Validating the Probability of Raid Annihilation Testbed Using a Statistical Approach

Dean Thomas and Rebecca Dickinson

THE PROBLEM

Modeling and Simulation (M&S) often provides essential information in evaluations of operational effectiveness, suitability, and survivability, especially in cases where end-to-end missions cannot be assessed because of safety, cost, or test range restrictions. Before M&S is used, analysts should validate the model to ensure that it reasonably represents the real world. Unfortunately, in operational testing it is often the case that only limited data are available for validation.

Live test events of new weapon systems are often expensive, and only a limited number of test events can be conducted. A well-designed test will intelligently distribute such events across the operational envelope. Nonetheless, when only limited data are available, there will be holes in our understanding of system performance. M&S can be used to extend the test results throughout the operational envelope. Validation is the process of determining the extent to which the M&S adequately represents the real world for its intended use. Thus, a question that testers often ask is how to best use a small number of live test results to validate that the M&S is providing meaningful results.

The Navy's Air Warfare (AW) Ship Self-Defense (SSD) Enterprise is an overarching test methodology that examines the ability of shipboard combat systems to defend a ship against a cruise missile attack. The primary metric is Probability of Raid Annihilation (PRA), which is the probability of defeating the entire raid of cruise missiles through a combination of reduced ship signature, missile and gun systems, and decoys and countermeasures. The AW SSD Enterprise uses a combination of live test results from a fleet ship, live test results from an unmanned, remote-controlled test ship,[1] and a model, the PRA Testbed, to assess performance. Analysts use the PRA Testbed to extend the results of live testing to threats that are

The test community has struggled with how to compare a few data points from live testing to the potentially hundreds of data points from the PRA Testbed, and, once that comparison occurs, how to conclude whether the PRA Testbed reasonably represents what was observed in live testing.

[1] The unmanned Self-Defense Test Ship (SDTS) conducts tests that are too risky on a manned ship. The test community has divided cruise missile threats into six categories. Safety restrictions preclude testing against most of these threats on a manned ship. In fact, short-range self-defense systems on manned ships can be tested against only one of the six categories, and there are restrictions even for that category. To understand performance against the threat, the unmanned SDTS, which has fewer safety restrictions, is used to test against a larger set of threat categories.

not available on test ranges and to other environmental conditions that may affect ship performance.

The test community has always understood that only a limited number of live test events would be available for validation of the PRA Testbed. Many scenarios – for example, USS *America* (LHA 6) defending itself against a maneuvering supersonic cruise missile raid – will be examined in only one live test event. The PRA Testbed, however, can simulate that same event tens or even hundreds of times. Consequently, the test community has struggled with how to compare a few data points from live testing to the potentially hundreds of data points from the PRA Testbed, and, once that comparison occurs, how to conclude whether the PRA Testbed reasonably represents what was observed in live testing.

This article outlines an approach IDA developed as part of our support to the Director, Operational Test and Evaluation, who oversees and approves the Navy's test strategies and plans. The statistical approach we developed can be used to formally compare results from the PRA Testbed runs to live test shots. The literature describes various methods for validating models, including graphical comparisons between live and simulation outcomes, hypothesis tests to compare means, and Fisher's combined probability test to compare distributions. These methods, however, do not address potential correlation in the test results, described below, that may occur in PRA scenarios.

PRA TESTBED OVERVIEW

The PRA Testbed is a complex federation of models. The individual federates model elements of the ship's combat system plus the environment and the threat. For example, to model USS *America's* combat system, the PRA Testbed includes federates for each of the ship's air defense radars (SPS-48, SPS-49, and SPQ-9B), each of the missile systems (Rolling Airframe Missile (RAM) and Evolved SeaSparrow Missile (ESSM)), the command and decision system (Ship Self-Defense System (SSDS)), and other combat system elements. The PRA Testbed also includes federates that model environmental conditions and specific incoming cruise missiles. The federates run simultaneously and interact with each other over a network. Consequently, the PRA Testbed inherently includes interactions between systems. For example, if a ship's self-defense decoy or countermeasure deceives an incoming cruise missile, the threat federate will alter the missile's trajectory, which is fed into the radar federates, which provide new positional updates to the tracker federate, which feeds a new track into the command and decision federate, which can then affect the scheduling of weapon launches.

