Proceedings of the INMM & ESARDA Joint Virtual Annual Meeting
August 23-26 & August 30-September 1, 2021

USING REPLICATIVE ASSESSMENT OF SPECTROSCOPIC EQUIPMENT (RASE) SOFTWARE TO DEVELOP VIRTUAL RADIATION DETECTION EQUIPMENT TESTING CAPABILITIES

A.S. Moore
PNNL

A. Kriss
PNNL

I. Garcia
PNNL

C. Simpson
PNNL

E. Gordon
PNNL

M. Woodring
PNNL

ABSTRACT A recent effort to validate the performance of the Replicative Assessment of Spectroscopic Equipment (RASE) software has been undertaken at Pacific Northwest National Laboratory by the Science and Engineering Team (SET) for the U.S. Department of Energy National Nuclear Security Administration's Office of Nuclear Smuggling Detection and Deterrence (NSDD). NSDD supports a large array of testing and evaluation campaigns to assess the performance of radiation detection instruments in various mission spaces before deployment. These campaigns are necessary to determine whether a system will fulfill the desired detection and performance capabilities of that mission space, yet they require large investments for the procurement of Systems Under Test (SUT), as well as scientist's time for in-laboratory testing and data analysis. RASE was developed to enable virtual testing of radiation detection instruments in a simulated environment. RASE down-samples and adjusts "base spectra" created from physically obtained source measurements to generate simulated spectra for a variety of source isotopes, background environments, dose/flux rates, and measurement durations. When combined with a manufacturer-provided "replay tool," these spectra can be processed using the instrument algorithm to obtain simulated real-world responses. The adoption of RASE into future test campaigns offers resiliency by using simulated testing capabilities rather than relying solely on in-field measurements, best serving the ongoing transition toward a more virtual and remote landscape. RASE allows test and evaluation teams to select only those instruments that pass predefined detection and/or identification thresholds while informing what in-field measurements would need to be conducted for the desired evaluation criteria. New algorithms and updates can also be tested in RASE before deployment. While RASE has the potential to reduce future testing costs, a broad validation effort had not yet been performed. This validation effort will be summarized, with results compared to experimental data from several previous in-field test campaigns.

INTRODUCTION

The U.S. Department of Energy National Nuclear Security Administration's Office of Nuclear Smuggling Detection and Deterrence (NSDD) commonly supports testing and evaluation campaigns to assess radiation detection equipment performance before its selection and deployment worldwide at border crossings to assist with its nuclear interdiction mission. These instruments include commercially available handheld systems, radioactive isotope identification devices (RIDs), wearable systems, Backpack Radiation Detectors (BRDs), portable area spectrometers (ASs), radiation portal monitors (RPMs), as well as some personal radiation detectors (PRDs).

Some mission applications allow for identification measurements performed by operators using handhelds. Others may require the detection and localization of radiation at a site. In some scenarios detection equipment must be capable of simultaneously detecting, locating, and identifying a radioactive source. Additionally, as nuclear medicine continues to expand and incorporate additional sources for treatment, these signatures must be considered in these environments as well to evaluate their impact on false alarms. New mission spaces are constantly evolving, as is the technology and associated detection algorithms developed and deployed to meet these, requiring routine evaluation.

When testing, these SUTs are often subjected to the same radiation environments they may encounter in the field and compared alongside each other to determine the relative performance criteria and optimal settings for a variety of metrics. Not only is demonstrating successful radiation detection capability performance crucial, so are considerations for the system's use, sustainability, and maintenance to support ease of use for the users, longevity of the inventory, and ease of upkeep in the field. This process requires the procurement of each SUT, sometimes multiple to test consistency, and requires numerous lab hours to design the testing plan as well as execute and analyze the results.

The RASE software has been developed by a team at Lawrence Livermore National Laboratory (LLNL) to enable virtual testing of some of these capabilities. RASE only allows virtual assessment of the identification functions of a gamma detector. Localization and detection are not able to be assessed in this software. A validation effort was undertaken using RASE to investigate its performance as well as indicate where it could be used to save time and effort in a comparison to lab measurements The different features of RASE will be discussed in this work, highlighting how they may support future virtual testing for NSDD, and other relevant missions, as a parallel to measurements conducted in the lab.

