INSTITUTE FOR DEFENSE ANALYSES

Testing Defense Systems

Laura J. Freeman, *Project Leader*
Thomas H. Johnson
Matthew R. Avery
Justace R. Clutter

1 June 2017

IDA NS D-8551

Log: H 2017-000346

INSTITUTE FOR DEFENSE ANALYSES
4850 Mark Center Drive
Alexandria, Virginia 22311-1882

The Institute for Defense Analyses is a non-profit corporation that operates three federally funded research and development centers to provide objective analyses of national security issues, particularly those requiring scientific and technical expertise, and conduct related research on other national challenges.

About This Publication

Defense systems are complex and contain a variety of software and hardware components. They consistently push the limits of scientific understanding, synergizing new and unique interfaces between software and hardware. They range from the traditional military system, such as fighter aircraft, to systems less likely to be associated with defense, like business and information technology systems. The complex, multi-functional nature of defense systems, along with the wide variety of system types, demands a structured but flexible analytical process for testing systems. Additionally, rigorous testing must ensure that representative users can effectively operate the system in a variety of environments and mission scenarios. This chapter highlights the core statistical methodologies that have proven useful in testing defense systems. Case studies illustrate the value of using statistical techniques in the design of tests and analysis of the resulting data.

For more information:
Laura J. Freeman, Project Leader
lfreeman@ida.org • (703) 845-2084

Robert R. Soule, Director, Operational Evaluation Division
rsoule@ida.org • (703) 845-2462

INSTITUTE FOR DEFENSE ANALYSES

IDA NS D-8551

Testing Defense Systems

Laura J. Freeman, *Project Leader*
Thomas H. Johnson
Matthew R. Avery
Justace R. Clutter

X

TESTING DEFENSE SYSTEMS

Laura J. Freeman, Thomas Johnson, Mathew Avery, V. Bram Lillard, Justace Clutter
Institute for Defense Analyses

Synopsis

Defense systems are complex and contain a variety of software and hardware components. They consistently push the limits of scientific understanding, synergizing new and unique interfaces between software and hardware. They range from the traditional military system, such as fighter aircraft, to systems less likely to be associated with defense, like business and information technology systems. The complex, multi-functional nature of defense systems, along with the wide variety of system types, demands a structured but flexible analytical process for testing systems. Additionally, rigorous testing must ensure that representative users can effectively operate the system in a variety of environments and mission scenarios. This chapter highlights the core statistical methodologies that have proven useful in testing defense systems. Case studies illustrate the value of using statistical techniques in the design of tests and analysis of the resulting data.

X.1 Introduction to Defense System Testing

Testing and the evaluation of test outcomes is a critical aspect of the acquisition process for all defense systems. The types of systems that undergo testing in the United States Department of Defense (DoD) are extremely diverse and include systems such as submarines, aircraft carriers, fighter aircraft, cargo aircraft, radar, radar jammers, transport trucks, armored vehicles, chemical agent detectors, payroll systems, medical information systems, and identification cards. The test and evaluation process provides valuable information throughout the acquisition process.

Testing is an essential aspect of the systems engineering process. From system design to the consumer's decisions to purchase the product, testing and the resulting analyses can provide insight and inform decisions. Conducting tests and evaluating the results are essential elements of developing quality products that deliver the performance required by the end user.

A scientific approach to testing provides a structured, defensible process that ensures the right amount of testing is conducted to answer key questions. A scientifically planned test requires the input and expertise of all stakeholders including project management, engineering and scientific expertise, and statistical expertise.

X.1.1 Operational Testing Overview

Defense system testing is commonly divided into phases: contractor testing, developmental testing, live fire testing, and operational testing.

During contractor testing, the contractor developing the system's design tests to compare design prototypes, characterize critical component capabilities, verify the system can meet requirements, among dozens of other objectives.

Developmental testing evaluates system capabilities using prototypes of critical sub-systems and prototypes of full systems. Recent efforts in developmental testing within the DoD have focused on characterizing system performance, assessing reliability, ensuring interoperability with other systems, and understanding cybersecurity vulnerabilities. Developmental testing can be conducted at the component or sub-system level, or by using the full system. Developmental testing can occur in the laboratory, at test ranges, or under operational conditions.

Live fire testing evaluates the vulnerability of the systems to potential threats and, in the case of weapon systems, the lethality of the system against potential targets. Because of the cost and safety limitations associated with detonating live weapons, live fire testing typically has limited sample sizes and relies heavily on design analyses, modeling and simulation, and existing data from combat.

In this chapter we focus on statistical design methods for operational testing. Operational testing is the final test event in the DoD acquisition process and acts as a form of acceptance testing by the government. It is required by law before a program can proceed to full rate production or the system can be fielded.

Operational testing is conducted with operationally representative users on production representative systems. Operators and maintainers use the system in an operationally realistic environment to conduct operational missions. The two primary areas of evaluation from an operational test are operational effectiveness and suitability. Operational effectiveness captures the overall degree of mission accomplishment or success of a unit equipped with a system. Operational suitability is the degree to which a system can be satisfactorily used in the field, with consideration given to: availability, compatibility, transportability, interoperability, reliability, maintainability, safety, human factors, wartime usage rates, manpower supportability, logistics supportability, documentation, and training requirements.

For those unfamiliar with the DoD Acquisition System, the 1998 National Research Council publication, "Statistics, Testing, and Defense Acquisition: New Approaches and Methodological Improvements" provides a detailed overview of the process, useful terminology, and a more detailed history. In this chapter, we focus on the unique statistical challenges of designing operational tests, many of which can be attributed to the process, but some of which are inherent to the complexity of the systems and the missions system operators must complete.

X.1.2 Unique Statistical Challenges of Operational Testing

The statistical design and analyses techniques used in operational tests must be diverse because DoD systems are purchased and employed to address a diverse set of missions. It would be impossible to cover all possible methods used to test complex systems in one chapter of a textbook, especially considering the plethora of test objectives, including analyses of mission capability, performance, reliability, maintainability, human factors, logistics, etc. However, there

are common themes that emerge across defense testing. This chapter focuses on those common themes and their statistical implications.

Multiple-dimensional evaluation

The analysis of complex systems requires the measurement of many different types of variables. System performance is important, but how users interact with a system is also critical in determining whether units can use a system to complete operational missions. Reliability and maintainability measure the likelihood that a system will be available for use when called upon by operators. All of these variables, along with countless others, must be collected during testing. Additionally, designing a test that focuses solely on system performance might miss the larger context of the mission. Therefore, it is essential to think about the full operational context when designing tests.

Complex System/Complex Operational Space

Not only do test outcomes (dependent variables) span a multi-dimensional space, the inputs (independent variables) also cover a range of conditions. Frequently, defense systems are designed to be used in multiple missions and each of those missions might also cover a complex operating space. Being able to cover the full operating environment efficiently is a core challenge of planning a defensible operational test. As a result, this chapter focuses on the experimental design process for complex systems.

Testing is expensive

The space we wish to understand mission capability across is complex, and often timelines are short and test points are expensive. This typically limits sample sizes in an operational test to a few dozen missions. Additionally, the cost of testing in an operational environment requires that any proposed test be scrutinized to ensure that all data collected provide necessary and useful information in the most efficient manner. Gauging the right amount of testing is more than simply determining the number of test points; equally important is the placement of those points across the operational envelope. The placement of the points is the most important aspect of determining whether the testing will be adequate to support the goals of the analysis.

Humans are part of the system

Systems on their own do not accomplish missions, humans equipped with systems accomplish missions. This also means that we must consider human-system interactions in evaluating system effectiveness and suitability. Therefore, tracking those human interactions through both behavior metrics and surveys is a critical component of operational testing.

Focus on timeliness, accuracy

For many systems, variables that capture timeliness and accuracy are useful for assessing whether the system is performing as desired. These variables capture whether a new system allows operators to perform a task better and/or faster than a previous system. However, these variables often have left-skewed distributions. Statistical methods for analyzing this data must account for this appropriately.

Focus on prediction

Typically, it is interesting to know whether a factor considered in the test design (say conducting missions during the day versus at night) has an impact on performance, but a more

meaningful analysis for decisions makers and system operators informs them of how those factors translate into future performance. They need to understand how good or bad system performance might be based on the specifics of missions that they might encounter. Statistically, this means that hypothesis tests on model coefficients are not in themselves of interest. Instead, the focus is on translating model results to predictions of future mission performance with associated levels of precision.

In this chapter we summarize commonly used techniques in defense system testing and specific challenges imposed by the nature of defense system testing. Case studies illustrate the application of the methods.

X.1.3 The Statistical Science of Testing Complex Systems

As every system is different, the engineering and scientific expertise will vary based on the system or process being tested. However, statistics provides us with concrete methods and tools for generating the knowledge sought in testing. These statistical methods are often independent of the application, although the appropriate methods are influenced by the characteristics of the data. The same statistical techniques that are used to compare a generic drug to a brand-name counterpart can be used to compare the accuracy of a new guidance system in an air-to-ground missile to the original guidance system.

Design of Experiments (DOE) is a structured and purposeful methodology for test planning. In the complex operational mission space, it is often only feasible to test a limited number of points. DOE provides the method for selecting those points from among the many possibilities. Montgomery (2008) provides a comprehensive resource for planning experiments, though most defense system tests are more complex than the examples in Montgomery's text.

After the test is complete, there are many statistical methods to choose from for analyzing the resulting data. Empirical models produce objective conclusions based on the observed data. Parametric regression models maximize information gained from test data, while non-parametric methods provide a robust assessment of the data free from model assumptions. Bayesian methods provide avenues for integrating additional sources of information.

The analysis model should reflect the observed data and not the planning process. In designing tests we often assume a statistical model for the analysis. However, there is no limitation requiring the use of that model in the analysis. Often, qualities of the data observed (e.g., skewness, lurking variables, etc.) lead us to employ different analysis methods than originally planned; this is completely acceptable and can better inform test design on similar products and processes in the future.

X.2 Design of Experiments for Defense System Testing

X.2.1 DoD Historical Perspectives

Historically, the DoD has not always employed methods like DOE for test planning. Common test design approaches for operational testing in the past included specialized/singular combat scenarios, changing one test condition at a time, conducting case studies, and avoiding control over test conditions (termed "free play"). These approaches fail to ensure that testing is

both efficient and able to characterize operational mission capabilities across the full range of operating conditions. Moreover, for complex systems, performance often depends on interactions (commonly second order) of variables, and these historical test strategies are inadequate to support estimation of interactions.

In 1998, the National Research Council reviewed operational 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 "State-of-the-art methods for experimental design should be routinely used in designing operational tests."

The broad application of DOE to operational test and evaluation is a relatively new development. In 2009, the Service Operational Test Agencies (OTAs), in collaboration with the Director, Operational Test and Evaluation (DOT&E), endorsed the use of DOE methods in DoD testing. In 2010, DOT&E, outlined clear requirements for using DOE to plan operational tests. In 2012, the Deputy Assistant Secretary of Defense for Developmental Test and Evaluation (DASD DT&E) endorsed a scientific approach to testing by including DOE methods in the Scientific Test and Analysis Techniques (STAT) T&E Implementation Plan. Over the last few years, the testing community has developed best practices in applying the methods.

X.2.2 Overview

Experimental design is a defensible methodology for deciding what data should be collected to answer the experiment (test) objectives. Notably, the first chapter of the Montgomery (2008) focuses on the process. The essential elements of an experimental design process are:

(1) Identify the questions to be answered, also known as the goals or objectives of the test.
(2) Identify the quantitative metrics, also known as response variables or dependent variables that will be measured to address the key questions.
(3) Identify the factors that affect the response variables. Also known as independent variables, these factors frame the broad categories of test conditions that affect the outcome of the test. Identify the levels for each factor. The levels represent various subcategories between which analysts and engineers expect test outcomes to vary significantly. When performance is expected to vary linearly, two levels are used. Identifying nonlinear performance requires three or more levels.
(4) Identify applicable test design techniques. Examples include factorial designs, response surface methodology, and combinatorial designs. The applicable test design method depends on the question, metrics, types of factors (numeric or categorical), and available test resources. Identify which combinations of factors and levels will be addressed in each test period. If the test is to be a "one-shot" test with no follow-up planned, then a more robust test may be required. If testing can be sequential in nature, then smaller screening experiments aimed at determining the most important factors should precede more in-depth investigations. Determine how much testing is enough by using relevant statistical measures (e.g., power, prediction variance, correlation in the factor space, etc.).
(5) Conduct testing
(6) Analyze data
(7) Draw conclusions

Employing DOE in this holistic sense ensures efficient and adequate testing and also aids in determining how much testing is sufficient. As a result, we can determine the optimal allocation of constrained test resources and provide an analytical trade-space for test planning.

X.2.3 Statistical Test Objectives

A clear test objective is the first step in planning a defensible experimental design and is essential in deciding how much testing is enough to meet the stated objectives. Table X.1 summarizes common general classes of test goals, commonly associated test phases, and potentially useful experimental designs strategies for achieving the specified goal.

Experimental design is a rich scientific methodology, containing many design types. The specific tool that is employed depends on the question to be answered (step 1). The questions and objectives can change as the system under test matures. The choice of DOE technique (step 4) should reflect the objective. Table X.1 lists several common objectives and the corresponding designs one might select to satisfy the corresponding objective. This list is intended to show the breadth of tools that are included in DOE, but is far from exhaustive. The *NIST Engineering Statistics Handbook* discusses several goals of tests including prediction, characterization, and optimization.

