Hindawi Publishing Corporation
Journal of Sensors
Volume 2015, Article ID 267462, 10 pages
http://dx.doi.org/10.1155/2015/267462

Research Article

Uncertainty Quantification in Application of the Enrichment Meter Principle for Nondestructive Assay of Special Nuclear Material

Tom Burr,[1] Stephen Croft,[2] and Ken Jarman[3]

[1]*International Atomic Energy Agency, 1400 Vienna, Austria*
[2]*Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA*
[3]*Pacific Northwest National Laboratory, Richland, WA 99354, USA*

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

Received 5 May 2015; Accepted 18 June 2015

Academic Editor: Jesus Corres

Nondestructive assay (NDA) of special nuclear material (SNM) is used in nonproliferation applications, including identification of SNM at border crossings, and quantifying SNM at safeguarded facilities. No assay method is complete without "error bars," which provide one widely used way to express confidence in assay results. NDA specialists typically partition total uncertainty into "random" and "systematic" components so that, for example, an error bar can be developed for the SNM mass estimate in one item or for the total SNM mass estimate in multiple items. Uncertainty quantification (UQ) for NDA has always been important, but greater rigor is needed and achievable using modern statistical methods. To this end, we describe the extent to which the guideline for expressing uncertainty in measurements (GUM) can be used for NDA. Also, we describe possible extensions to the GUM by illustrating UQ challenges in NDA that it does not address, including calibration with errors in predictors, model error, and item-specific biases. A case study is presented using gamma spectra and applying the enrichment meter principle to estimate the ^{235}U mass in an item. The case study illustrates how to update the ASTM international standard test method for application of the enrichment meter principle using gamma spectra.

1. Introduction

As world reliance on nuclear energy increases, concerns about proliferation of materials that could be used for weapons also increase. Fissile nuclear materials can be detected and/or characterized by observing radiation released by fission, such as gamma-rays and neutrons [1]. Therefore, neutron and gamma detectors are deployed in many nonproliferation efforts such as cargo screening at border crossings and assay at facilities that process special nuclear material (SNM), which is the main application we consider.

Nondestructive assay (NDA) of items containing SNM uses calibration and modeling to infer item characteristics on the basis of detected radiation such as neutron and gamma emissions. For example, the amount of ^{235}U in an item can be estimated by using a measured net weight of uranium U in the compound and a measured ^{235}U enrichment (the ratio ^{235}U/U). Enrichment can be measured using the 185.7 keV gamma-rays emitted from ^{235}U by applying the enrichment meter principle (EMP), which we consider here as our case study [2].

Uncertainty quantification (UQ) for NDA has always been important, but currently it is recognized that greater rigor is needed and achievable using modern statistical methods and by letting UQ have a more prominent role in assay development and assessment. UQ is often difficult but, if done well, can lead to improving the assay procedure itself. Therefore, we describe the extent to which the guideline for expression of uncertainty in measurements (GUM) can be used for NDA [3–5]. Also, this paper takes steps toward better UQ for NDA by illustrating UQ challenges that are not addressed by the GUM. These challenges include item-specific biases, calibration with errors in predictors,

and model error, especially when the model is a key step in the assay. A case study is presented using low-resolution NaI spectra and applying the enrichment meter principle to estimate the ^{235}U mass in an item. The case study illustrates how to update the current international standard test method (ASTM) for application of the enrichment meter principle using gamma spectra from a NaI detector. The paper is organized as follows. Section 2 gives additional background on NDA and UQ for NDA. Section 3 describes the GUM. Section 4 is the EMP case study. Section 5 is a discussion and summary.

2. Background on NDA and UQ for NDA

NDA is widely used in nuclear nonproliferation because most detectors are rugged and portable and so can be brought to the location of the item for an in situ measurement [6–9]. In contrast, destructive analytical chemistry assay (DA) methods such as mass spectrometry require that a sample from the item be brought to the instrument. Typically, NDA has smaller sampling errors but also tends to have larger overall errors than DA, because there is no item preparation step, although there are exceptions. Overall error includes all types of random and systematic error and describes the total variation around the measurand's true value [10]. An error is systematic if it impacts or could impact more than one assay. For example, errors in estimated calibration model parameters lead to systematic errors for all assays made during the same calibration period with the same instrument. An error due to variation in the container thickness of an item is systematic to that item but is most likely to be random across items. We discuss container thickness as an example of item-specific systematic error in the EMP case study.

In NDA nonproliferation applications, items emit neutrons and/or gamma-rays that provide information about the source material, such as isotopic content. However, item properties such as density, which relates to neutron and/or gamma absorption behavior of the item, can partially obscure the relation between the detected radiation and the source material; this adds a source of uncertainty to the estimated amount of SNM in the item. One can express item-specific impacts on uncertainty using a model such as

$$\frac{\mathrm{CR}}{M} = f\left(X_1, X_2, \ldots, X_p\right) + R, \tag{1}$$

where CR is the item's neutron or gamma count rate, M is the item mass, and $X_1, X_2, \ldots, X_p$ are p auxiliary predictor variables such as item density, source SNM heterogeneity within the item, and container thickness, which will generally be estimated or measured with error and so are regarded as random variables [6]. Regarding notation, we use capital letters to denote random variables. The random error term R can include variation in background that cannot be perfectly adjusted for, Poisson counting statistics effects, and random effects related to estimating the counts in a spectral region that are associated with the particular source SNM. In the EMP case study, the spectral region centers on the 185.7 keV gamma full-energy peak that is the basis for estimating the ^{235}U enrichment in a sample.

In principle, the $X_1, X_2, \ldots, X_p$ could be estimated for each item as part of the assay protocol. However, there would still be modeling error because the function f must be chosen or somehow inferred, possibly using purely empirical data analysis applied to calibration data [6, 11], or physics-based radiation transport codes such as Monte-Carlo-n-particle (MCNP [12]). Typically, only some of $X_1, X_2, \ldots, X_p$ will be measured as part of the assay protocol, as we illustrate in the EMP case study.

Readers familiar with errors in variables (also known as errors in predictors) might wonder if errors in variables techniques will be needed [13–16]. We consider errors in variables in the EMP case study. Readers familiar with Bayesian data analysis might wonder if the true item mass M is to be regarded as a random variable, as would be done in a Bayesian approach. This paper regards the mass M as a random quantity, so we use capital M or capital T (for true value).

2.1. UQ for NDA. Perhaps surprisingly, a thorough approach for quantifying and reporting uncertainty does not yet exist even for the relatively simple and widely fielded EMP technique (which is our case study). As the complexity of the measurement system increases (such as instruments deploying multiple correction algorithms and operated in unattended mode [8]), exactly how to do effective UQ becomes less clear.