A typical PRA Testbed scenario includes multiple incoming cruise missiles and multiple decoys and self-defense missiles. A notional scenario consists of two incoming threat cruise missiles with two RAM missiles launched against each cruise missile (four RAM total). Four scenarios,

examining four different threats, will be executed in live testing.

MODEL VALIDATION

Validation is the process of determining the degree to which a model and its associated data provide an accurate representation of the real world from the perspective of the intended uses of the model. The intended purpose of the PRA Testbed is to extend live test results to other environmental conditions and threats by first showing that the model can replicate the results of the live test events with known environmental conditions and threats. Many of the individual federates within the PRA Testbed have been used in previous studies, and consequently have been validated separately. However, the overall PRA Testbed that brings together all of the federates has not been validated in an end-to-end manner.

Our approach examines intermediate metrics to increase the amount of data available for the validation. Using PRA only would provide one data point per event – yes/no, the ship defeated the raid. Each of the continuous metrics, however, provides more than one response per event. For example, a single event (live test or PRA Testbed run) will yield two initial detection ranges (when the ship detects each cruise missile), four RAM miss distances, and four RAM intercept ranges.

A statistical model is built for each of the continuous metrics. For example, using initial detection range (IDR), the statistical model can be expressed as

$$IDR = \beta_0 + \beta_1 TestType + \beta_2 TestThreat + \beta_3 (TestType * TestThreat) + \epsilon. \quad (1)$$

The statistical model is a function of two categorical factors: Test Type and Test Threat. Test Type has two levels: live test or simulation run. Test Threat has four levels for the four threat categories presented during live testing. The model also includes the interaction term. If Test Type is not significant, the live tests and the PRA Testbed runs are providing statistically indistinguishable data. Previous testing shows that the initial detection range can vary substantially from one threat to the next, so the factor Test Threat should be statistically significant. The interaction term will indicate whether differences between live test shots and PRA Testbed runs depend on a specific test threat (e.g., the PRA Testbed is providing good results for only three of the four threats).

POWER CALCULATIONS

Statistical power is a useful tool for determining data requirements for validation. More data (e.g., more PRA Testbed runs or more live test events) result in higher probabilities of detecting differences between the PRA Testbed and live tests in the midst of variability in the data. In this example, the number of live test events is limited to one event per threat category, so the statistical power is used to select the number of PRA Testbed runs.

Because the observations within a single event may be correlated, IDA's analysis examined power curves for "best-case" and "worst-case" scenarios. The best-case scenario assumes that the two detection ranges within a single event are completely independent of each other. The worst-case scenario assumes that the two detection ranges within a single event are perfectly correlated. To illustrate this correlation, consider the two initial detection ranges from a single event (live event or simulation run). Since both threats in a scenario are identical and fly similar flight profiles, if a radar detects the lead threat at X nautical miles, it likely will detect the trail threat at about the same range.

Figure 1 shows power curves for the factor Test Type for the response initial detection range. Statistical power in this case measures the probability to correctly conclude that the PRA Testbed and live testing are providing different results when they truly are different. The curves in Figure 1 assume a signal-to-noise ratio of 1.[2] There were no historical data with which to determine an appropriate signal-to-noise ratio. Ultimately, a signal-to-noise ratio of 1 was selected because a smaller signal-to-noise ratio would imply that the model results and live results essentially overlap. If the two distributions completely or nearly completely overlap, then the