Table X.1. General Classes of Test Goals

Test Objective	Potentially useful Experimental Designs
Characterize performance across an operational envelope and determine whether a system meets requirements across a variety of operational conditions	Response surface designs, optimal designs, factorial designs, fractional factorial designs
Compare two or more systems across a variety of conditions	Factorial or fractional factorial designs, matched pairs optimal designs
Screen for important factors driving performance	Factorial or fractional factorial designs
Test for problem cases that degrade system performance	Combinatorial designs, Orthogonal Arrays, Space filling designs
Optimize system performance with respect to a set of conditions and inform system design	Response surface designs, optimal designs
Predict performance, reliability, or material properties at use conditions	Response Surface Designs, Optimal Designs, Accelerated life tests

Improve system reliability or performance by determining robust system configurations	Response surface designs, Taguchi designs (Robust Parameter Designs), Orthogonal Arrays

Characterize

Operational testing should be adequate to characterize system capabilities and shortfalls across all relevant operating conditions. Such full characterization ensures that fielding decisions are made with a clear understanding of system performance, since it is not cost effective to field weapon systems that do not work or provide no clear improvement over existing systems. Full characterization also enables testers to inform the warfighters, whose lives may depend upon these systems, about what these systems can and cannot do. It is important to note that if we are able to characterize performance with sufficient precision across a variety of conditions, then we are also able to determine whether the system meets a specified requirement at a similar level of precision across those same conditions. Multiple classes of test designs may be useful when characterization is the primary test goal, including factorial and fractional-factorial designs, response surface designs, and optimal test designs. The appropriate test design will depend on the complexity of the operational envelope and expected performance variation across the operational envelope. Some conditions (levels of the factors) might be difficult to obtain, making some test designs more suitable (e.g., optimal over factorial). In most cases, covering arrays and combinatorial test designs are inappropriate for characterization because they provide low (or no) power for detecting differences in performance across the operational envelope. Power is an important measure when the test goal is characterization.

Compare

Direct comparison between two or more systems is a common operational test goal. A variety of test designs are useful for comparing multiple systems. The best comparisons can be made using a matched design where the systems (or processes) being comparing are subjected to the same tests across all conditions. This approach controls for unwanted variability in the comparison. While it is always important to consider human factors in designing tests, it is especially important in comparison designs. When using a within-subject design that has users perform the same tasks on both systems, one must be sure to control for order effects. When using a between-subject design, where the same users do not use both systems, ensure enough testing is conducted to have high power for comparing the systems. Power for detecting performance differences among systems is particularly important for comparison tests. Low power tests could result in an inability to draw conclusions about differences in performance between systems after the testing is completed.

Screen

Screening is an important test goal prior to operational testing. An important part of an integrated test process is early identification of key factors that affect system performance. Identifying these key factors and screening out unimportant factors is essential to constructing a defensible operational test. Factorial designs and fractional factorial designs are extremely useful tools in screening for important factors. When the number of factors and levels under consideration is large, optimal designs and highly fractionated factorials also can be useful.

Problem Cases

All operational testing seeks to identify problems in performance. However, test designs specifically geared for quickly finding problem cases are unique to situations where the outcome of the test is deterministic. These experimental designs are common for software testing and covered extensively in Chapter XX (Software Chapter of this book). Phadke (1995) provides an overview of orthogonal arrays for problem identification. Combinatorial tests based on orthogonal arrays provide an efficient methodology for covering use cases such that faults caused by certain combinations of factors (two-way, three-way, etc.) can be detected quickly. Kuhn and Kacker (2008 and 2009) provide additional detail about combinatorial testing. Because the system's performance and test outcome is not stochastic, statistical power is not meaningful. Rather, the strength of the design is defined by the number of factor combinations (two-way, three-way, etc.) covered. Since the outcome is deterministic, a fault will be caused by a specific combination of factors, so a test that includes as many combinations as possible in as few runs as possible is desired.

Testing for problems is not limited to determinist outcomes. Often, when one develops a test plan consisting of the most stressing cases, a goal of the test is to look for problems. While this testing approach may find problems, overall it is a risky design strategy because it fails to provide understanding of system performance across a range of conditions. It typically results in confounding between factors, and limited ability to determine the cause of failures.

Optimize

Process optimization is not a common test goal of operational testing. However, it is useful in system design and manufacturing. Additionally, at the system level it can be useful in the development of tactics, techniques, and procedures (TTPs). Myers, Montgomery, and Anderson-Cook (2009) is a comprehensive resource on response surface methodologies for process optimization.

Predict

Two general classes of prediction are interpolation and extrapolation. In operational testing we often wish to predict performance in areas within the design spaced tested. In these cases an experimental design that provides flexible modeling options and low prediction variance is preferred. Such designs include response surface designs and optimal designs. The other class of prediction is based on extrapolation. Extrapolation is inherently riskier than interpolation. For example, many weapon and sensor systems become less accurate when used against targets at long ranges than at short ranges, but the relationship is not necessarily linear. If test data is only collected using targets at short ranges, an analyst may be tempted to use the resulting model to estimate accuracy against long range targets as well. This approach could fail if, for example, accuracy degrades at an increasing rate as targets get further and further away. When the relationship between the factor being extrapolated and the response variable being model is not well understood, extrapolation should be avoided. In cases where those relationships are well understood, extrapolation may provide useful information that would otherwise be inaccessible. Accelerated life tests for predicting reliability at use conditions are a type of extrapolation. In general, models used in operational testing must rigorously be validated and accredited using live test data collected across the full range of operational conditions.

Improve

Improve (unlike optimize) refers to tests that are specifically designed to make processes or systems robust to uncontrollable conditions. These types of experiments are used in designing systems to ensure robust performance across all operating conditions. Additionally, these designs are useful in design for reliability efforts. In these experiments the tester controls both controllable factors (that is factors that can be controlled in the manufacturing process) and uncontrollable factors (often referred to as noise factors, e.g. humidity, operating conditions, etc.). The goal of the test is to determine the settings of the controllable factors that result in robust performance across all levels of the noise factors. This test goal is important, but typically arises during the manufacturing process, rather than in OT where characterizing system performance across all conditions is the priority. Taguchi designs (Robust Parameter Designs) were originally developed to meet the "Improve" test objective. Phadke (1995) provides an introduction to these design methods. In the Taguchi thinking, interactions and statistical power are not important because the goal of the test is only to find the most robust setting of the controllable factors. However, the research community has identified many improvements over traditional Taguchi designs based on orthogonal arrays. Myers, Khuri, and Vining (1992) show how response surface methodology can be used to achieve similar objectives.

X.2.4 Common Design Methods

Full Factorial: factorial designs include at least two factors and examine all possible combinations of each level. This allows the evaluator to determine the impact of each factor as well as whether one factor influences the impact of another factor on the test outcome. They are highly efficient and informative designs, though potentially prohibitively costly when many factors are involved (e.g., more than four in most operational test cases). For this reason, factorial designs are more common in developmental than operational testing scenarios, or whenever there are relatively few factors in the design.

Fractional Factorial: a variation of full factorial designs that does not require all combinations of factor levels to be included in the test, and include only a fraction of the test points that would be in a full factorial. By trading off the ability to estimate of high-order interaction effects, fractional factorial designs can achieve large reductions in test points over full factorials. Fractional factorial designs are a popular choice for defense system testing due to the large numbers of factors.

Response Surface Designs: a collection of designs that spread test points to collect data throughout the experimental region such that a more detailed model of the pattern of responses can be ascertained. They are used to locate where in the design space (or under what conditions) responses are optimal, and often to "map" a system's performance across a variety of conditions. This often involves the inclusion of higher order effects that can estimate curves instead of monotonic linear effects. These designs also allow estimates of lack of fit and experimental error by adding center points, replications, and axial runs to factorial or fractional factorial base designs.

Optimal Designs: defined in terms of some optimality criteria typically related to minimizing prediction variance or model parameter standard errors. Based on a researcher-specified model and a fixed sample size, software algorithms identify the optimal test points to satisfy the chosen criteria. Optimal designs are most useful when the number of test points is constrained to preclude a factorial design.

Myers, Montgomery, Anderson-Cook provide an overview of common optimality criteria. Several methods exist for optimizing the test point coverage in optimal designs; these include, but are not limited to, D-, I-, and G-optimal criteria. D-optimal designs minimize the overall variance of the parameter estimates while also not letting the covariance between the parameter estimates get too large. I-optimal designs minimize the prediction variance. G-optimal designs minimize the *maximum* prediction error over the design space rather than the average prediction error. While many other design criteria have been proposed in the literature, D- and I- and G-optimal designs are perhaps the most popular and are available options in most statistical software packages capable of generating optimal designs.

Definitive Screening Designs: Jones and Nachtsheim (2011) developed definitive screening designs to estimate main effect and quadratic effects for a relatively large number of factors using a small design. These designs focus on quantitative variables and therefore have not gained much traction yet in operational testing.

Experimental Campaign Sequential Designs: Experimental campaigns (often simply referred to as sequential designs) stand in contrast to "one-shot" designs and employ a series or sequence of smaller tests. The goal is to learn from one test and modify subsequent tests based on this information. Results from preliminary tests may lead the evaluator to drop or add factors, modify the levels of a factor, and the creation of more precise response variable measures.

Point-by-Point Sequential Design: A form of sequential design used commonly in ballistic resistance testing. These designs sequentially (often after each data point) estimate the probability that a projectile will perforate a surface as a function of factors such as velocity and impact angle. Evaluators employing these tests commonly look for velocity at which a projectile has a specific probability (e.g., 50% probability) of penetrating a surface. Johnson et. al. (2014) provides a summary of several different sequential designs.

X.2.5 Statistical Model Supported (Model Resolution)

The statistical model supported by the test design is a primary consideration in determining test adequacy that is often overlooked. The statistical model is very important as it provides a snapshot of the knowledge gained about the behavior of the response across the operational envelope. The following types of statistical models are useful in thinking about test adequacy:

- First order models allow for the estimation of main effects only (shifts in the mean for categorical factors or linear relationships for continuous factors).
- Second order models allow for the estimation of main effects, two-way interaction effects, and quadratic terms for continuous factors.

Model complexity can extend to any order. Additionally, partial order models are possible; one example of a reduced second order model might contain main effects and two-way interactions, but not quadratic terms. Larger-order models result in more flexible modeling; allowing for a closer fit to the observed data. However, in operational testing, when the goal is to characterize performance, second order models tend to be adequate for describing major changes in performance across the operational envelope due to the principle of *sparsity of effects* (Myers, Montgomery, and Anderson-Cook, 2009), which states that most systems are dominated by a few main effects and low-order interaction effects. If the test goal is to screen for important factors a

lower-order model may be appropriate. For prediction of a complex response surface, higher order models may be necessary.

For two-level full and fractional factorial experiments, the order of the statistical model is often discussed in terms of their design "resolution." A design with greater resolution can accommodate higher order model terms than a design with lower resolution. The lower the resolution, the more terms in the model are confounded with other terms, making the cause of observed performance differences amongst the different test conditions difficult to resolve. Resolution III, IV and V designs are particularly important because they address second order models, which are common in operational testing. Definitions of these designs are shown below:

- **Resolution III designs**. Main effects may be indistinguishable from some two-factor interactions
- **Resolution IV designs**. All main effects can be estimated independently but some two-factor interactions may be indistinguishable from other two-factor interactions.
- **Resolution V designs**. All main effects and two-factor interactions can be estimated independently from each other.

X.2.6 Evaluation of Design Properties: Statistical Measures of Merit

A statistical measure of merit quantifies the quality of a specific aspect of a designed experiment. The cost of testing in defense applications requires that all test designs are defensible and statistical measures of merit provide the tools to quantify the quality of designs and compare between designs. No single statistical measure of merit, or group of measures, can completely characterize the quality of an experimental design. Certain qualities are difficult to quantify, such as the selection of factors or responses. In these situations, decisions are based on engineering and operator expertise. Table X.2 summarizes many different statistical measures of merit and their utilities. The appropriateness of a given measure in assessing the quality of a design is dependent on the goal of the test and the experimental design methodology used.

Table X.2. Statistical Measures of Merit

Statistical Measure of Merit	Experimental Design Utility	Goal
Confidence	The true negative rate (versus the corresponding risk, the false positive rate). Quantifies the likelihood in concluding a factor has no effect on the response variable when it really has no affect.	Maximize
Power	The true positive rate (versus the corresponding risk, the false negative rate). Quantifies the likelihood in concluding a factor has an effect on the response variable when it really does.	Maximize

Correlation Coefficients	The degree of linear relationship between individual factors.	Minimize correlation between factors
Variance Inflation Factor (VIF)	A one number summary describing the degree of collinearity with other factors in the model (provides less detail then the individual correlation coefficients).	1.0 is ideal, aim for less than 5.0
Scaled Prediction Variance	Gives the variance (i.e., precision) of the model prediction at a specified location in the design space (operational envelope).	Balance over regions of interest
Fraction of Design Space	Summarizes the scaled prediction variance across the entire design space (operational envelope).	Keep close to constant (horizontal line) for a large fraction of the design space

Confidence

In an operational test we typically have one or more research questions in mind. "Is the new system more survivable than the old system?" or "Does the reliability of the system meet the threshold?" or most importantly, "How does system performance vary across the operational envelope?" When answering these questions with statistical rigor, the concept of confidence is crucial. Confidence is best understood in the context of a specific hypothesis test. For example, consider a new information technology system designed to process repair requests quickly. To test whether this system has adequate performance across a range of operationally relevant request sizes, we use the following hypotheses

$$H_0: Request\ size\ does\ \mathbf{not}\ significantly\ impact\ the\ time\ to\ process\ requests$$

$$H_a: Request\ size\ does\ significantly\ impact\ the\ time\ to\ process\ requests$$

Note that if processing time varies across the operational envelope, we will *reject* this null hypothesis. While not specifically stated in the hypothesis, if we reject the null hypothesis we also will be able to estimate the effect that the request size has on processing time and compare that to requirements across the operational envelope. The confidence level of our test dictates how much we believe our result in the case we reject. For example, if it turns out that the request size does not impact the processing time for operationally representative files sizes, but we reject the null hypothesis, the result of the test is a false positive or Type I error. A desired Type I error rate of α can be achieved by using a test with $100(1-\alpha)\%$ confidence.

However, this doesn't necessarily mean we should always choose a high level of confidence for our test. The tradeoff for a high level of confidence is that we will fail to reject the null hypothesis more often, even when we should reject. Suppose now that the request size does significantly impact processing, but only by a small percentage over the range of operationally relevant file sizes. Without conducting a large test, it is likely that we will fail to reject the null hypothesis, resulting in a false negative or Type II error.