By comparison to UQ for NDA, UQ for DA is more mature. Space constraints do not permit a full comparison of how UQ is done in DA versus in NDA. However, one needs to ensure that any such comparison is between similar quantities. For example, when samples are collected and DA results are reported, sampling uncertainty (how representative the samples are of the entire item) is often not carried throughout the entire calculation. In other cases, some uncertainties may not be adequately evaluated to propagate through to the total measurement uncertainty. For example, there are uncertainties due to the fact that gamma measurements only sample the surface of an item because the sample itself attenuates and absorbs gamma-rays emitted from the central region of the item. Differences in DA and NDA arise primarily from the fact that, in DA, the sample is often modified (chemically treated) to match the analysis technique, allowing for more control of the measurement conditions, while, in NDA, the analysis technique is modified to match the item and measurement conditions. This means that NDA requires process knowledge in order to determine some components of the uncertainty, while DA requires process knowledge in order to prepare and measure the sample. Also, standards used in DA are much closer to the sample being measured than in NDA because it is not feasible to prepare a set of standards (isotopics, matrix, packaging, etc.) to fit all NDA measurement regimes. The DA and NDA communities both estimate uncertainty associated with certified standards [17]. And, both communities endorse sample exchange programs in which multiple laboratories measure the same measurand, providing data for a "top-down" approach to UQ. This paper is concerned with a "bottom-up" approach to UQ for NDA, where each estimated quantity in the assay procedure is

assessed for its contribution to the estimate of the overall uncertainty.

NDA is used for many material types, including well-characterized and consistent product material, and poorly characterized and inconsistent scrap and waste. Particularly for the less well-characterized and/or inconsistent material types, some type of model is used to adjust radiation count rates as in (1). Therefore, uncertainty in the model itself (how well the item conforms to the model assumptions) can be an important and difficult-to-characterize source of uncertainty.

UQ for NDA typically needs to allow for both an overall bias and for an item-specific bias, as well as include the catch-all "random" error. Therefore, one of the simplest but most useful error models is

$$M_{ij} = T_i + S + S_i + R_{ij}, \tag{2}$$

where M_{ij} is the jth measurement on the ith item, T_i is the unknown true value, S is the overall bias, S_i is the item-specific bias, and R_{ij} is the random error [10]. Although not shown explicitly, the GUM [5] endorses a reduced version of (2) in top-down UQ (not our focus here), given by $M_{ij} = T_i + S + R_{ij}$, which redefines R to include S_i and the R in (2). Because item-specific systematic error S_i propagates across items in the same way that random error does, this is sometimes adequate. However, it has been demonstrated that (2) is needed in some NDA settings [8, 10, 18]. To complete specification of the measurement model given by (2), one further assumes that S, and the variance of S_i, denoted σ_S^2, and the variance of R_{ij}, denoted σ_S^2, can be estimated from data [8, 10] in top-down UQ.

Model uncertainty impacts S and S_i in (2). One component of model error is model parameter estimation error, which is addressed in the EMP case study. Uncertainty in nuclear data, such as attenuation coefficients and emission intensities, is a special case of model parameter estimation error, which is addressed in [19]. Also addressed in [19] is model error itself. For example, model-based estimates of detector response functions derived, for example, from a radiation transport model such as MCNP [12], are used in one option to infer the relative abundance of the isotopes of Plutonium in samples using gamma spectroscopy. References [6, 11] also consider model error in the simpler setting of fitting multiple candidate models to the same calibration data.

3. The GUM

In metrology, uncertainty is a parameter that characterizes the dispersion of the *estimates* of a true quantity known as the measurand, and the GUM describes one main approach to estimate uncertainty. The GUM did not attempt to be comprehensive, and so, it is not surprising that subsequent specialized supplements have been developed, mostly influenced by UQ needs for DA. In the case of NDA, there are ASTM guides for every commonly used NDA method. The GUM, its eight technical appendices, and the supplements to the GUM are too lengthy to fully review here. But briefly, the main technical tool is a first-order Taylor approximation to the measurand equation

$$Y = f\left(X_1, X_2, \ldots, X_p\right), \tag{3}$$

which relates input quantities $X_1, X_2, \ldots, X_p$ to the measurand Y. Some of the input quantities can be estimates of other measurands, or of calibration parameters, so the measurand equation is quite general. Note that (3) does not include model error, which is sometimes needed [4, 11, 18, 19]. Also, note that (3) is aimed primarily at bottom-up UQ, using either steps in the assay method and uncertainties in the quantities $X_1, X_2, \ldots, X_p$ or using calibration data (see the EMP case study in Section 4). However, supplements to the GUM describe analysis of variance in the context of top-down UQ using measurement results from multiple laboratories and/or assay methods to measure the same measurand. The GUM does not explicitly present any measurement error models such as (2) but only considers the model for the measurand (3). However, the GUM endorses the notion of a measurement error model such as (2) in its top-down UQ. Note that (1) can be expressed as a model for the measurand, $M = \mathrm{CR}/f(X_1, X_2, \ldots, X_p) + R$, by algebraic rearrangement and redefining R so that it can be included among the X_i and also including the measured CR among the X_i. Although it is beyond our scope here, one could conceivably impose the effects of model error and/or measurement bias in the probability distributions for some of the X_i and then reexpress (3) in terms of a measurement error model such as (2).

The purpose of a measurement is to provide information about the measurand, such as the SNM mass. Both frequentist and Bayesian viewpoints are used in estimating the measurand and in characterizing the estimate's uncertainty. Elster [4] and Willink [20] point out that the GUM invokes both Bayesian and frequentist approaches in a manner that is potentially confusing. To modify the GUM so that a consistent approach is taken for all types of uncertainty, [3] suggests an entirely frequentist approach while others suggest an entirely Bayesian approach. Bich [3] also points out confusion between frequentist and Bayesian terminology and approaches in the GUM, which is one reason he believes it would be useful to revise the GUM. No matter which approach is used, making it clear which quantities are viewed as random and which are viewed as unknown constants will avoid needless confusion. However, the real challenges involve choosing likelihood for the data, a model to express how the measurand is estimated, and a model to describe the measurement process. In NDA, a model is often used to adjust test items to calibration items. These challenges are present in both frequentist and Bayesian approaches.

