 https://doi.org/10.3011/ESARDA.IJNSNP.2022.2

Statistical error model-based and GUM-based analysis of measurement uncertainties in nuclear safeguards – a reconciliation

Oscar Alique[1], Yetunde Aregbe[2], Raffaele Bencardino[3], Robert Binner[4], Thomas Burr[5], Jeff. A. Chapman[6], Stephen Croft[6], Andy Fellerman[7], Thomas Krieger[8], Klaus Martin[9], Peter Mason[10], Claude Norman[4*], Thomas Prohaska[11], Divyesh Trivedi[12], Steve Walsh[13], Dariusz Wegrzynek[14, 15], Ben Wright[7], Jan Wüster[4]

[1]European Commission - Directorate Translation – Evaluation and Analysis, BECH C3, Rue Alphonse Weicker 5, 2721, Luxembourg, Luxembourg, [2]European Commission – Joint Research Centre (EC-JRC), Directorate G – Nuclear Safety & Security Unit G.2 - Standards for Nuclear Safety, Security & Safeguards, Retieseweg 111, B-2440 Geel, Belgium, [3]European Commission – Directorate General for Energy – EURATOM Safeguards EUFO3260, 10 rue R. Stümper L-2557 Luxembourg, [4]International Atomic Energy Agency - Nuclear Fuel Cycle Information Analysis Section- Division of Information Management - Department of Safeguards, Vienna International Centre, PO Box 100, 1400 Vienna, Austria, [5]United States Department of Energy, Los Alamos National Laboratory, Los Alamos, United States of America, [6]United States Department of Energy, Oak Ridge National Laboratory, Oak Ridge, Tennessee, United States of America, [7]National Nuclear Laboratory (NNL), B170, Sellafield, Seascale, Cumbria CA20 1PG, United Kingdom, [8]Forschungszentrum Jülich, Institute of Energy and Climate Research (IEK), Jülich, Germany, [9]Trier, Germany, [10]United States Department of Energy, NBL Program Office, Oak Ridge, Tennessee, United States of America, [11]Montanuniversität Leoben - Allgemeine und Analytische Chemie, Franz-Josef-Strasse 18, 8700 Leoben, Austria, [12]National Nuclear Laboratory (NNL), 5th Floor, Chadwick House, Birchwood, Warrington, WA3 6AE, United Kingdom, [13]Montana, United States, [14]International Atomic Energy Agency- Coordination and Support Section-Office of Safeguards Analytical Services- Department of Safeguards, Vienna International Centre, PO Box 100, 1400 Vienna, Austria, [15]AGH University of Science and Technology, Faculty of Physics and Applied Computer Science, al.Mickiewicza 30, 30-059 Krakow, Poland

*corresponding author E-mail: c.norman@iaea.org

Abstract

At the occasion of the Consultants Group Meeting held to review the "International Target Values 2010 for Measurement Uncertainties in Safeguarding Nuclear Material" [3], discussions between experts highlighted the need to improve communication between different safeguards measurement communities, e.g. laboratory analysts, non-destructive assay specialists, safeguards data evaluators, and to reconcile their approaches to estimating measurement uncertainties. The purpose of this paper is to contribute to reaching a common understanding of the terminology and methodologies used by different professional groups in the field of uncertainty quantification.

Keywords: Statistical error model, GUM, terminology, uncertainty, measurement, nuclear safeguards

Introduction

Safeguards implementation requires a statistical evaluation of declared and verified amounts of nuclear material quantities to assess whether the differences can be explained by measurement errors or if they warrant further investigation. For this reason, the analysis of measurement error variances in the operator's and inspector's measurement systems and the modelling of their propagation into relevant evaluation statistics was researched within the safeguards community in the 1970s and 1980s. The approaches developed and the associated terminology have been consistently used since and are currently undergoing review and enhancement. Technical progress towards the end of the 20th century pushed the performance of measurement technologies in many disciplines to their practical or even theoretical limits of applicability, while increasing international cooperation, culminating in deep global supply chains, required a commonly understood way to communicate measurement results and associated uncertainties. The metrological community responded to these needs by defining an international standard, the 'Guide to the Expression of Uncertainty in Measurement' (GUM). Since its first publication in 1995, this standardized approach of first-principles ("bottom-up") uncertainty quantification (UQ) and the associated terminology have been adopted by an increasing number of laboratories, including those where nuclear material (NM) samples from the fuel cycle are regularly analyzed and those where instruments for destructive analysis (DA) and non-destructive assay

 Alique, O., Yetunde, A., Bencardino, R., Binner, R., Burr, T., Chapman, J.A., Croft, S., Fellerman, A., Krieger, T., Martin, K., Mason, P., Norman, C., Prohaska, T., Trivedi, D., Walsh, S., Wegrzynek, D., Wright, B., & Wüster, J. (2022, June). Statistical error model-based and GUM-based analysis of measurement uncertainties in nuclear safeguards – a reconciliation. ESARDA Bulletin - The International Journal of Nuclear Safeguards and Non-proliferation, 64(1), 10-29. https://doi.org/10.3011/ESARDA.IJNSNP.2022.2

(NDA) of NM are developed and calibrated. The adoption of the GUM is not complete but will eventually lead to every reputable laboratory measurement result being metrologically traceable and accompanied by a defensible uncertainty statement. On the other hand, the NDA measurement community faces specific UQ challenges, such as incompletely controlled measurement conditions and item-specific biases. Error variance propagation is a key component of UQ using both analytical and Monte Carlo approaches, however there is no general NDA UQ guide analogous to the GUM. The need for more comprehensive bottom-up UQ for NDA, including model-based adjustments of test items to calibration items is recognized. The GUM recommends and, in its capacity as an international guide published by the Joint Committee for Guides in Metrology (JCGM) of the Comité International des Poids et Mesures (CIPM) – i.e. the International Committee for Weights and Measures - also prescribes, a different terminology from the one traditionally employed by some of the authors and specialists in the analysis of measurement errors in safeguards, who typically are educated in the field of statistics. Further, certain sections of the GUM, especially in its earlier versions, explicitly discourage references to concepts that are central to safeguards practice, such as the concept of the true value of a measurand and the concept of an error model that explicitly distinguishes between random and systematic errors.

This dual difference, both terminological and conceptual, complicates communication between professional communities interested in measurement uncertainty, such as the safeguards laboratories and the safeguards statistical evaluation services. Attitudes have ranged from a desire to explain and to convince the other community to a more or less benign mutual neglect, but recently the communities took the opportunity to learn from each other. With this article the authors wish to overcome the undesirable impediments to communication between the relevant professional communities by reconciling the safeguards statistical error model with the GUM-based analysis of measurement uncertainties. The GUM is mostly known for bottom-up UQ used by metrologists, but also includes information regarding top-down UQ often used by statisticians. On the other hand, safeguards evaluators focus on top-down UQ by analysis of paired operator-inspector data, but also make use of bottom-up UQ. Properly interpreted these approaches are complementary rather than contradictory and hold the promise of mutual interdisciplinary cross-fertilization. Motivated by this perspective, the equivalence of top-down paired data analysis as applied by the IAEA with GUM-inspired repeatability and reproducibility analysis has recently been demonstrated. The approaches are not expected to be completely unified, because the underlying objectives are different, but the potential benefits of convergence in areas of overlap are identified and steps towards such convergence are recommended.

Starting in Section 1 with an historical review of UQ in the field of safeguards and the parallel development of the GUM, the paper describes in Section 2 and 3 how UQ methodologies are respectively used by safeguards data evaluators and the DA and NDA laboratories. The purpose of Section 4 is to systematically compare and reconcile the methodologies and terminology used by these communities while Section 5 focuses on the statistical bases of UQ methodologies. The complementarity of their purposes and the mutual benefits of communication and convergence between the professional communities involved are underlined in Sections 6 and 7 which support the paper's conclusion.

1. Historical Developments and current situation

1.1 Safeguards at the IAEA and EURATOM

The International Atomic Energy Agency (IAEA) was established in 1957 as an independent intergovernmental organization in the United Nations system. Article III of the IAEA's statute provides the IAEA with the authority to apply safeguards on nuclear material and other specified items. The IAEA's Department of Safeguards' primary role is to deter the proliferation of nuclear weapons by detecting early any misuse of nuclear material or technology, and by providing credible assurances that States are honouring the obligations stemming from their safeguards agreements.

Also in 1957, the European Atomic Energy Community, or EURATOM, was established and exists next to the European Union as a separate legal entity. Article 77 of the EURATOM Treaty provides the European Commission safeguards system with the authority to ensure that nuclear materials are not diverted from their intended civil uses, while complying with safeguards obligations concluded with third states and international organisations such as the IAEA.

In order to detect diversion of declared nuclear material, nuclear material accountancy (NMA) is used as the basic safeguards measure. NMA is that part of a nuclear material safeguards program that consists of procedures and systems to perform nuclear material measurements, prepare and maintain accounts and records, and perform data analyses. Statistical analysis is an essential element of effective NMA, and over the past 50 years, highly specialized statistical procedures have been developed to address unique problems encountered in NMA and associated verification activities [1]:

- Recognition of multiple error sources in a material balance (e.g., sampling, instrument, analyst, environmental conditions).
- Estimation of variance components associated with each error source.

- Reconciliation of measurement results from different measurement systems and different laboratories, often obtained from independent samples taken at different times.
- Assurance of independent verification of inventories and balances.

The safeguards mandate to independently evaluate material balance differences including the operator-declared *Material Unaccounted For (MUF[1])*, *the Shipper-Receiver Differences (SRD[2]),* and the projected Difference between Operator declaration and Inspector verification D statistic (D), fundamentally depends on the estimates of the measurement error uncertainties associated with all nuclear material quantities that enter the material balance [2]. Because the uncertainties associated with MUF, SRD, and D are obtained by error propagation methodologies applied to estimates of measurement error uncertainties, independent evaluation can only be accomplished if methods are available that allow safeguards evaluators to obtain such estimates.

Hence, early in the history of safeguards a need arose for specialized statistical procedures to estimate measurement uncertainties associated with nuclear material quantities, for both the operators' declared values and for the inspectors' verification results (for verifications by both DA and NDA). These estimates needed to be independent, i.e., not simply declared by the facility operator and accepted by the safeguards authority. Because independent information is gathered by the safeguards authority in the form of verification measurement results on a sample of items, this information needed to be utilized, in conjunction with the corresponding operator's declarations, to obtain estimates of measurement uncertainty for both operator and inspector through the analysis of operator-inspector paired differences (or, in the absence of sufficient independent measurement data, through the use of international target values (ITV) [3], themselves partially derived from historical paired difference analysis and other information such as the evaluation of laboratory measurement capabilities [4].

