IDA | RESEARCH NOTES

Fall 2015

Test and Evaluation:

Statistical Methods for Better System Assessments

IDA is the Institute for Defense Analyses, a non-profit corporation operating in the public interest.

IDA's three federally funded research and development centers provide objective analyses of national security issues and related national challenges, particularly those requiring extraordinary scientific and technical expertise.

This edition of *IDA Research Notes* was edited by Dr. V. Bram Lillard and Dr. Laura J. Freeman, Assistant Directors of IDA's Operational Evaluation Division. All of the articles were written by members of that division's research staff. Address questions about the specific research topics or related issues to:

Operational Evaluation Division (OED) Mr. Robert R. Soule, Director (703.845.2482, rsoule@ida.org)

Institute for Defense Analyses
4850 Mark Center Drive Alexandria,
Virginia 22311 www.ida.org

The current fiscal climate demands now, more than ever, that test and evaluation (T&E) provide relevant and credible characterization of system capabilities and shortfalls across all relevant operational conditions as efficiently as possible. In determining the answer to the question, "How much testing is enough?" it is imperative that the T&E community use methodologies focused on that issue.

This issue of *IDA Research Notes* showcases IDA's response to that challenge, centered on the Design of Experiments (DOE) methods championed by the Director, Operational Test and Evaluation in the Department of Defense (DoD), for whom much of our work was undertaken. DOE methods provide structured, efficient, and scientific approaches to test planning. They provide ***objective methods for assessing test adequacy*** not only by providing a quantitative basis for assessing how much testing is enough, but also by identifying where in the operational space the test points should be placed. DOE methods also provide an ***analytical trade-space between test resources and risk***, ensuring that tests are adequate to answer important questions. IDA has developed case studies, training materials, and publications that have been instrumental in shaping the T&E community's approach to applying DOE to operational testing.

DOE methods provide not only a scientific methodology for test planning, but also a roadmap for conducting post-test statistical analyses. Using DOE and corresponding statistical analysis methods instead of conventional approaches to data analysis, we have been able to learn more from tests without necessarily increasing their size. Statistical data-driven empirical models provide inferential weight to decision-makers about how systems will actually perform when deployed, and enable a fuller characterization of system performance across the variety of conditions in which the systems will operate. These methods ensure that ***robust and objective conclusions*** are drawn from test data. Statistical analysis techniques also help in "doing more with what you have" by providing methodologies to ***maximize information*** gained from test data through empirical statistical models, and by ***incorporating all relevant evidence*** (including previous test data and engineering expertise) into analyses.

The case studies described herein show that advanced design and analysis methodologies ensure that testing is both statistically rigorous and operationally meaningful, and capable of extracting the most information from operational test data. The articles include examples from multiple Services, and reveal how these methodologies are applicable across a wide variety of testing scenarios. This issue also highlights other emerging areas of research including Bayesian analysis methods to leverage all available information and statistical methods for validating complex models and simulations.

Assessing Submarine Sonar Performance Using Statistically Designed Tests shows how IDA was able to use advanced design techniques to develop a test design that incorporated operational, planning, and statistical objectives. It is an example of how advanced statistical analysis techniques can maximize information gained from very limited data in only a few test conditions.

Applying Advanced Statistical Analyses to Helicopter Missile Targeting Systems, illustrates how the combination of experimental design and logistic regression was able to extract essential information from test results that was not readily apparent by direct observation alone. The case study illustrates the importance of covering the full operational space and avoiding simple averages when analyzing test data.

Tackling Complex Problems: Analysis of the AN/TPQ-53 Counterfire Radar illustrates how advanced statistical analysis techniques can be used for tests where, in order to preserve operational realism, testers are unable to control all of the operational factors likely to affect performance. In this case study, regression techniques were used to determine causes of performance degradations across multiple operating modes, even with highly unbalanced data.

Improving Reliability Estimates with Bayesian Hierarchical Models, illustrates how Bayesian methods can be used to incorporate information from multiple phases of testing without biasing operational estimates of reliability. The results are robust estimates of system reliability, even in cases where only limited operational test reliability data are available.

Managing Risks: Statistically Principled Approaches to Combat Helmet Testing, shows the evolution of testing protocols for helmets, and how essential it is to apply statistical rigor to testing; without such rigor, the risks of accepting poorly performing helmets for use in combat could be unknown or, worse, could be unacceptably high, thereby placing soldiers' lives at risk.

Probability of Raid Annihilation (PRA) Testbed Validation, highlights an emerging area of emphasis for OT&E. Complex models and simulations, such as the Navy's PRA Testbed, provide valuable information to operational evaluations, especially in cases where safety concerns or range limitations severely constrain live testing. However, the outcomes from these complex system-of-systems models must be validated and evaluated to determine their operational value. The PRA Testbed article provides an example of how defensible statistical approaches to validation can provide a basis for understanding the value of complex simulations in evaluations of operational tests.

Assessing Submarine Sonar Performance Using Statistically Designed Tests

George M. Khoury, Justace R. Clutter, and V. Bram Lillard

THE PROBLEM

Historical at-sea methods for determining Anti-Submarine Warfare performance of the Navy's submarine sonar system are unable to characterize performance across a range of operational conditions and yield statistically significant results.

The Acoustic Rapid Commercial-off-the-Shelf Insertion (A-RCI) is the Navy's newest submarine sonar processing system. It provides hardware and software to process data from the submarine's sonar arrays and display those data to the sonar operators. A-RCI uses a spiral development model to procure new, commercial off-the-shelf computing hardware every two years. Buying new computing hardware over time capitalizes on the decreasing cost of processing power and ensures that an acceptable balance between obsolescent and modern hardware is maintained. To take advantage of the ever-improving processing power from hardware upgrades, a new version of A-RCI software, denoted an Advanced Processing

To address the shortcomings of A-RCI at-sea testing, IDA proposed augmenting the at-sea operational test events with so-called Operator-In-the-Loop (OIL) laboratory tests.

Figure 1. Four A-RCI Sonar Consoles aboard a Submarine

Build (APB), is developed every other year. Each APB incorporates feedback from Fleet users, fixes bugs discovered in previous versions, and adds new algorithms developed by industry and academia.

The primary role for A-RCI is to manage the large amount of information coming from the sonar arrays and display it to the operator so that he can make sense of it. To understand the scale of the operator's problem, consider that a *Virginia*-class submarine uses six sonar arrays for submarine searches, each providing information on all bearings, multiple elevation angles, and a range of frequencies. The sonar operator must monitor this multi-dimensional search space constantly, and it is impossible to display all of the information simultaneously. A-RCI provides displays and automation to help the operators manage this search space and help them detect contacts as quickly as possible.

The Navy's primary metric with which to evaluate A-RCI performance in the Anti-Submarine Warfare (ASW) mission is denoted ΔT. It is defined as the median time it takes for an operator to detect a submarine contact once that submarine's signal becomes available for display on sonar system screens. Although ΔT is not a measure of the submarine's overall ASW capability, it does quantify A-RCI's role in the detection process. The ongoing goal of A-RCI processing improvements is to minimize the time needed to find target signatures.

At-sea tests of A-RCI consist of two submarines searching for each other in a specified area. Although this technique provides an operationally realistic environment, it suffers from several drawbacks. Most notably, at-sea testing has never been able to show a statistically significant improvement in A-RCI performance over the course of a decade, during which time many software and hardware upgrades were fielded to the Fleet. A comparison has been impossible because two software versions are never compared in the same at-sea event, and the results of a test can depend on target and environmental characteristics that are impossible to control. Additionally, at-sea testing uses a single target and a single operational environment, which limits the assessment of performance of the new APB to only a small portion of the operational envelope. Finally, the cost and variability of at-sea testing have resulted in poor quantification of APB performance in the conditions tested.

To address the shortcomings of A-RCI at-sea testing, IDA proposed augmenting the at-sea operational test events with so-called Operator-In-the-Loop (OIL) laboratory tests. In an OIL test, a Fleet operator sits at a laboratory mockup of the A-RCI sonar system. The laboratory then plays back a recorded at-sea encounter between two submarines, and the operator declares when he has detected the threat submarine.[1] The laboratory allows the same encounter to be replayed on different

[1] U.S. submarines are capable of recording raw sonar data, that is, the voltage recorded by the individual hydrophones that make up the sonar arrays. Because these raw data are recorded before they are processed by A-RCI, it is possible to process the recorded data on any version of A-RCI.

versions of A-RCI, which perfectly controls for environmental and target variability; the only difference between the two presentations is the software used to process the data. The primary limitation of the OIL testing is that it only allows for a single array to be processed at one time. Therefore, the sonar array to be processed needs to be a controlled test factor, whereas in real operations all arrays operate simultaneously.

For several years, the Navy has used a similar laboratory test method to compare new versions of A-RCI to old versions, but typically used only a few submarine encounters for each comparison, and published the results long after the software was fielded. As part of our support to DOT&E, IDA proposed expanding the scope of such tests to include a wider variety of test conditions, as shown in Table 1, and to employ Design of Experiments methodologies to generate a more robust test that would characterize performance across a range of operational conditions. The primary goal of the test was to compare the latest version of the sonar system, denoted APB-11, with the previous version, APB-09. To characterize the systems, the test used operators of varying proficiency and controlled for characteristics of the target and the array being used.

FACTOR LEVELS HYPOTHESIZED EFFECT

IDA developed a 120-run, D-optimal, split-plot test design, with the distribution of runs as shown in Figure 2. A "run" consists of a single operator viewing a single recorded encounter, and a "Null" run is one in which no target is present. The

Table 1. Factors and Levels Used in the OIL Testing Analysis

Factor	Levels	Hypothesized Effect
Target Type	SSN, SSK	SSNs and SSKs exhibit different acoustic signatures. SSNs typically have more discrete tonal information because of the machinery associated with the nuclear reactor.
Array Type	A, B	Array type A typically detects targets at longer ranges, which would be expected to generate larger ΔTs.
Target Noise	Loud, Quiet	Loud targets are detected at longer ranges, which could lead to longer ΔTs. Conversely, loud targets typically have more discrete tonal information and are easier to identify, which could result in shorter ΔTs.
APB Version	APB-09, APB-11	The primary goal of the test was to compare the latest version of the sonar system, APB-11, with the previous version, APB-09.
Operator Proficiency	1 to 20	More proficient operators will detect a submarine more quickly. The numeric scale was developed by the Naval Undersea Warfare Center and is based on an operator's experience with the A-RCI system.

	Target Type		Array B	Array A
APB-11	SSK	Quiet	6	6
		Loud	6	12
	SSN	Quiet	6	12
		Loud	6	12
	Null		6	

	Target Type		Array B	Array A
APB-09	SSK	Quiet	4	4
		Loud	4	8
	SSN	Quiet	4	8
		Loud	4	8
	Null		4	

Total Number of Runs

72

48

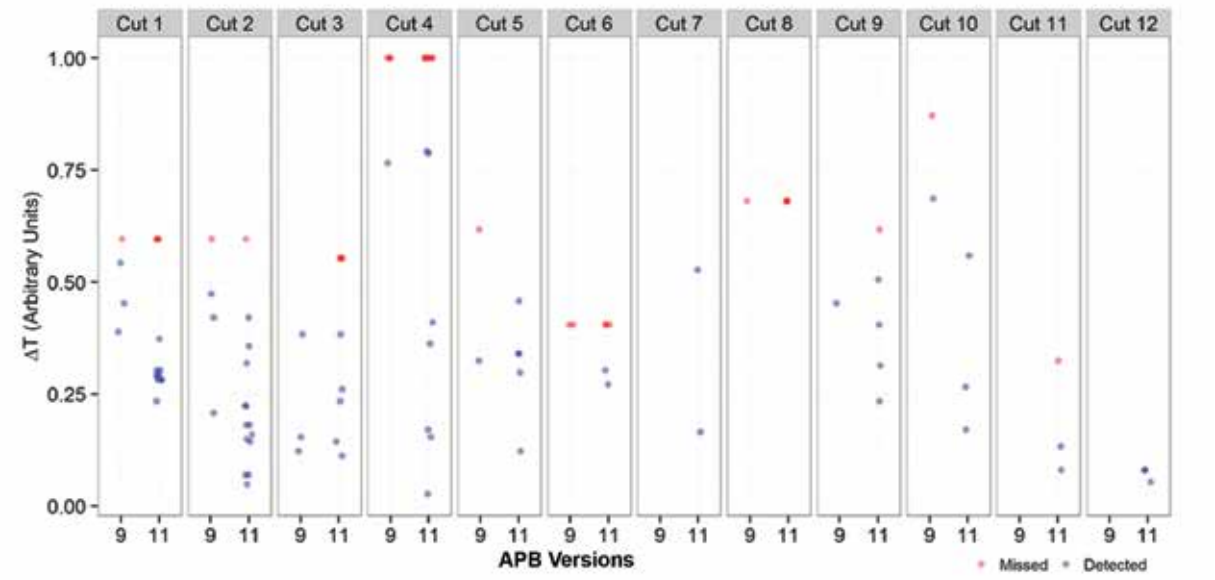
120

Figure 2. OIL Test Design Matrix

split-plot structure was used to limit the number of changes between the APB versions, as each change of APB required approximately 12 hours. A considerable amount of replication was built into the design to account for the fact that operator proficiency was not explicitly controlled. Instead, operators were chosen at random, and their proficiency was recorded during the events, which ensured a balanced distribution of proficiencies. Each operator reviewed up to six tapes, including a blank tape to check for false alarm rate. Finally, the Navy desired to focus the testing on APB-11, which resulted in the asymmetric test design shown; while this was not optimal for determining whether a significant APB difference existed, it provided a more precise understanding of performance for APB-11 (tighter confidence intervals).

TEST RESULTS

Figure 3 shows the raw results of the test. Each panel shows the results for a recorded encounter, with APB-09 results on the left and APB-11 results on the right. The blue dots are detection times; the red dots indicate runs in which the operator never detected the target before the



Each panel (Cut 1, Cut 2, ...) shows the results for a single recording. Blue points indicate detection times (arbitrary units). Red points indicate runs in which the operator did not detect the target; the time in these cases is the length of the recording.

Figure 3. Raw Results from A-RCI OIL Testing

recording finished. The location of the red dot indicates how long the target was on tape and not detected.

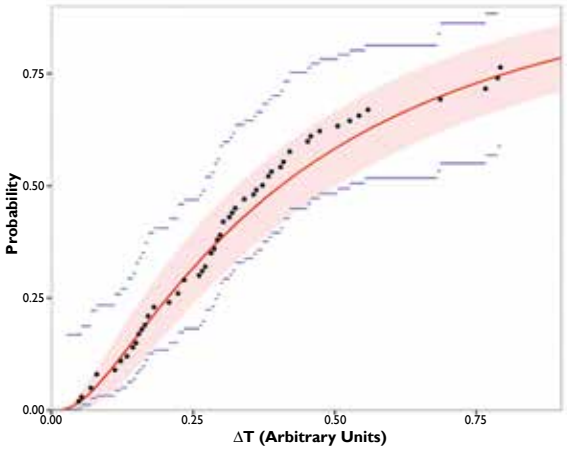
The advantage of examining the results by recording is that recordings control all aspects of the encounter; the environment and target are exactly the same for each playback, so any difference in performance is due to either operator proficiency or the capability of the processing system. Since the test was well balanced in terms of operator proficiency, any observed differences are most likely due to the processing system. In general, APB-11 exhibited improved performance in almost all of the recorded encounters; in each panel, the dots are generally lower for APB-11 than they are for APB-09, reflecting shorter times to detect threat submarines. Therefore, even without statistical analysis, APB-11 appears to be an improvement over APB-09. Such a limited analysis does not, however, make use of all the available information; APB-11 appears to be better, but the improvement varies with recording and it is unclear why. The test was designed to determine which of the controlled factors affect A-RCI performance, and for that a statistical analysis is necessary.

We performed a regression analysis to better understand how the controlled factors affected A-RCI performance. Our analysis accounts for missed detections by treating them as censored data points; in these cases, we assumed that the operator would have detected the contact if given enough time, so the full recorded length of time the contact was on the display is used as a lower bound estimate for the ΔT. We assumed that the data followed a lognormal distribution, in which the probability of observing a detection time x is the following:

$$p(x \mid \mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln(x)-\mu)^2}{2\sigma^2}}$$

Here, μ is related to the median of the distribution, and σ is a measure of its spread. Making this assumption allowed us to incorporate the missed detections using standard censored data analysis techniques.