Ambiguities in the GUM arise for at least three reasons [3, 4, 20]: (1) The GUM divides the treatment of errors into those evaluated by type A evaluation (traditional data-based empirical assessment) and those addressed by type B evaluation (expert opinion, experience with other similar measurements). However, type B evaluations are primarily Bayesian (degree of belief) without explicitly stating so (and need not be), while type A evaluations are primarily

frequentist (and need not be). The jargon used in describing type B evaluations implies that the true value T has a variance (a Bayesian view based on quantification of our state of knowledge). The jargon used in describing type A evaluations is frequentist, with statements such as $P(X-T > k_1\sigma) = 0.05$, with the interpretation that X varies randomly around the fitted quantity T, where σ is the known measurement error standard deviation. We endorse either view, when clearly explained, but typically write $P(X - T > k_2\widehat{\sigma}) = 0.05$, where the hat notation conveys that the standard deviation is an unknown parameter that must be estimated, so $k_2 > k_1$. (2) The GUM uses the same symbol X for a measurement result and for a true value, which also confuses the frequentist and Bayesian views. (3) There is vague use of the term "quantity." And, although the GUM attempted to clarify confusion between "error" and "uncertainty," it did not clearly use the term "error" when measurement error (which has a sign, positive or negative) was meant. Willink [20] aims to resolve these ambiguities by paying attention to notation and jargon, being careful to separate Bayesian from frequentist views, and pointing out a confusion of true values with measurements of true values. A recent issue in Metrologia devoted several papers (see, e.g., [3, 4]) to the success of the GUM and to features of the next version(s) of the GUM. Clearly delineating what terms are random and what terms are fixed but unknown is important. Also, the GUM does not explicitly address calibration; however, because calibration is almost never a completely straightforward application of ordinary regression, we agree with [4] that UQ for calibration deserves attention, as we illustrate next.

4. Case Study: Enrichment Meter Principle

4.1. EMP Description. The enrichment meter principle (EMP) aims to infer the fraction (enrichment) of ^{235}U in U by measuring the count rate of the strongest-intensity direct (full-energy) gamma from decay of ^{235}U, which is emitted at 185.7 keV [2, 18]. The EMP assumes that the detector field of view into each item is identical to that in the calibration items, that the item must be homogeneous with respect to both the ^{235}U enrichment and chemical composition, and that the container attenuation of gamma-rays is identical or at least similar to that in the calibration items [2] so that empirical correction factors have modest impact and are reasonably effective. If these three assumptions are met, the known physics implies that the enrichment of ^{235}U in the U is directly proportional to the count rate of the 185.7 keV gamma-rays emitted from the item; and, it has been shown that, under good measurement conditions, the EMP can have a random error relative standard deviation of less than 0.5% and bias of less than 1%, depending on the detector resolution, stability, and extent of corrections needed to adjust items to calibration conditions.

4.2. EMP Calibration. Calibration is performed using standard certified reference materials of "known" enrichment. Here, "known" is in quotes because both NDA and DA communities provide uncertainty statements for primary standard reference materials [17]. Typically "uncertainty" is defined as the total error (random plus systematic) standard deviation, of an item (a test item or a standard reference material) and is sometimes expressed as $\sigma_{\text{Total}} = \sqrt{\sigma_R^2 + \sigma_S^2}$. Corrections are made for attenuating materials between the uranium-bearing material and the detector and for chemical compounds different from the reference materials used for calibration. Detectors of any resolution (such as scintillators or semiconductors) can be used, and the EMP method can be used for the entire range of ^{235}U fraction (enrichment) as a weight percent, from 0.2%^{235}U to 97.5%^{235}U.

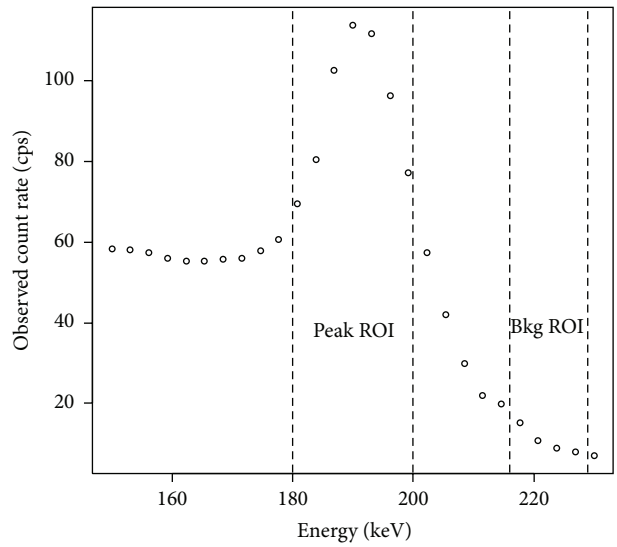


Figure 1: Example counts near the 185.7 keV line from a NaI detector.

There are several analysis options for EMP data. First, regarding the measurement itself, one must choose how to estimate the net peak area count rate associated with 185.7 keV gamma-rays. Then, one must choose how to deal with errors in the estimated count rate of 185.7 keV gamma-rays during both calibration and testing. Finally, one must determine the impact on uncertainty of model departures, such as variations in attenuation of the 185.7 keV gamma-rays due to variations in container thicknesses, for example.

4.3. EMP Examples. Regarding the measurement data, Figure 1 plots the counts in energy bins near the 185.7 keV energy from an item measured at Los Alamos National Laboratory using a relatively low-resolution hand-held NaI detector. Notice that the peak occurs at an apparent energy slightly higher than 185.7 keV. This is not unusual; it can occur because of energy calibration drift or because of interfering gamma energies from other isotopes slightly above 185.7 keV. One could use some type of peak fitting or background fitting to improve the estimate of the 185.7 keV count rate. In our analyses here, we use an option that fits the known enrichment in each of several standards to observed counts in a few energy channels near the 185.7 keV energy as the "peak" region and to the counts in a few energy channels just below

and just above the 185.7 keV energy to estimate background, expressed as

$$Y = \beta_1 X_1 + \beta_2 X_2 + R, \tag{4}$$

where Y is the enrichment, X_1 is the observed peak count rate near 185.7 keV, X_2 is the observed background count rate in a few neighboring energy channels near the 185.7 keV peak region, and R is random error (see Figure 1). Calibration data is used to estimate β_1 and β_2. One could constrain the estimates $\widehat{\beta}_1$ of β_1 and $\widehat{\beta}_2$ of β_2 to be equal in magnitude in the case where the same number of energy channels is used for both the peak and background, that would correspond to assuming a constant (nonsloping) background throughout the peak region, which does not appear to be appropriate for data such as in Figure 1. Therefore, in this example, we do not force the constraint $\widehat{\beta}_1 = -\widehat{\beta}_2$.

