CHANCE

ISSN: 0933-2480 (Print) 1867-2280 (Online) Journal homepage: www.tandfonline.com/journals/ucha20

Informing the Warfighter—Why Statistical Methods Matter in Defense Testing

Laura J. Freeman & Catherine Warner

To cite this article: Laura J. Freeman & Catherine Warner (2018) Informing the Warfighter—Why Statistical Methods Matter in Defense Testing, CHANCE, 31:2, 4-11, DOI: 10.1080/09332480.2018.1467627

To link to this article: https://doi.org/10.1080/09332480.2018.1467627

Published online: 27 Apr 2018.

Informing the Warfighter—Why Statistical Methods Matter in Defense Testing

Laura J. Freeman and Catherine Warner

The Department of Defense (DoD) acquires some of the world's most-complex systems. These systems push the limits of existing technology and span a wide range of domains. They provide the technology that enables the U.S. military to conduct operations. The diversity of system types includes fighter aircraft, submarines, stealth bombers, ground vehicles, transport planes, radars, radar jammers, satellites, data management tools, and enterprise management systems, among many others. They can be repairable or single-use; many are software-intensive.

Data and statistics are key to assessing these systems. Each one must be evaluated to determine whether it will be effective in enabling military users to accomplish missions. Early in a program's development, testing evaluates the performance of system components and sub-systems. An example of sub-system testing involves engine reliability for a new aircraft. Later, engineering prototypes of full systems are used to evaluate system performance and reliability and ensure they can operate with other systems, and to understand cybersecurity vulnerabilities. For example, a new aircraft may be taken on a simple flight to establish the range at which its onboard sensors can see a threat in the distance.

This testing of sub-systems and engineering prototypes is called developmental testing. The goals of developmental testing include both improving system design and informing decisions about production readiness.

Operational testing is the next phase of testing. As required by law, this phase must be conducted before a system can either start full-rate production or be fielded to military users. Operational testing involves military operators using the representative production system to create realistic combat scenarios. The users complete missions in operationally realistic environments—they test the system in an environment as close as possible to what it would experience in real combat. For example, new tanks are tested in the desert under high temperatures and in the presence of a realistic

Figure 1. The diversity of defense systems (clockwise from top left): amphibious assault vehicle; SSN 774 Virginia Class submarine; MV-22 Osprey; F-22A advanced tactical fighter.
Images obtained from the DOT&E Annual Report.

opposing force; conditions similar to combat in the Middle East.

The two primary areas of evaluation from an operational test are effectiveness and suitability. Operational effectiveness considers whether a unit equipped with a system can accomplish missions. Operational suitability is the degree to which a system can be satisfactorily placed in field use.

Operational testing is fundamentally different from developmental testing. Consider the example of testing a new family car model. Developmental testing might consist of a series of tests that assess engine reliability, fuel economy, radio system connectivity, Bluetooth connectivity, crash safety, etc. Operational testing, on the other hand, evaluates whether the family can use this car to get to the grocery store, work, and the beach for vacation. Can the family get to all these places reliably? Can Dad use the navigation system correctly to get to a new doctor?

While testing a new family car in this fashion might seem like an expensive endeavor, dozens of examples demonstrate how the operational context of a test was a key element in identifying problems before fielding military systems.

Examples include:

—*The environment can degrade performance (sometimes more than expected)*. Testing an airborne mine detection system provides a good example. It was designed to be employed from a helicopter to detect, classify, and localize shallow moored mines and floating mines on the water's surface using a laser. The Navy performed developmental testing in clear water in the Gulf of Mexico near Panama City and performance was acceptable.

The Navy then selected a location with poor water clarity for the operational test—conditions similar to the Persian Gulf. As expected, performance was worse than in clear water; more importantly, performance degraded far more than expected. This proved that the pre-test predictions and the tactics guides were inaccurate. Without testing in operational conditions, these limitations would not have been discovered until the system was used in combat.

—*Numbers and configurations of users can degrade performance.* Problems with networked radio systems frequently arise due to both the scale of operational units (number of users) and operational environments (terrain), which can limit line of sight between radios. Testing with operationally realistic units who are conducting actual missions not only helps to reveal these problems, but also provides the ability to measure their impact on the mission outcome.