Methodologies developed for this task as applied to IAEA safeguards were first formulated in 1977 by John Jaech in the *Safeguards Technical Manual (STM)* and further developed in subsequent STM volumes [5], mainly based on his earlier work for the U.S. Atomic Energy Commission [6]. As methodologies were further refined or newly developed, several revisions of the STM were published, with the final version (revision 5), re-named *Statistical Concepts and Techniques for IAEA Safeguards* [7], published in 1998. More recently, further extensions to the methodologies have been developed and tested (e.g. *Optanova,* a methodology and associated software for determining the optimal top-down estimators of the variances of random and systematic errors for paired and three-laboratory data [8]).

1.2 The Guide to the Expression of Uncertainty in Measurement (GUM)

In 1977, the International Committee for Weights and Measures (CIPM - Comité International des Poids et Mesures) asked the Bureau International des Poids et Mesures (BIPM) to address the problem of a lack of common agreement on expressing measurement uncertainties in order to facilitate comparison of laboratory results. The issue was addressed by the BIPM by convening a working group on the statement of uncertainties in 1980, including members from a number of national metrology institutes from around the world. The chair of the committee specifically stressed that the main goal of the working group was to develop clear and simple rules applicable to the determination of uncertainties, that these should be generally applicable to a large majority of users, and that it would be best to produce guidance that can be used at any level of metrology [9], [10].

While a major motivation was to address the significant issues being faced by the national metrology institutes evaluating measurements, which did not have a transparent or even comparable means of calculating and reporting measurement uncertainties, the principle of broad applicability was stressed from the outset. The result of the working group was recommendation INC-1, which was the progenitor of the modern *GUM*. The recommendations were approved by CIPM in 1981 and reaffirmed in 1986. At that time the CIPM asked the International Organization for Standardization (ISO) to work with a number of other standard setting bodies to develop a detailed guide based on the broad recommendations. The first full expression of the CIPM's recommendations was the GUM, published in 1993 [11] which has been periodically updated and is currently under the auspices of BIPM's Joint Committee on Guides to Metrology Working Group 1. The current GUM, published in 2008 [12], has been widely adopted in the analytical laboratory community.

1.3 Common Ground

At about the same time the GUM was developing, the IAEA established a set of expected measurement uncertainties associated with safeguards at nuclear fuel cycle facilities, but lacked specific details regarding the performance of measurement systems used for the determination of specific safeguarded nuclear materials. The Working Group on DA of ESARDA in 1979 presented a list of 'target values' for the uncertainty components in

1 In compliance with safeguards agreements and in application of the related NMA provisions, facility operators have to declare their balance and any MUF at the end of each material balance period (MBP), for each material balance area (MBA) and each nuclear material category. The MUF is defined as the difference between the physical inventory and the book inventory (accountancy ledger).

2 The difference between the quantity of nuclear material in a batch as stated by the shipping MBA and as measured at the receiving MBA.

nuclear material measurements [13]. A number of revisions were published [14,15] in consultation with laboratories and safeguards organizations, and eventually in 1993 the IAEA published a technical report detailing the collaborative effort [16], followed up in 2000 by the International Target Values 2000 (ITV-2000) [17]. The ITV-2000 document listed separately systematic and random components of uncertainty (which is essential for estimating the uncertainties associated with evaluating a material balance) for a number of measurement methods, and specifically stated that the developments related to GUM (referenced as the ISO, NIST and EURACHEM guides that were developing concurrently) involve uncertainty assessments in line with the developing GUM guidance. The 2010 ITV document [3] included for the first time a third column of uncertainty values, labelled "ITV", which are meant to document the total uncertainty associated to the methods listed in the tables.

In the following years, progress was made by both the laboratory and the safeguards communities in moving toward a better understanding of the methodologies used to estimate measurement uncertainties applicable to the respective needs of the two communities. However, misconceptions still exist, which were not entirely resolved during the consultations for establishing and updating the ITVs, and which this article strives to identify and clarify. These collaborative efforts between the safeguards and laboratory communities in establishing reference values for expected uncertainties associated with NMA and safeguards verification activities have increasingly highlighted the need for a common understanding of the statistical basis, terminology, and intended uses of uncertainty estimates applied in the course of performing evaluations of safeguards data. The "bottom-up" [3] approach to uncertainty estimation at the heart of GUM[4] (based on propagation of uncertainties determined for every component identified as influencing the outcome of a measurement) and the "top-down"[3] approach applied to safeguards verification data (based on ANOVA of operator-inspector differences) both arose out of historical need and serve their respective communities well[5].

3 A 'bottom-up" (first principles) approach starts from a measurement equation identifying all variables that influence the measurement results and propagates the corresponding uncertainty components to establish an uncertainty budget. A "top-down" (empirical) approach starts from a statistical measurement model and applies analysis of variance to data comparing measurement results with a reference such as quality control data or other measurement data, e.g., in the context of safeguards, declared data.

4 Note: the GUM also very briefly treats top-down UQ, but it is better known for bottom-up UQ.

5 The deliverable of any laboratory is a measurement result (expressed as a measured quantity value with uncertainty and traceability) of a measurand (analyte in the investigated matrix).The deliverable of the safeguards evaluator community is, inter alia, an assessment of the statistical significances of observed operator-inspector differences and their impact on drawing safeguards conclusions. While the common goal to evaluate measurement data is the same for the laboratory and the evaluator communities, the applied models will generally differ, based on the purpose of their construction and the nature of the deliverables.

However, there is much to be gained in 'reconciling', which means increasing the understanding among the safeguards evaluators, laboratories, nuclear facility operators and metrology experts, and in finding a shared language between the two approaches [18,19,20]. It is the intention of this article to bridge 'gaps' and facilitate this common understanding to the benefit of the safeguards measurement and metrology community, by describing how uncertainty quantification methodologies are used by safeguards data evaluators and the DA and NDA laboratories, comparing and reconciling the related methodologies and terminologies and the underlying complementarity of their purposes as well as the mutual benefits of communication and convergence between the professional communities involved.

2. Measurement uncertainty analysis in safeguards

One of the main purposes of safeguards verification activities is to detect in a timely manner and to deter the diversion of nuclear material from declared nuclear fuel cycle facilities. NMA is the basis for the detection of diversion of nuclear material by means of its keystone, material balance evaluation (MBE), which is performed for each material balance area (MBA), each material balance period (MBP) and each nuclear material category (e.g. depleted uranium, natural uranium, enriched uranium, the associated ^{235}U, plutonium, thorium). In bulk-handling facilities, where nuclear material is processed in loose forms such as gas, liquid or powder, and where nuclear material quantities are associated with process losses, where hold-up and waste have to be estimated, and where most accounting records are based on measurement results intrinsically subject to errors, MBE statistics such as MUF, SRD and the difference between the operator's declarations and the inspector's verification measurement results (operator-inspector differences, D) are necessarily non-zero. They have to be statistically tested to determine whether or not they can be explained by the operator's and inspector's measurement uncertainties.

Before statistical tests can be applied, measurement uncertainties must be estimated and propagated from the item level to the level of the MBA and MBP. The estimation of measurement uncertainties is one of the most demanding questions faced by statistical methodologies for safeguards. In some cases uncertainty estimates are documented by facility operators, analytical laboratories and/or instrument developers but to support credible conclusions regarding the absence of diversion, they need to be validated independently and their fitness for purpose needs to be assessed by safeguards analysts.

Estimating measurement error variances can be performed, for example, using a "bottom-up" approach via calibration certificates and the validation of nuclear

operators' data or a "top-down" approach by analysis of variance (ANOVA) of observed operator-inspector differences, i.e. paired-data. Because calibration-based uncertainty quantification (UQ) does not necessarily account for all sources of uncertainty, paired data analysis has traditionally been the method of choice at the IAEA. On the other hand, because paired-data based UQ does not only include the facility operators' measurement uncertainties, the analysis of operators' data has been the method of choice at the European Commission (EC) inspectorate.

In the top-down approach, the observed paired differences reflect the combined effect of the operator's and inspector's measurement errors, and form the basis for the estimation of the relative standard deviations (RSD) associated with these errors, which in turn are needed to obtain uncertainty estimates associated with MUF, SRD and D, to calculate verification sample sizes and to establish rejection limits for individual operator-inspector differences. Measurement uncertainty estimates are quantified by the absolute (relative) standard deviation of measurement errors, respectively denoted σ (δ). The propagation step requires, in addition to the separation of the operator's (σ_o or δ_o) and inspector's (σ_i or δ_i) error standard deviations, a further parsing of both of these into a random e.g. ($\sigma_{O,R}$ or $\delta_{O,R}$) and a (short-term) systematic e.g. ($\sigma_{O,S}$ or $\delta_{O,S}$) component, because the averaging process reduces the effect of random errors in multiple measurements while the effect of systematic errors is not reduced by averaging, and the different mode of propagation of these two error components into material balances makes it essential for safeguards analysts to obtain separate estimates for their respective standard deviations. It must be noted that the separation of error standard deviations into four components is required by the error propagation process regardless of the chosen UQ approach.

One of the main difficulties when applying ANOVA to paired data is to obtain separate estimates of the four different uncertainty components. This task is further complicated by the need to process outliers and to validate various assumptions (e.g. normally distributed random errors) that are necessary for the implementation of certain algorithms. As explained in section 1 above, the methods used by the IAEA to estimate measurement error variances were developed several decades ago (e.g. Grubbs Analysis, 1948 [21]) and are presently being refined.

When a bottom-up approach is applied to estimate the operator's error RSDs, the operator's UQ practice is audited and the resulting uncertainties are confirmed to comply with latest international standards. The uncertainty associated with the operator's declared MUF can then be computed by error propagation in order to perform a statistical test of the hypothesis that it can be explained by measurement errors.

3. GUM in the Laboratory

The chair of the BIPM committee tasked to develop what became the GUM specifically stressed that the main goal of the group was to develop clear and simple rules applicable to the determination of uncertainties, that these should be generally applicable to a large majority of users, and that it would be best to produce guidance that can be used at any level of metrology [9]. The idea of a standardized approach to uncertainty evaluation is to provide a method that is applicable to all types of measurements with results (including uncertainties) that are transparent and easily utilized by a variety of users (i.e. the value and uncertainty should be easily transferrable). While the effort and expertise necessary to produce an uncertainty evaluation is often not a simple task, the basic JCGM 100:2008 (GUM Guide) approach provides a stepwise and relatively easily taught and understood mechanism to uncertainty evaluation that has proved to be of great practical value to measurement practitioners [12]. While the requirements of ISO/IEC 17025:2005 and recently ISO/IEC 17025:2017 (and to some extent ISO/IEC 17043:2010 and ISO Guide 34:2009 and more recently ISO 17034:2016) have driven the utilization of the GUM, in workshops conducted throughout the nuclear measurement community, the use of the GUM approach has engendered much positive discussion and has been typically embraced by laboratory staff and management around the world [22,23,24,25]. In particular, laboratory professionals from bench technicians to measurement experts have indicated that the GUM gives them a usable framework for a better understanding of their measurements, helps them to identify potential problem areas, and provides them with useful guidance on how to report measurement results in a transparent and organized manner.