Although there is no *a priori* reason why the data should follow a lognormal distribution, our initial assumption was well supported by the data. Figure 4 shows the empirical cumulative distribution function of the data, along with a lognormal fit in red, the confidence region on the lognormal fit in pink,



Red line shows a lognormal fit. Pink region shows the 80% confidence region on the lognormal fit. Blue lines indicate the 80% confidence region on a non-parametric fit to the distribution. The data are well described by a lognormal distribution.

Figure 4. Empirical Cumulative Distribution of the OIL Data

and the confidence region of a non-parametric fit in blue lines. The data appear to be well described by a lognormal distribution.

Next, we assigned each recording to the factors listed in Table 1, and then fit the data according to the following model:

$$x \sim lognormal(\mu,\sigma)$$

$$\sigma = \text{constant}$$

$$\mu = \beta_0 + \beta_1 OP + \beta_2 APB + \beta_3 Target + \beta_4 Noise + \beta_5 Array + \beta_6 Target * Noise + \beta_7 Target * Array + \beta_8 Noise * Array + \beta_9 Target * Noise * Array$$

That is, we assumed that the median detection time depends on the factors listed in Table 1, along with second and third order interactions, and that the σ parameter was constant. In fact, we examined many possible models, including those with variable σ, but this model resulted in the lowest Akaike's Information Criterion (AIC), a metric of model desirability. Table 2 shows the results of the final fit and describes the qualitative behavior of the coefficients. All of the first-order effects were highly significant. Notably, APB-11 performed significantly better than APB-09, holding all other effects equal – and the magnitude of the effect was

Table 2. Results of the Model Fit to the Data

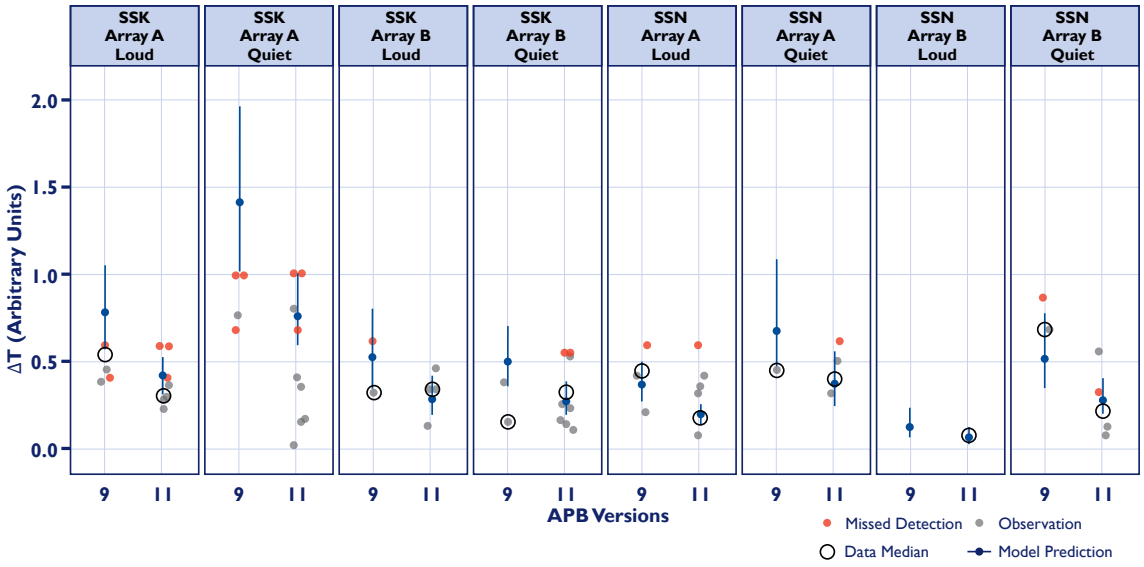
Term	Value†	Description of the Effect
β_1 (Operator experience level)	-0.074 ± 0.041	Increased operator proficiency results in shorter detection times. An increase in proficiency of one unit reduces median detection time by 7 percent.
β_2 (APB)	0.307 ± 0.129	Detection time is shorter for APB-11, **by 46 percent.**#
β_3 (Target)	0.359 ± 0.126	Detection time is shorter for SSN targets.
β_4 (Noise)	-0.324 ± 0.125	Detection time is shorter for loud targets.
β_5 (Array)	0.347 ± 0.125	Detection time is shorter for the Type B array.
β_6 (Target*Noise)	0.186 ± 0.126	Additional model terms added to improve predictions. The third-order interaction is marginally significant, so all second order interactions nested within the third order interaction were retained to preserve model hierarchy.
β_7(Target*Array)	0.011 ± 0.125	
β_8(Noise*Array)	0.021 ± 0.126	
β_9 (Target*Noise*Array)	-0.180 ± 0.125	

† Confidence interval is an 80% Wald interval
# APB-11 Provides a Statistically Significant Improvement.

substantial, as APB-11 detection times were 46 percent shorter than APB-11 times. Also, APB had no interaction with the other factors, which means APB-11 produced the same improvement regardless of the other factors. It did not matter whether the target was loud or quiet, SSN or SSK; switching from APB-09 to APB-11 reduced the median detection time by approximately 46 percent. This was the first time operational testing of A-RCI had shown a statistically significant improvement in an APB.

Figure 5 shows the results of the model fit (blue dots, with 80% confidence intervals shown as vertical lines), along with the actual median detection times in each group (black) and the raw detection times (light blue and red, as before). The model predictions generally agree with the median in each bin, indicating that our relatively simple model provides a good fit to the data. There is, however, notable disagreement between the data median and the model prediction for one bin: quiet SSK targets with array type B in APB-09. The difference is due to sparse data, rather than a poorly fitting model. The data median in this case is based on only three data points and is therefore highly variable, making it a poor estimator of the true performance in that bin. We believe the model estimate predicts the performance that would be observed if additional runs were conducted with APB-09.

Our analysis provides several benefits over the less sophisticated analysis based solely on individual recordings. First, differences in performance are now attributable to operationally relevant factors, such as target type or array type. In contrast to the naïve analysis by recording, our statistical analysis shows that APB-11



The model fits the data well and indicates that APB-11 outperforms APB-09 in all conditions. Data medians were omitted when the data in the bin were inadequate to support the estimation of a median value (e.g., too few data points). This illustrates another advantage of using the empirical statistical model, since it can estimate median performance in every bin whereas traditional data analysis methods might not be able to provide a robust estimate.

Figure 5. Model Predictions (Blue), along with the Median Detection Time Observed in Each Bin

outperforms APB-09 by 46 percent on average across all conditions. Second, our analysis allows us to extrapolate to areas where the data are limited. A few of the experimental configurations presented in Figure 5 do not have an observed data median for comparison with the model prediction, either because there were few data points or because there was an excess of censored values. An analysis using a simpler technique would not have been able to estimate performance in regions where the data were inadequate to produce an estimate of performance.

CONCLUSIONS

Operator-in-the-Loop testing has proven to be an effective way to compare the performance of different versions of the sonar processing system and to discover how performance varies across a variety of operationally important factors. By playing back recorded data from real-world submarine encounters, OIL testing controls for target and environmental variability in a way that traditional at-sea testing cannot. It provides more data at a lower cost, which has enabled IDA to show a statistically significant improvement in A-RCI for the first time, and it has allowed us to quantify the operational factors that affect the improvement. Laboratory testing will not soon replace all at-sea testing, but it is a valuable complement.

Dr. Khoury is a Research Staff Member in IDA's Operational Evaluation Division. He holds a Doctor of Philosophy in physics from the University of California, Santa Barbara.

Dr. Clutter is a Research Staff Member in IDA's Operational Evaluation Division. He holds a Doctor of Philosophy in physics from the University of Kansas.

Dr. Lillard is an Assistant Director in IDA's Operational Evaluation Division. He holds a Doctor of Philosophy in physics from the University of Maryland.

Applying Advanced Statistical Analyses to Helicopter Missile Targeting Systems

Howard C. Keese and Steven A. Rabinowitz

THE PROBLEM

Advanced analytical methods often extract the essential information from test results that may not have been readily apparent by direct observation alone. When testing the effectiveness of naval helicopters to defend a carrier group from surface attacks, the use of sophisticated statistical methods can provide operators with a greater understanding of both system capabilities and limitations during real-world employment.

DEFENDING THE CARRIER STRIKE GROUP

The United States Navy's Carrier Strike Groups are critical components of our national defense infrastructure. They are also prominent targets for potential U.S. enemies, and are subject to multidimensional threats from air, surface, and subsurface attacks. Consequently, the Navy dedicates significant resources to protecting the aircraft carrier and other high value units at sea.

The end of the Cold War led to a shift in the strategic paradigm for the Navy. It could no longer focus on a single, monolithic threat. In the 21st century, the Navy must be able to adapt to a wide array of disparate regional threats and operating environments. Instead of being primarily concerned with blue-water, open ocean combat, the Navy now also must be prepared to operate in the littorals – in close proximity to the shoreline, which, in turn, exposes U.S. ships to a multitude of new threats. Also important is the radically different mindset of some adversaries. Instead of possessing at least a passing concern with "living to fight another day," some enemies now attack with a suicidal determination. A driven enemy with no regard for personal survival poses a different challenge. One such asymmetric threat is the small boat suicide attack. The grave nature of this threat was dramatically illustrated by the October 2000 attack on USS *Cole*. Although this suicide bombing occurred in port, the Navy is equally concerned about possible small boat attacks at sea involving small arms, missiles, and torpedoes. To counter this threat, the fleet employs a layered defense, with fixed-wing aircraft providing longer range standoff engagements and the ships defending themselves close in. Between these ranges, embarked helicopters provide another defensive layer.

IDA analysts identified and helped construct the test design for the Multispectral Targeting System [MTS] for the operational testing conducted in 2014...[and] identified the dominant factors expected to affect system performance.

NAVAL HELICOPTERS

The Navy deploys two medium-lift, tactical, rotary-wing aircraft aboard carriers and surface combatants: the MH-60R and the MH-60S multi-mission helicopters. Both of these Sikorsky aircraft are derived from the Army's UH-60 Blackhawk, but the construction of each is uniquely tailored to operate in the maritime environment in support of Navy missions. With both radar and sonar sensors, the MH-60R is optimized for antisubmarine warfare. The MH-60S fills combat search and rescue and airborne logistics roles. Both aircraft contribute to surface warfare, providing strike capabilities against small surface targets. Recently, the Navy has been testing various weapons systems aboard these aircraft, including 50mm machine guns, 2.75-inch rockets, and guided missiles.

MTS AND HELLFIRE WEAPON SYSTEM

Both helicopters were designed to employ laser-guided AGM-114 Hellfire missiles (Figure 1). While the Hellfire missile originally was designed for anti-armor land warfare by the Army, the Hellfire's size, range, and lethality have proven useful for a variety of warfare areas. The original plan was to use the MH-60R/S's Hellfire missiles against enemy surface combatant ships, but the missiles can also be employed against small boat targets. Each MH-60R/S can carry eight missiles.

Recently the Navy upgraded the Forward-Looking Infrared (FLIR) system on both aircraft. The new system is known as the AN/AAS-44C(V) Multispectral Targeting System (MTS) imaging system. The MTS uses advanced electro-optic technologies to support navigation, search, and surveillance activities. It can also detect, track, and range surface threats, and its laser designator can illuminate targets to guide Hellfire missiles. The system also has a Day TV capability. When combined with the FLIR camera, the MTS provides imaging from the visible through far-infrared spectrum under all lighting conditions. The MTS features an Automatic Video Tracking (AVT) software algorithm designed to maintain a consistent track on the target and keep the laser designator

Figure 1. MH-60R (left) and MH-60S (right) Helicopters Employing AGM-114 Hellfire Missiles

beam accurately positioned on the target, allowing the Hellfire missile's laser seeker to guide the missile all the way to target impact. Determining the MTS's ability to enable accurate Hellfire employments, therefore, was the focus of the operational test. The critical issue was the MTS's ability to establish a solid engagement-quality lock on the intended target, maintaining laser illumination on the aim point throughout the weapon's time of flight. Therefore, the testing focused on measuring MTS targeting effectiveness across a variety of operational conditions.

TEST DESCRIPTION

In support of the Director, Operational Test and Evaluation, IDA researchers have been employing increasingly sophisticated statistical methods to plan and analyze field tests of critical defense systems. Using the Design of Experiments (DOE) methodology to develop the test plan, analysts identify specific factors that are expected to affect system performance in an operational setting. In the case of a weapon system, these might include the individual and relative motions of the launch platform and target, environmental factors, and weapon-specific data such as firing mode. The ultimate goal is to rigorously characterize weapon performance across the entire operational envelope as a function of those factors, singly and in combinations, rather than simply rolling up the data into an aggregate result. Consequently, instead of reporting out a single overall hit percentage, analysts can demonstrate how particular circumstances and their combined effects may increase or decrease a system's overall effectiveness. Design of Experiments techniques can generate an optimal run plan that provides statistically significant coverage of the various factors without requiring explicit testing of every possible combination. This allows testers to make the most efficient use of limited resources. During the planning phase, IDA analysts work with Service test personnel to determine how to control test scenarios in a manner that provides sufficient data to support factor analysis while preserving operational realism.

Working with the Navy and DOT&E, IDA analysts identified and helped construct the test design for the MTS for the operational testing conducted in 2014. Using both engineering judgment and tactical experience, IDA researchers identified the dominant factors expected to affect system performance. These included target size (large/small), the target's speed (fast/slow), and whether it is maneuvering (yes/no), all of which can be controlled by the run plan. Additionally, target aspect, lighting conditions (day/night), Hellfire targeting mode (target lock before launch/target lock after launch), and airframe (MH-60R/S) were considered in the analysis. Calculating every possible combination of the two levels for each of the seven factors (known as a full-factorial DOE) generated 128 total configurations (2^7), which established the basic data collection requirement for the test.

Although the Navy has not identified specific performance thresholds for the MTS, the helicopter requirements documents specify that the aircraft must be able to fire air-to-ground missiles capable of disabling or destroying a small boat at a standoff range that is beyond the threat of small arms fire and man-portable anti-aircraft missiles that might be carried on the threat boat. Testing of the MH-60R/S with MTS was conducted in the Chesapeake Bay area from August 2012 through January 2013, in two parts. The first phase was a simulated fire period where the helicopters acquired and tracked targets but did not launch any actual missiles. Instead, they carried a Captive Air Training Missile (CATM), which is a specially built Hellfire missile body without a rocket motor. The CATM replicates all Hellfire missile activity up to the point of missile launch, allowing testers to examine the crew's use of the MTS to track and lase the target. During the first phase of testing the targets were emulated by two different types of small fast manned boats: the 26-foot High-Speed Maneuvering Surface Target (HSMST) and the 50-foot Fast Attack Craft Trainer (FACT) (shown in Figure 2).

During the second phase of testing, the helicopters launched five actual Hellfire missiles to demonstrate end-to-end functionality against towed surface targets.

ANALYSIS

In general, data collected from testing are rich with information despite the fact that the results are a simple series of successes and failures (1's and 0's). Several methods exist to analyze such data. One traditional analysis method is to simply tally the successes, divide by the number of trials, and determine the overall success rate. Although this lends itself to simplicity in reporting ("overall MTS is successful X percent of the time"), it may be misleading because it is dependent on the specific allocation of test conditions conducted and might not be representative of a global average of system performance across a variety of future combat conditions. Furthermore, it hides important information about system performance.

Figure 2. The High-Speed Maneuvering Surface Target (HSMST) (top) and the Fast Attack Craft Trainer (FACT) (bottom)

Another methodology groups the 128 data points according to general categories of test conditions. Thus, based on technical experience, analysts might split the data according to threat, calculating the

probability of a successful engagement using 64 observations for each of the two target sizes. Next, they might decide to divide the data sample further by target speed, producing four separate results with sample sizes of 32 each, or simply report the average success rate for all fast speed targets and all slow speed targets. This approach may continue for additional factors. Figure 3 plots the data as grouped by three different factors (to avoid revealing classified details the results are normalized and the names of the specific factors are aliased).