Note that (4) expresses Y as a function of the measured X values and thus avoids some of the issues that arise in the errors in predictors literature [14–16, 18]. In contrast, one could write a model using the corresponding true, unknown values, $Y = \alpha_1\mu_1 + \alpha_2\mu_2 + R$, where μ_1 is the true peak count rate and μ_2 is the true background count rate. Note that we replaced parameters β_1 and β_2 in (4) with parameters α_1 and α_2, because the model parameters are different when there is error in the predictors. Then, if there was interest in the calibration parameters α_1 and α_2, the errors in X_1 and X_2 would be included in the analysis to estimate α_1 and α_2. However, if the main interest is estimating calibration parameters to best estimate Y, then (4) is the preferred approach, as we use here (see Section 4.4).

In order to investigate how to deal with errors in the estimated count rates, data were collected at Los Alamos National Laboratory using a NaI detector for each of five standards having nominal true values of $Y = 0.3206$, 0.7209, 1.9664, 2.9857, and 4.5168 weight percent, each with assigned standard deviation (due to uncertainty in the nominal values) σ_Y (denoted σ_{Total} above) of approximately 0.00148. The actual standard deviations varied slightly around 0.00148, but to simplify here, we assume each standard's nominal value has the same standard deviation, 0.00148. Two 300-second count time repeats were made of each standard, with peak count rates in counts per second of {60.29, 59.51}, {72.81, 73.26}, {112.29, 113.47}, {143.53, 144.32}, and {194.79, 194.52}, respectively, and background ("continuum") count rates of {33.36, 33.80}, {33.48, 32.91}, {33.17, 33.26}, {32.82, 32.88}, and {32.88, 33.44}, respectively.

Using the average of each of the 5 pairs of observed count rates based on 300-second count times and ignoring errors in these count rates, the ordinary least squares estimates of model parameters are $\widehat{\beta}_1 = 0.031$ and $\widehat{\beta}_2 = -0.046$. Recall that if the background shapes were flat near the peak, we would expect $\widehat{\beta}_1$ to be more nearly equal to $\widehat{\beta}_2$ in magnitude. There is no evidence of non-Poisson behavior in these data; the ratio of the variance to the mean of the counts is not statistically significantly different from 1 on the basis of a t-test. However, the RMSE from the fit to (4) is 0.02, which is much larger than the observed standard deviation (0.00148), indicating a lack of fit to the model for Y if the only error source was uncertainty in the nominal values of the five standards.

To confirm that an RMSE of 0.02 is evidence of a lack of fit to the assumed model, we first simulated 10^4 realizations of data from (4) with no error in X_1 or X_2 and found that the 99% quantile for the simulated RSDs is 0.0028, which is much smaller than the observed RMSE of 0.02. An additional set of 10^4 simulations confirmed that the lack of fit is at least partly explainable by estimation error in X_1 and X_2. We simulated the impact of counting for various times, and in counting for 300 seconds, the 0.99 quantile for the RMSE increases substantially, from 0.0028 to 0.048. In counting for 600 seconds, the 0.99 quantile for the RMSE is 0.034. There could be other sources of lack of fit in the real data, such as nonlinearity, but when we replace the observed Y with fitted Y values plus random errors with standard deviation equal to the observed standard deviation of 0.00148 (in a separate set of 10^4 realizations), forcing exact linearity as implied by (4), the RMSE values are not changed significantly. Therefore, the simulation shows that counting for 300 or 600 seconds is not sufficiently long for the errors in X_1 and X_2 to be negligible. From a practical standpoint, these count times are typical in EMP calibration. All simulations and analyses were performed in R [21].

Figure 2 plots the RMSE in estimating y for a range of count times (using the same count times in training and testing, for illustration) for each of three estimation methods. The RMSE in Figure 2 is the average root mean squared error in estimating y over a grid of 100 equally spaced test y values spanning the calibration range. The bottom plot (b) is the same as (a) but covers a wider range of count times. Estimation Method 1 is the observed RMSE in the 10^4 simulations, so it is essentially the "true" RMSE. As explained below, the assay has item-specific bias due to uncertainty in $\widehat{\beta}_1$ and $\widehat{\beta}_2$, but Method 1 estimate of the RMSE has negligible error. Methods 2 and 3 are more fully described in Section 4.4, but briefly, Method 2 is the estimated RMSE based on variance propagation that accounts for the covariance $\text{cov}(\widehat{\beta}_1, \widehat{\beta}_2)$, and Method 3 is the estimated RMSE, also based on variance propagation, but assuming, as does [2], that the covariance $\text{cov}(\widehat{\beta}_1, \widehat{\beta}_2)$ is the 0 matrix (i.e., ignoring error in $\widehat{\beta}_1$ and $\widehat{\beta}_2$, but considering only the error in X_1 and X_2) Method 3 clearly underestimates the true RMSE. Method 2 is highly accurate for the true RMSE at count times of approximately 10 seconds or longer; although the expression $(X^T X)^{-1}\sigma^2$ underestimates $\text{cov}(\widehat{\beta}_1, \widehat{\beta}_2)$ because of the effect of errors in X, Method 2 overestimated the true RMSE at lower count times (10 seconds or less) in all cases that we considered.

4.4. EMP Calibration Count Times. Considering the numerical results in Section 4.3, a practical question is whether 300 seconds is a reasonable count time. The shape of the RMSE curve in Figure 2(b) suggests that we should require approximately 300 seconds of count time or more, so that the RMSE is acceptably low and cannot be made much lower and so that Method 2 reliably estimates the true RMSE. Because there are not many calibration standards,