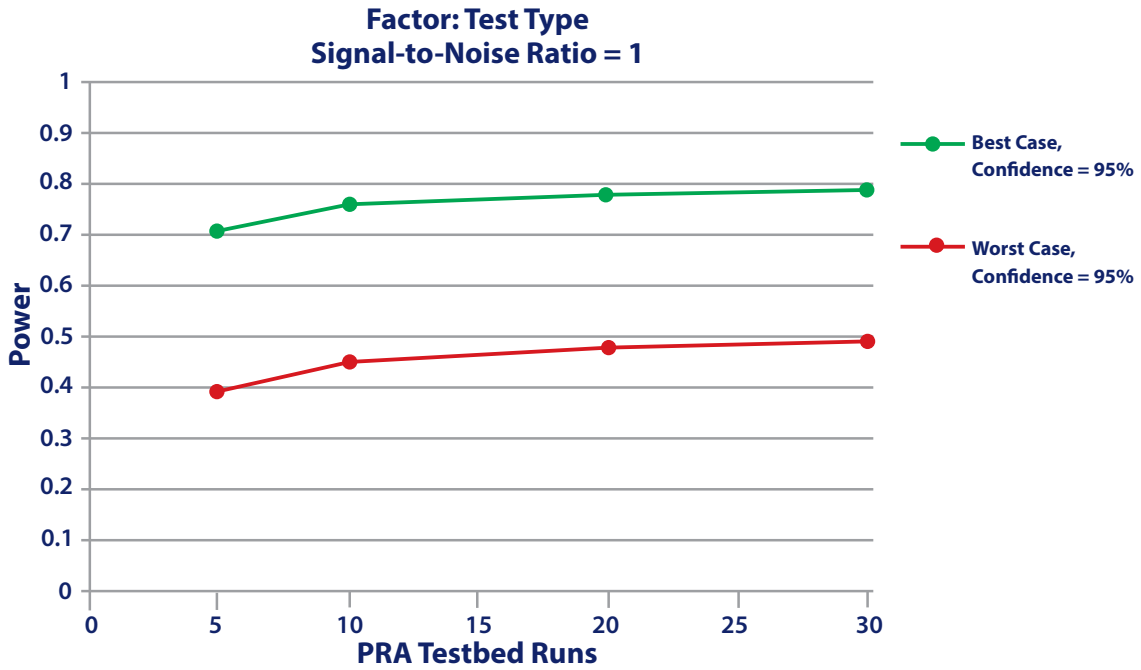

Figure 1. The power curve, assuming a confidence level of 95 percent and using a signal-to-noise ratio of 1, for the factor Test Type and the response initial detection range. The best-case scenario assumes detection ranges within a single event are completely independent; the worst-case scenario assumes that detection ranges within a single event are perfectly correlated.

[2] The signal-to-noise ratio is a ratio of the signal, which is the desired detectable change in the response variable, and the noise, which is the magnitude of the inherent system variability.

PRA Testbed provides a reasonable representation of the real world. If the signal-to-noise ratio is larger than 1, the two distributions are separated enough to conclude that the PRA Testbed does not provide a reasonable representation of the real world. Figure 2 illustrates this point, showing the separation between normal distributions for three different signal-to-noise ratios.

Figures 3 and 4 show the power curves for the factor Test Threat and the Test Type x Test Threat interaction for the response initial detection range. In Figure 3, the power curves are based on a larger signal-to-noise ratio of 2 because past operational testing indicates that combat system performance varies significantly between different threats. Consequently, large differences in the results should occur that are easy to detect. In Figure 4, the power curves using a signal-to-noise ratio of both 1 and 2 are shown to cover a wider range of possibilities.

The various power curves exhibit similar behavior, and all curves show only incremental gains in power after just 10 PRA Testbed runs. Similar behavior is seen with other continuous metrics such as missile miss distance. The small gains in power are attributable to the fact that there will be only one live test event per test threat. Overall, the figures show that this approach has reasonable power (0.61 to 0.91 at 20 runs) to detect differences between threats and marginal power (0.49 to 0.79 at 20 runs) to detect differences between the model and live test results when aggregating over all threats. Unfortunately, the only way to improve the ability to detect differences between the model and live testing, especially for a given threat (Figure 4), is by adding expensive live tests; in the case of LHA 6, no