A simple example demonstrating the practical use and benefit of the GUM arose during an introductory workshop at a US national laboratory when staff assigned to perform assay measurement of a plutonium storage tank provided their measurement method for modelling and evaluation by GUM. The tank was measured for accountability purposes on a semi-annual basis and results were submitted to the material accountancy organization. The procedure specified the use of a random error RSD of 4% for the distance of the detector from the tank (76 cm +/- 3 cm), to account for imprecision in reproducible placement of the detector. Employing the GUM methodology for this procedure was straightforward, with the result that the error RSD ascribed to the distance of the detector from the tank contributed nearly 80% to the overall RSD of the measurement. The entire exercise took about half an hour to perform. The technicians were particularly surprised at the influence of the distance uncertainty on the overall RSD and awareness was created among the technicians, lab manager and statistical staff of the laboratory.

The point of this simple example is to illustrate that the basic principles of the GUM are relatively easily grasped and implemented, that the GUM provides an accessible tool to measurement practitioners of varying expertise, and that these attributes provide a useful means for those performing and/or using measurement results to better understand their measurement processes. Similar positive results have arisen in a number of situations during workshops and discussions among laboratory staff.

More complex examples that arose during GUM uncertainty evaluations included demonstrations emphasizing the use of isotopic ratios rather than abundances in uncertainty determinations, the correlation of mass bias (K-factors) in thermal ionization mass spectrometry (TIMS) leading to unexpectedly small uncertainties for minor isotopes, and identifying unexpected significant contributors to certified reference materials (CRM) production efforts. In the case of the production of CRMs, the creation of GUM uncertainty budgets prior to any analytical effort has been invaluable in deciding the scope and breadth of effort required for the production and certification of a variety of uranium and plutonium CRMs, identifying key contributors to the final product's uncertainty [26]. For example the contributions to the combined standard uncertainty of a measurement result of $n(^{236}U)/n(^{238}U)$ in IRMM-2022 by TIMS are shown in Figure 1 [27,28].

In 2006, the EC-JRC conducted a proficiency testing exercise that included 71 laboratories from 26 countries performing uranium isotopic measurements [29]. A wide range of laboratories participated, active in research, environmental radioactivity measurements, monitoring of nuclear facilities, medical applications and safeguards. Of the 71 laboratories that submitted measurement results, 30% reported they held ISO 17025 accreditation. Nearly half of all laboratories reported that their uncertainties were calculated according to the GUM. While compliance with ISO 17025 requires GUM uncertainty evaluations, many labs in 2006 were already utilizing the GUM without the accreditation. Usage of the GUM has no doubt further increased in the ten years since this exercise was completed, with laboratories finding benefit in utilization of the GUM.

3.1 GUM in Proficiency Testing and Laboratory Self-Evaluation

Laboratories, particularly in safeguards, have to demonstrate their performance over short and long terms by means of conformity assessment and quality control tools [4]. They are required to have measures in place to ensure that the measurement process is stable and in control. The GUM-based uncertainty values in the ITV 2010 document are being utilized by laboratories and proficiency testing providers as benchmarks. The REIMEP-17 inter-laboratory comparison (ILC), reported in 2015, utilized the prescribed ISO 13528:2005 and ISO 17043:2010 statistics for evaluating laboratory results' agreement with the reference values for the distributed materials [30,31]. The laboratories in REIMEP 17 were thus evaluated against the ITV-2010 GUM-based uncertainties to compare their performance to the state of the practice for measurements as determined by the ITV-2010 document. The New Brunswick Laboratory Safeguards Measurement Evaluation Program (SME) utilizes a similar approach, The 2011 NBL SME Report utilized five different fuel-cycle materials, with 23 laboratories participating. The submitted results were compared to the reference values and also the ITV-2010 GUM-based target

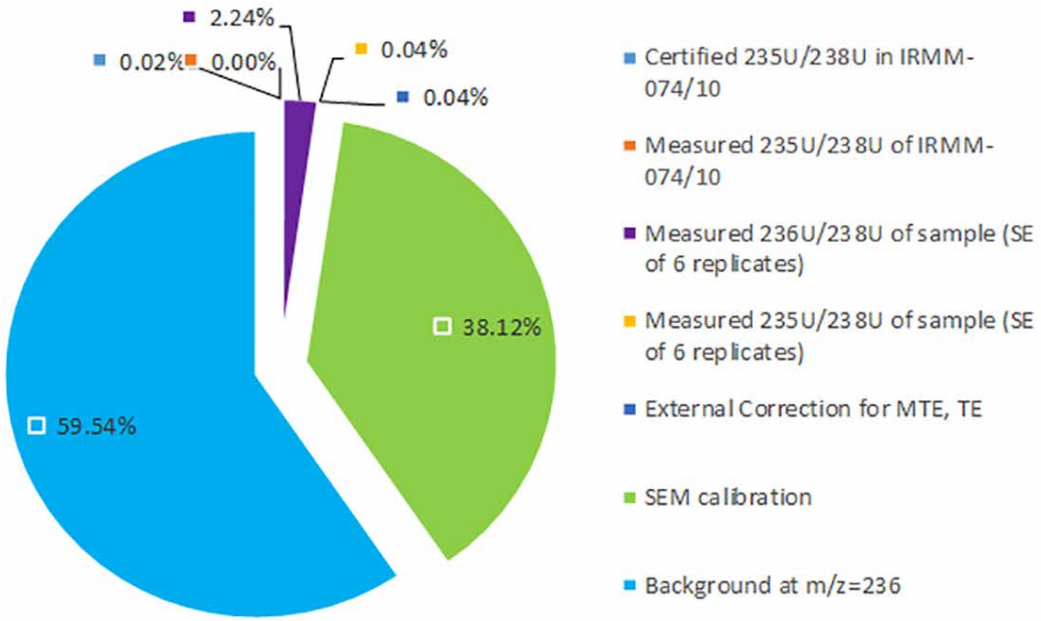

Figure 1: Uncertainty contributions for the measurement of n(236U)/n(238U) in IRMM-2022

values [32]. In most instances, the CRMs used in ILCs are produced by specialized laboratories using state-of-the-art methods and produced painstakingly to yield the smallest possible uncertainties at the highest metrological standards. Laboratories participating in proficiency testing exercises typically perform analyses using standard procedures, at state-of-the-practice levels and in many cases employing analytical methods yielding RSDs two-to-tenfold larger than the certified RSDs of the material. While the ISO 13528 and 17043 evaluation methods are useful for assessing agreement with the certified value in these cases, the additional comparison utilizing the ITV-2010 GUM-based uncertainties is of particular practical value in comparing laboratory and method performance.

Given the increasing use and importance of the ITV-2010 GUM-based RSDs, the refinement of the GUM based values could be addressed in a future revision of the ITV document, more accurately representing GUM-compliant uncertainties as determined in the field. The use of ILC exercise results and individual laboratory reporting is vital to this effort [33,34]. In conjunction with the next revision of the ITVs, the GUM-based laboratory performance values may be issued in a separate document. This would distinguish them from the ITVs, which are derived from historical paired difference analysis and other information, but not exclusively from the evaluation of laboratory measurement capabilities. This will distinguish them from the ITVs, which are partially derived from historical paired difference analysis and other information but not exclusively from the evaluation of laboratory measurement capabilities.

3.2 Fit-for-purpose according to GUM

The purpose of performing a measurement is to provide a result with stated uncertainty and traceability of a measurand to be utilized by one or more users for various purposes. A measurement result lacking a value an uncertainty or traceability is meaningless and not useable. For the purpose of decision making or conformity assessment, the measured value and uncertainty must be transferrable and comparable by the end-users. The result would provide measurement producers and users with the ability to demonstrate fitness for purpose, demonstrate laboratory proficiency and provide assurance of laboratory capability [35]. The methodical and accessible approach to uncertainty evaluation provided for by the GUM has been embraced within the nuclear measurement community and has been established in a large variety of nuclear measurement laboratories worldwide. It is worth noting that, given the specific purpose described in section 2, the IAEA, which is using the top-down paired data approach for UQ, does not need to take reported measurement uncertainties from either operator or inspector laboratories into account.

3.3 Destructive analysis (DA) Laboratory

To achieve traceability, one must link measurand identity and quantity value to a stated reference (preferably via calibration standards and CRMs). To give an example, DA nuclear laboratories are routinely measuring the plutonium amount in a plutonium nitrate solution sample [36]. They need to provide an accurate and traceable measurement result within the respective ITV-2010 GUM-based uncertainties. Often the method of choice is Isotope Dilution -Thermal Ionisation Mass Spectrometry (ID-TIMS). In ID-TIMS, using a ^{242}Pu enriched material as the spike, the ^{239}Pu content in an unknown sample can be determined by isotope dilution, through a measurement of the isotope ratio $R(^{242}Pu/^{239}Pu, B)$ in the blend. Following the GUM's generic measurand equation

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

where Y denotes a measurand determined from *N* other quantities X_1, X_2, …,X_N through the functional relationship *f*.

the plutonium amount content can be calculated as follows [37]:

$$c(Pu,X) = \frac{R(^{242}Pu/^{239}Pu,Y) - R(^{242}Pu/^{239}Pu,B)}{R(^{242}Pu/^{239}Pu,B) - R(^{242}Pu/^{239}Pu,X)} \cdot \frac{\sum_m R(^{m}Pu/^{239}Pu,X)}{\sum_m R(^{m}Pu/^{239}Pu,Y)} \cdot \frac{m(Y)}{m(X)} \cdot c(Pu,Y) \tag{2}$$

where:

$R(^{m}Pu/^{239}Pu, X)$ = amount ratio $^{m}Pu/^{239}Pu$ in the unknown sample material X

$R(^{m}Pu/^{239}Pu, Y)$ = amount ratio $^{m}Pu/^{239}Pu$ in the known spike material Y

$R(^{m}Pu/^{239}Pu, B)$ = amount ratio $^{m}Pu/^{239}Pu$ in the measured blend material B

$m(X)$ = mass of the unknown sample used in the measurement

$m(Y)$ = mass of the spike solution used in the measurement

$c(^{239}Pu,X)$ = amount content (moles) of $^{239}Pu\ /\ g$ sample material

$c(^{242}Pu,Y)$ = amount content (moles) of $^{242}Pu\ /\ g$ spike solution

$c(Pu,X)$ = amount content of $Pu\ /\ kg$ sample material

$c(Pu,Y)$ = amount content of $Pu\ /\ kg$ spike solution

If any of these components (analyte, value, uncertainty, unit) is missing, the measurement is meaningless. Laboratories estimate the combined standard uncertainty of a measurement result by applying either the bottom-up or top-down approach. In the bottom-up approach the uncertainties of each factor in the measurement model are estimated and these individual uncertainties are combined

according to the law of error propagation applied to Eq. (1) [38]

$$u_c(y(x_1,...,x_n)) = \sqrt{\sum_{i=1,n}(\frac{\delta y}{\delta x_i})^2 \times u(x_i)^2} \quad (3)$$

where y denotes the estimate of Y and x_i denotes the estimate of X_i.