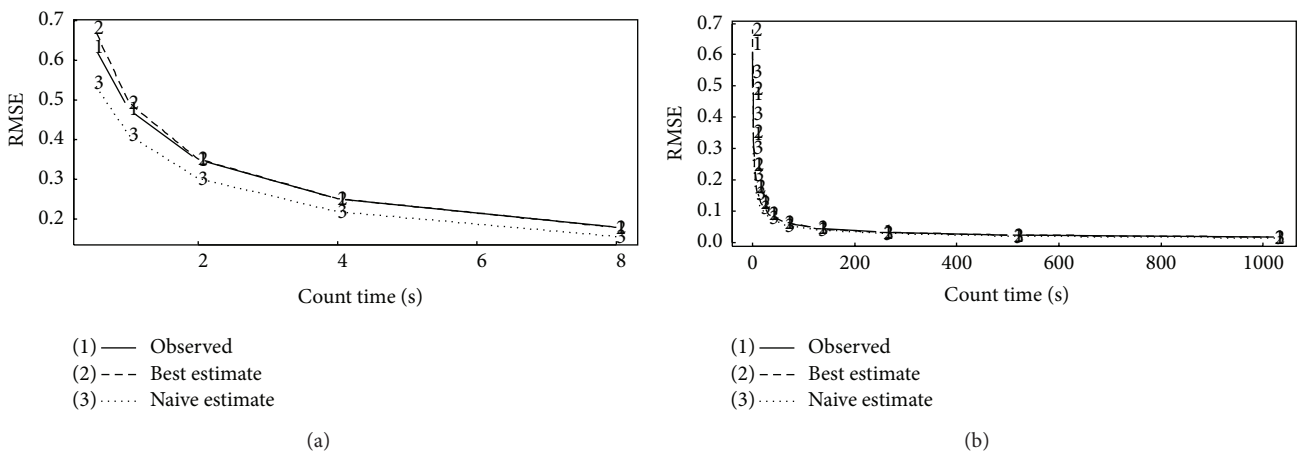


Figure 2: The RMSE versus count time.

one might expect to require the calibration protocol to count for sufficiently long time that the impact of errors in predictors is negligible during calibration. Although it would be desirable if estimation error in $\widehat{\beta}_1$ and $\widehat{\beta}_2$ was essentially all due to the limited number of standards used in the calibration (as in typical regression applications), Figure 2 and our analyses indicate that some estimation error in $\widehat{\beta}_1$ and $\widehat{\beta}_2$ will also be due to error in X_1 and X_2, even for reasonable count times such as 300 seconds per standard. However, in practice, one would use the RMSE in the estimated ys as calculated from the calibration data (0.02 in this example) to estimate the RMSE in estimating future y values (lower case y denotes the unknown true value); then, the estimated RMSE using estimation Method 2 will agree with the true RMSE, although there is a detectable difference from the zero-error-in-predictor situation (Figure 2).

Reference [2] does not provide guidance regarding what is a sufficiently long count time to ignore errors in X_1 and X_2. References [14, 15] provided rough guidance regarding when one can ignore errors in predictors. That guidance involves the ratio of the error standard deviation in any predictor compared to its range, but only in the case where the true model parameters β are the quantities of interest. Our recommendation consists of two parts. First, accept lack of fit compared to the zero-error-in-predictor case (RMSE in the fit of Y to X_1 and X_2 of 0.02 versus the expected RMSE of approximately 0.00148 if there was no model error). Second, update [2] using the following strategy to select count time for calibration items and for bottom-up UQ for EMP calibration:

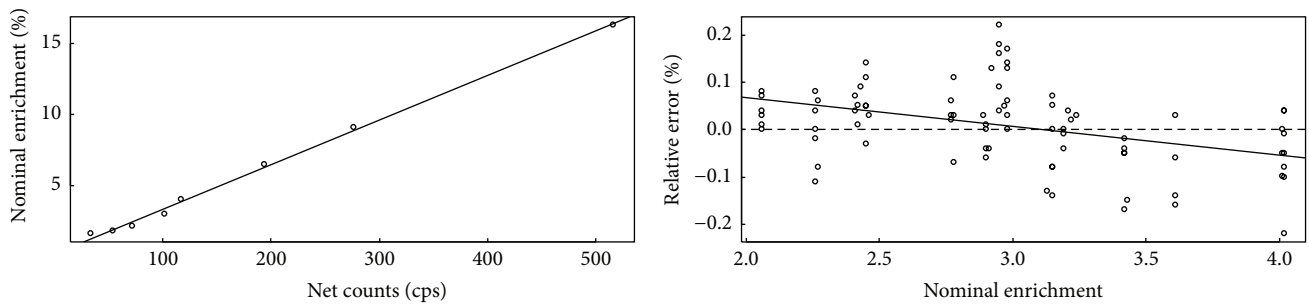
(1) Collect counts from each standard for approximately 300 seconds. Estimate μ_1 and μ_2 using the observed count rates, denoted $\widehat{\mu}_1$ and $\widehat{\mu}_2$, respectively.

(2) Assume $X_1 \sim \text{Poisson}(\widehat{\mu}_1)$ and $X_2 \sim \text{Poisson}(\widehat{\mu}_2)$.

(3) Compare the estimated variance based on $\text{var}(\widehat{\beta}_1 X_{1,\text{test}} + \widehat{\beta}_2 X_{2,\text{test}})$ to the observed variance in simulated Poisson data from step (2). Increase the count time CT_a in the calibration until the estimated variance is statistically indistinguishable from the observed variance in the simulated data. The ASTM standard for EMP [2] illustrates simple variance propagation to estimate $\text{var}(\widehat{\beta}_1 X_{1,\text{test}} + \widehat{\beta}_2 X_{2,\text{test}})$ but ignores the 2-by-2 covariance matrix $\text{cov}(\widehat{\beta}_1, \widehat{\beta}_2)$ given by the well-known result $\text{cov}(\widehat{\beta}_1, \widehat{\beta}_2) = (X^T X)^{-1}\sigma^2$, where X is the matrix with five rows and two columns containing the five X_1, X_2 pairs in calibration data. The expression for $\text{var}(\widehat{\beta}_1 X_{1,\text{test}} + \widehat{\beta}_2 X_{2,\text{test}})$ is known to be $\text{var}(\widehat{\beta}_1 X_{1,\text{test}}) + \text{var}(\widehat{\beta}_2 X_{2,\text{test}}) + 2\text{cov}(\widehat{\beta}_1 X_{1,\text{test}}, \widehat{\beta}_2 X_{2,\text{test}})$. The quantities $\widehat{\beta}_1$ and $X_{1,\text{test}}$ are independent random variables; therefore, $\text{var}(\widehat{\beta}_1 X_{1,\text{test}}) = \text{var}(\widehat{\beta}_1)\mu^2_{X_{1,\text{test}}} + \text{var}(X_1)\mu^2_{\widehat{\beta}_1} + \text{var}(\widehat{\beta}_1)\text{var}(X_1)$. Similarly, the other two terms in $\text{var}(\widehat{\beta}_1 X_{1,\text{test}} + \widehat{\beta}_2 X_{2,\text{test}})$ are easily calculated. Also, the formula to estimate the covariance between $\widehat{y}_1$ and $\widehat{y}_2$ from two test items is similar to that just described for $\text{var}(\widehat{\beta}_1 X_{1,\text{test}} + \widehat{\beta}_2 X_{2,\text{test}})$.