Clearly, with each division of the data, we learn more detailed information about system performance, but it comes at the price of statistical confidence. For a binomial (yes/no) response, the confidence intervals, sometimes known as error bars, for small sample sizes can be very large and therefore not particularly informative. In addition, examining the individual point estimate calculations for selected subsets of the data could mask important performance limitations that may exist.

A more rigorous approach uses logistic regression analysis. Logistic regression is similar to the more common linear regression, which predicts performance for a given input values using a linear relationship. Logistic regression employs the same techniques for finding the best fit to the data but is constructed using a more complex function to handle the binary nature of the data and predict the probabilistic outcomes. The logistic regression analysis can be extended to any number of factors (regressors)

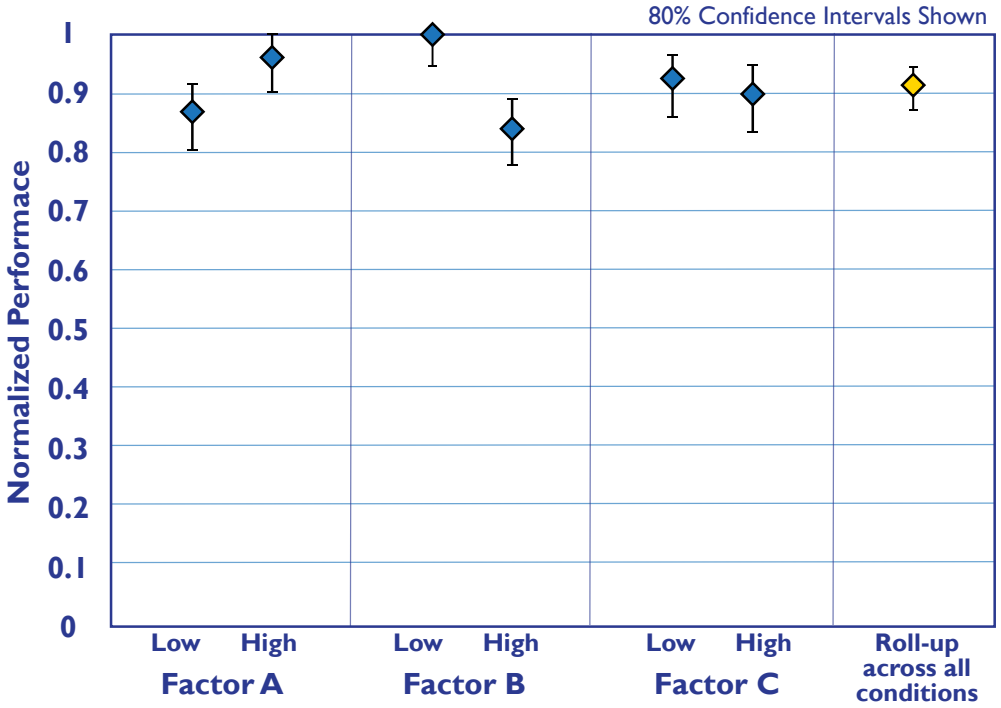


Figure 3. Binomial Point Estimates (Data Grouped by a Single Common Condition)

in order to produce a response surface. The most general form of the MTS logistic regression model is:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 Speed + \beta_2 Size \cdots + \beta_k Manuever + \beta_{12} Speed * Size + \cdots$$

where p is the probability of success, and the β_i's are linear coefficients linking the factors varied in the test to the probability of success. The analysis estimates each β_i. If β_i is not zero (more technically, is statistically significantly different from zero), then the factor or condition is important in explaining the probability of interest. This form of the equation is used because it shows that the factors and conditions impact the "log-odds" of the probability (the left-hand side of the equation) in an additive, linear way. We can rewrite this expression as:

$$p = \frac{\exp(\beta_0 + \beta_1 Speed + \beta_2 Size \cdots + \beta_k Manuever + \beta_{12} Speed*Size + \cdots)}{1 + \exp(\beta_0 + \beta_1 Speed + \beta_2 Size \cdots + \beta_k Manuever + \beta_{12} Speed*Size + \cdots)}$$

which gives a direct expression for the probability of interest.

In the case of the MTS analysis, IDA researchers utilized 128 data points to construct a regression model that includes all possible interactions between the factors. This technique readily identifies the combinations of factors that result in significant degradation of system performance that would not be easy to isolate through the manual data parsing method discussed above. IDA analysts were able to iteratively build and evaluate different regression models based on various combinations of factors and model terms. The ideal model is the simplest one that includes the most significant factors and their interactions while accurately predicting the system performance based on the data collected. In other words, the statistical analysis is formed and molded by the data alone.

Figure 4 shows the successful engagement predictions of the IDA regression model for MTS based on the data collected in the operational test. In order to mask the classified results, the specific factor names are aliased, and the probability is plotted on a normalized scale. The vertical bars on each performance estimate indicate the confidence intervals (error bars) for each combination of conditions. Note that while these bars are slightly larger than the ones shown in Figure 3, this is due to the finer binning. In fact, by using a regression model, the confidence intervals shown in this plot are smaller than they would be for calculating simple point estimates in each bin. This is because the regression model exploits information from across the data set, resulting in a more precise estimate of the statistical confidence. It is immediately obvious from the plot that the system's performance for most of the conditions is quite consistent. In other words, the MTS for some conditions is not likely to show much performance variation. However, for one particular combination of factors, circled in red, the probability of success is significantly lower. In this case, data exploration via regression modeling allowed IDA researchers to clearly identify a set of conditions that measurably degrade performance. Providing this information to the Navy might allow operators to make adjustments in their employment of

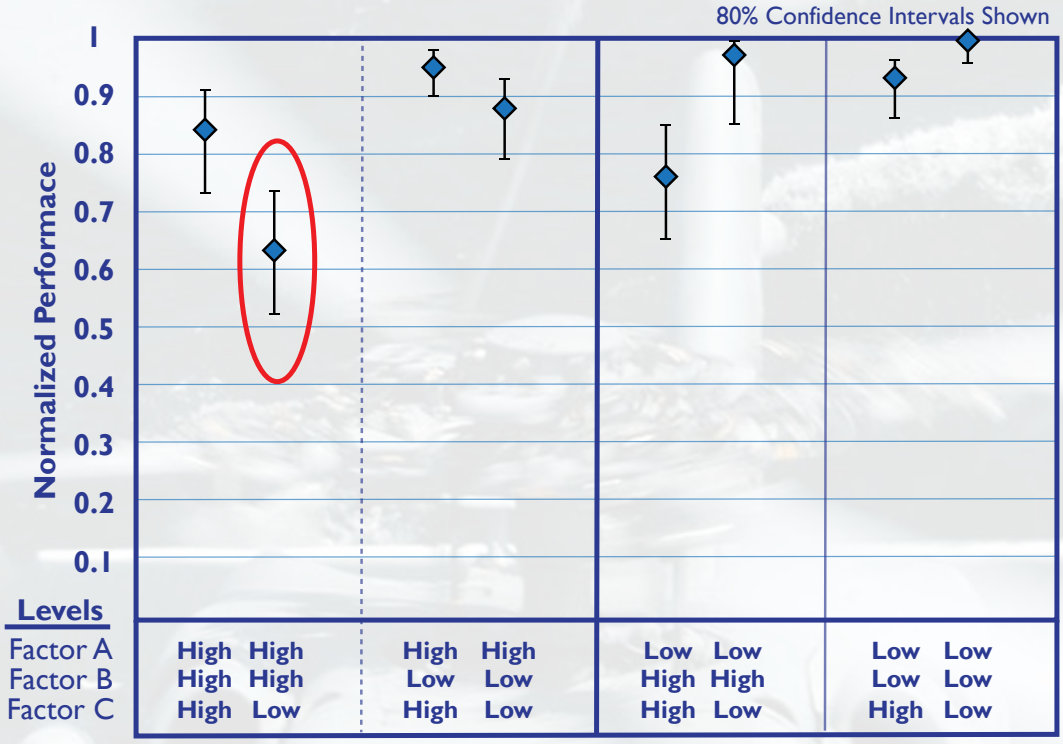


Figure 4. Logistic Regression Results

the MTS while helping the program manager and developers to focus resources on improving the system.

CONCLUSIONS

Properly evaluating system performance is critical to providing effective systems to our armed forces. IDA recently conducted an analysis of a new targeting system for Navy helicopters, applying rigorous statistical methods in order to discern key performance limitations. The resulting analysis provided the operational user with a more comprehensive understanding of their systems and highlighted key characteristics of operational performance that otherwise would not have been apparent. Armed with this knowledge, the Navy can develop the appropriate capabilities-based force structure and the most effective front line tactics, techniques, and procedures to counter the threat and safeguard our forces.

CAPT Keese (USN, ret.) is a Research Staff Member in IDA's Operational Evaluation Division. He holds a Master of Science in national resource strategy from the Industrial College of the Armed Forces, National Defense University and a Master of Science in information systems technology from the Naval Postgraduate School.

Dr. Rabinowitz is a Research Staff Member in IDA's Operational Evaluation Division. He holds a Doctor of Philosophy in physics from Columbia University.

Tackling Complex Problems: Analysis of the AN/TPQ-53 Counterfire Radar

Matthew R. Avery and Michael R. Shaw

THE PROBLEM

The performance of combat systems can be affected by a wide variety of operating conditions, threat types, system operating modes, and other physical factors. The character of the resulting multivariate test data can preclude simple or standard analysis methodologies. IDA's analysis methods rely on a variety of advanced statistical techniques to provide a better characterization of system capabilities than the techniques historically used to evaluate test results of combat systems.

IDA analyzed the target location error [TLE] data using a lognormal regression...to take the skewness of the data into account so that the fit has the same characteristics as the data.

BACKGROUND

Mortar, rocket, and artillery fire posed a significant threat to U.S. forces in Afghanistan and Iraq and will likely continue to pose a significant threat to ground troops in future conflicts. The AN/TPQ-53 Counterfire Radar (see Figure 1) is a ground-based radar designed to detect incoming mortar, artillery, and rocket projectiles; predict impact locations; and locate the threat geographically. Threat location information allows U.S. forces to return fire on the enemy location, and impact location information can be used to provide warnings to U.S. troops. The Q-53 is the next generation of counterfire radar, replacing the currently fielded AN/TPQ-36 and AN/TPQ-37 Firefinder. The Army conducted the Initial Operational Test and Evaluation (IOT&E) for the Q-53 in 2014, and the Army has

Note: The command and control vehicle is not shown.

Figure 1. Soldiers Emplacing the Q-53 Radar during the IOT&E

since made changes to the software and hardware designed to address discovered issues. The system had another IOT&E in June 2015. Because of urgent wartime requirements, the Army fielded 32 systems of an earlier version of the Q-53 radar. The Army plans to purchase an additional 136 Q-53s to allow every Army combat Brigade, Fires Brigade, and Divisional Artillery to have two Q-53 radars.

The Q-53 has a variety of operating modes designed to help optimize its search. The 360-degree mode searches for projectiles in all directions around the radar, while 90-degree search modes can be used to search for threats at longer ranges in a specific sector. In addition, the 90-degree mode has two sub-modes. In the 90-degree normal mode, the radar searches a 90-degree sector out to 60 kilometers. In the 90-degree Short Range Optimized Mode (SROM) mode, the radar focuses on short range threats, sacrificing some performance at longer ranges.

In addition to the various operating modes, the Q-53 radar's performance can vary depending on characteristics of incoming projectiles' trajectories and geometry relative to the radar's position. Determining how much the radar's performance varies across all these factors is essential to inform users of the capabilities and limitations of this system as well as to identify deficiencies in need of correction. Figure 2 outlines a standard fire mission for the Q-53. During a threat fire mission, the threat will fire projectiles at a target inside the search area of the Q-53. (Figure 2 shows a Q-53 operating

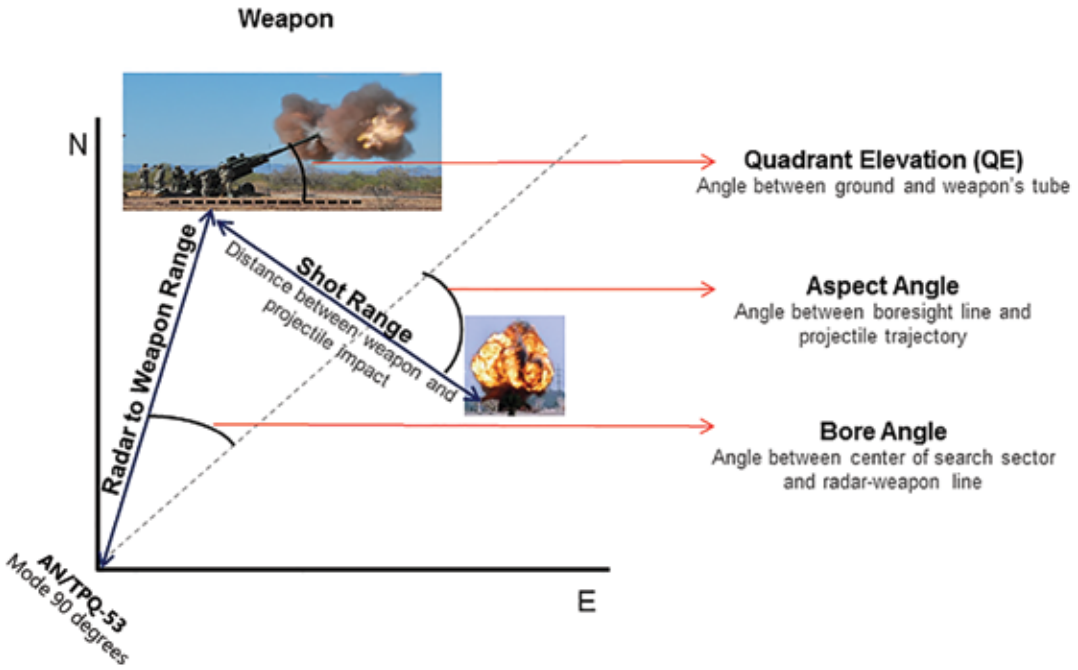


Figure 2. Example of a Fire Mission Including Relevant Geometric Factors Impacting Q-53 System Performance

in a 90-degree mode, so its search sector is limited to the area within the black bars.) The Q-53 must detect the projectile's trajectory and then estimate the position of the threat's weapon so U.S. forces can counter-attack. The specific geometry of the scenario will impact the Q-53's ability to track the projectile. Relevant factors include radar weapon range (the distance between the Q-53 and the weapon firing the projectile), quadrant elevation (the angle of the projectile's trajectory relative to the horizon), and shot range (the distance between the weapon and its target). When operating in 90-degree modes, the angle between the center of the radar's sector and the projectile's trajectory (bore angle) may also impact performance.

The key questions about system performance are: (1) Can the Q-53 detect shots with high probability? (2) Can the Q-53 locate a shot's origin with sufficient accuracy to provide an actionable counterfire grid location?

Q-53 OPERATIONAL TESTS

The June 2014 Q-53 IOT&E replicated typical Q-53 combat missions as much as possible given test constraints. Four radars (two Battalions) observed shots fired from a variety of weapons. Each Battalion decided how to employ the radar, within given test parameters, based on intelligence reports provided by the test team. Test personnel fired U.S. and threat weapons throughout four 72-hour test phases. During a single threat fire mission, test personnel fired projectiles (between 1 and 20, typically 10) from a single location using the same gun parameters, simulating a typical engagement that a Q-53 Battalion might encounter in a combat scenario. During a volley fire mission, test personnel fired projectiles from three weapons at the same time. Volley fire is a common technique used to increase the number of rounds hitting the target in a fire mission. Since the radar did not move during these missions, all of the factors in Figure 2 were held constant during each threat fire mission. Many missions were observed by two radars, enabling a single threat fire mission to be detected by two radars. Testers fired 2,873 projectiles, which resulted in 323 usable fire missions.

Figure 3 shows the raw probability of detection data. Each point represents a fire mission, with the size of the point determined by the number of shots taken in the fire mission, ranging from a single shot to as many as 20 projectiles. The percentage of those shots detected by the Q-53 counterfire radar is shown on the y-axis. The colors of the points show the munition, and different operating modes and fire rates are separated across the x-axis.