Power

One of the most direct ways to assess a test plan is answering the question, "Will this test plan tell me what I want to know?" In statistical analysis, a standard way to answer that question is by finding the power of the test plan. Power is the probability of finding a statistically significant relationship between the response variable and a particular factor. Recall that "statistically significant" depends on the choice of confidence level for your test; a smaller α makes it less likely that you will reject your null hypothesis and therefore decreases your power. The opposite is true for relatively larger values of α. Power in the above example reflects the probability of finding a statistically significant difference between the processing times due to the request size when a difference truly exists. The difference in time to process between a large and small request is the effect size. Larger effect sizes are easier to detect, so all else equal, your power will be higher if you have a larger effect size.

At the same time, if there is a lot of noise in the data, it becomes harder to identify these effects. When we execute 10 requests, we will get 10 different values for processing time, even if we have the same request size. This run-to-run variability (independent of our experimental factors) is called random error or noise. The more "noisy" the data, the less power we will have. We use the signal-to-noise ratio (SNR) to summarize these competing elements of power. SNR is calculated by taking the effect size and dividing by the standard deviation of the random error in the data. Larger signal-to-noise ratios result in higher power.

Another important consideration in calculating power is the sample size. In general, larger tests result in higher power for detecting significant relationships between experimental factors and the response variable. However, not all data points are equal. Careful consideration must be placed on selecting the points across the design region for maximizing power for the most important factors. The point placement across the operational envelope, the order in which runs were conducted, and the factor combinations selected all have a substantial impact on power. A general rule for optimizing power through experimental design is to balance the test across factor combinations. If we are trying to determine if there is a statistically significant difference in time to process between large and small requests, and we will have 20 runs, it is generally best to make sure that 10 of those runs are small requests and the other 10 are large requests. Rarely is a test design this simple (only a single, two-level factor), and as designs get more complex (additional factors, factors with more than two levels, restrictions such as blocking) it becomes more difficult to find an optimal experimental design. Software packages can be used to generate efficient designs and give you power estimates.

In summary, power is a function of the statistical confidence level, the effect size of interest, the variability in the outcomes, and the number and placement of test points. It not only describes the risk in concluding a factor does not have an effect on the response variable when it really does (Type II error), but also is directly related to the precision we will have in reporting results. This precision, which is related to the effect size, is key in the determination of test adequacy; without a measure of the precision we expect to obtain in the data analysis, we have no way of determining if the test will accurately characterize system performance across the operational envelope.

Collinearity

When designing an experiment, collinearity describes the degree of linear relationship between two or more factors. A well designed experiment minimizes the amount of collinearity between factors. Two or more factors are considered collinear if they move together linearly (as one increases, so does the other). Analysis of data containing highly collinear factors can be misleading, confusing, and imprecise. Variances of coefficient estimates become greatly inflated (making the precision of the test worse) when factors are highly collinear, leading to inflated non-significant p-values (Type II errors). Additionally, collinearity can lead to false positives. When a response is regressed on two highly collinear factors, an analysis of variance might report that both factors are significant. Yet, if only one factor is included in the model, the analysis of variance may indicate that the factor is not significant. Finally, using a model containing highly collinear factors to extrapolate or interpolate between design points can yield estimates with large uncertainty.

Figure 1 shows an example from operational testing in the use of multiple factors to describe the geometric location of an aircraft. An Apache helicopter engages an enemy tank at ten locations along two different profiles. The response variable is weapon accuracy, while the factors are slant range and altitude. The slant range is said to be collinear with altitude because there is a near-linear relationship between the two. The fit line, shown in the bottom left of Figure X.1, shows a positive linear relationship between the two factors as demonstrated by its positive slope; hence, altitude and slant range are, to some degree, collinear.

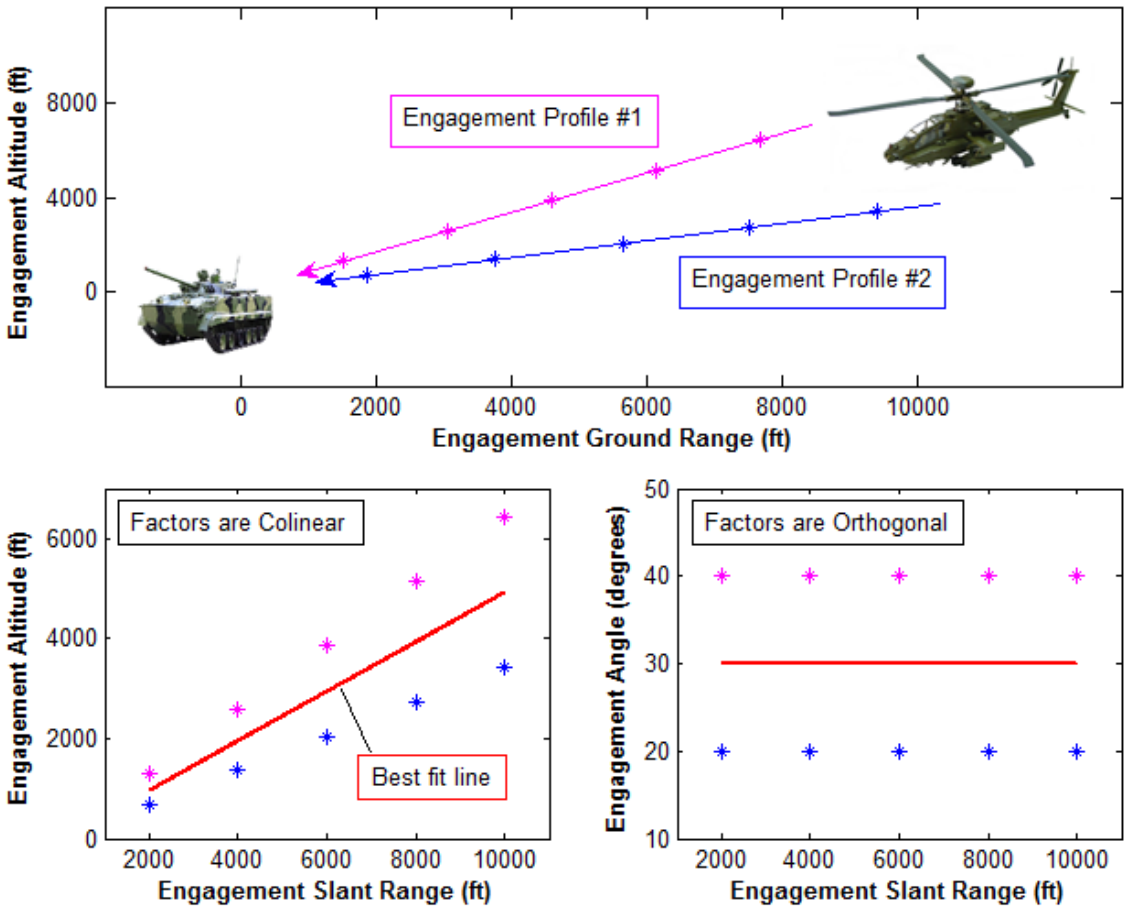


Figure X.1. Engagement Profiles and Factor Plots

While selecting a different flight profile could mitigate the collinearity between slant range and altitude, the factors are mathematically related. Therefore, we cannot completely eliminate the collinearity by adjusting the flight profile. What we can do to break the collinearity is replace altitude with engagement angle, as shown in the bottom right of Figure X.1. Using the same 10 points as before, the bottom right of Figure X.1 shows that we can create a similar experiment where the factors are not collinear. Notice that the best fit line is horizontal, indicating that there is no linear relationship between factors. That is, the factors are *orthogonal*.

Correlation coefficients and variance inflation factors (VIFs) are useful for assessing collinearity of a design. Both of these measures of merit are calculated and monitored prior to executing an experiment. They are functions of the number of runs, the factors and levels in an experiment and how those factors vary from run to run. They are not a function of the data collected from the test. They serve as a tool for establishing the merit of an experiment and can be used to compare DOEs.

The correlation coefficient is bounded between negative one and summarizes the linear dependence of two independent variables. This can be a confusing concept to those used to thinking of correlation a summary of a relationship between and independent and dependent variable. However, as shown above, design choices often result in correlations between factors. Good designs minimize this correlation between independent variables.

VIFs provide a one-number summary description of collinearity for each model term. For an experiment with multiple factors, the VIF associated with the ith factor reflects the increase in the variance of the estimated coefficient for that factor compared to if the factors were orthogonal, and is defined as $VIF_i = \frac{1}{1-\mathrm{R}_\mathrm{i}^2}$, where R_i^2 is the coefficient of determination of a regression model where the ith factor is treated as a response variable in the model with all of the other factors. VIF_i can range from one to infinity. Values equal to one imply orthogonality, while values greater than one indicate a degree of collinearity between factors. The square root of the variance inflation factor indicates how much larger the standard error is (and therefore, how much larger the confidence intervals will be), compared to a factor that is uncorrelated with the other factors. As a rule of thumb, values greater than 5 suggest that collinearity may be unduly influencing coefficient estimates.

Scaled Prediction Variance

In addition to coefficient estimation, one key reason we conduct experiments is to predict future performance of a system. *Prediction variance* describes the error involved with making a prediction using a regression model. Consider an operational test that consists of N runs and k factors. The corresponding first order regression model is $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$. The response variable $\boldsymbol{y}$ is size $[N \times 1]$, $\boldsymbol{X}$ is the $[N \times k]$ design matrix, $\boldsymbol{\beta}$ is a $[k \times 1]$ vector of coefficients, and $\boldsymbol{\epsilon}$ is a $[N \times 1]$ vector of random errors that has $E(\boldsymbol{\epsilon}) = 0$ and $var(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}_N$. The predicted value at any point in the design space is $\hat{\boldsymbol{y}} = \boldsymbol{x_0'}\hat{\boldsymbol{\beta}}$, where $\hat{\boldsymbol{\beta}} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y}$ is the maximum likelihood estimator of $\boldsymbol{\beta}$ and $\boldsymbol{x_0}$ is the vector corresponding to the prediction point in the design space, such as $\boldsymbol{x_0} = [1 \quad x_1 \quad x_2 \quad \ldots \quad x_k]$. The prediction variance at any point in the operational envelope, $\boldsymbol{x_0}$, is defined as $V = \sigma^2 \boldsymbol{x_0'}(\boldsymbol{X'X})^{-1}\boldsymbol{x_0}$. Thus, prediction variance is a function of the designed experiment $(\boldsymbol{X})$, the location in the design space where the prediction is made $(\boldsymbol{x_0})$, and the overall variance is the response (σ^2). Since σ^2 is unknown, prediction variance can be difficult to use for

evaluating the merit of an experimental design. Scaled prediction variance (SPV), on the other hand, normalizes the prediction variance by σ^2 so that SPV is a function of N, $\boldsymbol{X}$ and $\boldsymbol{x_0}$, that is

$$SPV = \frac{NV}{\sigma^2} = N\boldsymbol{x_0'}(\boldsymbol{X'X})^{-1}\boldsymbol{x_0} \quad .$$

The benefit of SPV is that it can be used to evaluate a designed experiment prior to running the test and collecting data. Multiple designed experiments can be postulated for a single test event and compared using SPV, allowing the best design to be selected. When assessing a design in this way, it is important to consider the full range of values each factor can take. For categorical factors, this is just a matter of considering prediction at each level of the relevant factors. For continuous variables, graphical methods such as contour plots are available.

To get a better understanding of SPV, consider a notional penetration test example. In this example, only 10 shots are available to characterize penetration depth of a small arms munition impacting an armor plate on a light combat vehicle as a function of muzzle velocity and range to target. Additionally, the test team plans to fit a second order regression model and has two candidate experimental designs shown in Table X.3. The levels of muzzle velocity and range are expressed in normalized units between minus one and one. The SPV contour plots for each experiment are shown in Figure X.2.

Table X.3. Two Candidate Experiments

Run Number	Design A		Design B	
	Muzzle Velocity	Range to Target	Muzzle Velocity	Range to Target
1	-1	-1	-1	-1
2	1	-1	1	-1
3	-1	1	-1	1
4	1	1	1	1
5	-1	0	0	1
6	1	0	1	0
7	0	-1	0	0
8	0	1	0	0
9	0	0	0	0
10	0	0	0	0

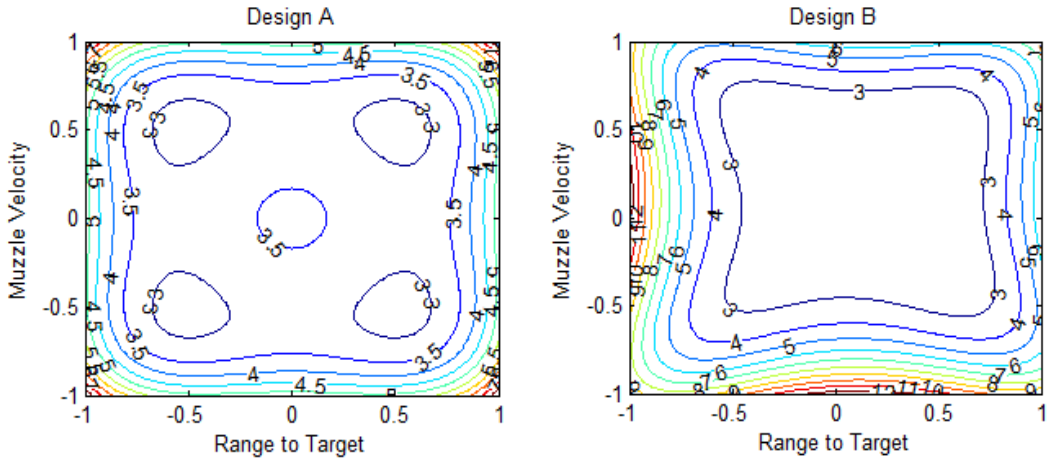


Figure X.2. Scaled Prediction Variance for Two Candidate Experimental Designs

Careful inspection of Figure 2 shows that while Design B has a larger region with the minimal SPV (less than 3.5), a greater portion of the design space for Design A has an SPV less than 4.0. Meanwhile, the SPV near the extremes is much greater for Design B. Based on these observations, Design A is preferred for the limited 10 shots.