TESTING AND EVALUATION OF VARIOUS RADIATION DETECTION INSTRUMENTATION FOR BORDER SECURITY

In a typical testing and evaluation campaign, a series of characterization measurements of a variety of systems are conducted. All SUT are subjected to the same conditions to compare performance as well as measures of sustainability and suitability. These conditions reflect radiation fields created by different levels of background, naturally occurring radioactive materials (NORM), or special nuclear material (SNM) sources, as well as operational considerations such as battery lifetime, ease of use, and weight. A test plan is designed to best align these laboratory tests to what SUTs would experience during their use for each appropriate deployment scenario.

These campaigns are useful for determining whether a system will fulfill the desired detection and performance capabilities of that mission space; they help to improve the understanding of deployed systems used by the global community when tested as the benchmark to new systems, as well as inform vendors of possible areas for improvement in their new systems or detection algorithms as they become available. The primary categories, and the objectives of, these tests are provided below. They require significant investments for the procurement of SUT, as well as considerable staff time for laboratory characterization, testing and data analysis.

P_{ID} Metrics

The Probability of Identification (P_{ID}) is an important metric to study. The P_{ID} is the probability of positive identification, found through the ratio of successful to total trials. A successful trial is dictated by the source being measured, and the objectives of the characterization and evaluation. For example, for some recent tests, in background a successful identification is no isotopes are present, or a background isotope identification such as K-40, Ra-226, Th-232 (or daughters), "background," or "NORM." For a source test case, the correct identification is that source, either on its own or in combination with a background source identification such as a K-40, Ra-226, Th-232, or a "NORM" identification or in combination with an additional non-SNM source . Additional non-SNM sources would not trigger the same type of response and adjudication process as would be needed for an SNM identification. Typically, the P_{ID} value is recorded for a range of SNM source fluence values, based off the primary photopeak of the isotope of interest, and used to produce a sigmoidal curve using a Boltzmann function, or "S-Curve" due to its shape (as shown in Figure 1). These S-Curves provide a comparable metric across SUT, with confidence intervals calculated for each point.

Correct identification rates are significant to the mission space as a system must be able to identify a source of interest (e.g., SNM) with a given level of confidence, and it must be able to do so within a realistic fluence range (related to distance from the source) that would be applicable during operation with associated shielding.

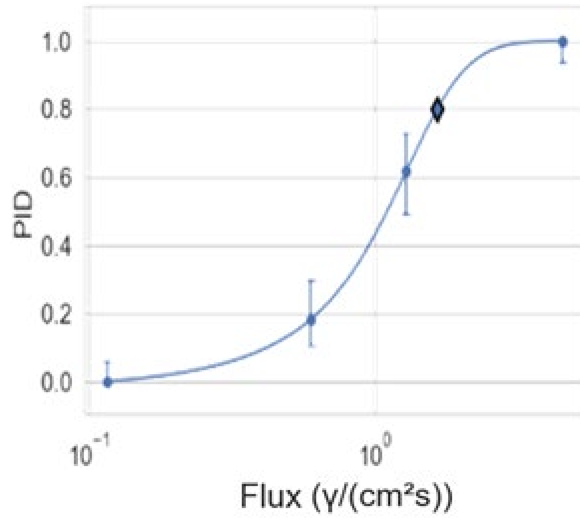

Figure 1. Example S-curve plot of the P_{ID} as a function of source fluence values.

P_{FID} Metrics

The Probability of False Identification (P_{FID}) is another significant metric to evaluate before selecting SUT for deployment. A false positive identification is operationally significant as the operator needs to adjudicate that the alarm is not caused by the presence of a source of interest, a process that affects the flow of passengers or commerce. In background conditions, a false positive result could be the identification of any non-background isotope or an "unknown," the intrinsic calibration source within the system, or a medical source. The identification of "background," "NORM," or a NORM isotope such as K-40, Th-232 (or daughters), or Ra-226, are not considered a false identification. In the presence of NORM (e.g., Ra-226 and Th-232 sources), a false positive identification then also includes any additional identification made alongside the present source, including an "unknown." Finally, a false positive in the presence of a source of interest like SNM would encompass an "unknown," an indication that no sources are present, no SNM sources are present, or identifying any additional SNM isotopes that are not present in the configuration. All of these false identifications would prompt the operator to seek further action to adjudicate. As part of the mission, it is crucial to prevent user fatigue, which could be caused by a high level of false identifications. Each false alarm compounds on this impact on operations.