As Figure 3 shows, there is substantial variability in probability of detection across different combinations of operating mode, munition, and rate of fire. There are geometric differences between operating modes, complicating the definition of a shot's geometry. For example, in 360-degree mode, there is no angular center and therefore no bore angle. As a result, the 90-degree modes must be analyzed separately from the 360-degree modes to ensure that bore angle is properly taken into account. Additionally, the data are heavily imbalanced. The

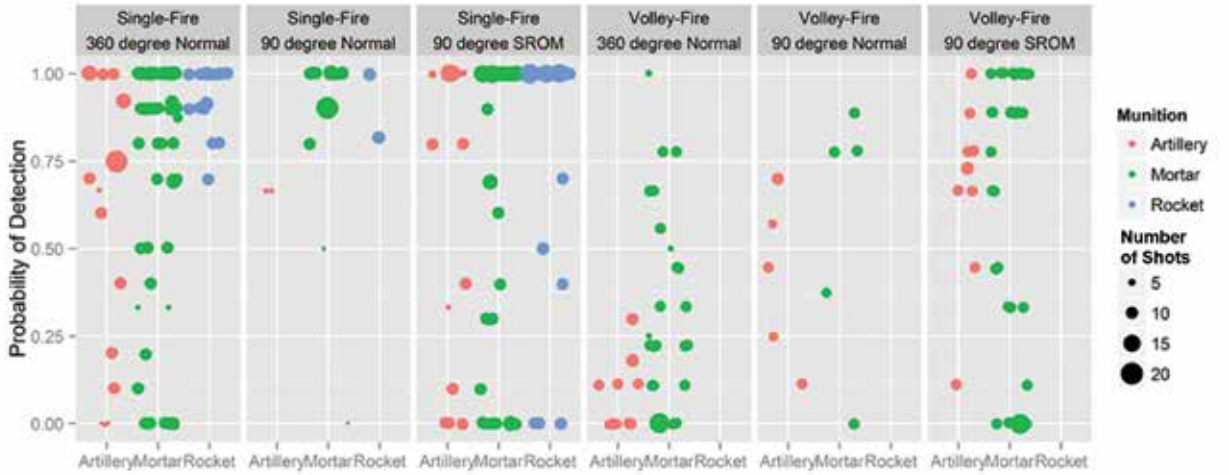


Figure 3. Detection Probabilities for 323 Fire Missions Conducted During the Q-53 IOT&E

choice of the 90-degree operating mode was left to the Brigades. They quickly learned that most of the threat missions were within SROM capabilities, so 90-degree Normal was used substantially less than 90-degree SROM. There are substantially fewer volley fire shots than single fire shots. (No volley fire rocket missions were undertaken because of test limitations.) Furthermore, many of the geometric factors described in Figure 2 were confounded with each other because of limited available firing points on the test range. As often happens in operational testing, the Q-53 test conditions resulted in imbalanced correlated data. The challenges in analyzing these types of data are best addressed with advanced analytical techniques.

LOGISTIC REGRESSION

When characterizing system performance, it is important to account for all factors that impact system performance. While Figure 3 shows some of the major factors that impact Q-53's ability to detect projectiles, the geometry of the shot (as shown in Figure 2) is not taken into account. Therefore, IDA employed a logistic regression analysis, a natural choice considering the complex nature of the problem. It allowed us to identify which factors were driving performance and to generate estimates of probability of detection for all combinations of factors. Most importantly, this approach allowed us to look at the impact of each factor, after accounting for the others, to determine which factors have the largest impact on performance. The general logistic regression equation is

$$log\left(\frac{p}{1-p}\right)=\beta_0+\beta_1 x_1+\cdots+\beta_N x_N.$$

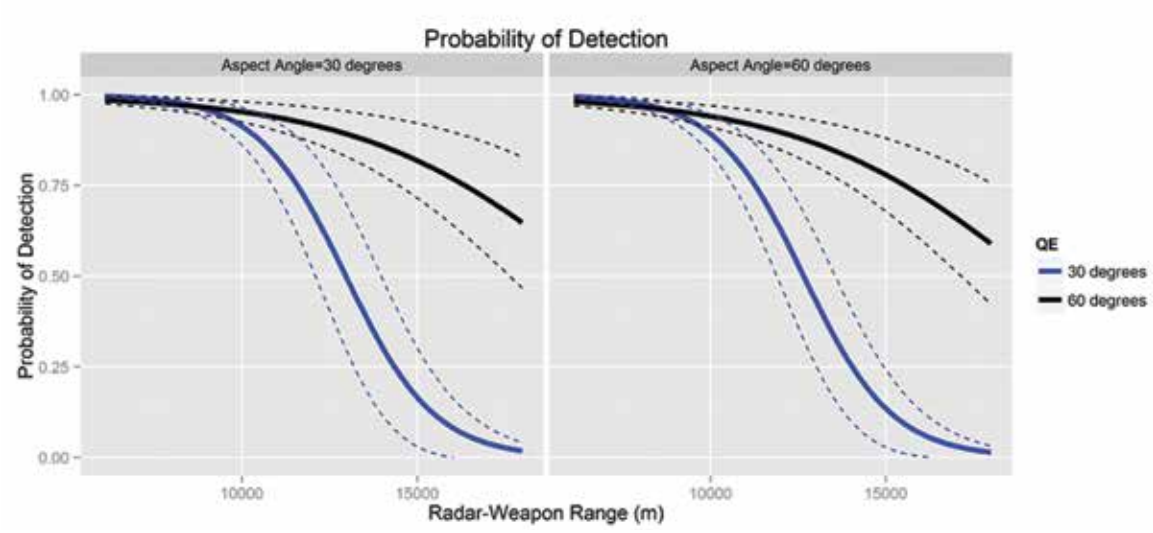
In our case, p is the probability of detection, and the x_i and β_i represent the factors and coefficients, respectively. This approach relates the log of the odds ratio of probability of detection to the various factors that impact the probability of detection. Unlike a more traditional approach that looks at factors one at a time, this method allows us to attribute changes in probability of detection to specific factors.

Importantly, this also allows us to identify which of our considered factors are not driving performance. Such factors can be eliminated from the statistical model, simplifying the final expression without surrendering its explanatory power.

ESTIMATING Q-53 DETECTION PERFORMANCE

The logistic regression model, once determined from the data, showed that – in addition to projectile type, operating mode, and rate of fire – radar weapon range, quadrant elevation (QE), aspect angle, and shot range had an impact on system performance. Figure 4 shows how the probability of detection changes as the distance between the weapon and the Q-53 counterfire radar increases when the system is in the 360-degree operating mode observing single-fire artillery engagements. The data also revealed that radar-weapon range and quadrant elevation affected Q-53's ability to detect incoming projectiles. These factors are linked to the time the projectile travels through the radar search sector. High arcing shots (larger values for quadrant elevation) are easier to see than shots with shallower trajectories that stay closer to the ground (low quadrant elevation) and are more likely to be masked by terrain. Longer shots (higher shot ranges) and shots with trajectories exposing larger cross-sections of the projectile to the radar (smaller aspect angles) were also easier for the Q-53 to detect, although the data showed these factors to be less important than radar-weapon range and quadrant elevation.

The logistic regression approach we employed also allows us to analyze the impacts of these factors simultaneously and observe how they interact. In Figure 4, as the radar-weapon range increases, the probability of detection drops sharply around 12,000 meters for shots with



QE – Quadrant Elevation

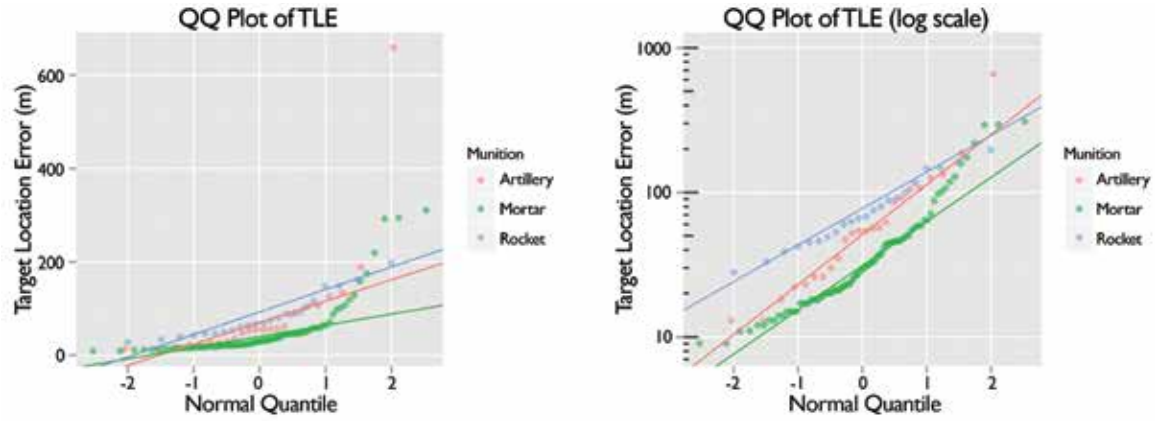
Figure 4. Probability of Detection for the Q-53 Counterfire Radar Using the 360-Degree Operating Mode Against Single-Fired Artillery

shallow shot trajectories (QE=30 degrees, shown with the blue lines). For the shots with more arc (QE=60 degrees, shown with black lines), the Q-53 is still able to detect with high probability at much longer ranges. While these factors have significant effects, other factors such as aspect angle have relatively minor effects on the probability of detection. Comparing the left and right panels of Figure 4, we can see that a 30-degree change in aspect angle results in a change in the probability of detection no greater than 7 percent. This logistic regression analysis allowed IDA to determine the relative impact of each factor on the probability of detection. While Figure 4 shows results for only a single combination of operating mode, munition, and projectile, IDA estimated the probability of detection across all factor levels. The Army could use this analysis to inform tactics for employing the system effectively in combat as well as identifying areas for future improvement of the system.

ESTIMATING THE THREAT'S LOCATION

In addition to detecting incoming projectiles, the Q-53 counterfire radar also estimates the location from which the detected projectiles were fired. The radar tracks the projectile through most of its flight and then backtracks the trajectory to estimate the threat's location (the point of origin of the trajectory). The distance between this point of origin and the location estimated by the Q-53 is referred to as target location error (TLE). The estimated location needs to be as accurate as possible, since it can become a target for counter-attack by U.S. forces. For this analysis, a single target location estimate was calculated for each fire mission, since all projectiles from a fire mission originated from the same location. As a result, there are fewer data for the TLE problem than the probability of detection problem. TLEs present an additional challenge, because these measurements are not normally distributed, which means standard analysis approaches will produce biased results. Figure 5 shows quantile plots of TLEs for the 360-degree operating mode, broken down by munition type. These quantile plots are arranged so data originating from a normal distribution will fall along the straight lines shown in the plot. The further away the data points fall from the straight line, the more the actual data distribution differs from a normal distribution. The chart on the left plots the raw data and reveals that they fall far from the straight lines. The plot on the right shows the same data on a log scale; the data fall much closer to the straight lines, which indicates that a lognormal distribution better represents the actual data distribution.

As a result, IDA analyzed the TLE data using a lognormal regression. This approach allows us to take the skewness of the data into account so that the fit has the same characteristics as the data. Figure 6 shows the results, with the figure on the left showing TLE for mortars and the figure on the right showing TLE for artillery and rockets. The green lines show the system's requirements, and the black lines show the estimated median TLE along with 80 percent confidence intervals. While TLE for mortars showed substantial variability,



If the data are normally distributed, the points should conform closely to the line. The plot for TLE shows that the largest observed TLEs far exceed the values expected from normally distributed data. By using the natural logarithm of the data (right plot), the data conform more closely to the normal distribution.

Figure 5. Quantile-quantile (QQ) Plots Used to Visually Assess Normality

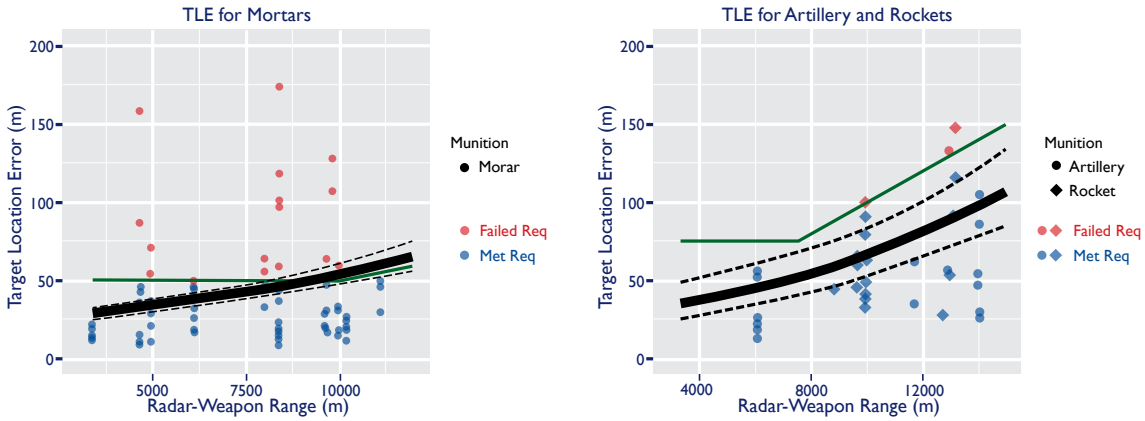


Figure 6. Q-53 Target Location Error for Estimated Weapon Locations

the large number of mortar fire missions allows us to make precise estimates of median TLE. The analysis revealed that the estimated median TLE tends to increase (get worse) as radar-weapon range increases. While the Q-53 is more accurate at estimating a mortar's location than the location of artillery and rocket weapons, the requirements for artillery and rockets were less stringent. As with probability of detection, IDA's regression approach accounts for the variety of factors impacting system performance, resulting in rigorous system evaluation.

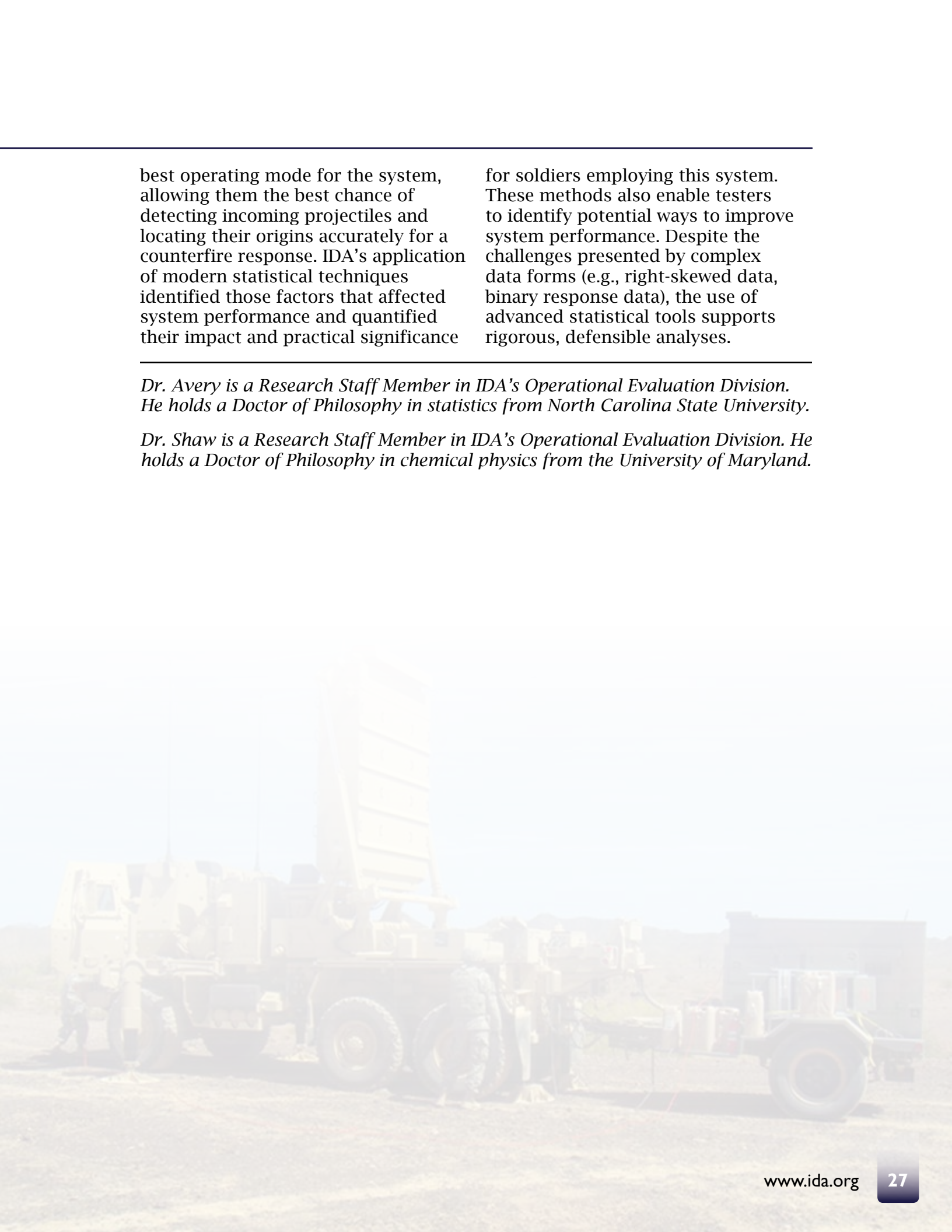
SUMMARY

IDA's analysis of the Q-53 Counterfire Radar illustrates the benefits of using more advanced data analysis techniques. Many factors, including physical factors related to the shot's geometry as well as threat and operating mode, affect Q-53 performance. Understanding the effects of these factors helps commanders in the field choose the

best operating mode for the system, allowing them the best chance of detecting incoming projectiles and locating their origins accurately for a counterfire response. IDA's application of modern statistical techniques identified those factors that affected system performance and quantified their impact and practical significance for soldiers employing this system. These methods also enable testers to identify potential ways to improve system performance. Despite the challenges presented by complex data forms (e.g., right-skewed data, binary response data), the use of advanced statistical tools supports rigorous, defensible analyses.