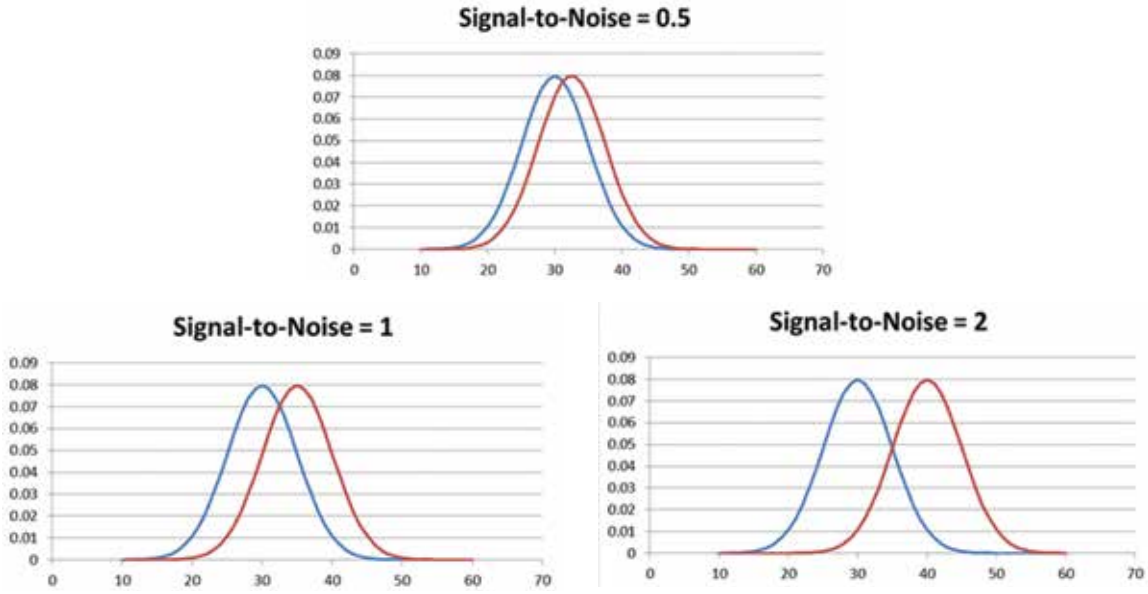

Figure 2. The separations between model and live test notional initial detection range distributions for signal-to-noise ratios of 0.5, 1, and 2.

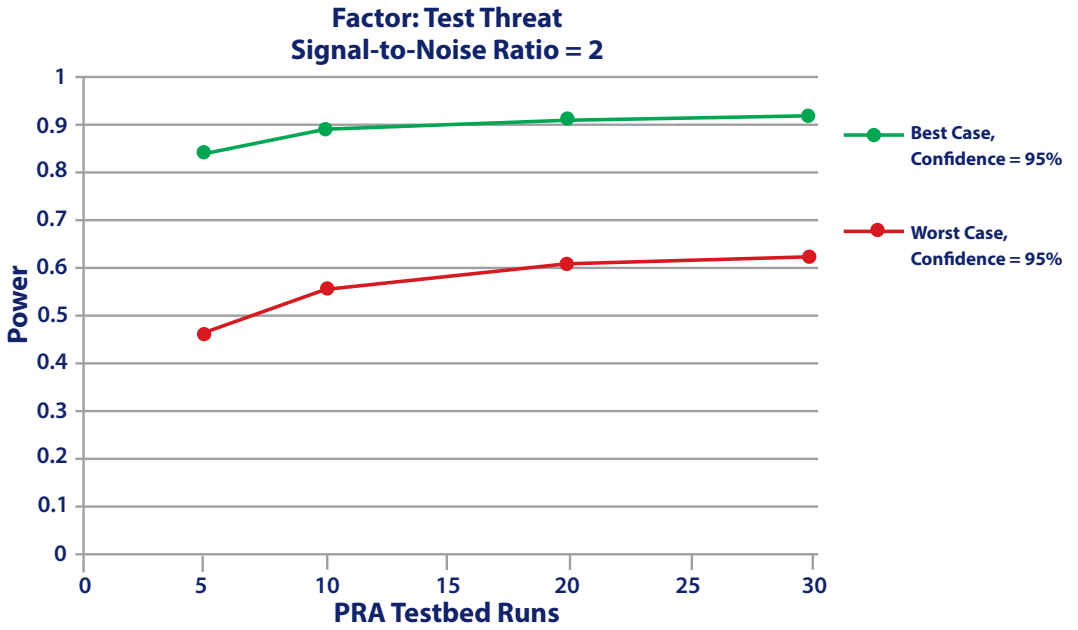

Figure 3. The power curve, assuming a confidence level of 95 percent and using a signal-to-noise ratio of 2, for the factor Test Threat and the response initial detection range. The best-case scenario assumes that detection ranges within a single event are completely independent; the worst-case scenario assumes that detection ranges within a single event are perfectly correlated.