In this example, it is reasonable to plot SPV because it is a two-dimensional problem. In cases where there are more than two factors, detailed plots are not straight forward. *Fraction of design space* (FDS) plots show the cumulative distribution of the SPV across the operational envelope (design space). An FDS shows the proportion of the design space with SPV less than or equal to a given value. For the previous example, Figure X.3 shows the FDS for Design A and B. This chart shows that nearly 80 percent of the Design A space has an SPV below 4.0, while roughly 55 percent of the Design B region has an SPV below 4.0. From this chart it is clear that the Army should choose Design A.

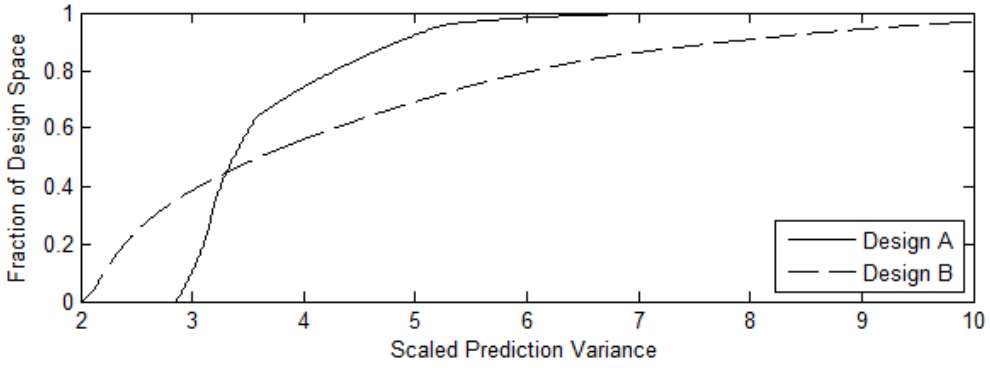


Figure X.3. FDS Graph for Candidate Experimental Designs

X.2.7 Best Practices for Using DOE in Operational Testing

As DOE has become standard practice for DoD operational tests, several best practices have been identified. First and foremost, it is essential to apply experimental design concepts early in the test planning process. Programs successful in using experimental design establish test

planning working groups that include all stakeholders, including the program manager, the requirements representative, developmental and operational testers, and subject matter experts in experimental design and analysis. All of these stakeholders are necessary to identify the key elements of test planning including: goals, response variables, factors, and analysis requirements. Ideally, any test strategy should be iterative in nature, accumulating evidence of system performance before and during operational testing.

Several specific analytical best practices have proven useful in ensuring efficient and effective testing while reducing the test resources required. These best practices include the following:

- Where possible, use continuous metrics as the primary measures of system performance as opposed to pass/fail probability-based metrics. Using continuous metrics has been shown to reduce test resource requirements by at least 30 to 50 percent for the same level of information. Cohen (1983) and Hamada (2002) provides examples and discussion on the advantages of continuous measurements in estimating conformance probabilities. However, many DoD requirements are specified in terms of probabilities. Even in these cases, testers should seek to recast requirements as continuous metrics.
- Similarly, use continuous factors when possible to cover the operational envelope. Identifying these continuous factors, or casting operational conditions in a continuous manner, enables the use of response surface design techniques specifically available for continuous factors. Using these techniques will also afford test efficiencies and provide more information-rich test results.
- Use sequential experimentation approaches to reduce required test resources in each test phase, while developing a comprehensive view of system performance.
- When employing DOE for test planning, focus on the factor-by-factor power calculations, vice a single "roll-up" power estimate. Historically, test planners have used one-sample hypothesis calculations to directly evaluate a single requirement, even in cases where multiple conditions were evaluated in testing. This approach is inadequate for ensuring performance is well-characterized in a test.
- Test goals should not be limited to verifying a single narrowly-defined requirement in a static condition. Rather, testing should aim to characterize performance of the unit when it is equipped with the system across all feasible and operationally realistic conditions.
- Include all relevant factors (cast as continuous where possible) in the test design. By selecting relevant test factors and forcing purposeful control of those factors we can ensure that the operational test covers conditions the system will encounter once fielded. Leveraging developmental test data is essential for narrowing the list of relevant factors and mitigating the risk of excluding important factors. Omitting known important factors from the test design results in holes in our knowledge of system performance. When resources are highly constrained we should leverage

advanced design techniques coupled with developmental testing to ensure we can incorporate as many factors as possible in the test design.

X.2.8 Survey Design and Human Factors

A key aspect of operational testing is observing the quality of human-systems interactions and their impact on mission accomplishment. Operators are a critical component of military systems. Hardware and software alone cannot accomplish missions. Systems that are too complex for operators to use compromise mission success by inducing system failures. Problems that arise because of poor interface design force the Services to invest in lengthy and expensive training programs to mitigate problems. It is critical to evaluate the usability of military systems as well as the workload, fatigue, and frustration that operators experience while employing the system. Surveys are often the only means to evaluate these issues, and proper scientific survey design must be done to ensure that the data collected to evaluate the quality of human-system interactions are valid and reliable.

Surveys capitalize on the thoughts and experience of the system operators to derive essential information for system evaluation. However, their use in operational testing has not always reflected the best practices of the human factors community. The resulting data have had limited utility in evaluations. Data from well written surveys are useful for (a) diagnosing why certain performance goals were not met (e.g., training, system design), and (b) empirically measuring human system integration (HSI) components such as workload and usability. Workload and usability ratings can also form the basis of a robust comparison between new and legacy systems. In nearly every case, data from well written and well administered surveys aid the evaluator in assessing effectiveness and suitability.

One of the most common survey mistakes is the inclusion of questions that ask whether the user thought the system's performance was effective, accurate, timely, or precise enough to complete the mission. Accurate measurement of performance, effectiveness, and situation awareness requires knowledge of ground truth for the test, which operators and maintainers typically do not have. Surveys are measures of thoughts that are highly affected by context and are therefore relative, whereas requirements and performance are absolute, and are better measured by the tester.

The following are some of the best practices, highlighted by the survey community, that we should consider when writing and administering surveys:

a. Neutrality in the questions: The goal of the survey is to obtain the respondent's thoughts. Phrasing questions in a manner that leads a respondent towards the tester's opinions will reduce the likelihood that the respondent provides unbiased answers.
b. Knowledge liability: Do not ask questions the respondent cannot answer (e.g., did the system provide accurate tracking information?).
c. User friendly: Reduce the respondent's effort by making questions brief and clear. Also, make sure that the order of the questions is logical to the respondent.
d. Singularity: Address only one topic in a question.
e. Minimal length: The perceived length of a survey and the actual time it takes to complete it affects data accuracy. Ask the minimum number of questions needed for the goal of the test.

f. Confidentiality: When respondents believe that their data will be kept confidential, they are more likely to provide their true thoughts. Names and other personally identifiable information should be kept separate from the actual survey.

X.3 Statistical Analyses for Complex Systems

X.3.1 Overview

Rigorous statistical analysis methodologies are crucial in testing complex military systems. A sound experimental design is less beneficial if we fail to employ the appropriate corresponding analysis techniques. In the past, analysis has been driven primarily by assessing summary requirements:

- The system should detect 90% of targets at 50 miles.
- The minimum detection range should be at least 50 miles.
- The minimum detection range for moving targets at night should be at least 40 miles.

Because these requirements do not include the operational conditions or provide overly narrow specification of a limited set of conditions they were historically interpreted as on average across the operational space (first two examples), or drove testing in a very specific set of conditions (3rd example), while excluding much of the relevant operational environment.

A better approach is to use experimental design and the regression techniques that characterize performance across the full operational space. Approaches like regression and linear models allow us to identify conditions that have an impact on system performance and determine the sets of conditions under which the system's requirements are met. This is achieved without requiring excessive replication under any one condition. These approaches maximize the information gained from each data point, resulting in efficient tests and defensible analyses.

X.3.2 Linear Models

Linear models (LM) provide an analysis framework for both continuous and categorical independent variables. The general form of the linear model is:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}$$

where $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{\sigma^2 I})$, $\boldsymbol{X}$ is the model matrix of size $n\ x\ p$, n is the number of runs, p is the number of model parameters, $\boldsymbol{y}$ is the $n\ x\ 1$ response variable of interest, and $\boldsymbol{\beta}$ is the vector of model parameters of size $p\ x\ 1$. The model is termed linear because it is linear in the model parameters ($\boldsymbol{\beta}$) and not necessarily in the variables. Therefore, the $\boldsymbol{X}$ matrix can contain quadratic terms or higher order polynomials.

Estimation of the linear model is done using maximum likelihood estimation, resulting in:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}' \cdot \boldsymbol{X})^{-1} \cdot \boldsymbol{X}' \cdot \boldsymbol{y}\ ,$$

where $\widehat{\boldsymbol{\beta}}$ are the estimated model coefficients. Many textbooks have discussed the general linear model so we will focus only on aspects for defense system testing here. Searle (2016) provides a comprehensive overview of linear models. In this section we highlight some aspects of modeling that are frequently used in defense analyses.

Dummy Variables/Variable Coding

One technique used frequently in defense analyses is dummy variable coding. When testing defense systems these variables are often categorical in nature. Examples include:

- Time of Day categorized into day or night. This factor is often included in operational test designs to ensure that systems can be operated both during the day and night without any degradation in performance. In some systems, (e.g., infrared sensors) it

might be worthwhile to capture illumination level continuously, but for many systems we are more worried about human factors considerations and day/night is adequate to capture any variability in outcomes.

- Vehicle types. Often when testing tracking systems we want to ensure they can track different types of vehicles successfully (e.g., tanks, trucks, boats). While it might be possible to capture the elements that differ between these vehicle types continuously, it may require more variables than types of vehicles resulting in the categorical variable being more efficient. Additionally, these continuous variables may not result in orthogonal independent variables introducing correlation in the design space.
- Presence/absence variables for defensive techniques, such as countermeasures and jamming.
- Operating mode. Often operators have the ability to select different operating modes that tune system settings. For example, a tracking system might have operating modes optimized to detection in a littoral environment, versus mountains, versus desert. The operator can only select one of the three modes, but we want a test design that covers all three.

Dummy variables allow for any categorical factor with k levels to be recoded as a k-1 indicator variables and used in regression analysis. For example for a vehicle type variable, we can define

- $x_1 = \begin{cases} 1, \ if \ wheeled \\ 0, \ if \ tracked \end{cases}$

Where x_1 represents the second column in $\boldsymbol{X}$ and the first column is a vector of ones for the intercept of the model. If there were actually three vehicle types of interest (tracked, wheeled, boat) then we could make 2 dummy variables to include them in the regression.

- So we define $x_1 = \begin{cases} 1, \ if \ wheeled \\ 0, \ if \ tracked \\ 0, \ if \ boat \end{cases}$
- And $x_2 = \begin{cases} 0, \ if \ wheeled \\ 1, \ if \ tracked \\ 0, \ if \ boat \end{cases}$

If both x_1 and x_2 equal zero for a given run, then a boat was the vehicle used in that run. Using dummy variable coding we capture both continuous and categorical factors. However, it is important to make sure the coefficients are properly interpreted. Additionally, one would not include interaction variables between levels of dummy variables. Focusing on model predictions instead of coefficient interpretations often helps in this regard.

Prediction

Another aspect of analysis that deserves additional discussion is the importance of prediction using linear models. In academic texts much of the interpretation of linear models focuses on the interpretation and testing of model coefficients. In a defense context, decision makers and operators are often most concerned with a characterization of how good/bad performance might be across a range of operational conditions and how that performance compares

to threshold requirements. Figure X.4 shows an example of a system characterization that focuses on the time to find an initial target using a regression analysis. In this example, the Apache helicopter was tasked to find targets in both low and high density combat scenarios. One goal of the test was to determine if adding Link-16 (a military tactical data network) to the helicopter improved performance. Additionally, the test sought to characterize search capability across a variety of operating conditions. The factors considered in the test design were:

- Time of day (day/night)
- Battlefield density (low/high)
- Link-16 data (yes/no).

The regression analysis revealed that both battlefield density and Link-16 information as well as their interaction resulted in different timelines to find initial targets. Figure X.4 shows the predicted time to find which could be compared to a requirement easily by decision makers.

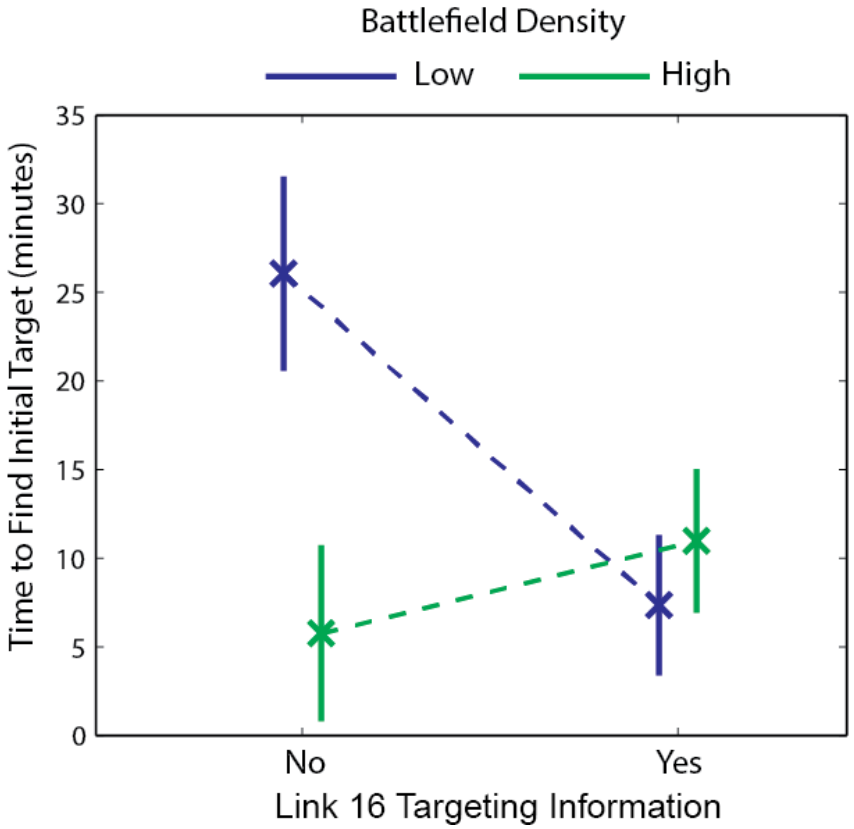


Figure X.4: Prediction plot for time to find initial target (minutes).