Another example is an aircraft tactical radio system. In this case, operational scenarios with multiple aircraft were necessary to discover problems, with multiple participants achieving a communication connection that was synchronized in real time. The operational test used a varied formation size. The problem occurred stochastically and as a function of the number of aircraft linking together. Previous testing of the link with only two aircraft did not reveal the problem; increasing the number of aircraft to operationally realistic configurations (four aircraft per mission) found that the issue occurred frequently.

—The operational environment induces reliability failure modes that previously were unknown. In the case of a Miniature Air Launch Decoy (MALD), storing an expendable air-launched decoy in operational conditions—rainstorms in Guam—resulted in water getting into the fuel bladder, which ultimately resulted in the MALD failing to deploy. The discovery of a previously unknown failure mode was the direct result of putting the system into operational conditions not previously captured in laboratory testing.

Implementing Statistical Design and Analysis in Operational Testing

In recent years, DoD guidance has directed that a statistical approach be taken when testing and evaluating DoD systems in operational testing. Specifically, the initial guidance focused on using design of experiments (DoE) for test planning (Guidance on the use of DoE in OT&E, 2010).

This is a relatively new approach to designing operational tests. Common test design approaches used in the past included specialized or singular combat scenarios, changing one test condition at a time, replicating a single condition specified by requirements, conducting case studies, and avoiding control over test conditions (termed "free play").

While these historical approaches produced tests that were operationally realistic, they lacked the scientific process needed to ensure that testing was both efficient and able to characterize system performance in a diverse range of conditions. For complex defense systems, performance often depends on interactions between independent variables. Statisticians know that historical test strategies typically are inadequate to support the estimation of second-order or higher interactions.

In 1998, the National Research Council reviewed test strategies of the time and concluded, "Current practices in defense testing and evaluation do not take full advantage of the benefits available from the use of state-of-the-art statistical methodology," and that "[s]tate-of-the-art methods for experimental design should be routinely used in designing operational tests."

However, it was not immediately obvious how DoE would apply to operational testing. A unique aspect of operational testing, especially when considered in combination with the application of DoE, is that there often are many uncontrollable variables. Operational users—human beings—introduce variability; the operational environment introduces variability—such as changes in the weather; and the evolving context of the mission introduces variability.

Another aspect of operational testing worth highlighting is the criticality of human-systems interactions and their impact on mission accomplishment. Because hardware and software cannot accomplish missions alone, operators are a critical component of military systems. Systems that are overly complex introduce failures and force the services to invest in lengthy training programs to mitigate problems that arise because of poor interface design.

To address the challenges of applying DoE to operational testing, DOT&E and IDA developed numerous case studies to illustrate the benefits. These case studies have shown that statistical methodologies are essential to constructing defensible test programs. In spite of the diversity of systems types, statistical tools are universally applicable to all systems in both developmental and operational testing.

Design of experiments provides the tools to span this complex area with a defensible approach. It is often a challenge to cover multiple missions and balance them with limited test resources. Statistical power analysis and other tools of assessing test adequacy have provided methods for ensuring that expensive tests will provide the information needed. Statistical models provide the tools to characterize outcomes across complex operational spaces and allow the data to inform that characterization.

Three examples show how statistical methods have provided:

1. Defensible rationales for test adequacy.
2. Efficient test plans.

3. The ability to characterize capability throughout operational conditions.

Example 1: Defensible Rationale for Test Adequacy—Long-range Anti-Ship Missile (LRASM)

A credible rationale for test adequacy is always important. Defense testing is expensive. Flying aircraft, firing missiles, and reserving space on test ranges all cost money. More than just the material costs, however, it also takes time and coordination to execute an operationally realistic test.

A heavy logistical burden is also associated with collecting every data point. Operational tests require the simultaneous coordination of test ranges, military service members who operate and maintain the systems, and the systems themselves. Therefore, once these resources come together, it is important to be sure that a test captures enough information to inform decision-makers on whether to acquire and field the system, while not over-testing. Researchers have to find the sweet spot of testing enough to inform the decision-maker while not spending any more resources than necessary, and must be able to find that spot consistently in hundreds of different systems.

Experimental design and statistical power calculations have proved to be essential tools for the DoD test community. In many cases, resources are extremely limited; using experimental design makes it possible to maximize the impact of those resources.