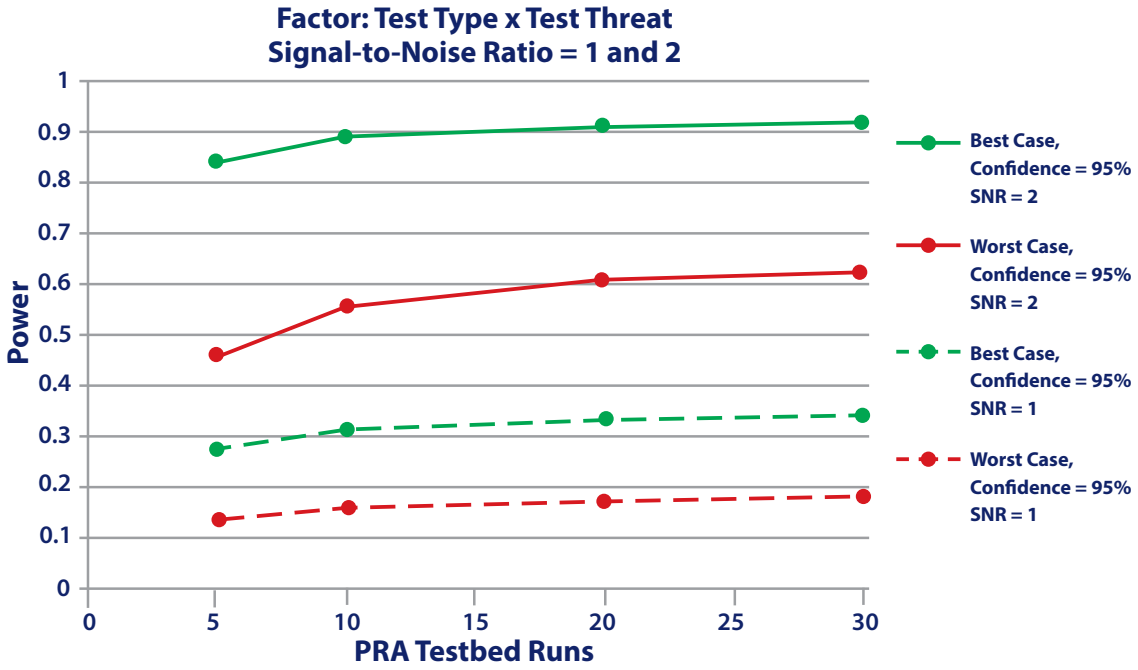

Figure 4. The power curve, assuming a confidence level of 95 percent and using the signal-to-noise ratios 1 and 2, for the interaction factor Test Type x Test Threat and the response initial detection range.

additional live tests can be added to the test program at this point.

STATISTICAL ANALYSIS

Once the data are collected, an analysis will need to be conducted to support validation. As noted earlier, one of the complications that the analysis will need to consider is possible correlation in the data. If complete independence among all of the data is assumed (or no correlation between responses from the same event), the statistical model is a standard linear model. For example, for initial detection range, the model is

$$IDR_i = \beta_0 + \beta_1 TestType_i + \beta_2 TestThreat_i + \beta_3 (TestType * TestThreat)_i + \epsilon_i \quad (2)$$

where $i=1,2,...,N$ is the total number of observations, and $\epsilon_i \sim N(0,\sigma^2)$ are the model errors. The model errors ϵ_i are assumed to follow a normal distribution with a mean of 0, a constant variance σ^2, and are independent of one another.