We emphasize the importance of prediction here, because while straight-forward for a linear model, correctly estimating confidence intervals can be challenging for more complex models. For a linear model the predicted values are given by:

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} \,,$$

where $\widehat{\boldsymbol{\beta}}$ are the estimated model coefficients and $\widehat{\boldsymbol{y}}$ are the predicted expected values, $\boldsymbol{X}$ is now the prediction matrix, containing the values of the predictor variables where predictions are required. Confidence intervals can be constructed for any realization of the input variables, $\boldsymbol{x}_0$, for $\widehat{\boldsymbol{y}}$ using:

$$\hat{y}(\boldsymbol{x}_0) \pm \boldsymbol{t}_{\boldsymbol{\alpha/2,N-p}}\widehat{\boldsymbol{SE}}(\hat{y}(\boldsymbol{x}_0))$$

Where $\widehat{\boldsymbol{SE}}(\widehat{\boldsymbol{y}}(\boldsymbol{x}_0))$ is the estimated standard error of the prediction at $\boldsymbol{x}_0$ and equal to:

$$\widehat{\boldsymbol{SE}}\big(\hat{y}(\boldsymbol{x}_0)\big) = \sqrt{\frac{\boldsymbol{y'y - b'X'y}}{\boldsymbol{(N-p)}}(\boldsymbol{x_0}' \cdot (\boldsymbol{X'} \cdot \boldsymbol{X})^{-\mathbf{1}} \cdot \boldsymbol{x_0})}$$

The confidence interval on the expected predicted value at $\boldsymbol{x}_0$ is notably different from a prediction interval which provides a confidence interval on a single future observation (y_0). The prediction interval and corresponding standard error are given by:

$$\hat{y}\ \ (\boldsymbol{x}_0) \pm \boldsymbol{t}_{\boldsymbol{\alpha/2,N-p}}\widehat{\boldsymbol{SE}}(\hat{y}\ \ (\boldsymbol{x}_0))$$

$$\widehat{\boldsymbol{SE}}\big(\hat{y}_0(\boldsymbol{x}_0)\big) = \sqrt{\frac{\boldsymbol{y'y - b'X'y}}{\boldsymbol{(N-p)}}(\mathbf{1} + \boldsymbol{x_0}' \cdot (\boldsymbol{X'} \cdot \boldsymbol{X})^{-\mathbf{1}} \cdot \boldsymbol{x_0})}$$

X.3.3 Lognormal Transformation

One of the primary limiting assumptions for linear models in defense testing is that the response variable is normally distributed. An effective defense system typically improves the accuracy and/or timeliness of processes or missions. Therefore, variables that focus on time and distance typically well reflect the outcome of operational missions. Examples of variables that might follow this pattern include times to repair, detection times, detection ranges, miss distances, and target location errors. While these variables are continuous, they are inherently right skewed because of the lower bound at zero. Arguably, in the Apache analysis a lognormal distribution would have better represented the data.

The lognormal distribution often provides a quick and easy solution for these cases due to the close relationship with the normal distribution. If the random variable T is lognormally distributed, then the random variable $Y = log(T)$ is normally distributed. Figure X.5 illustrates this concept visually. In Figure 1, notional mission completion time, t_i, on the left follow a lognormal distribution with parameters $\mu = 2$ and $\sigma = 0.5$, applying a log transformation the data on the right, $log(t_i)$, follow a normal distribution with the parameters $\mu = 2$ and $\sigma = 0.5$.

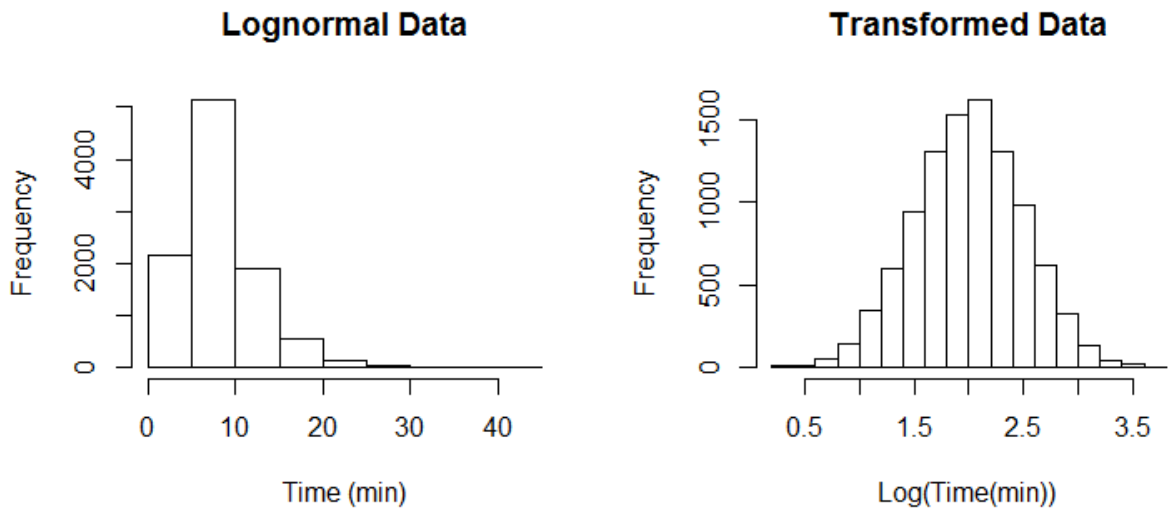


Figure X.5 Illustration of lognormal transformation on data.

Because the transformed data follow a normal distribution we know the expected value of the transformed data confidence intervals can be constructed using the *t*-distribution. However, no one is interested in the expected value of the mission completion times in log scale time. In order to compare the expected value to the requirement we must transform back to the original time units. The mean and confidence intervals in the original data units, however, cannot be constructed via a simple exponentiation of the mean and confidence intervals of the transformed data (i.e. the mean of the lognormal data is not found by $exp(\mu = 2)$. This common mistake has been made countless time in the analysis of skewed data. The correct expression for the mean in the original data units is estimated by:

$$\exp(\mu + \frac{\sigma^2}{2})$$

and a confidence interval for this quantity can be constructed using the multivariate delta method (i.e., propagation of error) to estimate the variance (see the next section for a derivation of this variance). Alternatively, if the median is of interest it can be calculated using $\exp(\mu)$.

Multivariate Delta Method

To illustrate the multivariate delta method for a simple lognormal regression, let $\boldsymbol{T}$ be the vector of mission completion times, and $\boldsymbol{y} = log(\boldsymbol{T})$. A simple linear regression model using the previous linear model notation is:

$$\boldsymbol{X} = [\overline{\mathbf{1}} \quad \boldsymbol{x}], \qquad \boldsymbol{\beta} = \begin{bmatrix}\beta_0 \\ \beta_1\end{bmatrix}$$

Where $\overline{\mathbf{1}}$ is a vector of ones, used to estimate the intercept, $\boldsymbol{x}$ is a vector of a single independent variable, β_0 is the intercept, and $\beta_1$is the slope of the regression equation relating $\boldsymbol{x}$ to log mission completion time.

We can use maximum likelihood estimation to calculate the estimates of $\hat{\beta}_0$, $\hat{\beta}_1$, and σ^2 in a linear model framework. Additionally, we will need the variance-covariance matrix for the

parameter estimates from the linear model. To construct confidence intervals around the expected mean mission completion time at a given value of the independent variable, x_0, let $X_0 = [1 \quad x_0]$ The value we are seeking then is:

$$E(T_0) = \exp\left(X_0^T\beta + {\sigma^2}/{2}\right) = \exp\left(\beta_0 + \beta_1 x_0 + {\sigma^2}/{2}\right)$$

It is not straightforward to calculate a confidence interval on this quantity. The multivariate delta-method provides one solution that results in a closed form expression for the confidence interval on the mean time at x_0.

The multivariate delta method states if one is interested in a function of maximum likelihood estimates $\boldsymbol{g}(\boldsymbol{\theta})$ where $\boldsymbol{\theta}$ is a vector of coefficients estimated using maximum likelihood estimation resulting in $\widehat{\boldsymbol{\theta}}$, then the function $\boldsymbol{g}(\widehat{\boldsymbol{\theta}})$ is approximately normally distributed with mean $\boldsymbol{g}(\boldsymbol{\theta})$ and variance-covariance matrix:

$$\Sigma_{\hat{g}} = \left[\frac{\partial \boldsymbol{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]^{\mathrm{T}} \boldsymbol{\Sigma}_{\widehat{\boldsymbol{\theta}}} \left[\frac{\partial \boldsymbol{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right]$$

where $\boldsymbol{\Sigma}_{\widehat{\boldsymbol{\theta}}}$ is the variance-covariance matrix for the original coefficients.

For our particular function of interest: $\exp\left(\beta_0 + \beta_1 r_0 + {\sigma^2}/{2}\right)$, we can calculate the vector of partial derivatives to be:

$$\left[\frac{\partial \boldsymbol{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right] = \begin{bmatrix} \frac{\partial \mathrm{T}_0}{\partial \beta_0} \\ \frac{\partial \mathrm{T}_0}{\partial \beta_1} \\ \frac{\partial \mathrm{T}_0}{\partial \sigma} \end{bmatrix} = \begin{bmatrix} \exp\left(\beta_0 + \beta_1 x_0 + {\sigma^2}/{2}\right) \\ x_0 \exp\left(\beta_0 + \beta_1 x_0 + {\sigma^2}/{2}\right) \\ \sigma \exp\left(\beta_0 + \beta_1 x_0 + {\sigma^2}/{2}\right) \end{bmatrix}$$

Now all that remains is to do the matrix multiplication and to an estimate of the variance of expected miss distance. Notice the dimension of the above vector is 3x1 for this example, and the variance-covariance matrix is a 3x3 matrix resulting in one value of variance for the miss distance. Once the matrix math is complete confidence intervals can be constructed using Wald's method assuming a normal distribution:

$$(1-\alpha)\%\ CI\text{: } \exp\left(\beta_0 + \beta_1 r_0 + {\sigma^2}/{2}\right) \pm \mathrm{z}_{1-\alpha/2}\Sigma_{\hat{g}} \quad ,$$

where $\mathrm{z}_{1-\alpha/2}$is the critical value of the normal distribution.

Notice that any function could be used here. Often for lognormal data people like to provide percentile estimates (e.g., the 50th percentile, or median, is often of interest). Confidence intervals for that expression can be calculated in a similar fashion. These confidence intervals tend to result in wider confidence intervals than the actual coverage would dictate. Bootstrapping methods can provide better coverage probabilities. Efron and Tibshirani (1994) provide a comprehensive overview of bootstrapping methods.

X.3.4 Discussion of Complex Analysis Methods

The combination of linear models with dummy variables and log transformations can be used to solve many operational test analysis problems. They are easy to implement and provide

straightforward approaches to prediction. Because they are parametric in nature, they are useful for the smaller data sets that often result from operational testing. However, they can lack the overall model complexity to represent the data earnestly.

One common response variable used in operational testing that immediately violates the assumptions of a linear model is binary (pass/fail) data. These measures are desirable because they are easy to measure. Additionally, for some systems, there may not be easily captured continuous variables or they may not directly translate into mission outcomes. For example, when testing a torpedo it might seem logical to capture miss distance as a continuous measure of torpedo effectiveness. However, torpedoes have point detonation fuses, meaning they must directly impact their target to have any effectiveness; a narrow miss is equivalent operationally to a wide miss, making miss distance an inadequate response variable to capture torpedo effectiveness.

Generalized linear models (GLM) provide an analysis solution for these more complex modeling needs. Developed by Nelder and Wedderburn in 1972, generalized linear models generalize the classical linear model by breaking the model into three parts – the random component, the systematic components, and the link function between the two. The random component allows for response variables that follow any distribution in the exponential family. The systematic component is the linear predictor that consists of a linear function of the independent variables. The link function relates the mean of the distribution (random component) to the linear predictor. Logistic regression is one common type of generalized linear models.

Mixed models are another complex methodology that may prove very useful to operational testing, but have not been widely implemented. In operational testing, observations occur within the context of missions and the mission often include clusters of observations. Mixed models contain both fixed and random effects. The random effect (unlike the random component of a GLM) is useful in accounting for mission variability and the fact that the missions executed are sample of all possible missions. A *mixed* model containing *fixed* or *population averaged* effects can be used to address both systematic variation across the missions due to the factors selected and *random* or *subject-specific* effects which address within mission variation. Mixed models for continuous normal outcomes were first presented by Laird and Ware (1982), and these models appear extensively in the literature. In non-normal data situations, these mixed models are commonly referred to as *generalized linear mixed models* (GLMM) and Myers, Montgomery, Vining and Robinson (2010) discuss applications of these models to engineering and industrial data.

X.4 Case Studies of Design and Analysis in Operational Testing

The following case studies provide representative illustrations of the designs and statistical techniques used in operational testing. The first example focuses on an experimental design for a new air-to-ground weapon and highlights the challenge of covering a complex operational space. The second example shows the full process, from planning to analysis, for a sonar software upgrade. The final example shows the evaluation of a counter-fire radar, but omits the design discussion.

It is worth noting that all of these designs are "one-shot" experiments as opposed to sequential experiments. It has long been a goal of DoD testing to conduct integrated testing, where data from earlier test phases could be used to either augment or inform later testing, but that often is unachievable due to completing objectives of tests and limited resources. Dickinson et. al. (2015) show the value of considering multiple phases of testing in the reliability analysis of the Stryker Family of Vehicles.