Regarding step (1), for completeness, one can demonstrate that we can ignore estimation error in $\widehat{\mu}_1$ and $\widehat{\mu}_2$ that are used in the subsequent simulations in steps (2) and (3) that produced Figure 2 by repeating steps (2) and (3) for another set of $\widehat{\mu}_1$ and $\widehat{\mu}_2$ values that are obtained from real data or from synthetic Poisson-generated data.

Next, calculate the observed variance in simulated Poisson data from step (2). Increase the count time to CT_b in the calibration until the estimated variance does not decrease substantially. Use the larger of this two count times, CT_a and CT_b. Current practice is to count each standard for at least 300 seconds. We do not know the origin of this suggested count time, but it appears to be quite reasonable in view of the criteria presented here. We emphasize that the chosen count time for each standard is related to collimation and detector dimensions; therefore, our example is specific to this particular data collection example, but the approach is general. Calibration following [2] would, in this example, use the average of each of the 5 pairs of 300-second count rates,



(a) Calibration data

(b) Field data

Figure 3: EMP data from [14]. Plot (a) is the nominal enrichment versus the net counts, with the fitted line shown. Plot (b) is the relative error versus the nominal enrichment for 82 test items. The fitted line is the least squares fit to the relative errors.

ignore errors in these count rates, and use ordinary least squares (or weighted least squares if the standards nominal values are assigned unequal variances) to estimate β_1 and β_2.

A simulation study such as just presented can be used to check whether 300 seconds is an adequately long count time. And, [2] should be modified to include the effect of $\mathrm{cov}(\widehat{\beta}_1, \widehat{\beta}_2)$ in estimating $\mathrm{var}(\widehat{\beta}_1 X_{1,\mathrm{test}} + \widehat{\beta}_2 X_{2,\mathrm{test}})$. This strategy cannot claim to result in a situation in which impact of errors in predictors is negligible during calibration, because the RMSE for y is inflated compared to the zero error in predictors case. However, errors in predictors are ignored in the sense that there is no adjustment of parameter estimates to account for errors in predictors [13–16, 18]. For completeness, we note that adjustment for errors in predictors is to choose values of $\widehat{x}_i$ (to estimate $x_{i,\mathrm{true}}$) and α_1, α_2 to minimize

$$\mathrm{RSS}_1 = \sum_{i=1}^{n_{\mathrm{train}}} \frac{(x_i - \widehat{x}_i)^2}{\sigma_{x_i}^2} + \frac{(y_i - \widehat{y}_i)^2}{\sigma_{y_i}^2}, \tag{5}$$

where $\widehat{y} = \widehat{\alpha}_1 \widehat{\mu}_1 + \widehat{\alpha}_2 \widehat{\mu}_2 = \widehat{\alpha}_1 \widehat{x}_{1,\mathrm{true}} + \widehat{\alpha}_2 \widehat{x}_{2,\mathrm{true}}$ is used to calculate $\widehat{y}_i = \widehat{y}_{i,\mathrm{true}}$ in the second term. But, in the EMP context, such an adjustment would actually increase the errors in the estimated y values, so is not recommended. If there is no adjustment for errors in predictors, then β_1, β_2 are chosen that minimize

$$\mathrm{RSS}_2 = \sum_{i=1}^{n_{\mathrm{train}}} \frac{(y_i - \widehat{y}_i)^2}{\sigma_{y_i}^2}, \tag{6}$$

where $y = \alpha_1 \mu_1 + \alpha_2 \mu_2 = \alpha_1 x_{1,\mathrm{true}} + \alpha_2 x_{2,\mathrm{true}}$. Also, here we assume that the weights $\sigma_{y_i}^2$ are known.

4.5. Item-Specific Bias in EMP Measurements. Next, we address item-specific bias. Container wall thickness attenuation correction is one source of item-specific bias. In addition, standard calibration using long enough count times to ignore errors in X_1 and X_2 in training data leads to item-specific bias due to the effect of fitting errors. This is easy to visualize in the case of fitting a slope and intercept to scalar data. Items at one end of the calibration range tend to be off in the same direction (positive or negative) from the true value. Therefore, trends are often observed in residuals associated with this effect. Figure 3 is an example using EMP data from 82 test items reported in [22], who fit $Y = \beta_1 + \beta_2 X + R$, where X was the net counts associated with the 185.7 keV gamma. Notice that this is a different model than the one we have used, $Y = \beta_1 X_1 + \beta_2 X_2 + R$ (note that we use the same symbol R for the random error term in both models but point out that the error distribution for R is not the same in both models). In Figure 3(b), there is a clear trend in the residuals, from positive to negative as the enrichment increases from 2 to 4 percent. The fact that error in the fitted parameters leads to such trends is a well-known statistical fact [13–16, 18].

Figure 4 plots the results of a simulation study that shows such trends in residuals can be explained to various degrees, depending on which model was fit, even without any errors in predictors, as was the case in the simulation. To visualize possible patterns in residuals from ordinary regression fitting, we used parameters from similar models fit to the data in Figure 3. Figures 4(a) and 4(b) assumed the true model had a slope and intercept and (a) fit a slope and intercept or (b) fit a slope only. Figures 4(c) and 4(d) assumed the true model had a slope but zero intercept, and (c) fit a slope and intercept or (d) fit a slope only. There are clear patterns in these residuals, and [22] fits a slope and intercept, so Figure 4(a) or Figure 4(c) is to be compared to Figure 3(b). There is item-specific variance, but it is entirely attributable to calibration. These types of residual patterns are predictable from ordinary regression theory without errors in predictors, but in our experience, it is helpful to visualize simulated data results that confirm our understanding.

From (2), one seeks to estimate $\sigma_S^2 = \mathrm{var}(S_i)$, the variance in the measurement errors in test items that arises because the items differ somewhat from the calibration items. In addition to the calibration-error-induced effects just illustrated, a common source of nuisance variation is varying container thickness, which leads to attenuation of the counts at different rates than in the calibration items. The attenuation correction factor depends on fundamental nuclear data (the attenuation coefficient of the item container), on the container thickness as measured by an ultrasonic gauge and on the form used for the correction to adjust test item container thickness to calibration item container thickness.