To account for the possibility that observations from the same event (or group) are correlated, a linear mixed model is employed. A mixed model allows for a wide variety of correlation patterns (or variance-covariance structures) to be explicitly modeled through an additional random effect, δ_i. For initial detection range, the mixed model is

$$IDR_{ij} = \beta_0 + \beta_1 TestType_i + \beta_2 TestThreat_i + \beta_3 (TestType*TestThreat)_i + \delta_i + \epsilon_{ij} \quad (3)$$

where $i=1,2,...n$ is the total number of events (live test and PRA Testbed runs), $j=1,2$ because there are two recorded IDRs per event, and $\beta_0,\ldots,\beta_3$ are the fixed effect model coefficients. The terms δ_i and ϵ_{ij} are random effects and represent two sources of variability, where

- δ_i represents the random error associated with the i^{th} test event and accounts for potential correlation between the results in a single test event, and
- ϵ_{ij} represents the random error associated with the j^{th} observation of the i^{th} test event and plays the same role as ϵ_i in Equation 2.

Because δ_i and ϵ_{ij} are random effects, they are represented by a distribution. It is common to assume that these effects are normally distributed ($\delta_i \sim N(0,\sigma_\delta^2)$ and $\epsilon_{ij} \sim N(0,\sigma^2)$) and that δ_i's and ϵ_{ij}'s are independent. These assumptions introduce the following variance-covariance matrix:

$$Var[\boldsymbol{IDR}] = \begin{bmatrix} \boldsymbol{\Sigma_1} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma_2} & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \boldsymbol{\Sigma_n} \end{bmatrix} \quad (4)$$

where the off-diagonal elements are 0 and the diagonal elements take the form

$$\boldsymbol{\Sigma_i} = Var[\boldsymbol{IDR_i}] = \begin{bmatrix} \sigma_\delta^2 + \sigma^2 & \sigma_\delta^2 \\ \sigma_\delta^2 & \sigma_\delta^2 + \sigma^2 \end{bmatrix} \quad (5)$$

This variance-covariance structure assumes that observations

in different groups are independent and that the correlation between two IDRs within a single event is constant.

$$\text{Corr}\left[IDR_{ij}, \text{IDR}_{ij'}\right] = \rho = \frac{\sigma_\delta^2}{\sigma_\delta^2+\sigma^2} \text{ where } j \neq j' \quad \textbf{(6)}$$

Notice that when there is little to no correlation between observations within a group (i.e., $\sigma_\delta^2 \approx 0$), the model and the analysis will simplify to the model and analysis presented for the complete independence case (see Equation 2).

It is important that the data analysis reflect the true nature of the data. Failure to account for the potential correlation could lead to a wrong conclusion. To demonstrate the importance of the analysis reflecting the true nature of the data, an example dataset was simulated.[3] In the simulated dataset, Test Threat will be significant, but Test Type is not. The example considers just two test threats and five PRA Testbed runs per threat, providing 24 observations.

Tables 1 and 2 provide the results of the analysis for the two modeling assumptions. Table 1 reports the fit of the linear regression model, which assumes that observations within a group are completely independent. Test Type and Test Threat are both found to be significant at the 95 percent confidence level.[4] Unfortunately, this conclusion is wrong because the data were generated assuming that Test Type was not significant. Table 2 reports the fit of the linear mixed model, which allows for

Table 1. Standard Linear Regression Model Results

Parameter Estimates				
Term	**Estimate**	**95% Confidence Interval**		**p-value**
		Lower Limit	**Upper Limit**	
Intercept (β_0)	35.87	35.21	36.53	**0.0001**
Test Type[Live] (β_1)	0.66	0.01	1.31	**0.0485**
Test Threat[A] (β_2)	-4.49	-5.14	-3.83	**0.0001**
Test Type[Live] x Test Threat[A] (β_3)	0.33	-0.32	0.98	0.3028

[3] The data set was generated using Equation 2 with the model settings $\beta_0=35, \beta_1=0, \beta_2=5, \beta_3=0$ and the variance components $\sigma_\delta^2=3$ and $\sigma^2=0.1$ (roughly 97 percent correlation).