When analyzing this performance, a box and whiskers plot is oftentimes used to show the P_{FID} and the upper and lower confidence intervals for a 95% confidence, indicating with 95% certainty that in independent samples, the statistical mean will fall within those bounds 95% of the time. So, if 100 sets of 1,500 trials were taken, it would be expected that 95 of the sample sets would exhibit a mean that falls within this confidence interval. Typically, this P_{FID} rate is compared across all SUT in the test campaign and weighted into the selection of the deployed system, selecting a system with a low P_{FID} rate. Here, the P_{FID} is instead compared between a set of original experimental data and the simulated data set using RASE with the same measurement parameters (rather than as an SUT-to-SUT comparison) (Figure 2).

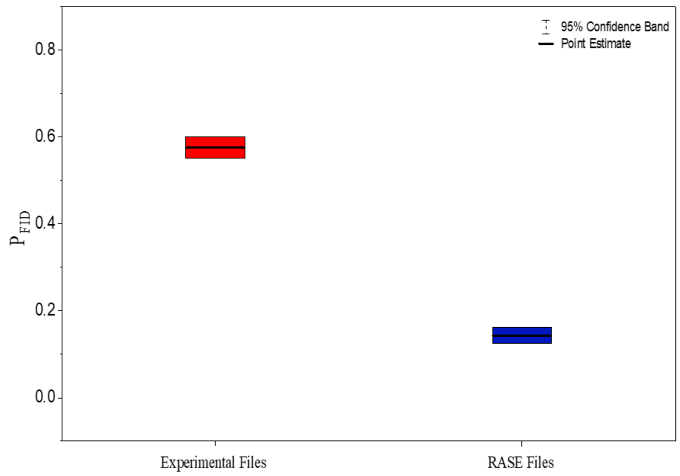

Figure 2. An example of a comparison of P_{FID} performance between original experimental files and RASE-simulated files with the same parameters.

Through these tests and comparisons, some SUTs are eliminated from future consideration, while others must be reworked by the vendor to improve them to a point of consideration. This modification could include new algorithm updates or updates to the nuclide libraries to reach this performance. These updates would need to be tested again to evaluate their impact on the system in an iterative process or a future campaign.

VIRTUAL TESTING AND EVALUATION

Various simulation software exists that can be used to model different source strengths, compositions, spectra, shielding, and distances from detectors to supplement physical characterization and testing. Additionally, software tools can be used to approximate the performance of a detector, based on the detector response function. Simulation software is commonly used for scenarios that may be complex and/or expensive to measure in the lab (e.g., a masking scenario, multiple source scenarios, various background conditions, rare SNM sources that may not be on-hand, etc.), enabling a broader understanding of the performance of an instrument under those conditions at very low costs. Or it can be used when the instrument is not readily available, but its performance has been benchmarked in the past, and new test cases need to be studied. One such tool is Gamma Detector Response and Analysis Software (GADRAS).

GADRAS

GADRAS [1] is a gamma-ray spectral analysis and modelling software toolset developed at Sandia National Laboratory for use in analyzing data from scintillators and semiconductors. GADRAS models include both source terms and detector response to match data acquired with characterized detection systems. GADRAS can also create spectra for use in cases that cannot easily be realized in experiments, such as stepped variation of SNM geometry, mass, or activity. GADRAS computes templates that are used to analyze gamma-ray spectra from detector response functions. The local environment around a detector can be altered via parameters that represent scattering from the environment surrounding a detector. This enables the user to generate spectra with correct features. GADRAS uses a full-spectrum analysis method for simulated gamma-ray spectra, where an entire spectrum is fit with one or more computed spectral templates. This approach differs from other analysis methods in which radionuclide concentrations are determined by finding the areas of characteristic photopeaks while ignoring the continuum regions. A list of GADRAS simulation capabilities is given below:

- Outputting data that can be used to generate instrument files in the correct format.
- Running in batch mode, thus being able to generate multiple examples of a single test case.
- Accounting for varying backgrounds between 5 to 20 μR/hr.
- Handling background suppression.
- Ability to simulate large, distributed sources together with point-like sources.
- Variation of source location within a cargo container.