LRASM is a long-range, precision-guided anti-ship missile. It uses multiple sensors to find a target, a data link to communicate with the aircraft that launched the missile, and an enhanced Global Positioning System to detect and destroy specific targets within a group of numerous ships at sea.

As can be imagined, each of these technologies is expensive and only a limited number of LRASMs will be produced. Missiles used for testing are not available for fielded use, making it important to only use the minimal number of missiles in testing to obtain important information.

A key operational contribution of LRASM is to be able to target a specific ship out of many at sea. To test that capability thoroughly requires multiple ships at sea, a range in which it is safe to launch missiles, and a launch platform for the weapon. This is no small feat to coordinate.

Due to the inherent costs and limitations of testing all weapons, modeling and simulation (M&S) often are used to fill gaps in knowledge from live testing. Sanchez (2017) discusses different design approaches for M&S in her article in this issue of *CHANCE*. It is also important here because the operational shots have a dual purpose: getting enough live data to both identify problems in operations and support the validation of the modeling and simulation.

The Navy was considering reducing the number of weapons for testing by half. Using experimental design techniques, the test team was able to clearly show that the proposed reduction excluded important aspects of the operational engagements that looked at different target ranges and aspect angles, which could affect success rates.

The reduction in shots could be mitigated by using M&S in those ranges and aspect angles, but doing so requires that the M&S be validated for operational evaluation. Specifically, the usefulness and limitations of the M&S should be characterized, and uncertainty should be quantified to the extent possible by comparing the M&S to live data.

Therefore, the test team conducted a statistical power calculation comparing free-flight data and M&S outcomes using a Fisher's combined probability test. Their analysis showed that the planned free-flight program provided enough information to validate the M&S at an acceptable level of risk, but cutting the shots in half would not provide adequate information (adequate statistical power) to detect differences between free-flight testing and the M&S. Reducing the number of shots risked mischaracterizing the performance of the weapon in the operational test space by using an inadequately validated model.

Through statistical analysis techniques, the test team was able to clearly discuss the trade-space between the two test designs and determine that the test team should not reduce the existing design because it provided a minimally adequate test for assessing weapon performance and validating the M&S.

Example 2: Efficient Test Plans that Cover the Operational Envelope—F-35 Joint Strike Fighter

Defense systems are often designed to be used in multiple missions, and each of those missions may cover a complex operating space. Being able to cover the full operating environment efficiently is a core challenge of planning a defensible operational test. Gauging the right amount of testing involves more than simply determining the number of test points; equally important is the placement of those points across the region of operability.

Placement of the points is the most-important aspect of determining whether the testing will be adequate to support the goals of

Figure 2. F-35 C in flight.
Source: DOT&E Annual Report.

the analysis. The goal of collecting the right data in the right locations is to understand where systems work and to what extent.

The F-35 is a multi-role fighter aircraft being produced in three variants for the Air Force, Marine Corps, and Navy. Its multi-role nature covers many diverse missions, including air-to-surface attack, aerial reconnaissance, close air support, offensive counter air, defensive counter air, destruction and suppression of enemy air defenses, anti-surface warfare, cruise missile defense, and combat search and rescue.

The three aircraft variants, a range of potential ground and air threats, various weapon loads, the need to test during both day and night, the movement of potential targets, and information quality provided to the aircraft all further compound this complexity. Arguably, it is the most-complex mission and environment that ever has been considered in a single operational test.

The philosophies and methods inherent to experimental design provided the necessary framework for covering such a complex operational space defensibly. The overarching test approach for the F-35 initial operational test, which remains to be conducted, was to create detailed test designs for evaluating each of the core mission areas by defining appropriate, measurable response variables that correspond to operational effectiveness of each mission area.

The test team divided the operational space—using DoE concepts—into factors that would affect the response variables, such as type of ground threat or number and types of air threat, and varied those factors to ensure coverage of where the F-35 may be used in combat.

The test team also used the principles behind DoE to span the operational space efficiently. For example, they blocked out the design by dividing the threat continuum into categories and then correlated the threat coverage blocks with appropriate mission areas. They ensured coverage of key capabilities by focusing each capability assessment in the most-relevant mission area.

Finding, tracking, and engaging moving ground targets are only covered in two of the mission areas, but the performance assessment of the radar in these two mission areas will enable developing inferences across all of the mission areas.