Dr. Avery is a Research Staff Member in IDA's Operational Evaluation Division. He holds a Doctor of Philosophy in statistics from North Carolina State University.

Dr. Shaw is a Research Staff Member in IDA's Operational Evaluation Division. He holds a Doctor of Philosophy in chemical physics from the University of Maryland.



Improving Reliability Estimates with Bayesian Hierarchical Models

Kassie Fronczyk, Rebecca Dickinson, and Laura Freeman

THE PROBLEM

The reliability of a weapon system is an essential component of its suitability for operational deployment. Yet, in an era of reduced budgets and limited testing, verifying that reliability requirements have been met can be challenging, particularly using traditional analysis methods that depend on a single set of data coming from a single test phase.

Short test periods, high reliability requirements, or few observed failures can result in little confidence in the reliability estimates... DoD [should] employ statistical approaches to capitalize on all available data from multiple test periods and not limit the reliability analysis to a single test period.

In the Department of Defense (DoD), test data are often collected in several phases. The two broad types of testing are developmental testing and operational testing. The primary goal of a developmental test (DT) is to verify that a system meets its design specifications. This testing can occur as contractor testing, government testing, or a mixture of both and is usually carried out in a controlled environment that often lacks the realism of combat scenarios and trained users. The purpose of an operational test (OT), on the other hand, is to determine whether the system is effective and suitable in a combat scenario. OT data are collected under test conditions that replicate, as much as possible, field use.

Reliability is one of the primary aspects of a system's operational suitability. It is important that a system perform as intended under realistic operating conditions for a specified period of time without failure. Reliability requirements for ground vehicles are often based on the mean number of miles between failures. A serious equipment failure that occurs during mission execution and results in the abort or termination of a mission is scored as an Operational Mission Failure (OMF). A less critical failure of a mission-essential component is scored as an Essential Function Failure (EFF). For example, an engine failure would be scored as an EFF if a vehicle took multiple attempts to start but eventually succeeded. If the vehicle could not be started, it would be scored as an OMF.

Requirements are typically written in terms of OMFs. Verifying whether the reliability requirements of a system have been met by looking at only a single test phase, however, can be challenging. Short test periods, high reliability requirements, or few observed failures can result in little confidence in the reliability estimates. The National Academies, in three

separate studies (National Research Council 1998,[1] 2004,[2] and 2015[3]), have recommended that DoD employ statistical approaches to capitalize on all available data from multiple test periods and not limit the reliability analysis to a single test period. Despite these recommendations, nearly every published analysis of a major weapon system's reliability limits the assessment to the last test phase, typically because that phase examined the most representative system configuration. In support of the Director, Operational Test and Evaluation (DOT&E), IDA has begun to explore improved techniques for estimating reliability using data from multiple test periods.

BAYESIAN PARADIGM

When combining information from multiple test periods, models need to be carefully selected and evaluated to ensure that they accurately reflect the data and the underlying physical processes. The Bayesian paradigm is tailor-made for these situations because it allows the combination of multiple sources of data and variability to obtain more robust reliability estimates and quantify properly the uncertainty and precision of the estimates. The use of Bayesian methods is becoming increasingly popular because leveraging all of the available information when making decisions under uncertainty makes practical sense. This article uses reliability data from two families of vehicles tracked through multiple phases of testing to illustrate the Bayesian approach of combining information. Applying these methods results in better estimates of system reliability and more precise inferences.

The first case study uses reliability data from the Stryker family of vehicles (FoV), which are armored combat vehicles built for the U.S. Army. The FoV includes 10 system configurations, with two main versions: the Infantry Carrier Vehicle (ICV) (see Figure 1) and the Mobile Gun System (MGS). Our study focuses on the ICV, which provides protected transport and supporting fire for its two-man crew and squad of nine

Source: M1126 Infantry Carrier Vehicle
The ICV serves as the base vehicle for eight additional system configurations.

Figure 1. Stryker Infantry Carrier Vehicle (ICV)

1 National Research Council. 1998. Statistics, Testing, and Defense Acquisition: New Approaches and Methodological Improvements. Washington, DC: The National Academies Press.

2 National Research Council. 2004. Improved Operational Testing and Evaluation Methods of Combining Test Information for the Stryker Family of Vehicles and Related Army Systems: Phase II Report. Washington, DC: The National Academies Press.

3 National Research Council. 2015. Reliability Growth: Enhancing Defense System Reliability. Washington, DC: The National Academies Press.

infantry soldiers. The ICV serves as the base vehicle for the eight remaining system configurations.[4] The vehicles share a common chassis and are outfitted with additional components specific to the mission of each vehicle. This analysis heavily leverages the common chassis of the vehicles to support combining information from all of the configurations. The reliability data, at the OMF level, used in this study come from two test phases: one DT and one OT.

The second case study is based on a notional future combat family of vehicles and data collected from multiple testing phases, as would be common for a program like Stryker. For this example, we will assume a family of vehicles similar to Stryker with four vehicles of various configurations that go through a series of three test phases with a corrective action period between each phase. Unlike the Stryker case study, for this notional example, we assume more detailed failure data are available, specifically EFFs and OMFs, as opposed to only OMFs. Because all OMFs are, by definition, EFFs, using all failures in the analysis provides a more robust reliability estimate.

For both cases, the goal is to characterize the reliability of the entire family of vehicles. In the Stryker study, we have OT data, but these data are limited; therefore, we need to leverage the commonalities of the vehicles and the DT data. For the notional future combat vehicle, we have assumed detailed information is available about failures from the three phases of testing and can pool information across the phases and four vehicles. A Bayesian framework that requires only slight modifications from one FoV to the other provides a mechanism to make the most use of this additional information.

STATISTICAL MODELS FOR COMBINING DATA: BAYESIAN RELIABILITY

A standard reliability analysis employed by the DoD test community considers each test phase (and each system configuration, such as vehicle type) independently and uses the exponential distribution to empirically model the miles between failures. Reliability is expressed in terms of the mean number of miles between a failure (MMBF), and is estimated as

$$\widehat{\text{MMBF}} = \frac{\text{Total Miles Driven}}{\text{\# of Failures}}.$$

Although this approach is standard for nearly every ground vehicle program in the Department, it ignores valuable information on individual vehicles in different phases of testing. Although frequentist statistical methods similar to the standard reliability analysis described previously (and illustrated in the Stryker analysis) could be used, a Bayesian approach provides a natural framework for combining multiple sources and types of information.

[4] The Antitank Guided Missile Vehicle (ATGMV), Commander's Vehicle (CV), Engineer Squad Vehicle (ESV), Fire Support Vehicle (FSV), Medical Evacuation Vehicle (MEV), Mortar Carrier Vehicle (MCV), Reconnaissance Vehicle (RV), and the Nuclear, Biological and Chemical Reconnaissance Vehicle (NBC RV). The NBC RV was excluded from the study because of its different acquisition timeline.

Bayesian methods are valuable for their logical integration of prior information and their practical convenience for modeling and estimation. Before looking at the test data, we construct a prior distribution – or starting assessment – for the parameters in the empirical model that we plan to construct. We use the data to revise our starting assessment and derive the updated assessment (i.e., the posterior distribution) for the parameters in the empirical model.

The reliability of the FoV is defined as a function of the failure rate parameter, λ (i.e., the mean time between failure (MTBF) or MMBF is $1/\lambda$). The exponential distribution is often used as the underlying assumption for the data's distribution, and a common choice of a prior distribution to describe the possible values of λ is the gamma distribution. The gamma distribution restricts the value of the failure rate to positive values and provides computational ease. Table 1 shows the Bayesian models for the Stryker and notional future combat vehicle FoV side by side to highlight the similarities and differences. For the Stryker analysis, we construct the statistical model such that each vehicle variant has its own failure rate, which is estimated by the data, and a single parameter to capture a common downgrade across vehicles from DT to OT. On the other hand, in the future combat vehicle example, the statistical model is written to capture the fact that the program has the ability to fix specific failure modes between phases (i.e., the corrective action periods). The statistical model, therefore, includes a separate estimate for each failure mode in each of the postulated test phases and fix effectiveness factors specific to each failure mode.

STRYKER FOV: ANALYSIS AND RESULTS

The reliability requirement for Stryker is that each vehicle has a mean of at least 1,000 miles between OMFs. Frequentist and Bayesian inference techniques were both employed to compare and contrast different approaches to combining

Table 1. Bayesian Reliability Models for Stryker and Future Combat Vehicle

Stryker	Future Combat Vehicle
$t_{DT} \sim exp(\lambda_i)$ $t_{OT} \sim exp(\lambda_i/\eta)$	$t_{T_1} \sim exp(\lambda_{ij})$ $t_{T_2} \sim exp(\lambda_{ij}(1-\rho_{1_j}))$ $t_{T_3} \sim exp(\lambda_{ij}(1-\rho_{1_j})(1-\rho_{2_j}))$
i=1, 2, 3, 4 (vehicle variants)	i = 1, 2, 3, 4(vehicle variants) j = 1, 2, ..., 26(failure modes)
$\lambda_i \sim gamma(a, b)$	$\lambda_{ij} \sim gamma(a, b)$
$\eta \sim beta(1, 1)$	$\rho_1 \sim beta(1, 1)$ $\rho_2 \sim beta(1, 1)$
$a \sim gamma(.001, .001)$	$a \sim gamma(.001, .001)$
$b \sim gamma(.001, .001)$	$b \sim gamma(.001, .001)$

DT and OT data. Figure 2 illustrates the results of the traditional analysis, the frequentist analysis,[5] and the Bayesian analysis. All three analyses use an exponential distribution to model the miles between failures, as discussed previously.

The mean miles between operational mission failures (MMBOMF) estimates reported in Figure 2 under the traditional analysis do not leverage the DT data or the relationships among the various types of Stryker vehicles. Notice that the CV vehicle stands out as potentially having an optimistically high MMBOMF of 8,494 miles. This estimate is based on a single failure and a combination of all the individual operating distances for each of the six CV vehicles. None of the six CV vehicles, however, used in OT traveled more than 2,000 miles. To claim that any one vehicle's MMBOMF is greater than 8,000 miles when no single vehicle traveled that far is questionable. Furthermore, if we consider that the estimate of MMBOMF in DT for the CV was less than 2,200 miles, we can conclude that it is highly unlikely that we would see such large improvements in the reliability between late DT and OT since no major changes were made to the system configuration. The MMBOMF estimate based on the traditional analysis approach is therefore highly suspect.

Confidence intervals for the FSV and the RV are also extremely wide because of the limited number of failures observed in OT. Because no failures were recorded for the MEV in OT, only a lower confidence bound can be estimated.

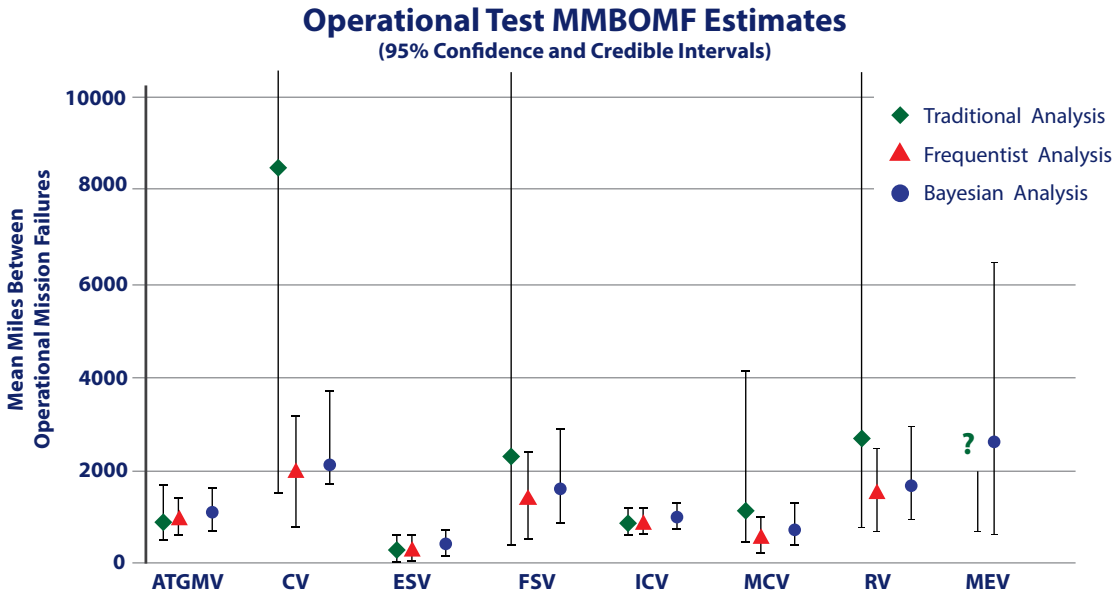


Figure 2. Stryker FoV: Comparisons of the OT MMBOMF Vehicle Variant Estimates for the Traditional Analysis, Frequentist Analysis, and Bayesian Analysis Using the Exponential Distribution

[5] An exponential regression model was used. Test phase and vehicle variants were included as explanatory variables so that individual reliability estimates for each of the vehicles within each test phase could be estimated.

Using a statistical model to formally account for differences in performance across test phases and vehicle variant has a large practical impact on the reliability results. In Figure 2, the two model-based analyses (i.e., frequentist analysis and Bayesian analysis) provide a more realistic estimate of CV reliability and improve the overall precision of the estimates of system reliability for the vehicles that exhibited a small number of failures in OT. The tighter confidence intervals are obtained by leveraging the failure information from the other variants and DT data. One clear advantage of using the Bayesian analysis in this example is that we can obtain a point estimate for the reliability of the MEV. The reliability estimate for the MEV is driven by the information that we have for the seven other vehicles.

The Stryker example demonstrates that when we combine the available information across two test phases, the reliability estimates are more accurate and precise than estimates based solely on OT data. We also obtain inferences for vehicles on which no OT data are available. The analysis considers only OMFs since this analysis allows for a direct comparison between standard DoD analysis and the analysis that combines information across the DT and OT phases. However, further improvements in reliability estimates might be achieved by leveraging information from EFFs and/or failure modes. In the following example, we leverage information from OMFs and EFFs.