In the top-down approach, combined effects covering several factors - also unknowns - are estimated using uncertainties due to repeatability (u_{rep}), intermediate precision (u_{ip}) and "trueness" (u_t), as established by means of a CRM, combined with uncertainties for calibration (u_{cal}) [39].

$$u_c = \sqrt{u_{rep}^2 + u_{ip}^2 + u_t^2 + u_{cal}^2} \quad (4)$$

If performed accurately and documented properly, so that it is possible for an external auditor to reproduce how the combined standard uncertainty was estimated, the laboratory is fully compliant with ISO standards independently of the chosen approach. One main advantage of a bottom-up approach is that it yields detailed information for method improvement, whereas such information is not revealed by a top-down approach.

3.4 Non-destructive assay (NDA) Laboratory

NDA of items containing nuclear material uses calibration and modelling to infer item characteristics such as nuclear material mass on the basis of detected radiation such as neutron and gamma emissions. Three specific issues in UQ for NDA are as follows.

NDA is often applied in challenging settings because the detector is brought to the facility where ambient conditions can vary over time, and because the items to be assayed are often heterogeneous in some way. Because of such challenges, dark uncertainty [33] can be large, as is evident whenever bottom-up UQ predicts smaller uncertainty than is observed in empirical (top-down) UQ [34] (by "uncertainty" we mean the reproducibility standard deviation as quantified, for example, in an ILC) [39].

NDA is widely applied in situations where the items subject to measurement differ substantially from the calibration items; therefore, the concept of item-specific bias has long been recognized [40,41].

Currently, there is no general UQ guide for NDA that is analogous to the GUM. But, the GUM is typically followed for the error variance propagation steps in UQ, and each NDA method has a specific and documented implementation of UQ, for example, ASTM C1514 for the enrichment meter principle (EMP) as discussed in full detail elsewhere [42]. However, this NDA example needs to be presented here in a consolidated manner to follow the reasoning towards reconciliation of complementary approaches as discussed later on in sections 5 and 6

Example: Enrichment Meter Principle (EMP) for gamma spectroscopy

This sub-section provides an example that involves calibration of gamma spectroscopy in order to describe some of the statistical aspects of bottom-up UQ. The amount of ^{235}U in an item can be estimated by using a measured net weight of uranium U in the item 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 EMP. The 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 [43,44,45]. The EMP assumes that the detector field of view into each item is identical to that in the calibration items (the "infinite thickness" assumption), 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 equal or similar to that in the calibration items, 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. It has been shown empirically that under good measurement conditions, the EMP can have a random error RSD of less than 0.5 % and a long term bias of less than 1 %, depending on the detector resolution, stability, and extent of corrections needed to adjust items to calibration conditions. However, in some EMP applications, the random error RSD can be larger than bottom-up UQ predicts (see next paragraph) and larger than the 0.5% target random RSD. For example, assay of the ^{235}U mass in UO_2 drums suggests that there is larger-than-anticipated random RSD in some deployments of the EMP.

To investigate UQ for the EMP, Burr et al. [46] fit 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 somewhere below and above the 185.7 keV energy but outside the peak area to estimate background (two-region EMP method), expressed as

$$Y = \beta_1 X_1 + \beta_2 X_2 + R \quad (5)$$

where Y is the enrichment, X_1 is the peak count rate near 185.7keV, X_2 is the background count rate in neighbouring energy channels near the 185.7keV peak region, and R is random error. Figure 2 is an example low-resolution (NaI detector) gamma spectrum near the 185.7keV. The two background ROI counts can be combined into one count, resulting in two predictors as in Eq. (5): X_1 is the peak ROI counts and X_2 is the background ROI counts to be used to predict enrichment E in Eq. (5) using least squares

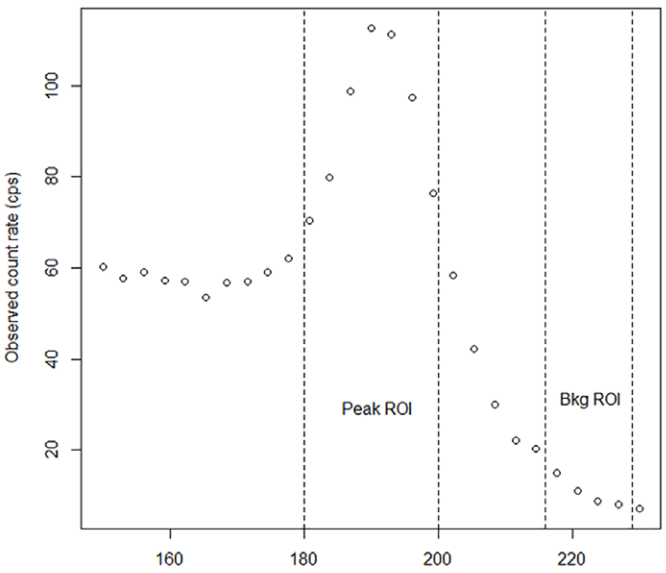

Figure 2: Example low-resolution (NaI detector) gamma spectrum near the 185.7keV peak with two background regions (one region below the 185.7 keV peak and one region above the 185.7 keV peak)

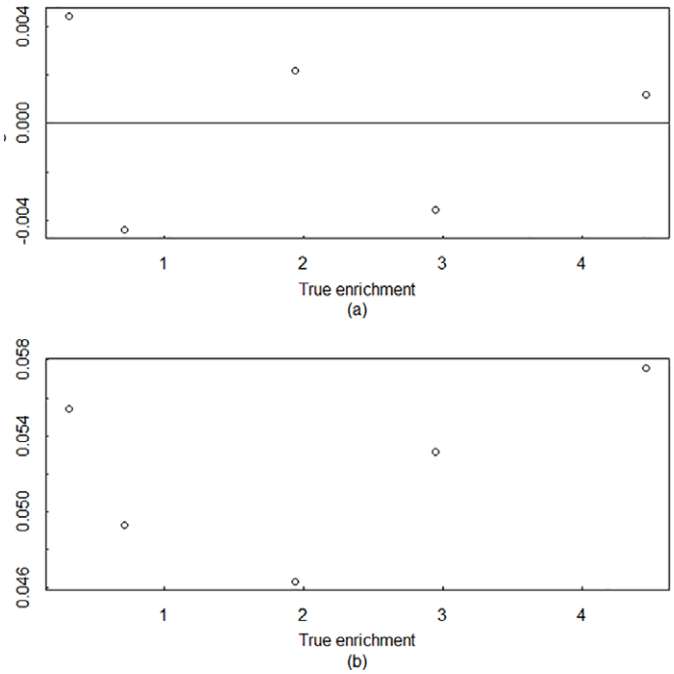

Figure 3: The average residual (a) and the MSE (b) in testing data, using the same values for testing and training. Results are based on 10^5 simulations so simulation error is negligible. The data, collected during 2015 at Oak Ridge National Laboratory are enrichment Y = {0.3166, 0.7119, 1.9420, 2.9492, 4.4623}, and $X1$ = {5616,10298,25093,37103, 55178}, $X2$ = {1803, 1815,1914,1984,2132}. The assumed absolute error standard deviations were 0.0035 in Y and 1% of the range of $X1$ in $X1$ and 1% of the range of $X2$ in $X2$.

regression. There will be measurement errors in X_1 and X_2 and there will often be correction factors applied to X_1 and X_2, for example, to adjust test item container thickness to calibration item container thickness. Calibration data is used to produce estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ of the two model parameters, β_1 and β_2 . The covariance matrix of the random variables $(\hat{\beta}_1, \hat{\beta}_2)$ is not necessarily well approximated by the usual least squares expression because of errors in X_1 and X_2. Therefore, [44,45] suggest that the mean squared error (MSE) in $\hat{Y}$ be estimated using simulation of the calibration procedure, which easily allows for errors in X_1 and X_2 arising from Poisson counting statistics, and also arising from other sources, such as container thickness (with or without an adjustment for the measured container thickness) varying among test items. Errors in X_1 and X_2 due to imperfect adjustment for container thickness can manifest as item-specific bias. The simulation strategy in [44,45] and the summary sub-section below illustrate how item-specific bias can be understood and estimated. The MSE in $\hat{Y}$ is defined as usual, as

$$E((\hat{Y} - Y_{true})^2) = E(\hat{Y} - E(\hat{Y})^2) + (E\hat{Y} - Y_{true})^2 = variance + bias^2.$$

We can express the simple calibration Eq. (5) as in Eq. (1), where we identify X_1 as $\hat{\beta}_1$, X_2 as $\hat{\beta}_2$, X_3 as X_1, and X_4 as X_2, respectively, with cov $(\hat{\beta}_1, \hat{\beta}_2)$ estimated by simulation, so with some effort, GUM's Eq. (1) could be used to estimate var $(\hat{Y}_1)$ and cov $(\hat{Y}_1, \hat{Y}_2)$, although Elster [47] points out that GUM's Eq. (1) is not actually designed to be applied to calibration applications, regardless of whether there are errors in the predictors X_1 and X_2 (which complicates the data analysis). Some of the numerical bottom-up UQ examples in [44,45] have estimated random error RSD ranging from less than its 0.5% target to approximately 1.0% (because of item-specific biases arising due to container thickness variations and other effects,) but less than the 1.81% reported from empirical (top-down) UQ of the UO_2 drums example by Walsh et al. [48].