[4] P-values are used to determine the outcome of a statistical hypothesis test, and they represent the probability of the outcome occurring by chance alone. The smaller the p-value, the higher the statistical confidence in the conclusion. The p-value for Test Type is 0.0485 and for Test Threat it is 0.001, seen in Table 1. Both p-values are less than the cutoff value of 0.05, which corresponds to significance at the 95 percent confidence level.

Table 2. Linear Mixed Regression Model Results with a Random Group Effect To Account for Correlation Between Observations in the Same Event

Parameter Estimates				
Term	**Estimate**	**95% Confidence Interval**		**p-value**
		Lower Limit	**Upper Limit**	
Intercept (β_0)	35.87	34.76	36.98	**0.0001**
Test Type[Live] (β_1)	0.66	-0.44	1.77	0.2072
Test Threat[A] (β_2)	-4.49	-5.59	-3.38	**0.0001**
Test Type[Live] x Test Threat[A] (β_3)	0.33	-0.78	1.44	0.5092
Random Effect	**Variance Component**	**95% Confidence Interval**		**Percent of Total**
		Lower Limit	**Upper Limit**	
Group	1.48	0.65	5.89	**91.49**†
Residual	0.14	0.07	0.37	8.51
Total	1.62	0.76	5.52	100

† Estimation of correlation.

the assumption that observations within a group are correlated and, in fact, reports that the estimate of correlation is roughly 92 percent. The only factor found to be significant at the 95 percent confidence level is Test Threat.[5] This conclusion is consistent with the assumption that was made when generating the data. This clear difference between the two approaches demonstrates the need for using the linear mixed model analysis to account for potential correlation within the data. The linear mixed model provides an estimate of the correlation using the data and does not require any guesswork by the analyst or subject matter expert.

CONCLUSION

Overall, the approach outlined above provides a straightforward method for validating a simulation for which a limited number of live test events are available. By using a statistical model, results from the PRA Testbed runs can be formally compared to the live test events. The model allows analysts to test for a Test Type effect, a Test Threat effect, and an interaction effect. If the Test Type effect is not statistically significant, then the PRA Testbed runs are providing meaningful data.

The power curves help analysts understand how many PRA Testbed

[5] The p-value for Test Type is 0.2072 and for Test Threat is 0.001 (also see Table 2). Only the p-value for Test Threat is less than the cutoff value of 0.05, which corresponds to significance at the 95 percent confidence level.

runs are needed for validation. Because so few live test events are planned, only small gains in power after 10 PRA Testbed runs per scenario are observed. The AW SSD Enterprise effort is planning to execute 30 runs per scenario to exercise the simulation and to discover any bugs. Consequently, sufficient PRA Testbed data for the comparison should be available.

The proposed validation approach has several limitations. Normally, one constructs a test to determine whether two items are different. The approach is to assume that they are the same (the null hypothesis) and prove that they are statistically different by rejecting the null hypothesis. However, this approach does the opposite, which provides a weaker claim. Furthermore, due to the fact that there will be just one live shot per threat condition, the analysis will not be able to adequately differentiate between problems with bias versus variance in the model. The limited live testing in this example limits the usefulness of the experimental design approach.

More research is needed to determine appropriate methods for selecting what live points within the operational space should be chosen for an optimal ability to validate the model. Design of experiments is a potential path toward better model validation. A combined experimental design and analysis approach will allow for sizing the number of live tests to detect meaningful differences, strategic replication to address variance/bias, and a parametric analysis to incorporate sensitivity and prediction analyses.

Despite the limitations of few live data, this approach illustrates how more rigorous statistical methods provide the testing and acquisition communities more robust and objective conclusions from both M&S and live test data. IDA, in support of DOT&E, will continue to lead the way in advocating for and researching new statistical methods for test and evaluation in the Department of Defense.

Dr. Thomas is an Assistant Director in IDA's Operational Evaluation Division. He holds a Doctor of Philosophy in physics from the State University of New York (SUNY), Stony Brook.

Dr. Dickinson is a Research Staff Member in IDA's Operational Evaluation Division. She holds a Doctor of Philosophy in statistics from the Virginia Polytechnic Institute and State University.