FUTURE COMBAT VEHICLE: ILLUSTRATION OF ANALYSIS AND RESULTS

For the notional future combat vehicle example, we assume very detailed failure information exists for the four vehicles tested in three test phases. In other words, the data for EFFs are available in addition to OMFs, and each EFF and OMF is attributed to a specific failure mode (e.g., brakes, fuel system, and suspension). The Bayesian model in Table 1 allows for a separate reliability estimate for each observed failure mode that arises across the test phases. Also, by using the information learned in the analysis about the individual failure modes, we can estimate the reliability for each vehicle. This reliability estimate provides a much richer source of information than the estimate derived in the equation on page 30, which simply takes the total number of miles driven by all four vehicles in each phase and divides by the total number of failures from the phase to determine reliability for the FoV.

In the Department of Defense, reliability requirements are typically written at the family level for these types of programs and in the language of OMFs. However, this analysis focuses at the vehicle level and includes all EFFs. By analyzing all EFFs and capitalizing on the information that is known about each of the failure modes, we are more likely to identify a larger portion of failures that cause system downtime, which will lead to greater improvements in reliability, availability, and maintainability and reduced operating and maintenance

costs. Furthermore, by breaking out the failures by vehicle, we more accurately determine the reliability of the FoV.

Figure 3 shows the estimated mean miles between essential function failures (MMBEFF) for each vehicle across the three test phases. The statistical model links the reliability estimates through each of the three phases by estimating the reliability as a function of the successful fixes between phases. If a program is using the corrective action period to fix some fraction of the observed failure modes, then the MMBEFF across the three phases of test should increase. In fact, the model assumes that this increase must occur (see Table 1). As seen in Figure 3, for the four vehicles, the MMBEFF increases from around 50 miles to around 60 miles from Phase 1 to Phase 2 and then gains another 30 miles from Phase 2 to Phase 3.

Similar to the Stryker example, we investigate the gain over a traditional analysis. Figure 4 shows MMBEFF estimates and intervals for the four vehicles across the three test phases using the Bayesian hierarchical model and the traditional exponential analysis, separated by vehicle and phase. The Bayesian analysis always provides a tighter interval estimate, meaning those results are more certain and precise, which is a direct result of leveraging information from all vehicles and all phases of test. The Bayesian analysis also shows distinct growth for each of the four vehicles, while the traditional analysis reveals growth in reliability across phases for only two of the four vehicles: vehicle 3 and vehicle 4.

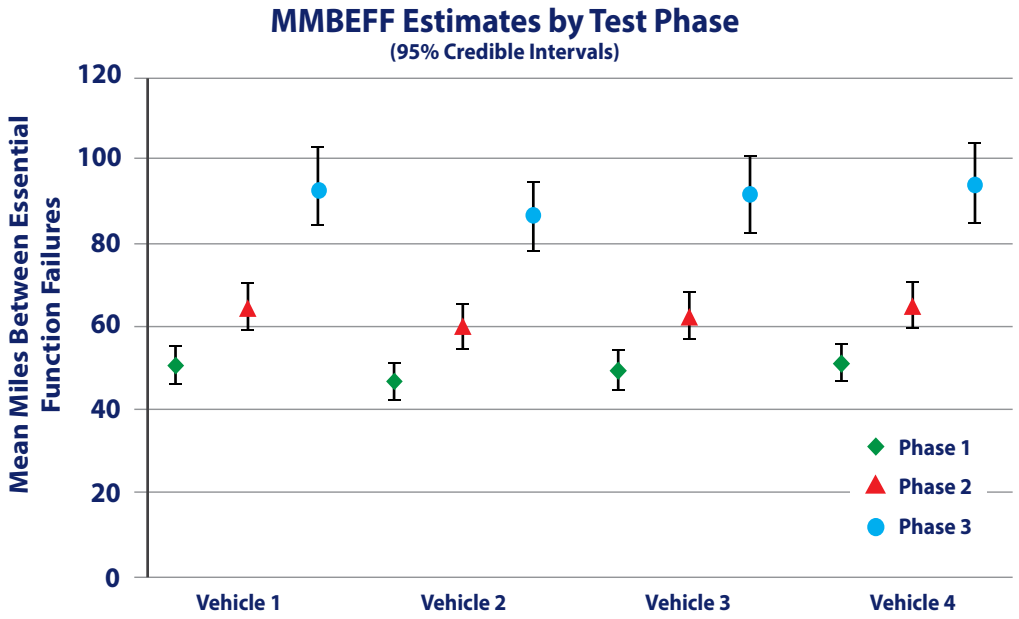


Figure 3. MMBEFF Estimates and 95 Percent Credible Intervals for Each Future Combat Vehicle and All Three Phases of Test

In Figure 4, the traditional analysis does not show any evidence of growth in the first two phases of testing for three of the four vehicles (only vehicle 3 shows some marginal growth). The Bayesian analysis assumes that growth will occur between each phase as a result of the model specification, so the results are more definitive with respect to growth between periods; however, this assumption is not required. Future sensitivity analyses of the results on the model specification will be important for understanding the influence of the model and assumptions used. Nevertheless, these results reveal the strength of these methods for analyzing data and capitalizing on all the data available to provide more accurate insight into system reliability over time.

CONCLUSIONS

The Bayesian approach to reliability analysis provides a formal framework to combine information from multiple sources and attain appropriate uncertainty quantification. The two examples discussed in this article illustrate the advantages of

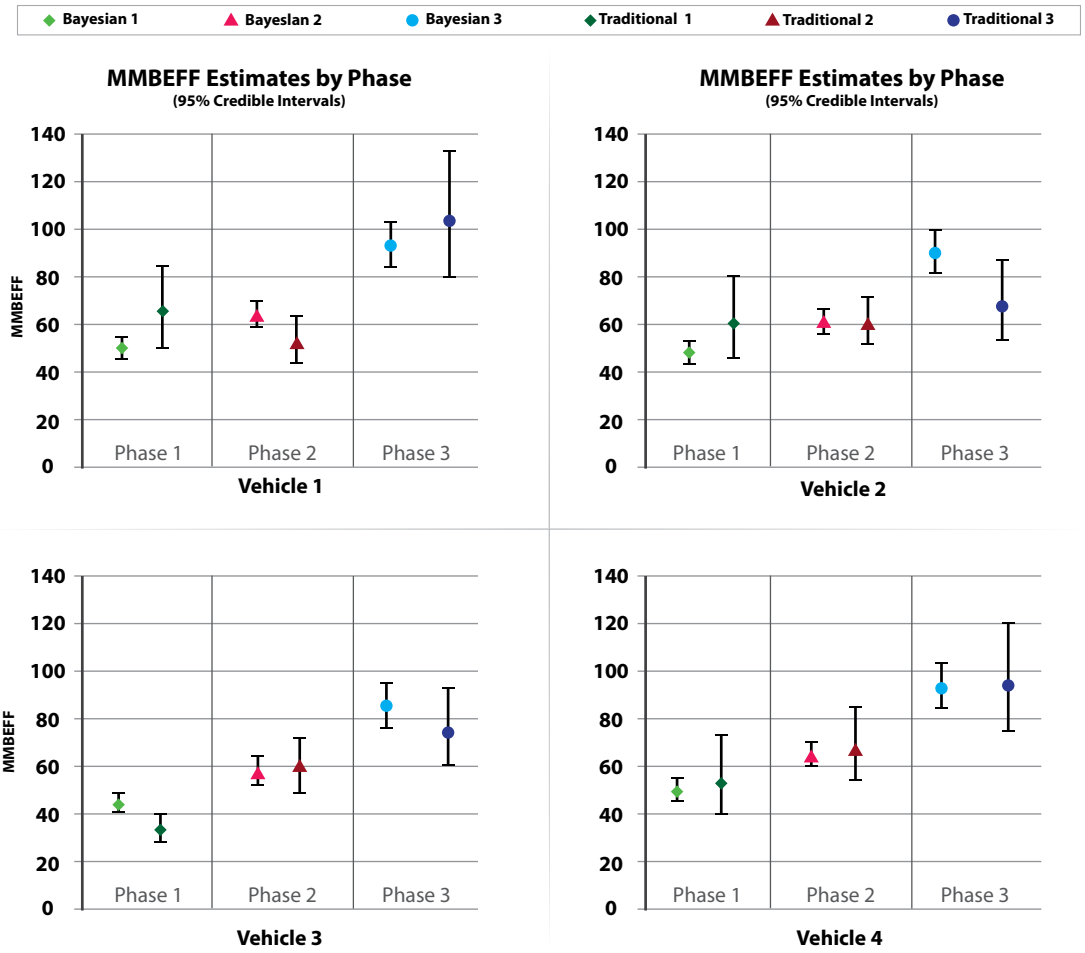


Figure 4. Comparisons of the MMBEFF for Four Vehicles Across the Three Phases of Test, for the Bayesian Analysis and Traditional Analysis Using the Exponential Distribution

using data from multiple phases of testing and leveraging data from systems with common infrastructures. The results are better estimates of system reliability and more precise inferences. Further improvements in reliability estimates are achieved by leveraging information from EFFs. By exploiting all available information and tools, we can obtain rich inferences for very complex problems.

Dr. Fronczyk is a Research Staff Member in IDA's Operational Evaluation Division. She holds a Doctor of Philosophy in statistics and stochastic modeling from the University of California, Santa Cruz.

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.

Dr. Freeman is an Assistant Director in IDA's Operational Evaluation Division. She holds a Doctor of Philosophy in statistics from the Virginia Polytechnic Institute and State University.

Managing Risks: Statistically Principled Approaches to Combat Helmet Testing

Janice Hester, Thomas Johnson, and Laura Freeman

PROBLEM

Combat helmets protect troops against artillery rounds, mines, and small caliber bullets. Helmet designers strive to achieve high ballistic protection with lightweight helmets. Modern combat helmets are made from dynamic materials such as aramid and ultra-high molecular weight polyethylene fibers, which show more variability in performance than simpler armors. The Services conduct acceptance tests to evaluate the ballistic performance of each helmet design and production lot. The challenge to testers is to construct efficient tests to determine whether these helmets meet performance criteria.

BACKGROUND

Combat helmet designs are driven by the balance between increasing ballistic protection and decreasing weight. Starting in World War I, troops wore steel helmets to protect against artillery rounds. In 1985, the Personnel Armor System for Ground Troops (PASGT) helmet was fielded. The PASGT helmet was made from a laminate of ballistic material with aramid fibers, and it improved protection against fragments. In 2002, the U.S. Army replaced the PASGT helmet with the lighter weight Advanced Combat Helmet (ACH). The ACH and similar helmets are currently the most common helmets worn by U.S. troops. Recently, the U.S. Marine Corps developed the Enhanced Combat Helmet (ECH), which has a ballistic laminate of ultra-high molecular weight polyethylene fibers and provides some limited protection against small caliber bullets. Helmet designs continue to evolve, and the U.S. Army is pursuing two new helmet types – one that provides the protection of the ACH but is lighter weight and another that provides the protection of the ECH but is lighter weight. Figure 1 shows the evolution of combat helmets through the years.

IDA's analyses were central to the development of the most recent version of the improved, statistically principled acceptance test protocols for combat helmets.

Beginning in 2007, congressional concern about the accuracy and consistency of body armor testing led to increased involvement in personal protective equipment by the Director, Operational Test and Evaluation (DOT&E). To address the concerns of Congress, DOT&E worked with the Services to develop test protocols for the ballistic components of First Article Testing (FAT) and Lot Acceptance Testing (LAT) for both body armor and combat helmets. In 2009, DOT&E asked IDA to expand its support for live fire

Figure 1. Evolution of DoD Combat Helmets

test and evaluation to include personal protective equipment. IDA's analyses were central to the development of the most recent version of the improved, statistically principled acceptance test protocols for combat helmets.

Helmets must protect against multiple ballistic threats; this article focuses on IDA's work on testing for resistance to penetration and ballistic limit estimation. Our work on evaluating resistance to penetration using statistically principled testing has led to an improved FAT protocol for aramid-based helmets. Our related research comparing newer design methods for fragment testing suggests that additional improvements to the protocols are possible for the estimation of ballistic limits. The statistical work discussed in this article is supported by frequent observations of helmet testing and continual analysis of helmet test data, which together ensure that the statistical studies are relevant to helmet testing.

RESISTANCE TO PENETRATION

Combat helmets must demonstrate a high probability of stopping perforation from a 9mm handgun round, and some designs must also prevent perforation from another specified small arms round. Each helmet design comes in at least four sizes, and during FAT they are shot at five locations on the helmet and subjected to four separate environmental conditioning treatments. The FAT must provide confidence that all helmet sizes have acceptable performance under all test conditions. The primary statistical challenge for this component of testing is to design an efficient test that provides this confidence while still achieving a low risk of rejecting helmets with good performance.

The response of a combat helmet to a threat impact is stochastic, so resistance to penetration is characterized as the probability of a projectile completely penetrating through the helmet. This probability should be very low. The probability of penetration can be measured with increasing precision as the number of test shots increases, but helmet testing is expensive and destroys the tested helmets. Accordingly, tests should be efficient in the number of test articles they require.

Acceptance test designs should balance the risks of wrongly accepting a product that performs poorly and of rejecting a product that performs well. One important statistical tool for comparing acceptance test design is an operating characteristic (OC) curve, which shows the probability of accepting a helmet (passing the test) as a function of the true probability of penetration. Figure 2 shows OC curves for three notional tests that range in size from 75 to 450 test shots. The numbers of shots and allowable penetrations determine the shape of the curve, including the government's risk of accepting helmets with low performance and the manufacturer's risk that helmets with high performance will be rejected. The curves in Figure 2 all have the same manufacturer's risk; increasing the test size results in a steeper OC curve and decreases the government risk.

A helmet design's resistance to penetration can vary among the helmet sizes or across test conditions. A test with a single acceptance criterion on the helmet design's performance across all sizes and test conditions is therefore not sufficient. Instead, helmets must demonstrate performance across all sizes and conditions, which tends to increase the probability of incorrectly concluding that a helmet does not meet performance criteria.

IDA developed an analytical framework for resistance to

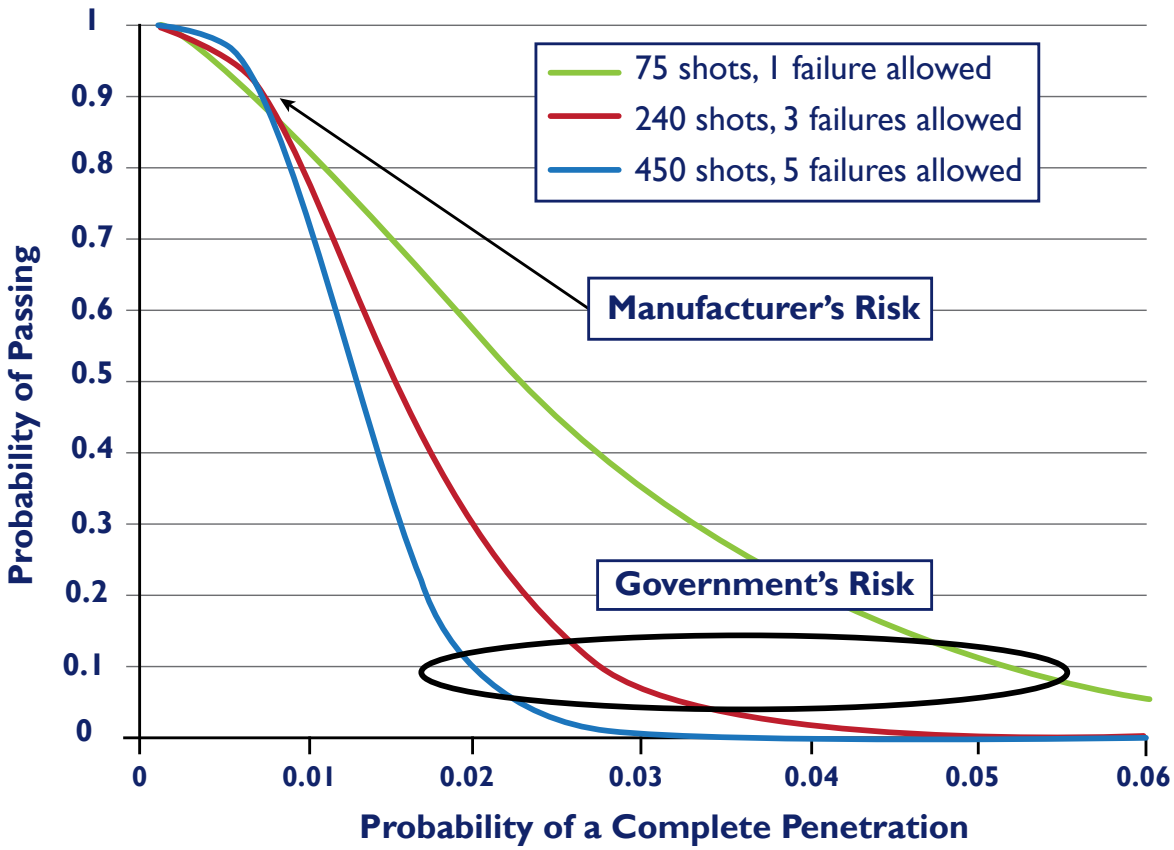


Figure 2. Operating Characteristic Curve for Test Sizes Ranging from 75 to 450 Shots

penetration testing during FAT that captures the tradeoff between the risks of accepting a helmet that performs poorly in one condition and of rejecting a helmet that performs uniformly well across all conditions. This is in contrast to simpler tests, like those shown in Figure 2, for which government and manufacturer risks are set for two different performance levels of the same characteristic (e.g., a single, aggregated probability of a complete penetration).