Experimental design enabled the test team to adequately cover nine core mission areas, multiple operational capabilities, and multiple factors within each mission area in a combined total of 110 trials. While a very large test, this will provide information on F-35 capabilities and how those capabilities translate into operational outcomes.

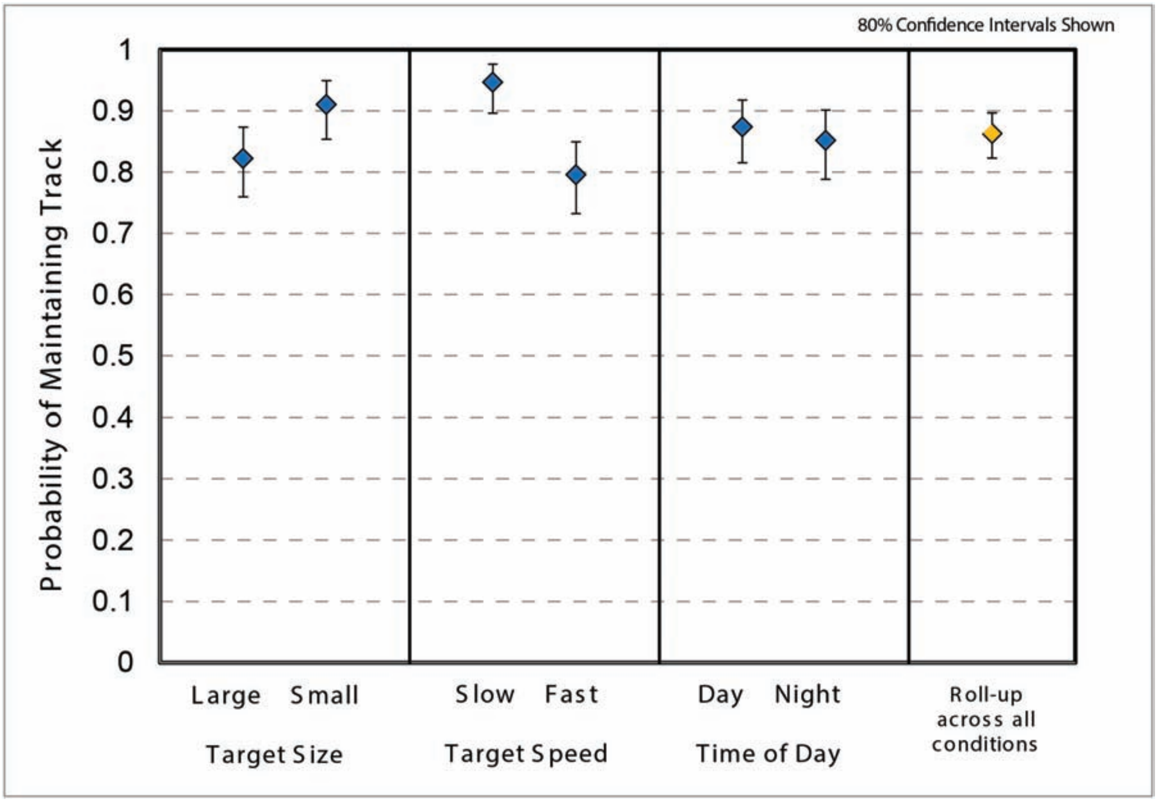

Figure 3. A historical analysis approach: calculating average performance across each condition, or a global average.

By treating the variant of the F-35 as a factor (independent variable) in the test design, the testers also were able to leverage relevant performance and mission level data across variants, resulting in a reduction in required sorties when compared to previous test designs on legacy platforms.

Example 3: Statistical Analysis to Characterize Capability in the Operational Envelope—Tracking a Moving Target

It is imperative not only to cover the operational space in testing, but also to analyze the resulting data appropriately to understand where systems work, to what extent, and how much precision there is in the conclusions. After conducting the test, statistical analysis methods provide a defensible data analysis approach.

These empirical models allow for objective conclusions based on observed data. Parametric regression methods allow maximizing information gained from test data, while non-parametric methods can provide a robust assessment of the data that is free from model assumptions. Bayesian methods provide avenues for integrating additional sources of information.