To evaluate the impact of attenuation effect, we multiply $X_{1,\mathrm{test}}$ and $X_{2,\mathrm{test}}$ by a random scale factor f (the peak

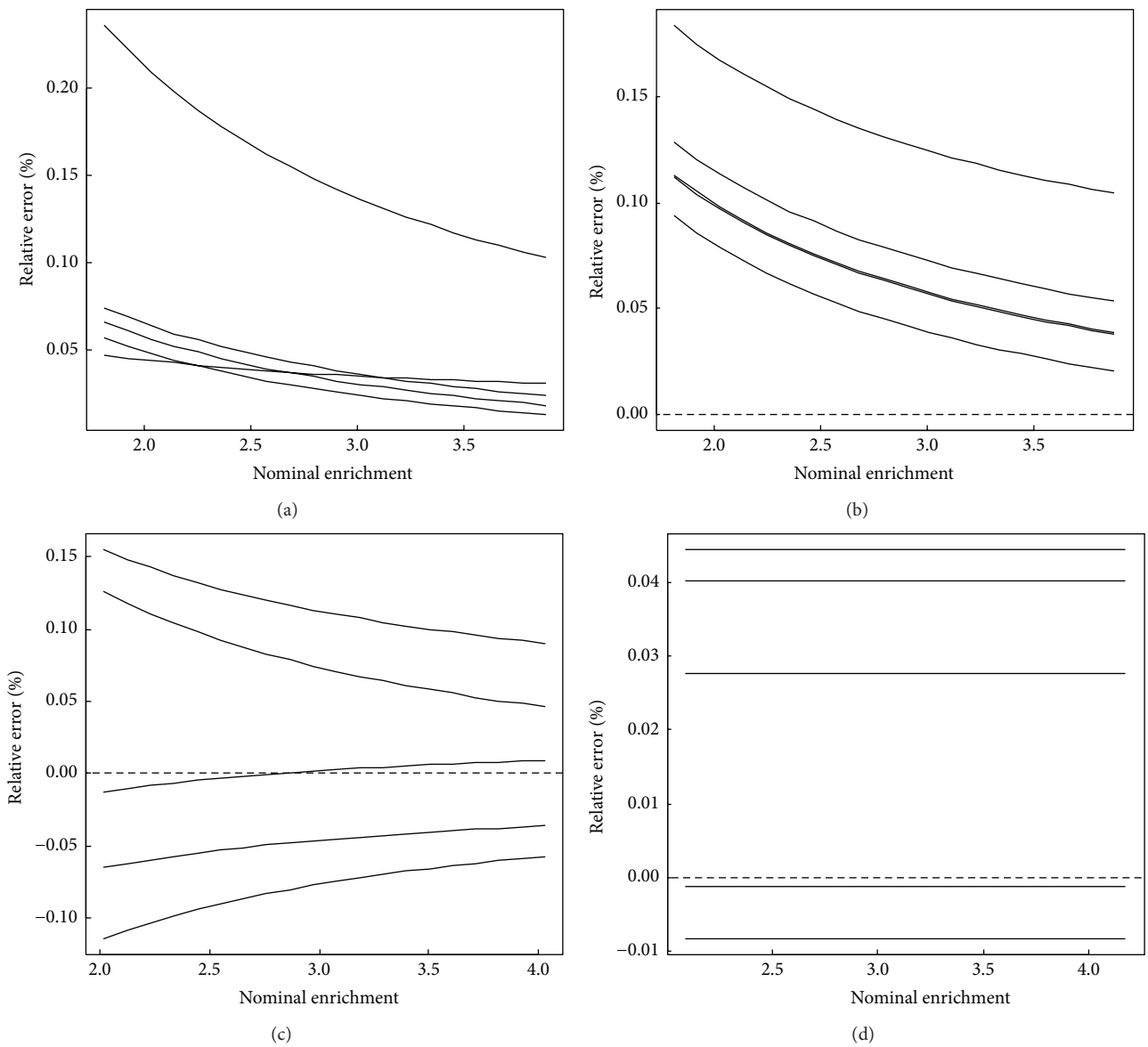


Figure 4: Fits to residuals. Top: the correct model is intercept and slope and (a) fit has intercept and slope or (b) fit has slope only. Bottom: the correct model is slope only, and (c) fit has intercept and slope, and (d) fit has slope only.

and background energy regions are sufficiently close that we assume both have the same scale factor f). The random scale factor f has a mean value of 1 and standard deviation determined by the quality of the attenuation correction factor. We then repeat the simulation strategy described above, using $X_{1,\text{test}} * f$ and $X_{2,\text{test}} * f$ in place of $X_{1,\text{test}}$ and $X_{2,\text{test}}$, respectively. Figure 5 plots example measurement errors, in an example using the previous calibration results applied to both the calibration data and to held-out test data that spans the enrichment range to nearly 100%. The test data included UF_6 cylinders and U_3O_8 cans in [2]. The solid vertical lines extend to $\pm 2\text{RMSE}_1$, where RMSE_1 is calculated as described previously (using Method 2). The dashed vertical lines extend to $\pm 2\text{RMSE}_2$, where RMSE_2 is also calculated as described previously (and verified by simulation), but the variance of X_1 and X_2 is increased to account for 20% estimation error in f (a nominal value, used here just as an example). The RMSE_2 is larger than RMSE_1 but is not much larger in this example, even with the relatively large 20% estimation error standard deviation in the estimate of f. Notice that 3 of the 13 test items exceed the larger limits of $\pm 2\text{RMSE}_2$. Evidently there are other unknown error sources.

4.6. EMP Case Study Summary. To summarize the EMP case study, for a test item one can write the measurand equation as $Y_{\text{Test}} = f(\widehat{\beta}_1, \widehat{\beta}_2, X_{1,\text{test}}, X_{2,\text{test}}) = \widehat{\beta}_1 X_{1,\text{test}} + \widehat{\beta}_2 X_{2,\text{test}}$. Elster [4] points out that the GUM's measurand equation does not allow for exploration of other models (such as using only one calibration parameter multiplied by the net estimate 186 keV

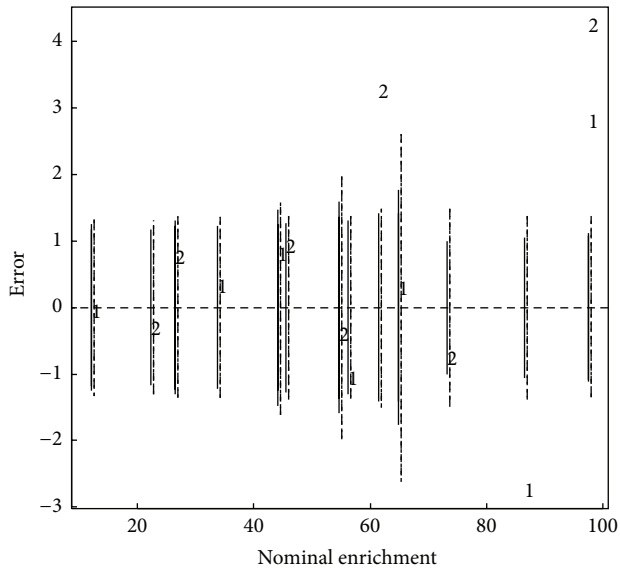


Figure 5: Results on EMP test data that included UF_6 cylinders (5A cylinders) and U_3O_8 cans (Z cans) in [2]. The solid vertical lines extend to $\pm 2\text{RMSE}_1$. The dashed vertical lines extend to $\pm 2\text{RMSE}_2$.