RASE CAPABILITIES

RASE "is a semi-empirical approach to generating synthetic gamma-ray spectra for injection into a radionuclide identification algorithm of a vendor-provided radiological detection system" [2]. It down-samples and adjusts "base spectra" created from physically obtained source measurements, to generate simulated spectra for a variety of source isotopes, dose/fluence rates, and measurement durations. When combined with a manufacturer-provided "replay tool," these spectra can be processed using the instrument algorithm to obtain simulated real-world responses. The software could ultimately be used as an alternative to some aspects of labor-intensive measurement campaigns. GADRAS-simulated spectra are also directly compatible with RASE to enable fully virtual testing for various scenarios.

The RASE user interface is a GUI [graphic user interface] with buttons and menu dialog windows that the user can select and edit their inputs within. First, the base spectra must be created. This can be done using the Base Spectra Creation tool where the user loads in the measured .n42 files of the foreground measurement of each desired source. The associated fluence or exposure rate at which each spectrum was acquired in must be entered, a description can be used to distinguish files, and a background file must be indicated to form the proper background subtraction for each source file. The tool will then output the base spectra for use in RASE. Within this new .n42 file, a RASE sensitivity factor, or RASE flux sensitivity factor, will be generated that considers the fluence or dose of the relevant photopeak as well as the background and foreground count rates.

The user must also specify an instrument, where they input the name, select the replay tool, assign base spectra, and can define any influences. Finally, scenarios must be created that reflect the radiation environment. Within this menu dialog the user has the ability to combine several foreground sources from the available base spectra for that instrument, with a background configuration, and indicate each of their fluence values. The number of replications (mirroring the number of trials measured in the lab) and the acquisition time can be set. This can be done for a range of fluence values or for a variety of source environments. Then, RASE will produce the P_{FID} and P_{ID} metrics in a table format that should simulate the detector performance if that software version were to be tested in the lab. Additionally, S-Curves can be produced in RASE with associated confidence intervals of the P_{ID} based on the user's selection. All of these data can be exported and saved for further analysis and reference.

Different replay tool versions can be tested across the same scenarios for an algorithm comparison, different background conditions may be tested, and different acquisition durations and number of replications can be studied for their impact on the final results.

Limitations

RASE on its own does not perform start to finish simulations where the detector, base spectra, and sensitivity factors can be modeled internally; these spectra and replay tools must be available. Dynamic identification scenarios are not currently supported by RASE at the time this paper was written but it is under development. Time variations on data, such as drift, are also not currently supported by RASE at the time of this paper. Results cannot be of higher quality than the incoming long dwell measurement used to build the base spectra; therefore, the selected data should be high fidelity long dwell static measurements for each source and SUT to ensure the desired statistics on the simulated results.

RASE also needs the applicable vendor-provided replay tool, a proper .n42 template, and a data file translator to operate. Replay tools are provided by the manufacturer of the algorithm in-place on a detector. Version capabilities must be consistent between the software/firmware used on the instrument during data collection, and the version in place within the replay tool; the compatible algorithm version of the tool must be determined to ensure agreement between the original experimental files and the replayed data files before RASE execution to prevent bias in the results. This is dependent on whether or not large changes were made by the vendors to the detection and identification algorithms between versions. Additionally, it is possible that the same algorithm version replay tool may not produce identical results to those measured in the lab when deployed on the system. Finally, the current version of RASE cannot be used to vary background conditions for instruments with intrinsic calibration sources.

VALIDATION OF THE SOFTWARE PERFORMANCE

While RASE has the potential to reduce future testing costs, a broad validation effort had not yet been performed. Data from four previous PNNL test campaigns were used in an effort to evaluate the performance of RASE for four commonly tested SUTs, selected based on the availability of compatible replay tools and result translators. Experimental data taken using three RIDs and one BRD were evaluated against RASE simulations for matching these test conditions with SNM, background, and NORM sources. Additional RID, BRD, AS and RPM systems have also been widely tested in previous campaigns, however they did not have the necessary replay tools available from the manufacturer in order to execute RASE.