Figure 3: plots the average residual versus the true enrichment in fitting Y as a function of X_1 and X_2 (Eq. (5)). Because 10^5 simulations were used, simulation error is negligible [49]. Figure 3: (b) is an example of a simulation-based bottom-up prediction of uncertainty due to calibration errors. The caption of Figure 3: lists the data and measurement error standard deviations in Y, X_1, and X_2 in training (calibration) and testing, which can be modified to mimic the effect of varying container thickness, with or without an adjustment for container thickness being different in training than in testing items.

Burr et al. [50] compare simulation-based UQ to analytical approximations of UQ for calibration data. If the operator uses some other method, such as DA, then the operator's DA measurement can be assessed using separate simulation.

Discussion

Generally in NDA 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, or the distribution of radiation-absorbing isotopes, which relate to neutron and/or gamma absorption behaviour of the item, can partially obscure the relation between detected radiation and the source material; this adds a source of uncertainty to the estimated amount of SNM (Special Nuclear Material) in the item. One can express item-specific impacts on uncertainty using a model such as

$$CR/M = g(X_1, X_2, ..., X_N), \quad (6)$$

where CR is the item's neutron or gamma count rate, M is the item SNM mass, g is a known function, and $X_1, X_2, \ldots, X_N$ are N auxiliary predictor variables such as item density, source SNM heterogeneity, and container thickness, which will generally be estimated or measured with error and so are regarded as random variables. To map Eq. (4), to GUM's Eq. (2), write

$$M = CR / g(X_1, X_2, ..., X_N) = h(X_1, X_2, ..., X_M) \quad (7)$$

where the measured CR is now among the $M = N+1$ inputs. Note that Eq. (6) is the same as Eq. (1), but some of the X_i account for item-specific departures from reference items used for calibration.

Top-down UQ used in MBE estimates the random and short-term systematic standard deviations σ_R and σ_S, which are estimated from data sets that have items measured by each of two or more assay methods. The net random error can include variation in background that cannot be perfectly adjusted for, Poisson counting statistics effects, item-specific biases, and other random effects. In principle, each of $X_1, X_2, \ldots, X_N$ could be estimated for each item as part of the assay protocol. However, there would still be modelling error because the function f must be chosen or somehow inferred, possibly using purely empirical data mining applied to calibration data [51] or physics-based radiation transport codes. Typically, only some of $X_1, X_2, \ldots, X_N$ will be measured as part of the assay protocol. Most assay methods rely on a calibration step [52]; as mentioned, calibration is not fully addressed in the GUM [53, 47, 54, 44, 45, 46] but one GUM supplement in progress will address calibration; and, the GUM is currently being revised to include more detail on calibration [47, 53].

4. 'Bridging the gap' - Reconciliation of terminological and methodological differences

Common or mapped terminology and a common mathematical basis are prerequisites for any reconciliation process. The limited understanding between the laboratory and evaluator communities has a mathematical/logical component and also a paradigmatic component. This article aims to establish a common language between the communities by translating and mapping terms used by the two communities. 'Mapping' means here to list recurring perceived differences stemming from semantics, including a sociological component in using terminology in a particular manner within a community, ('apparent differences' – Table 1) as well as differences originating from the application of metrological/statistical concepts ('conceptual differences' – Table 2). The first can lead to 'bridging the gap' towards harmonisation of terminology and formalism between communities. The latter can lead to 'bridging the gap' towards mutual understanding, complementarity, and convergence, see sections 6, and 7, while maintaining those differences of approach that are rooted in different requirements in the problem domain. Wherever necessary, additional notes and references to publications or other sections in this article are in the third column of the Tables. 'Reconciled' in Table 1 means that there is a consensus between the laboratory and evaluator communities about the listed differences not being conceptual but rooted in terminology. Specific terms or concepts that cause recurring misunderstandings between the communities are discussed in more detail see section 4.2. Although changes as a result of the on-going revision of the GUM are not anticipated and beyond the scope of this article, the notes to some of the entries in the mapping tables are recommending a revised version of the GUM when deemed necessary [47].

1. Terminology - Apparent Differences

Terminology used by evaluators in safeguards	Terminology used by laboratories in safeguards	Notes
Observable, measurand	Measurand - quantity intended to be measured [VIM 2.3], analyte	[39]
Measured value	Quantity value representing a measurement result [VIM 2.10]	Measurement results are quantitative probabilistic statements on the measurand.
Estimate of the true value of the measurand, associated with measurement errors (random, short-term systematic, bias)	Measurement result associated with an interval of reasonable values of the measurand; best estimate of the measurand, along with an associated measurement uncertainty [VIM 2.19]	The true value of the measurand is a fixed and unknowable constant; the result of the measurement of the measurand can be quantified. The concept of true value is inseparable from the definition of the particular quantity to be measured, see section 4.2
Measurement error standard deviation	Standard measurement uncertainty [VIM 2.26]	The expression “measurement error” may be wrongly used instead of measurement error standard deviation, see section 4.2. The absolute error standard deviation is usually denoted σ.
Relative measurement error standard deviation (RSD)	Relative standard deviation (RSD)	RSD denotes a relative standard deviation (δ), i.e. the standard deviation divided by the absolute value of the mean.
Standard deviation associated with the value of a standard	Combined standard uncertainty of a reference value [VIM 5.18]	Error of a standard may be wrongly used as a synonym of uncertainty.
Total measurement error standard deviation, propagated measurement error standard deviation	Combined standard uncertainty [VIM 2.31]	
Confidence interval: a range of values that contains the true (unknown) value of a parameter, e.g. the measurand, with a given probability referred to as the confidence level; (adopting the frequentist view that in a collection of such intervals the percentage that contain the true value of the measurand should tend toward the stated confidence level as their number increases)	Coverage factor k: a multiplication factor defining the width of the coverage interval of reasonable values of the measurand. The choice of k depends on the level of confidence required for the measurement result, usually expressed as $\hat{y} \pm k u_c(\hat{y})$ k is that value satisfying the probability statement in Eq. (G.1a) in JCGM 100 and $u_c(\hat{y})$ is an estimate of the standard deviation of $(\hat{y})$	Confidence level is isomorphic but not equivalent to coverage factor: k defines an interval corresponding to a certain confidence level, see JCGM 100:2008 G 6.1 [55].
Total error standard deviation	Expanded uncertainty *U*	
Consistency of estimates: the difference between the values of these estimates is smaller than a multiple of the standard deviation of the difference. The level of consistency is related to the chosen multiple of the standard deviation	Metrological compatibility of measurement results: the difference between two measurement results is smaller than the expanded uncertainty of their difference [VIM 2.47], see sction 4.2	

Table 1: Mapping of 'reconciled' terminological differences

4.1 Differences in approaches

Approaches used by evaluators in safeguards	Approaches used by laboratories in safeguards	Difference in approach
Principal objective of uncertainty quantification: to determine the significance of observed differences between two independent determinations of a quantity or combined quantities through statistical error propagation.	Principal objective of uncertainty quantification: to completely qualify a single measurement result	It is recognized by both groups that a complete expression of a measurement result consists of a quantity value, a statement of its uncertainty and a metrological traceability statement [VIM 2.41]. However, safeguards data evaluation typically deals with algebraic combinations of measurements such as differences, not with single measurements.
Statistical measurement error model	Measurement model - rule for converting a quantity value into the corresponding value of the measurand [JCGM 104:2009, 3.10]	Measurement results are quantitative probabilistic statements on the measurand.
The preferred error model allows for long-term systematic error (=bias), short-term systematic error and random error. Measurement error model (simplified) $Y = \mu + b + R + S$ Y denotes the measured value μ: denotes the true (but unknowable) value b: denotes a bias (long-term systematic error) R: denotes a random error of expectation E(R)=0 and of standard deviation denoted $\sigma_{Y,R}$ S: denotes a short term systematic error of expectation E(S)=0 and of standard deviation denoted $\sigma_{Y,S}$ The total uncertainty (standard deviation) associated to Y is given by: $\sigma_Y = \sqrt{\sigma_{Y,R}^2 + \sigma_{Y,S}^2}$	The preferred error model allows for Type A errors (can be reduced by repetition of measurements) and for Type B errors (can be reduced by other means). In the expression of uncertainty the effects of both types of error are combined Measurement model (simplified) $Y = \hat{y} \pm k u_c(\hat{y})$ Y denotes the measurand $\hat{y}$ denotes the estimate of the measurand Y ; Y includes a correction factor taking into account the measurement bias (the measurement model does not distinguish between bias and short-term systematic error) k: denotes a coverage factor. $u_c(\hat{y})$ denotes the estimate of the combined standard uncertainty $u_c(\hat{y})$ (Type A, Type B uncertainties propagated from all input quantities in $\hat{y}$ including the standard uncertainty u_b associated to the correction factor for bias)	The terms in the evaluator's error model are introduced to describe: S: a fluctuating error component (of random nature) often seen in the data, for example between inspections or between calibrations, which is superposed to random fluctuations between individual measurements and b: a possible long-term bias, which is not of random nature.
When combining standard deviations, random and short-term systematic components are propagated differently.	The uncertainty being related to a single value, the random and systematic components are combined in a single uncertainty estimate	SG data analysis must take the different behaviour of measurement error components through combination of measurement results into account, whereas labs deliver a single measurement result and strive to minimize any short-term systematic component.
The main method of uncertainty quantification is analysis of variance based on paired (or multiple) data from independent measurement methods	Standard methods of uncertainty quantification are repetition under controlled conditions and quality control (QC) with certified standard materials.	Labs can control the measurement conditions, can perform as many repetitions as needed and have certified reference materials available. Safeguards data evaluators on the other hand analyze operator and inspector measurements of the same items performed in conditions that they do not control.

Table 2: Mapping of conceptual differences

4.2 Discussion of recurrent terms of misunderstanding

True quantity value

The primary objective of the safeguards evaluators, approach is not to estimate the true value of a measurand but to estimate random and systematic error variances by means of ANOVA applied to operator-inspector differences as described in section 5 and in [48] or by propagation of known operator's and inspector's measurement uncertainties. The objective of a measurement following the GUM's bottom-up approach as understood by the laboratory community, (see section 3), is neither to determine a true value nor to produce separate estimates of random and systematic error from paired data, but rather to determine an interval of reasonable values of the measurand, based on the assumption that no mistakes have been made in performing the measurement [39] and that the measurement conditions have been adequately controlled. In terms of reconciliation, both approaches rely on the concept of a true value to make a measurement meaningful and characterize its performance ([56], GUM D.3.5). In that sense the GUM's measurement objective is to establish an interval of values within which the true value of the measurand (with sufficiently small intrinsic uncertainty, GUM D.3.4) is believed to lie, with a given degree of belief, based on the available information from the measurement and possibly possibly from other sources [48,57].

