantile is considerably lower for case A than for case B. Other values for the parameters (δ_{RO}, δ_{SO}, δ_{RI}, δ_{SI}, $\overline{\mu}$, σ_μ, the number of

falsified items, and the amount falsified per item) lead to different conclusions about uncertainty as a function of n_2 in how the DP decreases as a function of n_2. For example, if we reduce $\overline{\mu} = 15$ to $\overline{\mu} = 1$ in this example, then the confidence interval lengths are very short for both case A and case B.

For this same example, we can also compute the DP in using the D statistic to detect the loss (which the operator attempts to mask by falsifying the data). For the example just described (for which simulation results are shown in Fig. 8), the true DP in using the D statistic (using an alarm threshold of σ_D and $n_2 = 30$ using Eq. (2)) is 0.65. The corresponding true DP for one-at-a-time testing is 0.27. Therefore, in this example, with 10 of 200 items falsified, each by an amount of 8 units, the D statistic has lower DP than the $n_2 = 30$ one-at-a-time tests. In other examples, the D statistic will have higher DP, particularly when there are many falsified items in the population. For example, if we increase the number of defectives in this example from 10 of 200 to 20, 30, or 40 of 200, then the DPs are (0.17, 0.17), (0.08, 0.15), and (0.06, 0.14) for one-at-a-time testing and for the D statistic, respectively. These are low DPs, largely because the measurement error variances are large in this example. One can also assess the sensitivity of the estimated DP in using the D statistic to the uncertainty in the estimated variances; for brevity, we do not show that here.

5 Discussion and summary

This study was motivated by three considerations. First, there is an ongoing need to improve UQ for error variance estimation. For example, some applications involve characterizing items for long-term storage and the measurement error behaviour for the items is not well known, so an initial metrology study with to-be-determined sample sizes is required. Second, we recently provided the capability to allow for multiplicative error models in evaluating the D statistic (Eq. (2)) [4,5]. Third, we recently provided the capability to allow for both random and systematic errors in one-at-a-time item testing (Eq. (3)).

We presented a simulation study that assumed error variances are estimated using an initial metrology study characterized by g measurement groups and n paired operator, inspector measurements per group. Not surprisingly, both one-item-at-a-time testing and pattern testing using the D statistic, it appears that $g = 2$ and $n = 2$ is too small for effective variance estimation.

Therefore, the sample sizes in the previous and current inspections will impact the estimated DP and FAP, as is illustrated by numerical examples. The numerical examples include application of the new expression for the variance of the D statistic assuming the measurement error model is multiplicative (Eq. (2)) is used in a simulation study and new application of both random and systematic error variances in one-item-at-a-time testing (Eq. (3)).

Future work will evaluate the impact of larger values of product variability, σ_μ^2 on the standard Grubbs' estimator; this study used a very small value of σ_μ^2, which is adequate in some contexts, such as product streams. The value of σ_μ^2 could be considerably larger in some NM streams, particularly waste streams. Therefore, this study also evaluated the relative differences $d_j = (o_j - i_j)/o_j$ to estimate the aggregated quantities needed in equations (2) and (3), $\delta_R = \sqrt{\delta_{RO}^2 + \delta_{RI}^2}, \delta_S = \sqrt{\delta_{SO}^2 + \delta_{SI}^2}$, using a modified Grubbs' estimation, to mitigate the impact of noise in estimation of σ_μ. Because σ_μ^2 is a source of noise in estimating the individual measurement error variances [15], a Bayesian alternative is under investigation to reduce its impact [16]. Also, one could base a statistical test for data falsification based on the relative differences between operator and inspector measurements $d = (o - i)/o$ in which case an alternate expression to equation (2) for σ_D that does not involve the product variability σ_μ^2 would be used.

5.1 Implications and influences

This study was motivated by three considerations, each of which have implications for future work. First, there is an ongoing need to improve UQ for error variance estimation. For example, some applications involve characterizing items for long-term storage and the measurement error behaviour might not be well known for the items, so an initial metrology study with to-be-determined sample sizes is required. Second, we recently provided the capability to allow for multiplicative error models in evaluating the D statistic (Eq. (2) in Sect. 3) [4,5]. Third, we recently provided the capability to allow for both random and systematic errors in one-at-a-time item testing (Eq. (3) in Sect. 3). Previous to this work, the variance of the D statistic was estimated by assuming measurement error models are additive rather than multiplicative, and one-at-a-time item testing assumed that all measurement errors were purely random.

The authors acknowledge CETAMA for hosting the November 17–19, 2015 conference on sampling and characterizing where this paper was first presented.

References

1. R. Avenhaus, M. Canty, *Compliance Quantified* (Cambridge University Press, 1996)
2. T. Burr, M.S. Hamada, Revisiting statistical aspects of nuclear material accounting, Sci. Technol. Nucl. Install. **2013**, 961360 (2013)
3. T. Burr, M.S. Hamada, Bayesian updating of material balances covariance matrices using training data, Int. J. Prognost. Health Monitor. **5**, 6 (2014)
4. E. Bonner, T. Burr, T. Guzzardo, T. Krieger, C. Norman, K. Zhao, D.H. Beddingfield, W. Geist, M. Laughter, T. Lee, Ensuring the effectiveness of safeguards through comprehensive uncertainty quantification, J. Nucl. Mater. Manage. **44**, 53 (2016)
5. T. Burr, T. Krieger, K. Zhao, Grubbs' estimators in multiplicative error models, IAEA report, 2015
6. F. Grubbs, On estimating precision of measuring instruments and product variability, J. Am. Stat. Assoc. **43**, 243 (1948)

7. K. Martin, A. Böckenhoff, Analysis of short-term systematic measurement error variance for the difference of paired data without repetition of measurement, Adv. Stat. Anal. **91**, 291 (2007)
8. R. Miller, *Beyond ANOVA: Basics of Applied Statistics* (Chapman & Hall, 1998)
9. C. Norman, Measurement errors and their propagation, Internal IAEA Document, 2014
10. G. Marsaglia, Ratios of normal variables, J. Stat. Softw. **16**, 2 (2006)
11. T. Burr, T. Krieger, K. Zhao, Variations of the *D* statistics for additive and multiplicative error models, IAEA report, 2015
12. Guide to the Expression of Uncertainty in Measurement, JCGM 100: www.bipm.org (2008)
13. R Core Team R, *A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria, 2012): www.R-project.org
14. M. Freimer, G. Mudholkar, G. Kollia, C. Lin, A study of the generalized Tukey Lambda family, Commun. Stat. Theor. Methods **17**, 3547 (1988)
15. F. Lombard, C. Potgieter, Another look at Grubbs' estimators, Chemom. Intell. Lab. Syst. **110**, 74 (2012)
16. C. Elster, Bayesian uncertainty analysis compared to the application of the gum and its supplements, Metrologia **51**, S159 (2014)

Cite this article as: Tom Burr, Thomas Krieger, Claude Norman, Ke Zhao, The impact of metrology study sample size on uncertainty in IAEA safeguards calculations, EPJ Nuclear Sci. Technol. **2**, 36 (2016)