Instead of selecting a single pass/fail criterion, we select a set of pass/fail criteria that specify a maximum acceptable number of complete penetrations across all shots taken and a maximum acceptable number of complete penetrations within the shots taken on each individual test condition. For example, in the new protocol for aramid-based helmets, no more than three penetrations for the 9mm round are allowed across all sizes, environments, and locations (240 shots total). Of those three penetrations, no more than two can be in any one size. Similar criteria exist for environment and shot location.

Figure 3 shows the operating characteristic curves for the protocol for aramid-based helmets. The dotted blue curve shows the probability of passing the aggregate criterion (three allowed penetrations across all 240 shots) as a function of the aggregate probability of a complete penetration; the solid green curve shows the probability for each helmet size of passing the criterion on the individual size (two allowed on any single size) as a function of the probability of complete penetration for that helmet size; and the solid red curve shows the OC curve for passing all of the multiple test criteria simultaneously for the simple case in which the probability of a complete penetration does not vary among the helmet sizes or test conditions. Figure 3 illustrates how the statistical methodology IDA developed provides acceptable risk points both when all helmets have uniformly high performance and when one helmet size is different.

The key element of a hierarchical test is that if a helmet has uniform performance across the conditions, then the risk points for the full hierarchical test closely match the risk points for the aggregate criterion alone. The criteria on the individual conditions are selected such that, for a helmet with uniform performance, simultaneously passing the aggregate criteria and failing an individual criterion through random chance are unlikely. The benefit of this approach is that the FAT results are diagnostic and easy to interpret. If a helmet design's aggregate performance is low but uniform across the test conditions, then failing for the aggregate criterion is more likely than failing for one of the criteria on the individual conditions. On the other hand, if a helmet has high aggregate performance but a single low performing condition, then failing the pass/fail criterion on that condition is the most likely result. One drawback to this approach is that the aggregate and individual criteria cannot be specified independently. Finer control over these risk points is possible with more complex test

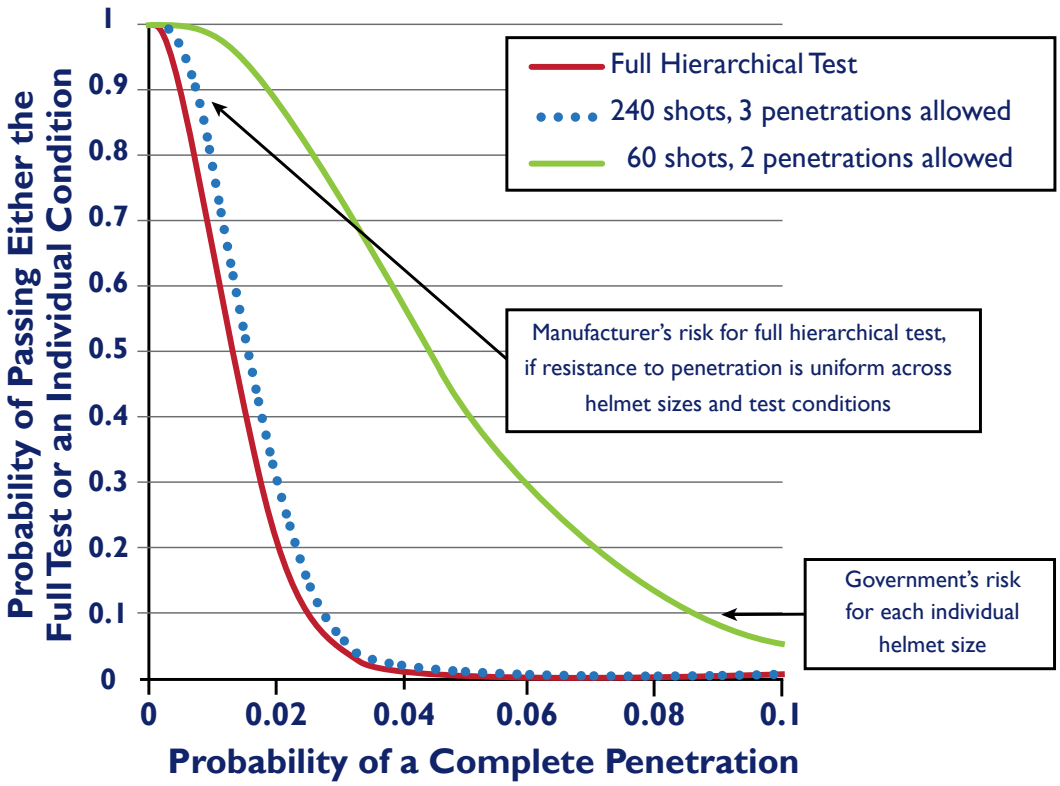


Figure 3. Operating Characteristic Curves for the DOT&E Aramid Helmet Protocol

designs that include the possibility for multiple rounds of testing.

BEHIND HELMET BLUNT TRAUMA

The rapid deformation of a combat helmet following a ballistic impact creates a potential for blunt trauma injury even when the projectile does not completely penetrate the helmet; the deforming helmet shell can impact the wearer's head. To mitigate this risk, the helmet's deformation following an impact with the 9mm test round is measured during testing and compared to established upper limits. Figure 4 shows the image of a head form filled with clay before a shot is taken (left) and after (right); the maximum deformation is measured from the deepest location in the clay indent.

The FAT and LAT protocols include a procedure for assessing the measured deformations against the established upper limits. Deformation requires a different

Figure 4. Clay Helmet Head Form Before the Shot (left) and After the Shot (right) Illustrating the Helmet Deformation into the Clay Channel

analysis method than resistance to penetration, because deformation is a continuous metric (a measured value) rather than a binomial (success/failure) metric. IDA used simulation studies to investigate the best approach to writing a protocol for deformation that accounts for the multiple test conditions. We showed that Analysis of Variance (ANOVA) can be applied within a FAT to test individual conditions while controlling overall risks.

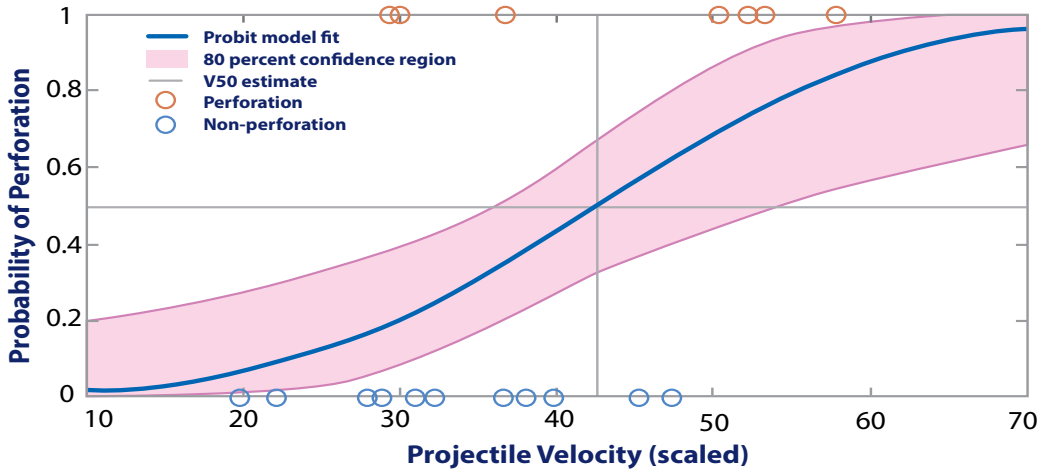
FRAGMENT THREATS AND BALLISTIC LIMIT ESTIMATION

The Services set helmet performance requirements for the minimum ballistic limits against several standard fragment simulants; these limits are incorporated into the DOT&E protocols. The ballistic limit is the velocity at which a projectile completely penetrates the helmet 50 percent of the time. While the velocity corresponding to a lower probability of penetration would be a better measurement of ballistic protection, the 50th percentile has been used historically because it is the percentile that is measured with the greatest precision.

The test and academic communities have developed several different procedures for determining the ballistic limit of armor through testing. IDA performed a simulation study to determine which of six published procedures would be the most efficient and accurate if used for helmet testing. Each procedure combines a set of rules for selecting shot velocities, terminating testing, and calculating the ballistic limit. To ensure that the simulation results were relevant, the simulation incorporated historical helmet performance data.

To estimate the ballistic limit, testers vary the velocity of the test fragment between shots in a prescribed manner with the goal of finding a velocity range in which there is a mix of penetrations (failures) and non-penetrations (successes). The orange and blue circles in Figure 5 are example data for a ballistic limit test; they illustrate the spread in helmet performance for velocities near the ballistic limit.

Under the current test procedures, which are known as the "up-down method," testers select each shot velocity by increasing or decreasing the velocity based on the previous shot's outcome. Once testers achieve a predetermined equal number of complete penetrations and helmet successes within a fixed velocity range, they stop the test and estimate the ballistic limit as the arithmetic mean of this set of shots. The up-down method is not statistically rigorous for multiple reasons, but in particular it frequently does not use all of the data to determine the ballistic limit. For example, if eight shots are required to get three successes and three failures in the required velocity range, the analysis throws away the other two data points. Newer test design and analysis methods use generalized empirical model fits based on all the data to both determine the next shot in the test sequence and characterize the probability of penetration as a function of the projectile's velocity.



Orange circles indicate shots that perforated the helmet while blue circles indicate shots that were stopped by the helmet. Note that there is a velocity range in which the results vary.

Figure 5. Example Data from a Ballistic Limit Test

The most important result of IDA's simulation study was that, regardless of the test design method used to select shot velocities, using model fitting and maximum likelihood estimation along with the associated criteria for stopping the test resulted in a more efficient test than the up-down method. Figure 5 shows a probit model fit to the example test data; this is an example of an empirical model fit that uses maximum likelihood estimation to assess the probability of penetration as a function of velocity. The measured ballistic limit is an estimate of the true ballistic limit, but also includes error due to variability in the helmet's performance near the ballistic limit. By using maximum likelihood estimation, the ballistic limit can be estimated with fewer shots on average without increasing either the bias (the difference between the measured and the true value) or the variance in the estimate.

Misestimating helmet performance prior to testing can lead to poor choices of fragment velocity during testing, which can increase both the dispersion of and the bias in the ballistic limit estimate. Our simulation study demonstrated that test designs that use generalized linear modeling (i.e., three-phase optimal design (3POD) and Neyer's Method) to select the shot velocities are less sensitive to initial misestimates of helmet ballistic limit than the up-down method. Figure 6 shows the median bias and interquartile range (25th and 75th percentiles shown as the lines extending from each marker) of the ballistic limit estimates for each method for a range of initial misestimates in the variance; the most desirable result is a bias of zero with a narrow interquartile range. Note that the starting assumptions about helmet performance were intentionally misestimated to show the robustness of each method to having limited knowledge of the actual performance variability around the ballistic limit for the helmet design under test.

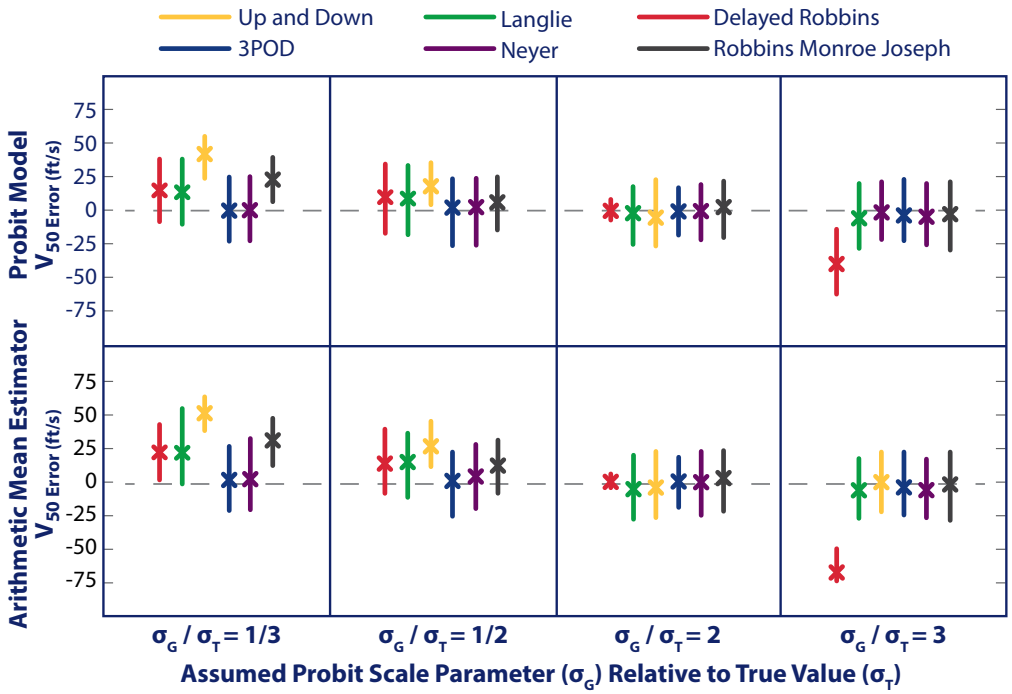


Figure 6. Simulation Results Comparing the Maximum Likelihood Estimate from the Probit Model and the Arithmetic Mean Estimator for Various Test Strategies

CONCLUSION

The Services will continue to pursue lighter helmets and improved ballistic performance. The application of statistically principled test designs will help ensure that new combat helmets have acceptable ballistic performance. IDA has developed innovative design methods that have improved helmet testing protocols for resistance to penetration, while balancing risks to both government and manufacturer across multiple conditions. IDA's research on ballistic limit design and analysis methods shows that further improvements can be made to existing protocols. Making these improvements will ultimately provide a better understanding of helmet ballistic performance, resulting in better equipment for our soldiers.

Dr. Hester is an Adjunct Research Staff member in IDA's Operational Evaluation Division. She holds a Doctor of Philosophy in physics from Princeton University.

Dr. Johnson is a Research Staff member in IDA's Operational Evaluation Division. He holds a Doctor of Philosophy in aerospace engineering from Old Dominion University.

Dr. Freeman is an Assistant Director in IDA's Operational Evaluation Division. She holds a Doctor of Philosophy in statistics from the Virginia Polytechnic Institute and State University.