Metrological compatibility

Metrological compatibility of measurement results is the property of a set of measurement results for a specified measurand, for which the absolute value of the difference of any pair of measured quantity values from two different measurement results is smaller than a chosen multiple of the standard measurement uncertainty of that difference [VIM 2.47,58]. Safeguards evaluation addresses a similar but different problem domain concerned with the evaluation of mass differences to determine if they are explicable by measurement uncertainties, considering detection probabilities of nuclear material diversion and the risk of false alarm. Note: the IAEA (at present) relies mostly on top-down UQ based on ANOVA independently from any uncertainties reported by the laboratories, to which the top-down uncertainty estimates can be compared to identify the existence of sources of uncertainty outside of or unaccounted for by the laboratories. On the other hand, the EC makes use of both operator's and inspector's measurement results, including the reported and validated uncertainties, to build a bottom-up uncertainty budget.

Measurment trueness

Measurement trueness is not a quantity and thus cannot be expressed numerically, but a 'trueness check' is part of a laboratory's method validation [39]. This means to compare the measured value of a measurand associated to a certified (matrix) reference material (x_m) to its certified value (x_{CRM}) and to assess their metrological compatibility in order to exclude any significant bias (Δ_m).

$$\Delta_m = |x_m - x_{CRM}| \tag{8}$$

The standard measurement uncertainty for Δ_m is given by:

$$u_\Delta = \sqrt{u_m^2 + u_{CRM}^2} \tag{9}$$

If $abs(\Delta_m) \leq 2u_\Delta$, there is no evidence that the measured and the certified value are incompatible (a hypothesis test that the bias is zero would not be rejected at the 0.05 significance level). Thus, there is no significant bias, no correction is needed and u_m is is used in subsequent data evaluation. If there is a significant difference, the laboratory preferably improves the method or, in case this is not possible must correct the measurement model for the quantified bias and propagate the uncertainty introduced by the correction.

Systematic errors and measurement bias.

A laboratory can tailor its effort depending on the available resources to do method validation, including performing measurements under repeatability and reproducibility conditions, and to check for trueness, aiming to establish reasonable combined measurement uncertainties to provide a fit-for-purpose measurement result. In safeguards verification, opportunities to perform measurements under repeatability and reproducibility conditions are severely limited by inspection schedules and practicalities - see sections 2 and 4. In addition, the safeguards authority does not determine the measurement procedures used by the operators. Therefore, one source of misunderstanding between the two communities lies in the following:

For the evaluator community, the terms 'bias' (or long-term systematic error), 'short-term systematic error', and 'random error', are integral parts of the statistical model of measurement error. When dealing with operator-inspector differences, many sources of error remain unknown. Thus one must allow for the presence of both bias and short-term systematic error. In MBE, a short-term systematic error is a random variable with expectation zero that is constant for a group of measurements (for example, a group can be a time period such as a 1-week inspection period each year) and is a component of the total error that cannot be reduced by averaging over all measurements in a group. The average (expectation) of short term systematic errors observed over a long period (a large number of shorter periods each corresponding to one systematic error observation) tends to zero.

Although the laboratory community uses a similar terminology, there are clear differences in approaches, particularly also because the GUM-based ITVs-2010 are values for

uncertainties associated with a single determination result [3]. In the measurement model as described in GUM there is no notion of 'time' or measurement group. From the laboratory viewpoint, a measurement bias is not related to a timeframe (short term – long term). In GUM a measurement bias is stated to be an estimate of a systematic measurement error; however, we anticipate that the next version of GUM will define a bias to be a true unknown quantity, not an estimate [59]. But regardless, a reference quantity value is required to quantify bias, and if the bias is significant, a correction factor with a combined standard uncertainty can be applied to the measurement to take this bias into account. Both the laboratory and MBE communities accept the possibility of performing bias adjustments; however, the laboratory preferred method if the bias is statistically significant is to return to first principles and remove or reduce the source of bias. Depending on the MBP, which is commonly one year, one can consider that the assessment by a laboratory of a bias by means of a QC chart based on a certified reference material [38] can correspond to a short term or long term systematic error according to MBE terminology. In MBE, the term bias is used to denote a long-term systematic error, a fixed effect, not modelled as a random variable, to distinguish it from the short-term systematic error, which is modelled as a random variable fluctuating with the measurement conditions. A method is unbiased if the long-term systematic error is zero. Short-term systematic and random errors always exist and propagate differently. This is a clear example where the same term 'bias' is used by the two communities but with different meanings, causing misconceptions and misunderstanding because it is related to different effects. The safeguards statistical data evaluators partition error variance into random and short-term systematic in their approach to assess whether a bias is significant. From a pure measurement point of view, a long term systematic error can only be assessed via a series of measurements during a certain time frame [25]. This approach cannot be easily implemented in MBE because metrological conditions can change across balance periods. However, laboratories can demonstrate long term measurement performance via regular participation in ILCs with independent and traceable reference values in compliance with ISO 13528:2005 [60, 61 ,62 ,63]. A recurring measurement bias of an operator or safeguards laboratory could be translated into an indication for a long-term systematic error in MBE [4, 61]. In the case of an operator measurement bias, this could also be interpreted as an indicaton of possible diversion.

5. The statistical basis of different approaches to quantification of measurement uncertainty

Reconciling GUM-based UQ and UQ via the IAEA error model empirical estimation was approached by reviewing the design basis and corresponding mathematical/statistical formalisms of each. The full scope of this investigation can be found in [48].

Empirical approaches to UQ, such as estimation of variance components by an appropriate ANOVA, are applied in metrology to estimate a specified error variance parameter of a measurement method. When estimating variance parameters in an empirical approach, the precision conditions under which the data are collected must be clearly specified – this includes statements regarding the degree to which the sample replicates are true measurement replicates, as well as an acknowledgement of what measurement conditions may have changed when measuring the set of items (e.g. day, analyst, calibration, instrument, etc.). The statistical model and corresponding estimation approach in conjunction with the conditions under which the replicate measurements were collected imply how the resulting variance estimates are to be interpreted and used in subsequent UQ exercises.

Ideally, the term 'top-down' uncertainty should only be used in conjunction with the specific empirical approach of reproducibility studies that deal with measurement replication across many participant analytical systems, thereby covering a wide range of varying environmental conditions (this is described in ISO 21748 [64]). Therefore 'top-down' UQ involves explicit estimation of the reproducibility standard deviation as defined in ISO 5725 [65]. Estimates of variance obtained in reproducibility studies comprise a theoretically and empirically justified benchmarking of an important component of true uncertainty in a measurement method – i.e. the reproducibility standard deviation is the primary empirically derived parameter for estimating the uncertainty of a measurement method. Because of this, estimates of the reproducibility standard deviation are used to assess the correctness of an analytical laboratory's uncertainty evaluated for a measurement method.

The current 'best practices' approach for UQ of analytical methods is the GUM, JCGM:100 2008 [12]. The GUM is often referred to as a 'bottom-up' approach. The measurement method is described by a model equation where all input quantities comprising the final measurement result are stated. Each input quantity is assigned an uncertainty either through experimentation and appealing to the appropriate estimation procedure and often application of ANOVA or variance components estimation (this is referred to as Type A evaluation), or via other sources including expert knowledge, published data, reference material certificates, physics based limits, etc. (this is referred to as Type B evaluation).

The IAEA The IAEA uses many similar methods for MBE. The same statistical approaches are appealed to (most notably variance components estimation by ANOVA). The fundamental error model assumed for UQ of measurements taken for safeguards purposes includes variances

accounting for product variability, and also random and systematic error variances. The systematic error variance is historically modelled to represent the aggregate 'between inspection' shifts which can be due to many factors, including: changes in calibration, inspectors, background, and any other effects. The random error variance has been demonstrated to be the combination of pure random error (variance due to the repeatability of the measurement method) plus item specific bias because test samples are not true replicates (they are different sampled items from the facility).

Walsh et al [48] studied in detail one approach among an ensemble used by the IAEA to produce ITVs and uncertainty estimates for use in MBE (Grubbs' ANOVA applied to paired (operator, inspector) verification data obtained over multiple inspection periods) and revealed that the estimate of random error standard deviation can be almost interpreted as the inverse of method repeatability precision as defined in the international vocabulary of metrology, except for being larger by item-specific bias because test samples are not true replicates. The short-term systematic error variance estimate can be used in error propagation for MBE. MBE requires separately 4 variance components, i.e. the random and systematic error variance estimates of the operator and inspector measurement systems since (1) standard assumptions of the variance components imply that the random and systematic error variances propagate differently through an MBE calculation and (2) MBE comprises three statistical evaluations: operator's MUF, the D Statistics, and the *Inspector's estimate of Material Unaccounted For* (IMUF).

6. Complementarity of Approaches

Integrating competences across academic disciplines is a creative approach to reaching effective solutions. Going beyond disciplinarity might be remedial to problematic epistemological and political effects of excessive specialization. Therefore, disciplinary and interdisciplinary approaches should not be seen as mutually exclusive, but possibly complementary.

As discussed above, the authors are seeking reconciliation of terminological and conceptual differences in nuclear safeguards quantification of measurement uncertainties. The disciplinary UQ approaches involved are those of the measurement laboratories, metrology institutes, nuclear operators, and safeguards evaluators. These fields originally developed their approach for different specific purposes, and today the need for reconciliation is addressed.

The discussion over conceptual and terminological differences in section 4 showed that a language-mapping table could improve the understanding between communities adopting different approaches. Although the two main ways (top-down and bottom-up) of estimating uncertainty in measurements are not contradictory, and the GUM and the standard statistical error theory are consistent in the probabilistic UQ modelling, in a number of areas they differ substantially.

A review of statistical models and computational methods [57] highlighted the value they bring to the evaluation of measurement uncertainty. A number of problems still beyond the reach of GUM such as some aspects of calibration uncertainty, multi-dimensional absorption spectra, ILCs, and attribute testing, can be addressed using top-down UQ by long-standing observation equations (measurement error models) and statistical analysis [10, 55]. The bottom-up GUM approach is therefore complemented, rather than contradicted, by top-down statistical models and associated ANOVA-based variance component estimation and by Monte Carlo methods.

Contributors[6] influential to the GUM revision, propose to regard a measurement result as a *degree of belief* probability distribution for the measurand. Descending from the pattern of dispersion of values, as well as from uncertainties estimated by expert judgement [64], the probability distribution reveals the true character of the measurement uncertainty. The distribution could then be represented, for communication purposes only, by simpler summaries, such as the mean, mode, or others, and the standard deviation.