Original experimental results for: false alarm and identification rates in background, false identification rates in the presence of NORM, and probability of identification vs. source fluence rates in a static scanning mode from each test campaign were compared to RASE simulated results with focus on P_{ID} and P_{FID} metrics. Five scenarios were used as part of an overall validation of RASE against previously measured data. The combination of these validation efforts represents a subset of scenarios designed and conducted in a physical measurement test campaign, illustrating how RASE may be used in a realistic way for instrument evaluation.

From these efforts, numerous comparisons were made between the experimentally measured SUT performance and the simulated performance across different sources. Some examples of these findings are provided in Figure 3 through Figure 8, highlighting both SUT-source configurations that agreed, and others that were not. It was shown that it is possible to replicate a SUT's performance through simulation with RASE, mirroring responses virtually that could be measured in the lab. However, this effort revealed that RASE and the replay tools could benefit from additional development to make them easier for less experienced users to reach agreement across all SUT-source configurations.

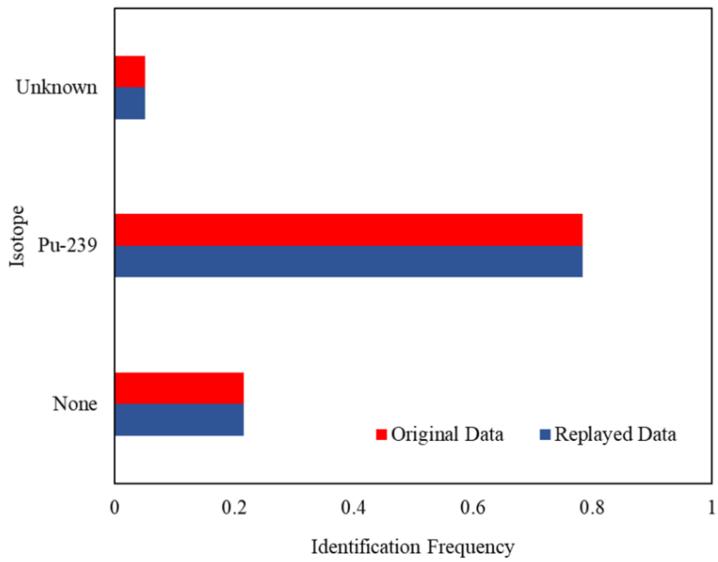

Figure 3. The isotopic frequency comparison of original experimental results to the replayed analysis results for SUT 1, highlighting their full agreement.

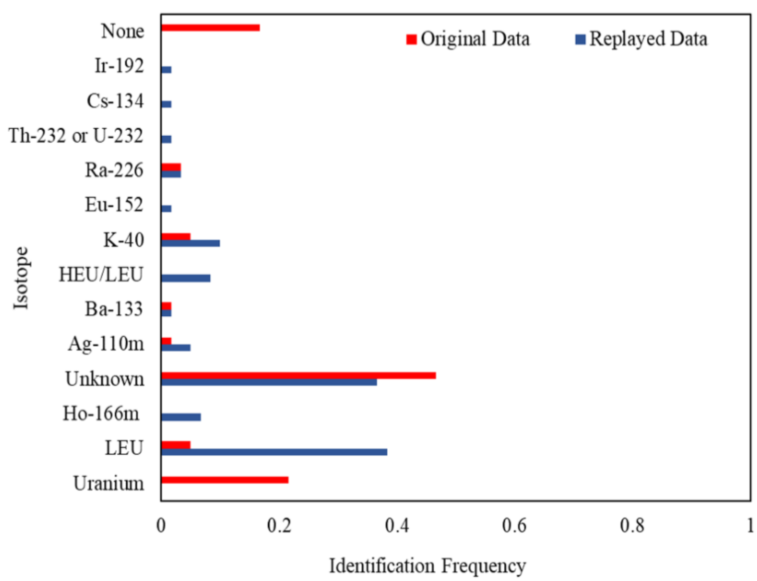

Figure 4. The isotopic frequency comparison of original experimental results to the replayed analysis results for SUT 2, highlighting their disagreement.