X.4.1 New Air-to-Ground Weapon Design Case Study

A new air-to-ground weapon ready for operational testing has three different methods of finding and targeting ground enemies. It is designed to launch from several different types of aircraft and to find both fixed and moving targets. There are three targeting methods: coordinate attack, laser attack, and new attack mode. Coordinate attack uses a fixed set of GPS coordinates to guide the weapon to the target. Laser attack uses a laser designation from either the aircraft or a ground based source to find the target. The new attack uses both millimeter wave and infrared targeting methods. Of the three targeting technologies, coordinate attack and laser attack are well understood and have been implemented on many legacy systems. The new attack methodology however is new technology for this weapon.

The goal of the operational test is to characterize the weapon's ability to find, fix, target, track, and engage a variety of operationally realistic targets. The test team identified several response variables of interest in assessing the system including track accuracy, target location error, and miss distance of the weapon upon firing. One aspect of developing operational test programs is that the designs more often than not do not reflect a single response variable, instead, many response variables are typically necessary to fully characterize the system. Even in system like an air-to-ground missile which has a fairly specific mission, several variables are necessary to fully understand the operational effectiveness of the system.

The design strategy breaks the test into three designs: one for each attack mode that focuses solely on the factors that are expected to affect weapon performance for that attack mode. The three designs are then stacked together and matched with variables from an overarching design. The overarching design ensures coverage across a range of launch conditions. While these overarching factors are not expected to impact performance they all represent operationally realistic conditions that operators might encounter. The blocking methodology used in the overarching design ensures that attack mode's designs are not confounded with the factors in the overarching design.

In all of the designs the primary response variables is miss distance. Miss distance provides continuous information about end-to-end weapon engagement capability, but it may be measured differently for each of the attack modes. For example, miss distance for coordinate attack is expected to be the difference between the hit location and the GPS coordinates provided to the

weapon; while in new attack mode, miss distance may be the difference between the target centroid and the hit location.

Overarching Design

Table X.4 summarizes the factors and levels in the overarching design strategy. The first three factors (airspeed, altitude, range) represent the acceptable launch region for the weapon. Time of day was included to ensure that test points would be collected both during the day and at night. This is an example of a factor that is actually not expected to impact the test outcome, but was included as a precaution. Previous systems tests have shown that sometimes lighting considerations can lead to potentially catastrophic events due to human factors considerations, for example displays are not clearly visible during the day (due to glare).

The five overarching factors in Table 1 were used to generate a 32-run full factorial, which was then partially replicated based on the required sizes for the test design for each of the attack modes for a total of 50 runs.

Table X.4. Overarching Open Air Test – Test Design Summary

Factor	Levels
Release Range	Far, Near
Release Altitude	High, Low
Release Airspeed	Fast, Slow
In-Flight Target Update (IFTU) Rate	Fast, Slow
Time of Day (TOD)	Day, Night

Figure X.6 provides the power analysis for the overarching design for identifying difference between these factors that span all three attack modes. The power calculations are provided for generic signal to noise ratios (SNR). The SNR reflects the difference we wish to detect (signal) due to a change in the factor relative to the unexplained variability (noise). Generally, a value of two is considered a large effect size, while a SNR of one is a relatively small effect size. It should be noted that these power calculations assume independence from all of the factors nested in the attack mode designs, which may not be reasonable. Therefore, the power calculations in Figure X.6 do not provide a strong understanding of test adequacy; rather they provide an indication of the coverage of the space. It is notable that if differences exist in performance between attack modes, the overarching design provides high power for detecting those differences, shown by the attack model curve in Figure X.6.

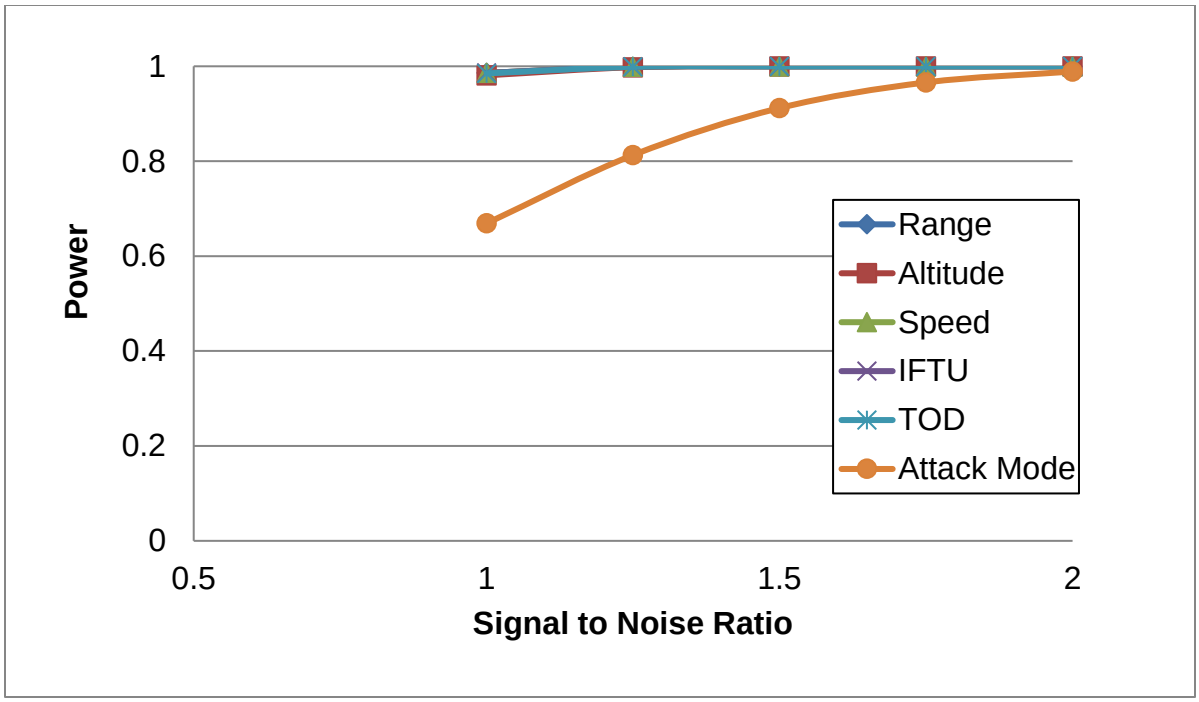


Figure X.6. Overarching Design Power at 80 Percent Confidence Level

GPS Coordinate Attack

Coordinate attack is a well understood legacy capability. Additionally, as many historical tests have shown, very few factors impact the accuracy of a coordinate attack. Table X.5 shows the factors and design approach for coordinate attack. A replicated full factorial design with eight runs was selected because this is a relatively simple targeting mode that requires few factors to characterize performance across the employment space. Additionally, since this capability is the legacy capability, we are only interested in detecting large changes in performance.

Table X.5. GPS Coordinate Attack – Test Design Summary

Factor	Levels
GPS	Degraded, Full
Impact Angle	High, Low

Figure X.7 shows the power analysis for main effects in the coordinate attack design. The second order interaction also has the same power because this is a full factorial design. The eight-run design will be able to detect only large changes in performance across the factor space. However, this is acceptable due to the relatively straightforward nature of the coordinate attack targeting mode.

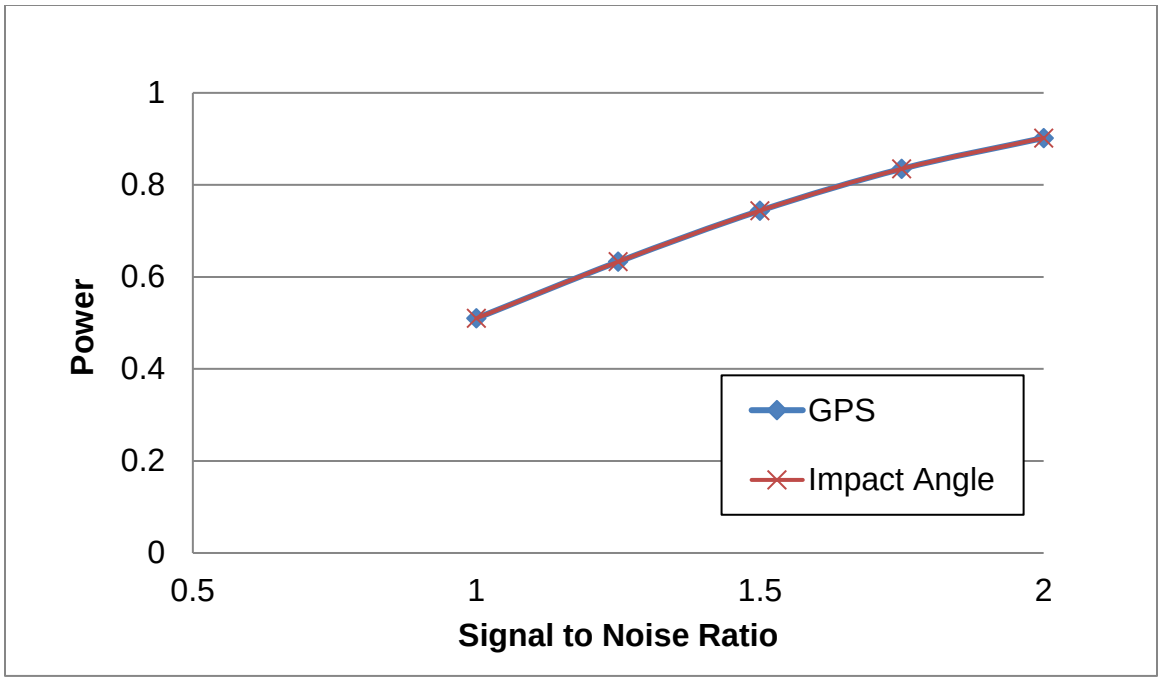


Figure X.7. Coordinate Attack Design Power at 80 Percent Confidence Level

Laser Attack

Table X.6 summarizes the anticipated factors for laser attack and the design. Historical analyses of laser guided tests have shown that the lasing source, target speed, and target aspect are the primary factors that impact the effectiveness of a laser guided bomb. Again, a relatively small eight run full factorial design was selected because we only wish to determine large changes in performance for this straightforward targeting method.

Table X.6. Laser Attack – Test Design Summary

Factor	All Levels
Target Aspect	Head/Tail, 90 degrees
Laser Source	Self, Other
Target Speed	High, Low

Since there was some concern from the test team that time of day could impact the weapon's ability to acquire the laser spot, the laser attack design was matched with the overarching design in such a way that allows for the estimation of the effects of time of day. Figure X.8 shows the power calculations for main effects for the laser attack design factors (all the same) and time of day. Again, the relatively low power for this design is acceptable because we only wish to determine large changes in system performance for the laser targeting method. The power is slightly lower for time of day due to small correlations between time of day and the other factors.

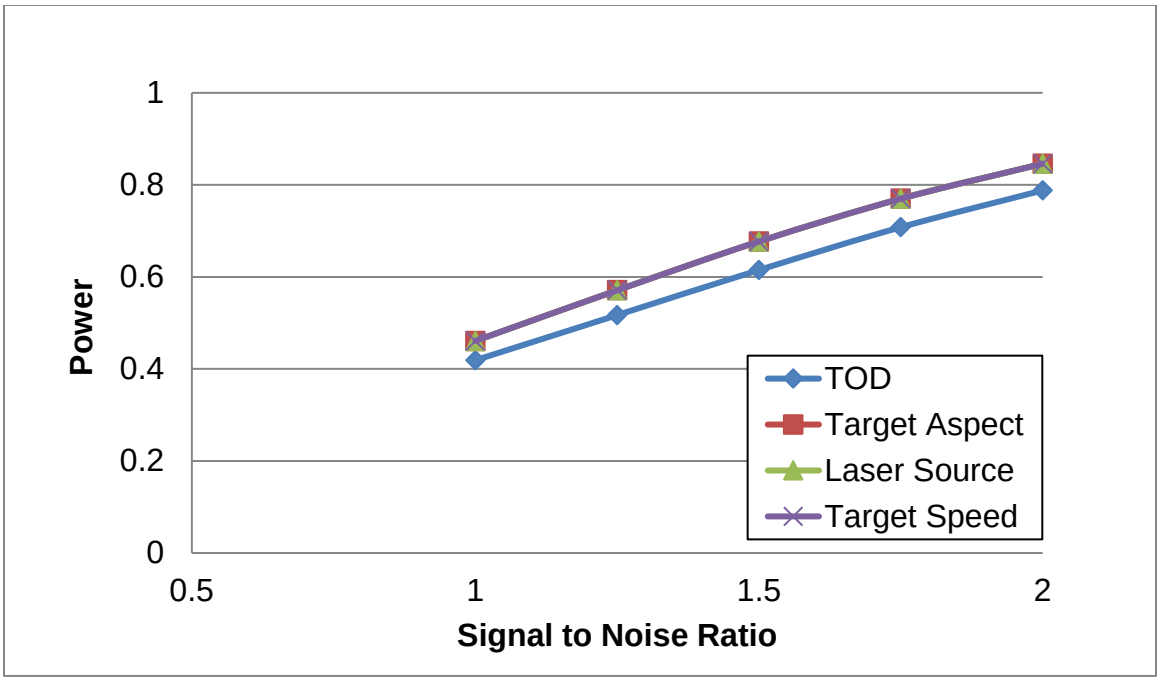


Figure X.8. Laser Attack Design Power at 80 Percent Confidence Level

New Attack Mode

The New Attack is the most complex attack mode and the primary reason one would employ new weapons compared to a less expensive legacy laser guided bomb or GPS guided bomb. The new attack uses both millimeter wave and infrared targeting methods. Table X.7 describes the factors that could impact the performance of the new attack mode.

Table X.7. New Attack Mode– Test Design Summary

Factor	All Levels
Target Type	Wheeled, Tracked
Target Speed	Stopped, Fast, Slow
Target Aspect	Head/Tail, 90 degrees
Infrared Countermeasures (IR CM)	No, Yes
Millimeter Wave Countermeasures (MMW CM)	No, Yes
Confusers	Yes, None

A 34-run D-optimal test design supports all main effects and all two-way interactions for the design factors. Additionally, the update rate and TOD could be important in normal attack performance. Figure X.9 below shows the power analysis for all main effects. The power is high for all main effects. However, to estimate all of the main effects and two-factor interactions for the full design space would require a design with at least 45 points. Instead, the test team selected a design that minimizes correlations between two-factor interactions.