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

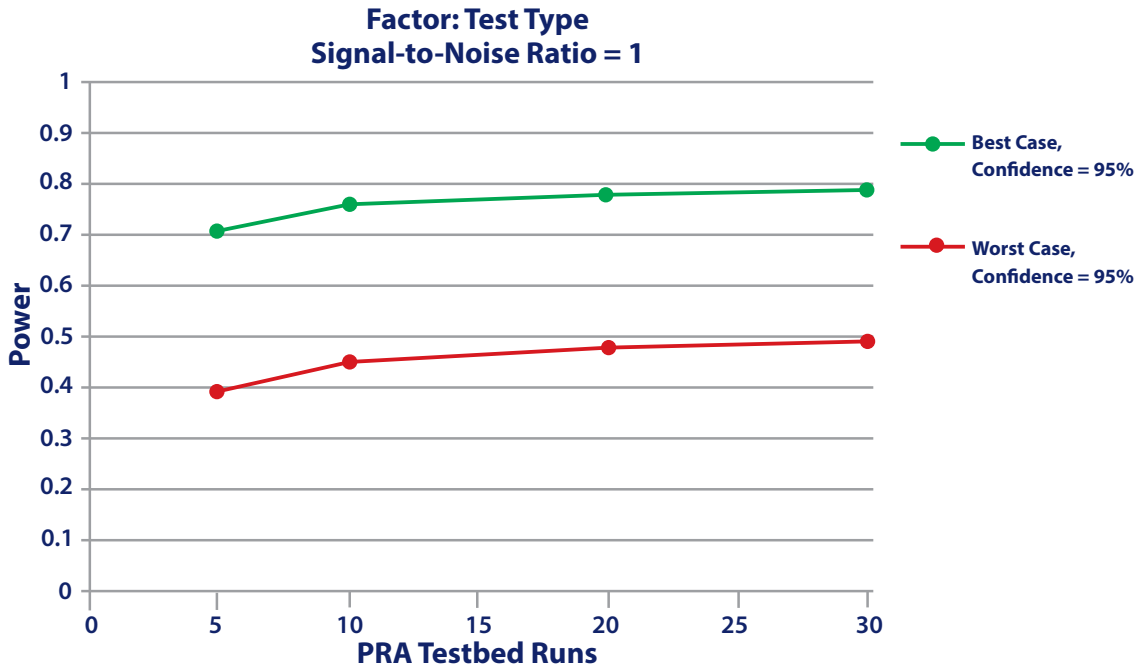


Figure 1. The power curve, assuming a confidence level of 95 percent and using a signal-to-noise ratio of 1, for the factor Test Type and the response initial detection range. The best-case scenario assumes detection ranges within a single event are completely independent; the worst-case scenario assumes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

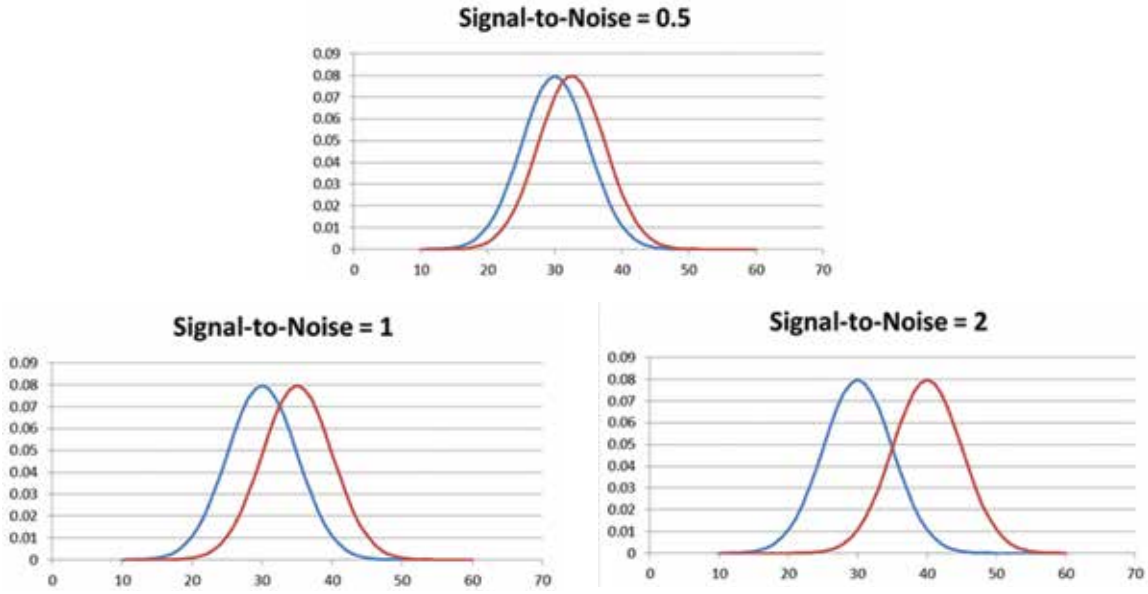


Figure 2. The separations between model and live test notional initial detection range distributions for signal-to-noise ratios of 0.5, 1, and 2.

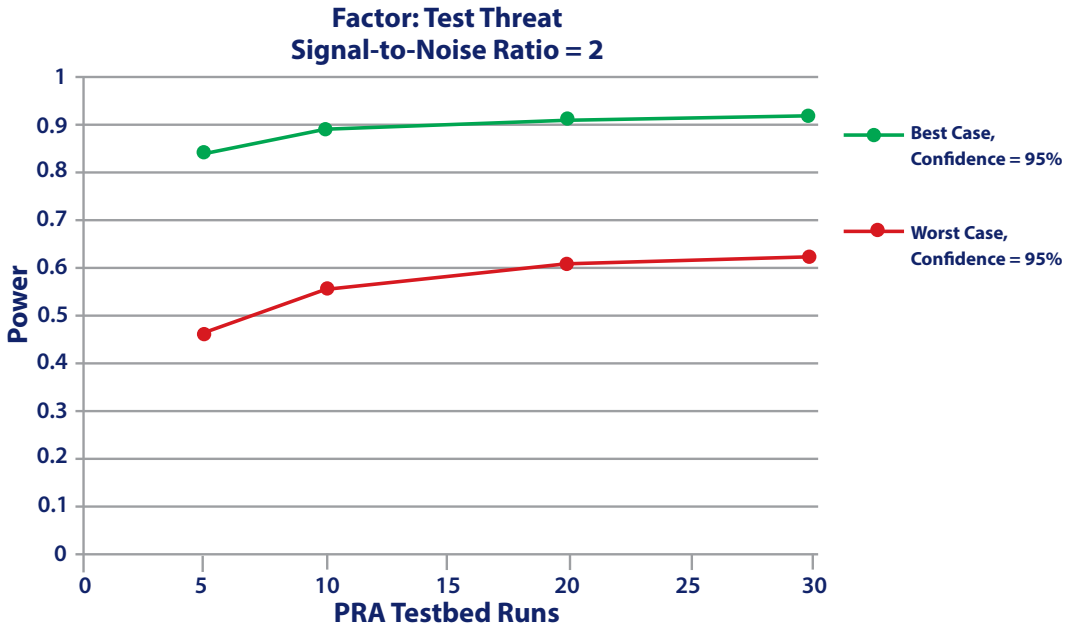


Figure 3. The power curve, assuming a confidence level of 95 percent and using a signal-to-noise ratio of 2, for the factor Test Threat and the response initial detection range. The best-case scenario assumes that detection ranges within a single event are completely independent; the worst-case scenario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

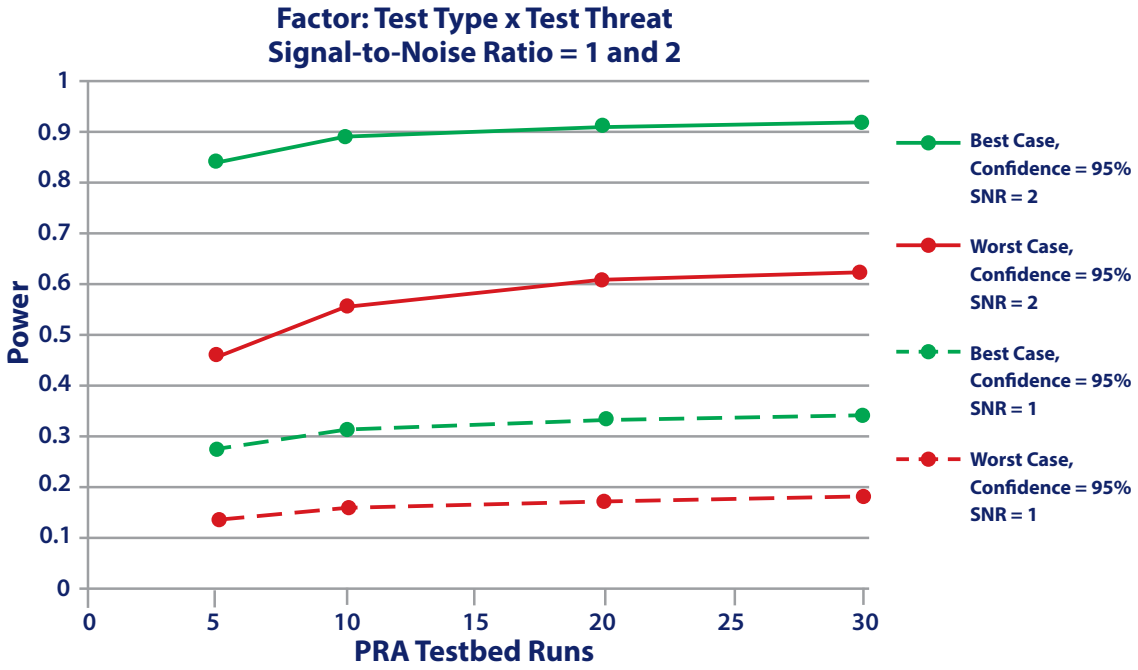


Figure 4. The power curve, assuming a confidence level of 95 percent and using the signal-to-noise ratios 1 and 2, for the interaction factor Test Type x Test Threat and the response initial detection range.

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.

Past Issues

Technological Innovation for National Security

- Acquisition in a Global Technology Environment
- Lessons on Defense Research and Development Management
- Commercial Industry Research and Development Best Practices
- Strengthening Department of Defense Laboratories
- Policies of Federal Security Laboratories
- The Civilian Science and Engineering Workforce in Defense Laboratories
- Technology Transfer: Practices from the Department of Defense

Acquisition

- Defining Acquisition Trade Space Through "DERIVE"
- Supporting Acquisition Decisions in Air Mobility
- Assessing Reliability with Limited Flight Testing
- Promise, Reality, and Limitations of Software Defined Radios
- Implications of Contractor Working Capital on Contract Pricing and Financing
- The Mechanisms and Value of Competition
- Early Management of Acquisition Programs

Security in Africa

- Trends in Africa Provide Reasons for Optimism
- China's Soft Power Strategy in Africa
- Sudan on a Precipice
- A New Threat: Radicalized Somali-American Youth
- Chinese Arms Sales to Africa
- Potential of Engagement Networks in Africa
- Defense Environmental Cooperation with South Africa

Challenges in Cyberspace

- Cyberspace – The Fifth and Dominant Operational Domain
- Transitioning to Secure Web-Based Standards
- Information Assurance Assessments for Fielded Systems During Combatant Command Exercises
- Supplier-Supply Chain Risk Management
- Internet-Derived Targeting: Trends and Technology Forecasting
- Training the DoD Cybersecurity Workforce

Today's Security Challenges

- A Framework for Irregular Warfare Capabilities
- Bridging the Interagency Gap for Stability Operations
- Developing an Adaptability Training Strategy
- Force Sizing for Stability Operations
- Planning Forces for Steady State Foreign Internal Defense and Counterinsurgency
- Test and Evaluation for Rapid-Fielding Programs
- Understanding Security Threats in East Africa
- Supporting Warfighting Commands
- Detecting Improvised Explosive Devices
- Building Partner Capacity
- Combating the Trans-South Atlantic Drug Trade
- Countering Transnational Criminal Insurgents
- Using Economic and Financial Leverage
- Understanding the Conflict in Sudan

Resource Analyses

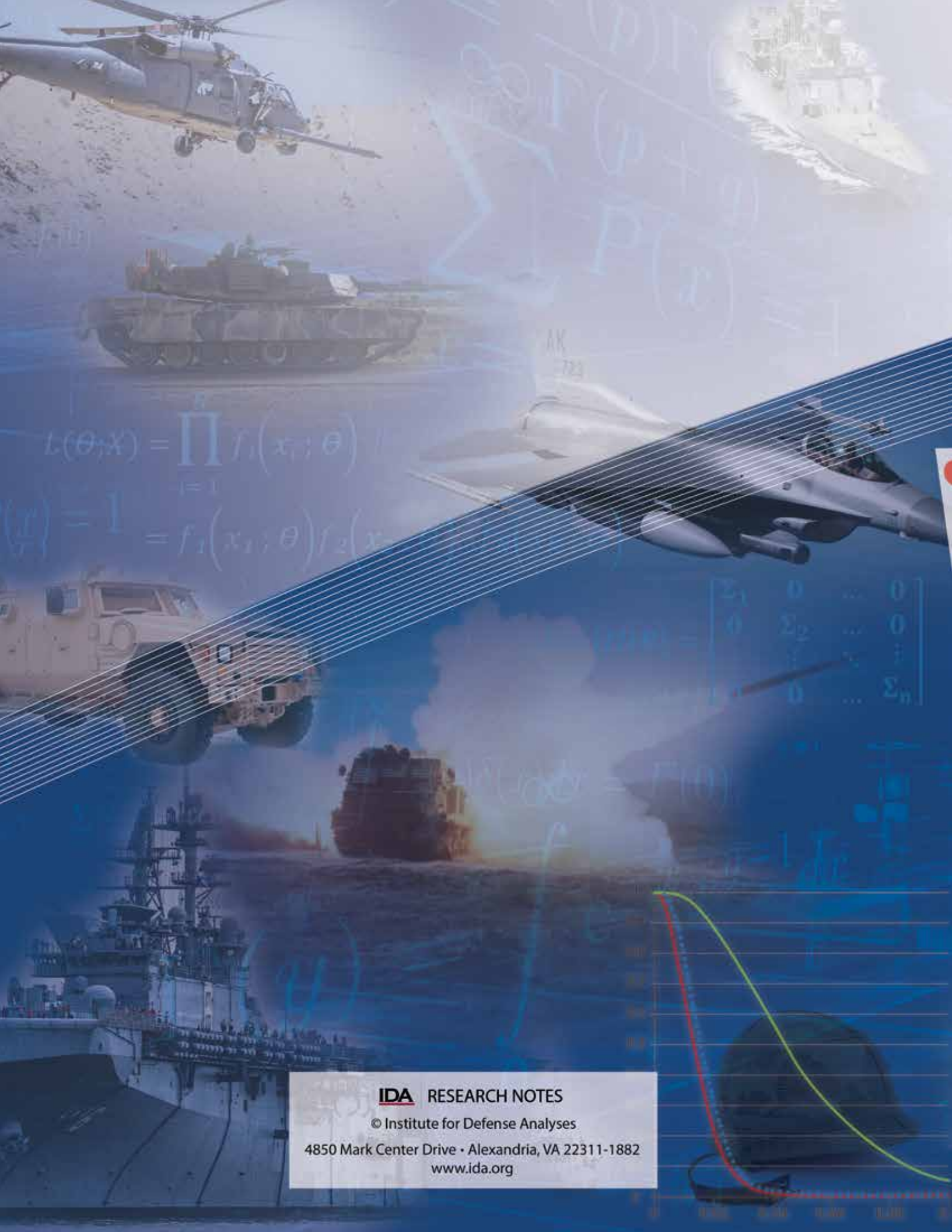
- Evaluating the Costs and Benefits of Competition for Joint Strike Fighter Engines
- Analysis and Forecasts of TRICARE Costs
- Cost Savings from the Post-Cold War Consolidation of the Defense Industrial Base
- Effects of Reserve Mobilization on Employers
- Does DoD Profit Policy Motivate Contractors?
- Auctions in Military Compensation

Focusing on the Asia-Pacific Region

- Making Security Partners Better Resource Managers
- Collaborating with Singapore
- Intellectual Outreach to the Muslim World
- Inside North Korea
- Red Teaming for Terminal Fury
- Promoting Interagency Cooperation in Shaping U.S.-China Relations
- Extending Trilateral Cooperation for Disasters
- Developing Human Capital in China

Homeland Security

- Port Vulnerability
- Assessing the EMP Threat
- Homeland Defense Scenarios
- Transport and Dispersion Models
- IT Security



IDA RESEARCH NOTES

4850 Mark Center Drive • Alexandria, VA 22311-1882
www.ida.org