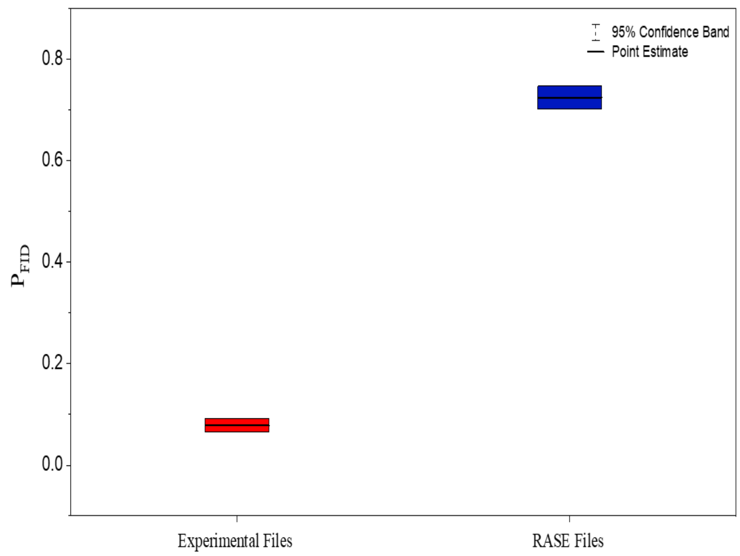

Figure 5. The P_{FID} performance comparison between the original experimental file analysis and the RASE-simulated file analysis for SUT 2. There is a significantly greater P_{FID} for the simulated files for this SUT-source combination.

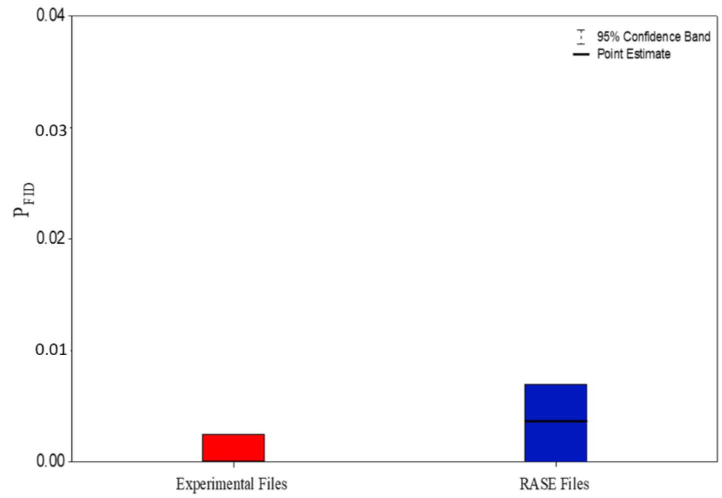

Figure 6. The P_{FID} performance comparison between the original experimental file analysis and the RASE-simulated file analysis for SUT 3. There is better agreement in the P_{FID} between the original results and the simulated files for this SUT-source combination.

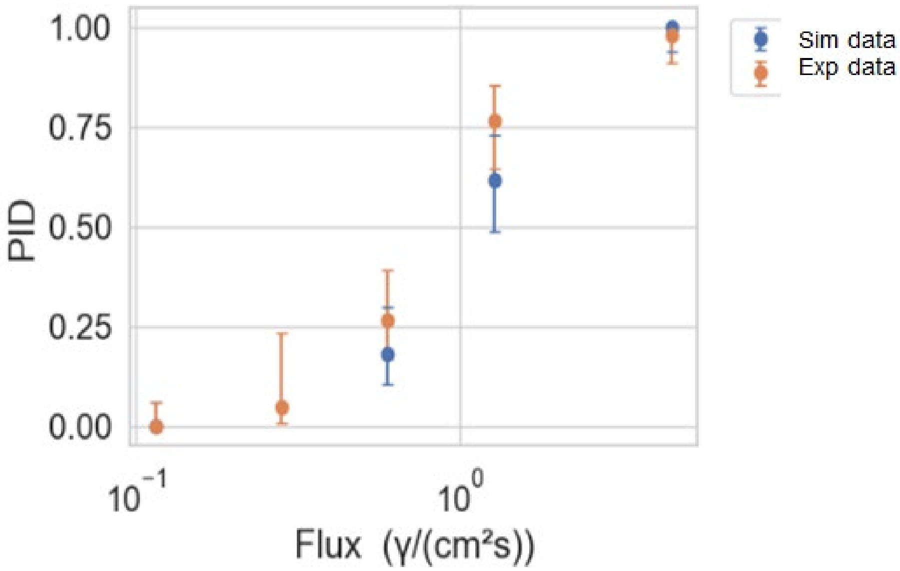

Figure 7. The S-curve comparison between original experimental data and the RASE-simulated data results for an HEU configuration, with the same measurement parameters for SUT 1. This figure highlights their agreement.