Multidisciplinary approaches to UQ thus complement disciplinary ones, introducing elements of elicitation and prior knowledge to the distribution of measurement results. This way of thinking is formalised in the Bayesian approach to inference, and it is identified as an area of potential expansion of GUM, both for bottom-up and top-down UQ, to address the challenges that measurement science will be facing in the years to come [47, 42].

In this article, deductive and inductive logical processes are addressed respectively in the top-down and bottom-up approaches. A deductive approach to processing information focuses to the most general first, and then narrows it down to the more specific. Conversely, inductive reasoning starts with specific observations of input quantities and then broadens the concepts up to generalisations and theories. It has to be noted that statistical methods stemming from inductive reasoning are intrinsic to both, the top-down as well as the bottom-up approach [66].

In practical terms, both statistical inference and probability theory are used in the metrological approach to UQ. Nevertheless, the input quantities that form the basis of the error model will drive the determination of the overall uncertainty. To illustrate this, the uncertainties associated with UO^2 drums measurements reported in section 3 are quantified focusing on the individual input quantities Xi from Eq. 1.

6 International conference that celebrated the twentieth anniversary of the GUM publication.

The uncertainties associated to the *Xi* are then propagated bottom-up to estimate the overall uncertainty on *Y*. Conversely, the same example is presented in section 5 by Walsh et al [48]. concentrating on the performance of the complete method. The reproducibility standard deviation, equivalent to the overall uncertainty on *Y* [48], is derived in top-down fashion by estimators applied to the ANOVA on paired measurement results.

As mentioned in section 4.3, MBE requires independent estimates of random and systematic components for the measurement uncertainties affecting the material balance [48]. Both bottom-up and top-down approaches can provide these estimates, but not without specific weaknesses, pointing to the possible advantage of a hybrid approach combining the two UQ methods.

1. The bottom-up approach does not necessarily model variation in all the effects influencing the measurement result [33, 34]. For example, in the UO_2 drums example that uses the EMP, variation in drum container thickness and self-absorption due to elemental matrix and its density are not fully accounted for. As discussed in section 3, calibration items differ from measured items, because calibrations could only be performed using drums containing reference material distributed differently than in measured drums. Hence, factoring expert knowledge into Monte Carlo simulations proved to be remedial. Failing to identify significant variation of input quantities could lead to uncertainty underestimation, which in safeguards terms translates into unnecessarily high false alarm rates.
2. The top-down approach assumes that all the variances associated with the input of the mathematical model vary representatively across the reproducibility study. However, variations associated with item-specific features and/or spectrum background cannot be perfectly accounted for based on the measurement method only, and would benefit from the expert judgement of an analyst to assess their impact on the overall measurement process. Failing to ensure representative variations in the course of the 3-year exercise discussed in section 5 [48], has the potential to lead to uncertainty overestimation. Thus, uncertainties potentially tuned to conceal nuclear material diversion could be deemed acceptable by the safeguards evaluators.

Comparing the two estimates discussed in section 3 and 5 is useful at this stage to assess the completeness of the UO^2-drums measurement model. The bottom-up UQ random RSD values cited in section 3 range from less than its 0.5% target to more than 1.0%, but less than the 1.81% RSD calculated in the top-down UQ approach applied in Section 5 [48]. For practical uncertainty estimates, therefore, it is recommended to appropriately use elements of both methods, in a hybrid, interdisciplinary way of thinking.

7. Benefits of Convergence

Professionals in Safeguards work in applied science, where scientific methods are developed to solve specific problems effectively, and then operationally optimized to make them fit for efficient production (be it performing measurements, analysing samples or evaluating data). Development and rehearsal of patterns of thought and their associated notation, terminology and jargon is part of the optimization process. Other than in the realm of pure science, basic assumptions and theories underlying the practically applied methods are not continuously questioned. Professional exchanges tend to be with experts in one's own field, who "speak the same language" and follow the same thought patterns. Discussions with experts in adjacent fields tend to remain at a shallower level, because of a lack of adequate understanding of each other's problem space, preferred solutions and accepted terminology.

Nevertheless, adjacent professional groups, such as staff at an analytical laboratory and staff of a statistical data evaluation group, can successfully collaborate under the premise of mutual recognition of expertise and as long as organizational and technical interfaces (such as distribution of responsibilities for the various process steps, and data exchange formats) are well-defined and respected. Seeking deeper understanding inevitably costs additional effort and may create insecurity and friction, as longstanding practices are being examined and criticised by knowledgable outsiders.

There are, however, at least two weighty reasons for why it is worthwhile to make the effort required to understand one's professional neighbours and to make oneself understood:

The first reason is following a broad and accelerating societal trend: authority, including professional authority, no longer goes unquestioned. Institutional status, educational credentials and a slightly aloof attitude do not bestow the expert with credibility. Credibility flows from an openness to review and the willingness to explain. And the capability to review and receptiveness to learning is easiest to find in adjacent professional communities. For this reason, no group of specialized experts can nowadays afford to not reach out to their neighbours.

The second reason is the opportunity to improve one's own approaches and practices by

- accepting, seriously considering and where useful incorporating constructive criticism, of which adjacent professional groups are uniquely capable; and
- enriching one's methodological portfolio by testing and adopting methods developed by adjacent professional groups.

As an example we consider potential benefits from convergence between the bottom-up approach and the top-down approach to measurement uncertainty estimation within the framework of the evaluation of the material balance of a bulk-handling nuclear facility, for example a fuel fabrication plant or an enrichment plant. A correct and credible assessment of measurement uncertainties is critical, because the aggregated and propagated uncertainties determine the variances of the fundamental statistics MUF, D and IMUF. While the expected value of these measures is zero, their actual values for each material balance period are non-zero and the crucial question to be answered in the evaluation of the facility is, whether the deviation from zero can plausibly, i.e. with reasonable confidence, be explained by legitimate measurement errors. Should the answer be no, alternative explanations, including the possible diversion of nuclear material, would need to be considered.

Bottom-up strives to understand all sources of uncertainty from first principles, exact knowledge of measurement practices and metrological traceability, see section 4.

Top-down uses statistical analysis to estimate and allocate uncertainties from paired data analysis, see sections 4 and 5.

Bottom-up analysis usually understates the variances of the MUF, D and IMUF statistics, as only known causes of uncertainty are within scope of the analysis. Uncertainty arising from unknown causes is itself, however, far from being an unknown phenomenon; it has been termed “dark uncertainty” [33]. Such unknown causes can, for example, be sampling errors due to material heterogeneity, chemical changes to the material over time, human mistakes or uncontrollable measurement conditions. Alique et al.[67] have presented a bottom-up methodology for estimation of sigma-MUF; in his example the bottom-up sigma-MUF is a factor of 71 smaller than sigma-MUF based on the ITV 2010 [3]. In the evaluation of nuclear facilities, dark uncertainty can be a large fraction of total uncertainty, and a decision criterion built exclusively on bottom-up uncertainty will tend to result in unacceptably high false alarm rates and can be perceived as unrealistic.

Top-down analysis provides a more realistic approach as, by construction, it takes into account all sources of uncertainty. However this should not lead to accepting (purposefully or not) poor performance, which would decrease detection probability. Therefore, a top-down -approach to UQ must be combined with a comparison with ITVs and a close monitoring of trends. The causes for significant changes must be investigated by obtaining additional information about measurement conditions and procedures.

So joining forces is an attractive proposition: “Bottom-uppers” can use paired data analysis and three-lab analysis based on Grubbs [21] and subsequently improved methods to allocate uncertainty components, and to quantify dark uncertainty not yet covered in their uncertainty budgets. “Top-downers” can feed prior knowledge on the uncertainty of certain measurement methods into their analysis and use the method of uncertainty-budgeting to identify dark uncertainty. A jointly derived decision criterion would have a good chance of striking a better balance between the twin risks of non-detection and false alarms. Subsequently, both groups can collaborate in characterizing and reducing the existing unknown sources of uncertainty and thereby increase the effectiveness of safeguards. Motivated by this perspective, the mathematical equivalence of paired data analysis as practiced by IAEA with GUM-based methods has recently been demonstrated [48].

The dialogue between metrologists, statisticians, mathematicians and laboratory professionals is in full bloom today [55], and this attempt to foster a mutual understanding between the laboratory and evaluator community in safeguards is believed to be of interest also to other measurement and evaluation disciplines [68, 69, 70]. It is even considered as a potential contribution to the ongoing process of the GUM revision.

It is the authors’ wish that our article should motivate and facilitate this kind of fruitful collaboration.

8. Acknowledgements

The internal review of this article by Oscar Zurron Ciffentes was highly appreciated. Further, the authors would like to acknowledge feedback and comments from the EC-JRC colleagues, Roger Wellum, Evelyn Zuleger, Stefaan Pommé, Stephan Richter and Rožle Jakopič, from Guy Granier from the CEA/CETAMA and from John Howell, University of Glasgow (emeritus). We also would like to thank all the participants in the ESARDA Joint Workshop on Applied Metrology & Material Balance Evaluation and the IAEA International Technical Meetings on Statistical Methodologies for Safeguards.

9. References

[1] NUREG/CR-4604,Statistical Methods for Nuclear Material Management, PNL-5849, December 1988

[2] IAEA Safeguards Glossary 2001, Edition International Nuclear Verification Series No.3

[3] IAEA-STR-368 International Target Values for Measurement Uncertainties in Safeguarding Nuclear Materials, Vienna; November 2010

[4] Aregbe Y, Jakopič R, Richter S, Venchiarutti C; Conformity assessment in nuclear material and environmental sample analysis; Proceedings Symposium on International Safeguards Linking Strategy, Implementation and People; 2014; S13–01

[5] IAEA-TECDOC-261; IAEA Safeguards Technical Manual Part F: Statistical Concepts and Technique, Vol.3; Vienna; 1982

[6] Jaech J; Statistical Methods in Nuclear Material Control, TID-26298; National Technical Information Service, U. S. Department of Commerce, Springfield, Virginia, USA;1973

[7] IAEA/SG/SCT/5, Statistical Concepts and Techniques for IAEA Safeguards, Fifth Edition; IAEA, Vienna;1998

[8] Martin K; IAEA Technical Report Estimation Of Variances Of Random And Short-Term Systematic; Measurement Errors Based On Data From Two And Three Independent Measurement Methods; unpublished manuscript; 2016

[9] BIPM 1980; Report of the BIPM Working Group on the Statement of Uncertainties to the Comite International des Poids et Mesures; www.bipm.org/utils/common/pdf/WGUncertainties1980.pdf

[10] Cox M, Harris P; GUM anniversary issue; Metrologia 51; 2014; S141–S143

[11] ISO/IEC Guide 98:1993; Guide to the expression of uncertainty in measurement (GUM)

[12] ISO/IEC Guide 98-3:2008; Guide to the expression of uncertainty in measurement (GUM)

[13] ESARDA 1979; Target Values" for uncertainty components in DA methods presented to Euratom and IAEA

[14] ESARDA 1983; Target values for uncertainty components in fissile element and isotope assay – achievable uncertainties in destructive assay of nuclear material, 6th ESARDA Symposium, Venezia, Italy, May 14-18, 1984

[15] De Bièvre P, Baumann S, Gorgenyi T, Kuhn E, Deron S, Dalton J, Perrin R E, Pietri C, De Regge P; Target values for uncertainty components in fissile isotope and element assay; Journal of the Institute of Nuclear Materials Management; 15(4); 1987; p. 99-104

[16] Kuhn, E., et al.,1993 International target Values for Uncertainty Components in Fissile Isotope and Element Accountancy for the Effective Safeguarding of Nuclear Materials; IAEA STR-294; Vienna; 1994.