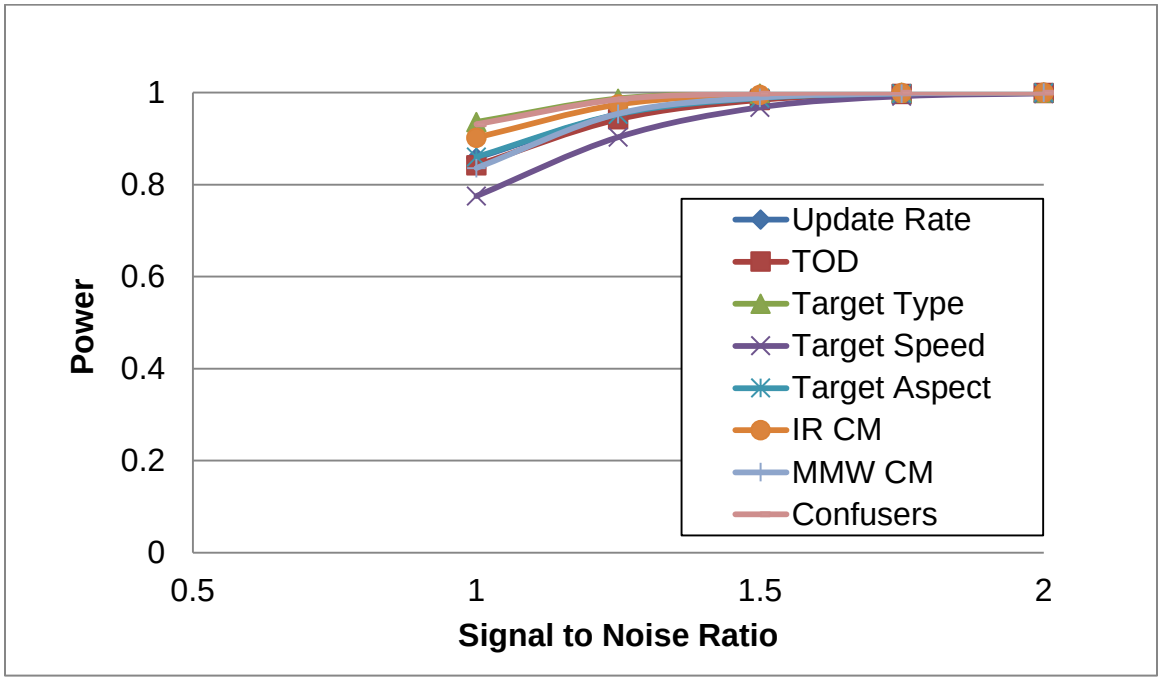


Figure X.9. Normal Attack Design Power at the 80 Percent Confidence Level

Figure X.10 shows the correlations between all factors and their two-way interactions. Using this design will allow the analysis to pull the most important factors from all possible two-factor interactions. It is possible that if we learn that factors are not important in earlier testing we can better optimize the test design to address the most important factors.

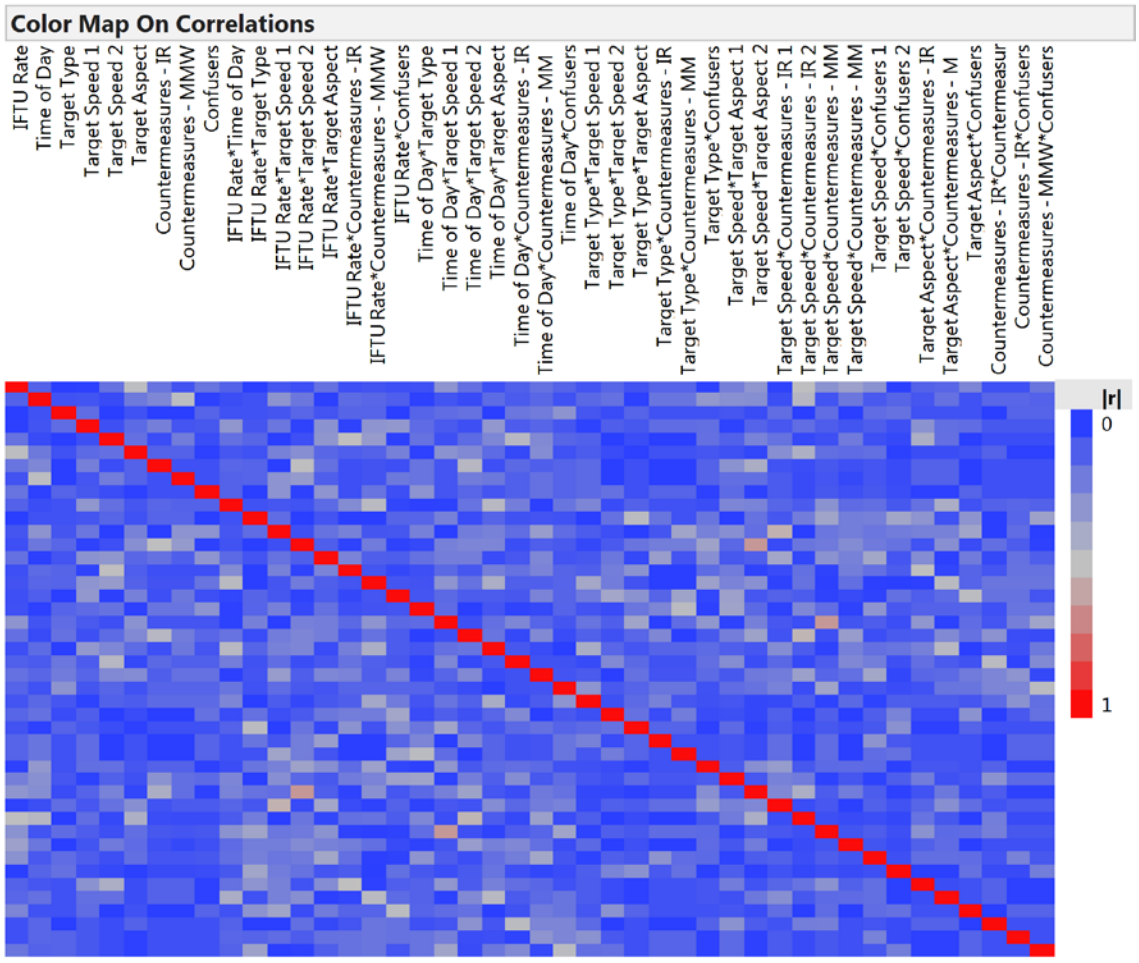


Figure X.10. Color Map of Correlations for all Factors and Two-way Interactions for Normal Attack Design

This illustrative example of using DOE in operational testing shows how different design techniques can be used to address multiple test objectives. Arguably, three separate designs could have been constructed, but breaking the overarching design into blocks ensures that the full operational space is covered across the three different attack modes. Power analysis provides a methodology for showing that 50 missions will be adequate to fully characterize the new attack mode as well as provide limited information on the two legacy modes.

X.4.2 Operational Testing Using Statistically Designed Laboratory Tests

Acoustic-Rapid COTS 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 ever-decreasing cost of computing hardware and ensures 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 Build (APB), is developed every other year and 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 operators 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 operator manage this search space and help him detect contacts as quickly as possible.

The Navy's primary metric used to evaluate A-RCI performance in the Anti-Submarine Warfare (ASW) mission is 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 (ΔT). This measure quantifies A-RCI's role in the detection process. The 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 never has 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 has resulted in poor quantification of APB performance (wide error bars) in the conditions tested.

To address the shortcomings of A-RCI at-sea testing, we use Operator-In-the-Loop (OIL) laboratory testing. In this testing, a fleet operator sits at a laboratory mockup of the sonar system. The laboratory then plays back a recorded encounter between two submarines, and the operator declares when he has detected the threat submarine. The laboratory allows the same encounter to be replayed on different versions of the software, 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 laboratory 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.

Table X.8 shows the factors considered in the experimental design. 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 better characterize the systems, the test used operators of varying proficiency and controlled for characteristics of the target and the array being used.

Table X.8. Factors and Levels used in the OIL Testing Analysis

Factor	Levels	Hypothesized Effect
Target Type	A, B	Different submarine types exhibit different acoustic signatures. Type A has more discrete tonal information because of the.
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.

Figure X.10 shows the original design proposed for this operational testing. The design is a 120-run factorial design with strategic replication. While not apparent in Figure 2, the design is a split-plot design. 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 split-plot structure was used to limit the number of changes of the APB version, as each software change required approximately 12 hours. The large 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 did provide a more precise (tighter confidence intervals) understanding on performance for APB-11.

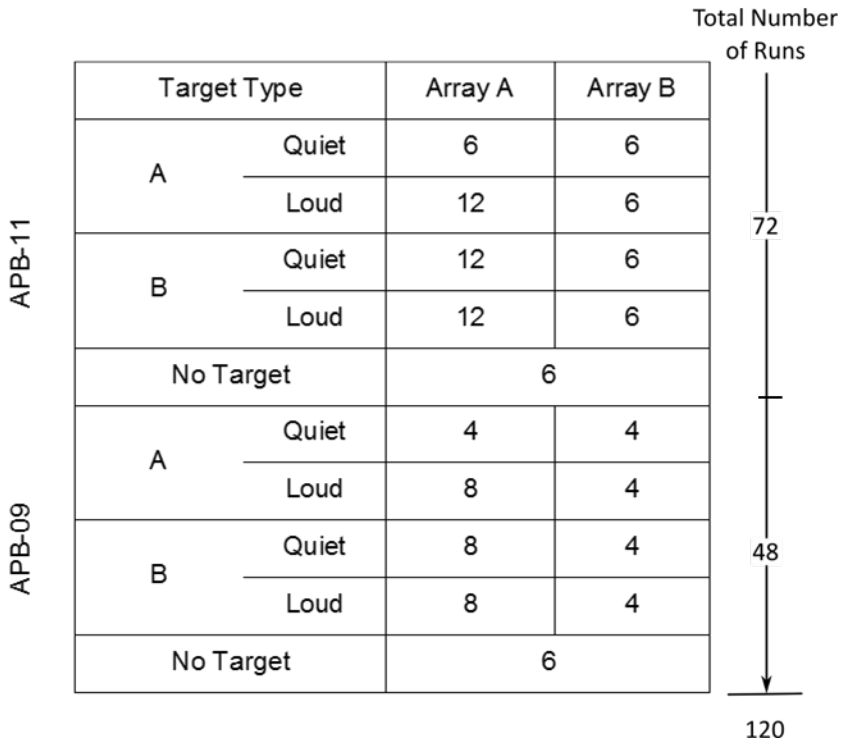


Figure X.10. OIL Test Design Matrix

Figure X.11 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, and the red dots indicate runs in which the operator never detected the target before the recording finished. The location of the red dot indicates how long the target was on tape and not detected. A close inspection of Figure X.11 reveals that the test points do not match the experimental design. It is not uncommon for test plans to change during operational testing where unanticipated execution challenges can arise. In this case, operator availability was not uniformly distributed across operator proficiency levels. Therefore, instead of preserving the original design, testers assigned operators to tapes to ensure a balance of proficiency levels across other factors.

Figure X.11 shows the results by both experimental design bin and individual cut of tape used in that bin. The advantage to 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. The test was well balanced in terms of operator proficiency, so 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. 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 more rigorous statistical analysis is necessary.

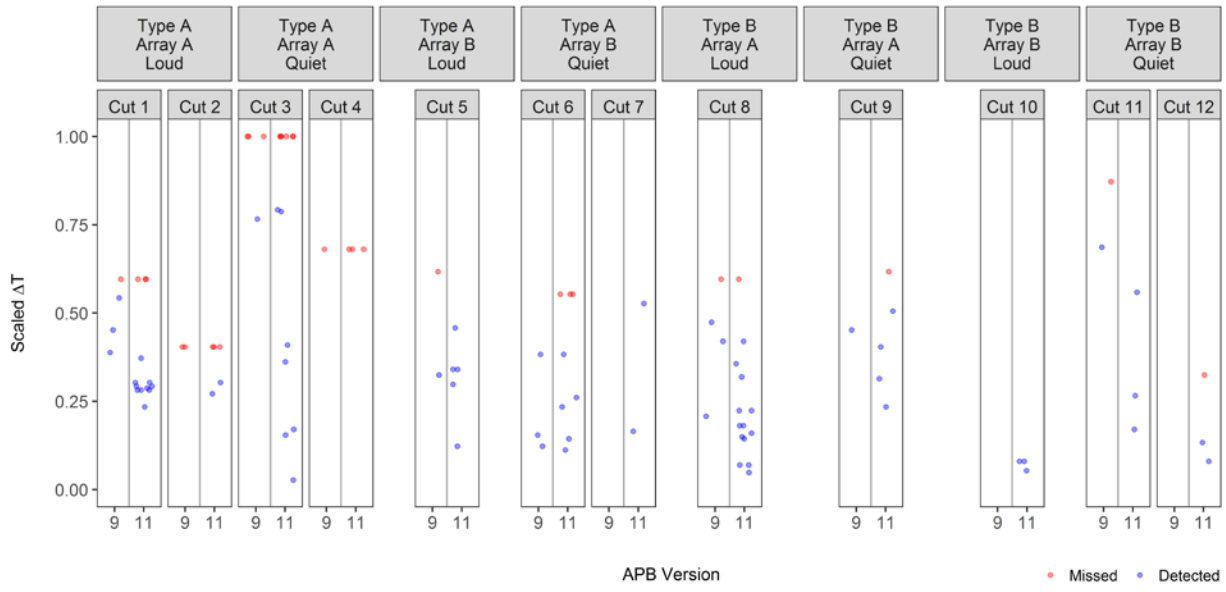


Figure X.11. Raw Results from the Operator in the Loop Testing.