counts) or the use of metrics other than minimized squared error to estimate $\widehat{\beta}_1$ and $\widehat{\beta}_2$. The GUM is in the process of being revised [3], so perhaps calibration will be treated more completely in the next GUM version. For comparison, see (3) for the GUM notation, where, for example, X_1 is a generic quantity with uncertainty that could be a derived quantity such as $\widehat{\beta}_1$ in our example. Once the user confirms that a sufficiently long count time was used in calibration using a simulation strategy such as that described, one can modify [2] by adding the impact of nonzero $\text{cov}(\widehat{\beta}_1, \widehat{\beta}_2)$ to evaluate $\text{var}(y_{\text{test}})$. Errors in $X_{1,\text{test}}$ and in $X_{2,\text{test}}$ will be present due to modest count times for test items and due, for example, to imperfect adjustments for container thickness. We recommend that an updated version of [2] describes how to account for $\text{cov}(\widehat{\beta}_1, \widehat{\beta}_2)$, how to use simulation to determine appropriate count times in calibration, and how to estimate item-specific systematic error variance using calibration and test data.

5. Conclusions and Outlook

This paper described challenges in UQ for NDA, some of which are addressable using the GUM's concept of a measurement equation. A case study was presented involving the EMP, which is among the simplest NDA techniques (the EMP has a proportionate response, and well-established first principles), yet the UQ portion of the ASTM [2] for the EMP needs to be updated, as we illustrated. The two main updates are to allow for uncertainty in the fitted calibration parameters and to provide practitioners with a strategy to choose a count time that is sufficiently long that it is acceptable to ignore the effect of measurement errors in predictors (peak and background gamma count rates) in the uncertainty in the fitted calibration parameters. References [3, 4, 20] all indicate ways that the GUM can be improved, some of which were discussed in Section 3. One need for improvement in the GUM was illustrated in the EMP case study. The EMP case study involves calibration with errors in predictors, which is not described in the current GUM.

Conflict of Interests

The authors declare no conflict of interests.

Authors' Contribution

Tom Burr analyzed the real data and the simulated data in Section 4. Tom Burr, Stephen Croft, and Ken Jarman wrote the paper. Stephen Croft and Ken Jarman contributed equally to this work.

Acknowledgments

The authors gratefully acknowledge the Office of Defense Nuclear Nonproliferation Research and Development in the National Nuclear Security Administration and the Atomic Energy Agency.

References

[1] R. T. Kouzes, E. R. Siciliano, J. H. Ely, P. E. Keller, and R. J. McConn, "Passive neutron detection for interdiction of nuclear material at borders," *Nuclear Instruments and Methods in Physics Research, Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, vol. 584, no. 2-3, pp. 383–400, 2008.

[2] *Standard Test Method for Measurement of 235U Fraction Using the Enrichment Meter Principle*, C 1514-08, ASTM.

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

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

[5] Guide to Expressing Uncertainty in Measurement, JCGM 100:2008, 2008, http://www.bipm.org/en/about-us/.

[6] T. L. Burr, M. M. Pickrell, T. H. Prettyman, P. M. Rinard, and T. R. Wenz, "Data mining: applications to nondestructive assay data," *Journal of the Institute of Nuclear Materials Management*, vol. 27, no. 2, pp. 40–47, 1999.

[7] T. Burr, H. Trellue, S. Tobin et al., "Integrated nondestructive assay systems to estimate plutonium in spent fuel assemblies," *Nuclear Science and Engineering*, vol. 179, no. 3, pp. 321–332, 2015.

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

[9] S. Croft and T. Burr, "Calibration of nondestructive assay instruments: an application of linear regression and propagation of variance," *Applied Mathematics*, vol. 5, no. 5, pp. 785–798, 2014.

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

[11] T. Burr, J. Dowell, T. Trellue, and S. Tobin, "Measuring the effects of data mining on inference," in *Encyclopedia of Information Science and Technology*, IGI Global, 3rd edition, 2015.

[12] *Monte Carlo N-Particle Code*, Los Alamos National Laboratory, Los Alamos, NM, USA, 2009, http://mcnp.lanl.gov/.

[13] R. Carroll, D. Ruppert, and L. Stefanski, *Measurement Error in Nonlinear Models*, Chapman & Hall, London, UK, 1996.

[14] R. B. Davies and B. Hutton, "The effect of errors in the independent variables in linear regression," *Biometrika*, vol. 62, no. 2, pp. 383–391, 1975.

[15] W. A. Fuller, *Measurement Error Models*, John Wiley & Sons, New York, NY, USA, 1987.

[16] T. Burr, S. Croft, and B. C. Reed, "Least-squares fitting with errors in the response and predictor," *International Journal of Metrology and Quality Engineering*, vol. 3, no. 2, pp. 117–123, 2012.

[17] A. M. H. van der Veen and J. Pauwels, "Uncertainty calculations in the certification of reference materials. 1. Principles of analysis of variance," *Accreditation and Quality Assurance*, vol. 5, no. 12, pp. 464–469, 2000.

[18] T. L. Burr and P. L. Knepper, "A study of the effect of measurement error in predictor variables in nondestructive assay," *Applied Radiation and Isotopes*, vol. 53, no. 4-5, pp. 547–555, 2000.

[19] T. Burr, S. Croft, A. Hoover, and M. Rabin, "Exploring the impact of nuclear data uncertainties in ultra-high resolution gamma spectroscopy for isotopic analysis using approximate Bayesian computation," *Nuclear Data Sheets*, vol. 123, pp. 140–145, 2015.

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

[21] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2012, http://www.r-project.org/.

[22] G. Gundersen and I. Cohen, "Enrichment measurement in low enriched 235U fuel pellets," in *Proceedings of the 13th Annual Institute of Nuclear Materials Management Meeting*, Boston, Mass, USA, June 1972.