The following example is for a system whose purpose is to maintain a lock on a moving target. If the system can maintain the track for the desired period of time, the test trial is scored as a success; if the system drops the track at any point, then the test trial is scored as a failure. The purpose of the test was to characterize the probability of maintaining track across all the operating conditions.

The factors that drive the probability that the system is able to successfully maintain track include:

- Time of day (day/night)
- Target size (small/large)
- Target speed (slow/fast)

Figure 3 shows a traditional historical analysis in which average proportions were calculated for all conditions (far right, roll-up) and for common conditions (e.g., all trials against large targets). The selected variables and levels for binning often are designated by test team expertise.

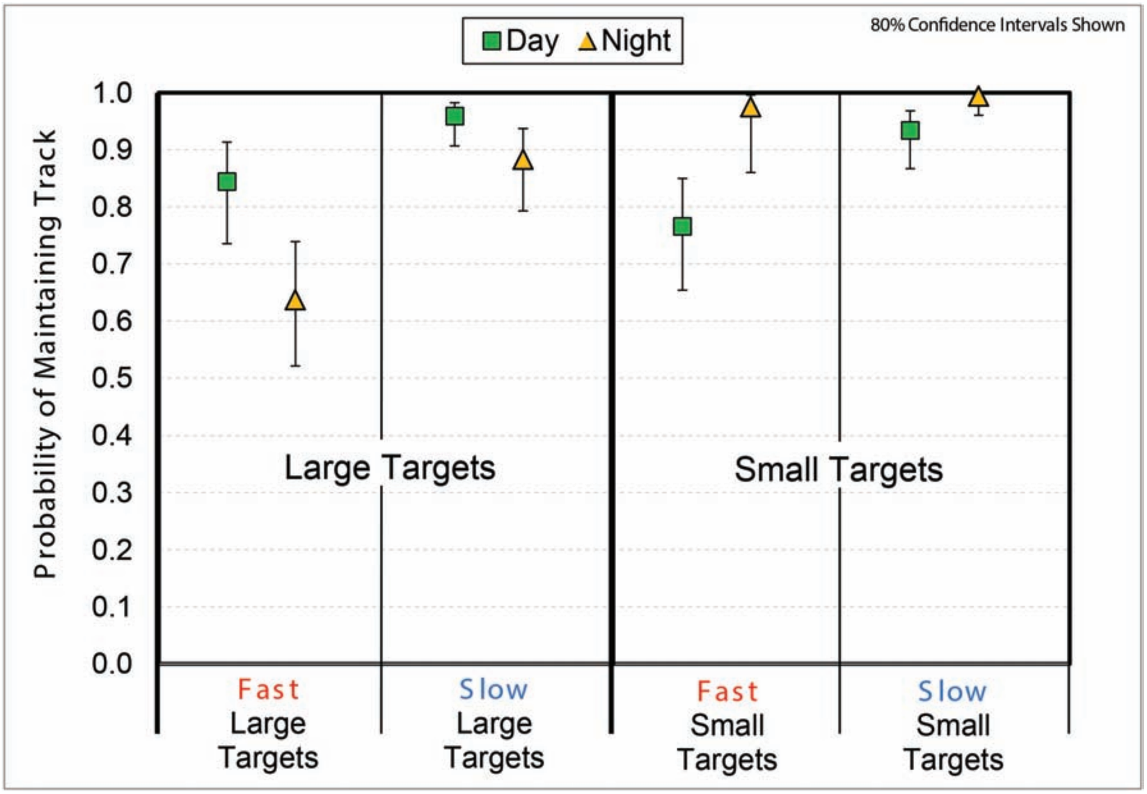

Figure 4. A powerful analysis approach: logistic regression enables a full characterization of system performance.

In Figure 4, the predicted probabilities of maintaining a track and corresponding confidence intervals are the output of a logistic regression analysis. Using logistic regression and model selection techniques lets testers distill the most-important results from the data. In the analysis of the tracking system, the logistic regression revealed a significant interaction between time of day and target size. This interaction results in poor performance in a specific set of conditions: large, fast targets at night.

It is clear from this analysis that the overall roll-up calculation provides little insight into system performance in operations. However, these global averages (and sometimes medians) are all too frequently used in defense analyses. It is important to highlight examples like these because they showcase why statistical analysis matters. Clearly, the traditional analysis of taking simple averages fails to identify an important performance degradation that occurs against large, fast-moving targets at night.