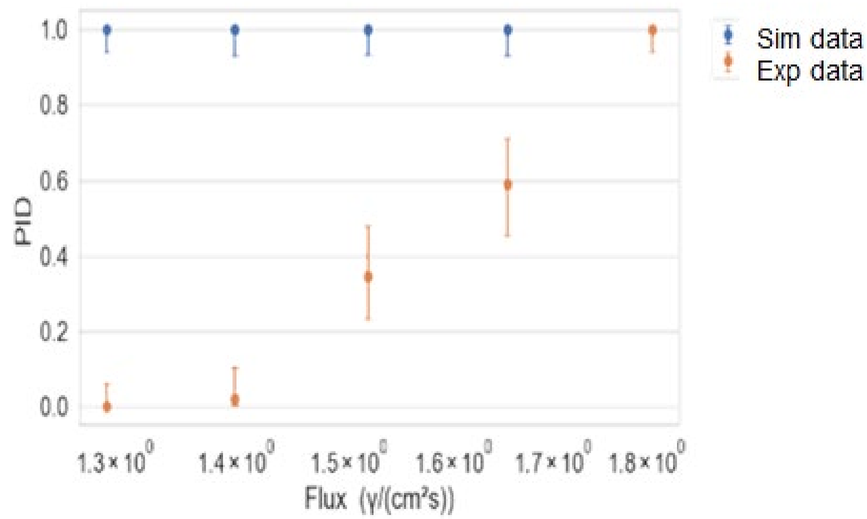

Figure 8. The S-curve comparison between original experimental data and the RASE-simulated data results for an HEU configuration, with the same measurement parameters for SUT 3. This figure highlights their disagreement.

FUTURE USE

RASE is a useful tool that will help with future testing efforts in several ways including: determining the SUT-to-source distances for P_{ID} vs fluence measurements to minimize the number of measurements needed to span this range through the utility of the S-curve plotting, evaluating the identification performance of different software versions, as well as evaluating the P_{FID} performance of the SUT in background and NORM conditions.

Once confidence has been established with the tool, RASE results can be used to reconfirm physical measurements while informing what in-field measurements would need to be conducted for the desired evaluation criteria, or validate/assess regression or improvement when manufacturers make changes to algorithms. This will aid in test planning. New algorithms and updates can also be tested in RASE before deployment.

However, there are certain aspects of virtual testing that will not be able to substitute for laboratory measurements. These include the maintenance and sustainability considerations that need to be evaluated by working with the system. Additionally, any time a new SUT or a new algorithm is introduced that has not been tested or integrated within RASE, it would be good practice to take several benchmark measurements in the lab to compare against the simulated results to ensure confidence in them. Ultimately, RASE can be used to decrease testing costs and timelines and will complement additional performance evaluation measurements conducted in the lab.

ACKNOWLEDGMENTS

This work was funded by the U.S. Department of Energy National Nuclear Security Administration's Office of Nuclear Smuggling Detection and Deterrence. The authors would also like to thank the RASE Developers at Lawrence Livermore National Laboratory for their help, patience, and useful information provided while working to perform these validation efforts.

REFERENCES

[1] Theisen, Lisa Anne; Mitchell, Dean James; Thoreson, Gregory G.; Harding, Lee T.; Horne, Steve; Bradley, Jon David et al. GADRAS-DRF user's manual., report, September 1, 2013; Albuquerque, New Mexico. (https://digital.library.unt.edu/ark:/67531/metadc841144/ : accessed July 20, 2021), University of North Texas Libraries, UNT Digital Library, https://digital.library.unt.edu; crediting UNT Libraries Government Documents Department.

[2] RASE Software User Manual, Release v2.0, LLNL, Oct. 12, 2020