[17] International Target Values 2000 for Measurement Uncertainties in Safeguarding Nuclear Materials; IAEA-SM-367/5/01

[18] https://esarda.jrc.ec.europa.eu/index.php?option=com_content&view=article&id=242:report-on-the-workshop-on-reference-material-needs-and-evaluation-of-uncertainties-in-da-and-nda&catid=91&Itemid=331

[19] IAEA 2013 Meeting Minutes, IAEA International Technical Meeting on Statistical Methodologies for Safeguards; 16-18 October 2013, Vienna

[20] 36th ESARDA Annual Meeting 2014; Report on the ESARDA Joint Workshop 'Applied Metrology & Material Balance Evaluation'; Luxembourg 12 May 2014

[21] Grubbs F E; On Estimating Precision of Measuring Instruments and Product Variability, Journal of the American Statistical Association 43; 1948; 243-264

[22] ISO/IEC 17025:2017; General requirements for the competence of testing and calibration laboratories, International Organization for Standardization; Geneva

[23] ISO/IEC 17043: 2010; Conformity assessment – general requirements for proficiency testing, International Organization for Standardization, Geneva

[24] ISO Guide 34:2009; General requirements for the competence of reference material producers, International Organization for Standardization, Geneva

[25] ISO 17034:2016, General requirements for the competence of reference material producers, International Organization for Standardization, Geneva

[26] Bürger S, Essex R M, Mathew K J, Richter S, Thomas R B; Implementation of Guide to the expression of Uncertainty in Measurement (GUM) to multi-collector TIMS uranium isotope ratio metrology; 2010 International Journal of Mass Spectrometry 294 65–76

[27] Richter S, Venchiarutti C, Hennessy C, Jacobsson U, Bujak R, Truyens J, Aregbe Y; Preparation and certification of the uranium nitrate solution reference materials series IRMM-2019 to IRMM-2029 for the isotopic composition; Journal of Radioanalytical and Nuclear Chemistry 318:1359–1368; 2018

[28] ASTM C1832 – 16 2016 Standard Test Method for Determination of Uranium Isotopic Composition by the Modified Total Evaporation (MTE) Method Using a Thermal Ionization Mass Spectrometer

[29] Richter S, Alonso A, Truyens J, Kühn H, Verbruggen A, Wellum R; REIMEP-18: Interlaboratory Comparison for the Measurement of Uranium Isotopic Rations in Nitric Acid Solution; Report EUR 22529; 2006

[30] Jakopic R, Bujak R, Aregbe Y, Richter S, Buda R, Zuleger E; REIMEP-17: plutonium and uranium amount content, and isotope amount ratios in synthetic input solution; Report EUR 26667 EN; 2014

[31] Jakopic R,Bujak R, Aregbe Y, Richter S, Buda R, Zuleger E; Results of the REIMEP-17 interlaboratory comparison for the measurement of the U and Pu amount content and isotope amount ratios in the synthetic dissolved spent nuclear fuel solution; Accred Qual Assur 20; 2015; 421–429

[32] Mason P; An Evaluation of Uranium Measurement Capabilities and Comparison to State-of-the-Practice Target Values, including an Examination of Historical Performance; NBL-2011-ME Annual Report; 2011

[33] Thompson M, Ellison S L R, Dark uncertainty; Accred Qual Assur 16; 2011; 483–487

[34] Walsh S J, Venzin A, Wegrzynek D, Mansoux C; Using reproducibility to test the adequacy of GUM based uncertainty quantification; American Nuclear Society Advances in Nuclear Non-proliferation Technology & Policy Conference 25-30 September, 2016, Santa Fe NM

[35] Pendrill L R; Using measurement uncertainty in decision-making and conformity assessment; Metrologia 51; S206–S218; 2014

[36] Clark D L, Geeson D A, Hanrahan R J; Plutonium Handbook, 2nd edition, Volume 4; American Nuclear Society; 2019

[37] ASTM C1672 – 17; Standard Test Method for Determination of Uranium or Plutonium Isotopic Composition or Concentration by the Total Evaporation Method Using a Thermal Ionization Mass Spectrometer; 2017

[38] Vogl J; Characterisation of reference materials by isotope dilution mass spectrometry; Journal of Analytical Atomic Spectrometry 22; 2007; 475–492

[39] JCGM 200:2012 International Vocabulary of Metrology – Basic and General Concepts and Associated Terms: https://jcgm.bipm.org/vim/en/index.html

[40] Burr T, Sampson T, Vo D; Statistical evaluation of FRAM γ-ray isotopic analysis data Applied Radiation and Isotopes 62; 2005; 931-940

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

[42] Burr T, Krieger T, Norman C; Approximate Bayesian Computation Applied to Nuclear Safeguards Metrology ESARDA BULLETIN No. 57; December 2018

[43] ASTM C1514; Standard Test Method for Measurement of 235U Fraction Using the Enrichment Meter Principle; 2008

[44] Burr T, Croft S, Jarman K, Uncertainty Quantification in Application of the Enrichment Meter Principle for Nondestructive Assay of Special Nuclear Material; Journal of Sensors 15, Article ID 267462; 2015

[45] 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; 2016; 49-57

[46] Burr T, Croft S, Dale D, Favalli A, Weaver B, Williams B.; Emerging Applications of Bottom-Up Uncertainty Quantification in Nondestructive Assay; ESARDA Bulletin 53; 2015; 54-61

[47] Elster C; Bayesian uncertainty analysis compared with the application of the GUM and its supplements; Metrologia 51; 2014; S159–S166

[48] Walsh, S, Burr T, Martin K; Discussion of the IAEA Error Approach to Producing Variance Estimates 4 for Use in Material Balance Evaluation and the International Target Values, and Comparison to Metrological Definitions of Precision; Journal of Nuclear Materials Management; Volume XLV; 2017; 4-14

[49] R Core Team; R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria; 2018; Available online at https://www.R-project.org/

[50] Burr T, Croft S, Jarman K, Nicholson A, Norman C, Walsh S; Improved uncertainty quantification in nondestructive assay for nonproliferation Chemometrics and Intelligent Laboratory Systems; 2016; 159 164-173

[51] Burr T, Pickrell M, Rinard P, Wenz T; Data mining: applications to nondestructive assay data. Journal of Nuclear Materials Management 27(2); 1999; 40-47

[52] Dufour J L, Pepin N, Deyglun, Weber A-L; Optimisation and uncertainty estimation of the enrichment meter measurement technique for UF6 cylinders; ESARDA BULLETIN No. 59; December 2019

[53] Bich W; Revision of the 'Guide to the Expression of Uncertainty in Measurement'. Why and how; Metrologia 51; 2014; S155–S158

[54] Bonner E,. Burr T, Krieger T, Martin K, Norman C; Comprehensive Uncertainty Quantification in Nuclear Safeguards Science and Technology of Nuclear Installations Hindawi Science and Technology of Nuclear Installations; Volume 2017; Article ID 2679243; 2017;

[55] Possolo A; Statistical models and computation to evaluate measurement uncertainty; Metrologia 51; 2014; S228–S236

[56] Ehrlich C; Terminological aspects of the Guide to the Expression of Uncertainty in Measurement (GUM); Metrologia 51; 2014; S145–S154

[57] Hagan A O; Eliciting and using expert knowledge in metrology; Metrologia 51; 2014; S237–S244

[58] EURACHEM; Terminology in Analytical Measurement – Introduction to VIM 3, first edition 2011, https://www.eurachem.org/index.php/publications/guides/48-gdtam11

[59] Willink R; Measurement Uncertainty and Probability, Cambridge University Press Cambridge;2013

[60] https://ec.europa.eu/jrc/en/interlaboratory-comparisons

[61] Crozet M, Roudil D, Rigaux C, Bertorello C, Picart S, Maillard C; EQRAIN: uranium and plutonium interlaboratory exercises from 1997 to 2016—comparison to ITVs-2010 Journal of Radioanalytical and Nuclear Chemistry 319; 2019; 1013–1021; https://cetama.partenaires.cea.fr/

[62] https://science.osti.gov/nbl/Programs/Measurement-Evaluation

[63] ISO 13528:2005: Statistical methods for use in proficiency testing by inter-laboratory comparisons

[64] ISO 21748:2017 Guidance for the use of repeatability, reproducibility and trueness estimates in measurement uncertainty evaluation

[65] ISO 5725: Accuracy (trueness and precision) of measurement methods and results; 1994

[66] Sprenger J; Statistics between inductive logic and empirical science Journal of Applied Logic 7; 2009; 239–250

[67] Alique O, Vaccaro S, Svedkauskaite J; The use of measurement uncertainty in nuclear material accountancy and verification; ESARDA BULLETIN, No. 52; 2015; 35-40

[68] ESARDA World Café Report – Stresa 2019; Implementation of the RG2019 Actions and Roadmap: https://esarda.jrc.ec.europa.eu/index.php?option=com_content&view=article&id=392:world-cafe-report&catid=23&Itemid=101, https://esarda.jrc.ec.europa.eu/images/files/miscellaneous/ESARDA%20World%20Caf%D0%92%20report_final.pdf

[69] Lindstrom R M; 2017 Believable statements of uncertainty and believable science J Radioanal Nucl Chem 311:1019 1022

[70] Niemeyer I, Dreicer M, Stein G; Nuclear Non-proliferation and Arms Control Verification; 2020; Springer https://doi.org/10.1007/978-3-030-29537-0