The statistical analysis also can provide insights for system development and future testing of the system. For example, developers might target future system changes to ensure that large, fast targets at night can be tracked equally as well as others. This also would be a focus of follow-up testing.

Advanced Analysis Methods to Meet Unique T&E Challenges

Balancing complex systems, environments, missions, and the need to account for human factors with the structured use of experimental design and statistical analysis results in many interesting research questions and advanced statistical analysis challenges. Statistical analyses that extend beyond the straightforward regression/linear model analysis are slowly gaining traction in the DoD.

Examples include:

- Bayesian analysis methods (especially in a reliability context) allow leveraging information from multiple phases of testing while ensuring the results still reflect the operational reliability.
- Survival analysis and mixture distributions for performance measures such as *detection range and time to detect* allow incorporating information from continuous measures in cases where traditional pass/fail metrics (e.g., probability of detect) would have been the only measures previously considered.
- Generalized linear models and mixed models allow flexible analysis methodologies that truly reflect the character of the data.

However, these analysis methods require more than an introductory level understanding of statistics. The DoD's testing professionals need the assistance of the statistical community in making these tools more accessible and recruiting individuals with strong statistical backgrounds into defense positions.

It is highly important for the statistical community to engage with defense testers to develop statistical analysis methods that meet the unique challenges of the operational environment.

As our systems become even more complex and leverage levels of autonomy, continuous and integrated testing will be necessary. The application of state-of-the-art statistical methodologies must continue to evolve to confront these new challenges. Seeing the importance of statistical methods in defense testing, elements of the DoD have started investing in statistical research through the Science of Test Research Consortium, which includes the Air Force Institute of Technology, Naval Postgraduate School, Arizona State University, Virginia Tech, North Carolina State University, Florida State University, University of Arkansas, and Rochester Institute of Technology.

We also have started a Statistical Engineering Collaboration with NASA to share best practices across organizations. We have designed a workshop to strengthen the community in statistical approaches to testing, evaluation, and modeling and simulation in defense and aerospace. The workshop also seeks to link practitioners and statisticians in a bi-directional exchange. The information exchange consists of practitioners providing challenging and interesting problems and statisticians providing training, consultation, and access to new ideas.

Finally, we are working to make statistics more accessible and specifically targeted to the DoD test community through an educational website (testscience.org).

We encourage statisticians to consider careers in defense. There are opportunities for statisticians in defense test agencies, working in all phases of testing. The challenges are complex and we need statistical thinking to help us solve them. C

Further Reading

DOT&E Website with Guidance Memos: *http://www.dote.osd.mil/guidance.html*.

Johnson, R.T., Hutto, G.T., Simpson, J.R., and Montgomery, D.C. 2012. Designed experiments for the defense community. *Quality Engineering 24*(1), 60–79.

Montgomery, D.C. 2008. *Design and Analysis of Experiments*. John Wiley & Sons.

National Research Council. 1998. *Statistics, Testing, and Defense Acquisition: New Approaches and Methodological Improvements*. Washington, DC.

Sanchez, S.M. 2017. Data Farming: Reaping Insights from Simulation Experiments. *CHANCE*, *31*(2).

Test Science Website: *http://testscience.org/*.

About the Authors

Laura Freeman is an assistant director of the Operational Evaluation Division at the Institute for Defense Analyses. She serves as the primary statistical advisor to the director of Operational Test and Evaluation (DOT&E) and leads the test science task, which is dedicated to expanding the use of statistical methods in the DoD test community. Freeman has a BS in aerospace engineering, MS in statistics, and PhD in statistics, all from Virginia Tech. Her PhD research was on design and analysis of experiments for reliability data.

Catherine Warner is director of the Centre for Maritime Research and Experimentation in the NATO Science and Technology Office. Previously she served as science advisor for the Director, Operational Test and Evaluation (DOT&E). She has been involved with operational testing and evaluation since 1991, when she became a research staff member at the Institute for Defense Analyses. Warner previously worked at the Lawrence Livermore National Laboratory. She earned both BS and MS degrees in chemistry from the University of New Mexico and San Jose State University, and both MA and PhD degrees in chemistry from Princeton University.