A regression analysis allows us to better understand how the controlled factors affected A-RCI performance. Our analysis however, must account for missed detections. We treated them as censored data points, a useful methodology from reliability analysis (see Meeker and Escobar 1998). In the case of a failed detection, we assumed that the operator would have detected the contact if given enough time. We assumed that the data followed a lognormal distribution, in which the probability of observing a detection time x is the following:

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

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 X.12 shows the empirical cumulative distribution function of the data, along with a lognormal fit and corresponding confidence region on the lognormal fit. The dashed lines represent the confidence region on the non-parametric fit. The data appear to be well described by a lognormal distribution.

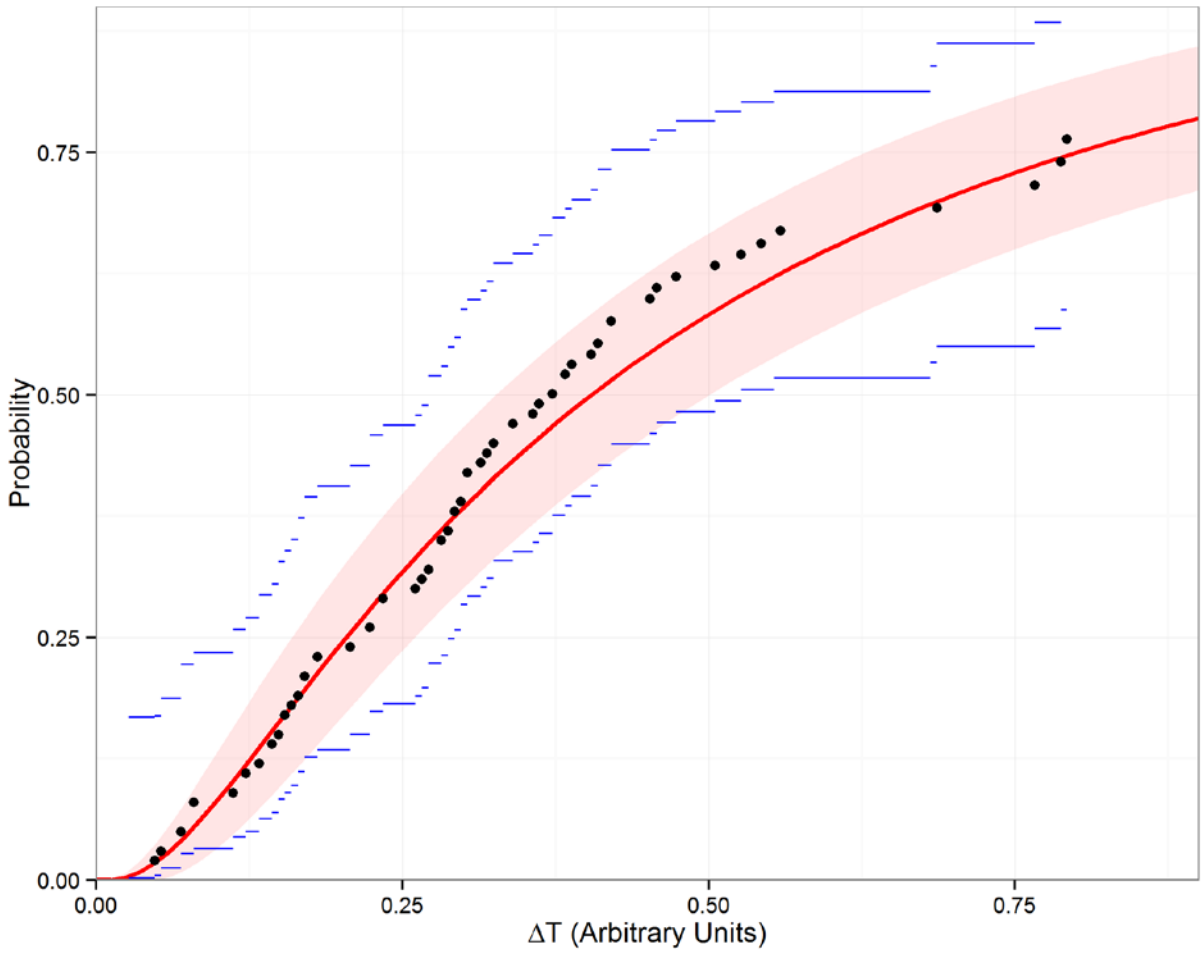


Figure X.12. Empirical Cumulative Distribution of the OIL Data.

After performing model selection we settled on the following model, which minimized the Akaike's Information Criterion (AIC):

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

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

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

Table X.9 shows the results of the final fit and describes the qualitative behavior of the coefficients. We found that the median detection time depends on the factors considered in the design, and that the σ parameter was constant. All of the first-order effects were highly significant. APB-11 is significantly better than APB-09. Also notable is the fact that APB had no interaction with the other factors, which means that APB-11 produced an 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.

Table X.9. Results of the model fit to the data. APB-11 provides a statistically significant improvement.

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.
β_3 (Target)	0.359 ± 0.126	Detection time is shorter for one target
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	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

Figure X.13 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. There is, however, notable disagreement between the data median and the model prediction for one bin: quiet, type A targets with array type B in APB-09. 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.

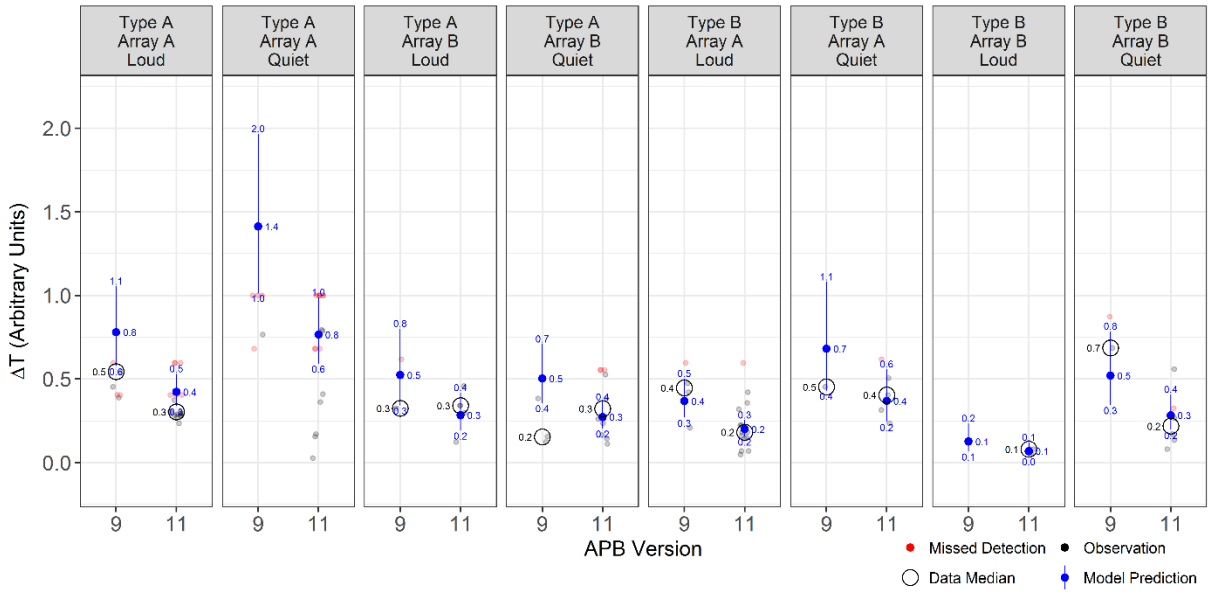


Figure X.13. Model predictions (blue), along with the median detection time observed in each bin.

The lognormal regression analysis provides several benefits over the naïve analysis based solely on individual recordings. First, differences in performance are now attributable to operationally relevant factors, such as target type, array type, etc. In contrast to the naïve analysis by recording, the statistical analysis shows that APB-11 outperforms APB-09 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 X.13 do not have an observed data median for comparison with the model prediction, either because there were few data points or 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.

X.4.3 Counterfire Radar Statistical Analysis

The performance of combat systems can be affected by a wide variety of operating conditions, threat types, system operating modes, and other physical factors. In this case study we look at how different statistical analyses can be used to summarize complex system behavior. The AN/TPQ-53 Counterfire Radar shown in Figure X.14 is a ground-based radar designed to detect incoming mortar, artillery, and rocket projectiles; predict impact locations; and locate the threat gun 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 friendly troops. The Q-53 is the next generation of counterfire radar, replacing the currently fielded radars.

Figure X.14 – Soldiers emplacing the Q-53 radar during operational testing.

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, while 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 trajectory 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. Figure X.15 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 X.15 shows a Q-53 operating 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 threat's position. 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.

Two metrics best answer the key questions about system performance:

(1) Can the Q-53 detect projectiles with high probability?

(2) Can the Q-53 locate a projectile's origin with sufficient accuracy to provide an actionable counterfire grid location?

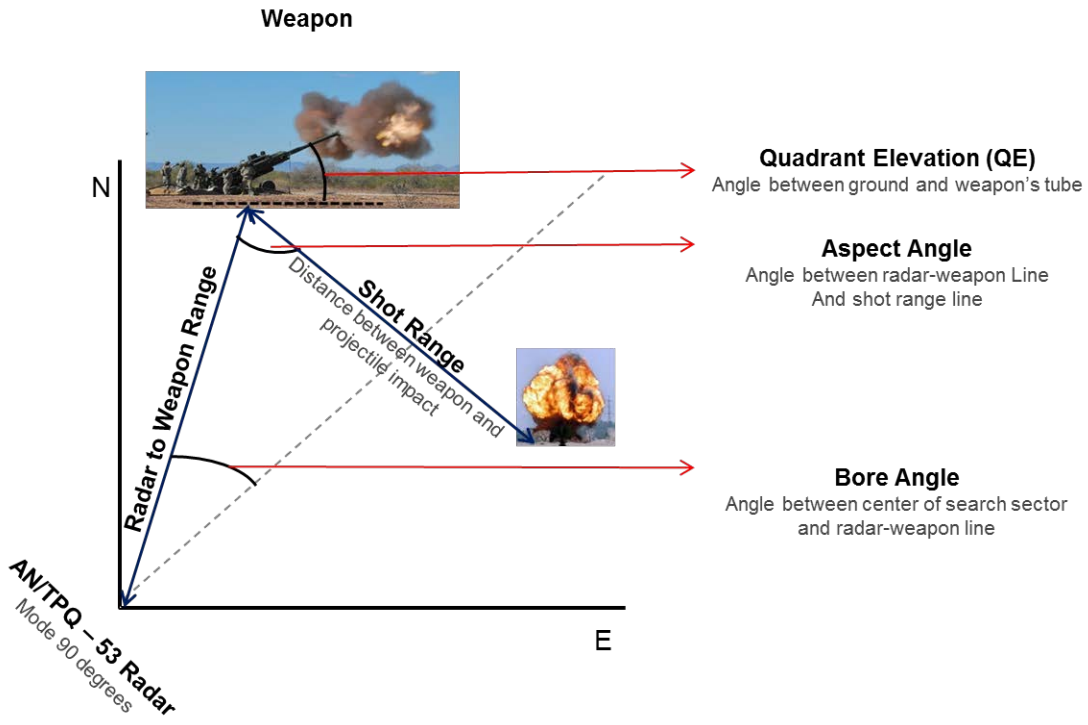


Figure X.15. Example fire mission including relevant geometric factors impacting Q-53 system performance.

The operational test of the Q-53 replicated typical 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 (typically ten) 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 fire mission. Since the radar did not move during these missions, all of the factors in Figure X.15 were held constant during each threat fire mission. Many missions were observed by two radars, so a single threat fire

mission could be detected by two radars. Testers fired 2,873 projectiles, which resulted in 323 usable fire missions.

Figure X.16 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 separated across the x-axis.

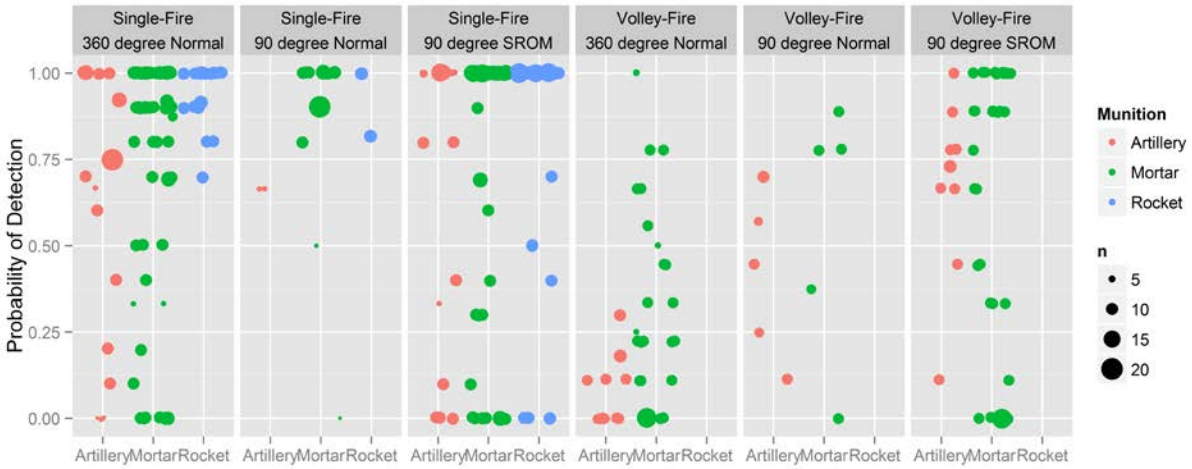


Figure X.16. Detection probabilities for 323 fire missions conducted during the Q-53 IOT&E.

As Figure X.16 shows, there is substantial variability in probability of hit 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. Bore angle is the angle between the weapon and the center of the radar's search sector. 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 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. Furthermore, many of the geometric factors 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 using statistical regression models.

When characterizing system performance, it is important to account for all factors that impact system performance. A logistic regression analysis was the natural analysis model choice considering the complex nature of the problem, allowing us to identify which factors were driving

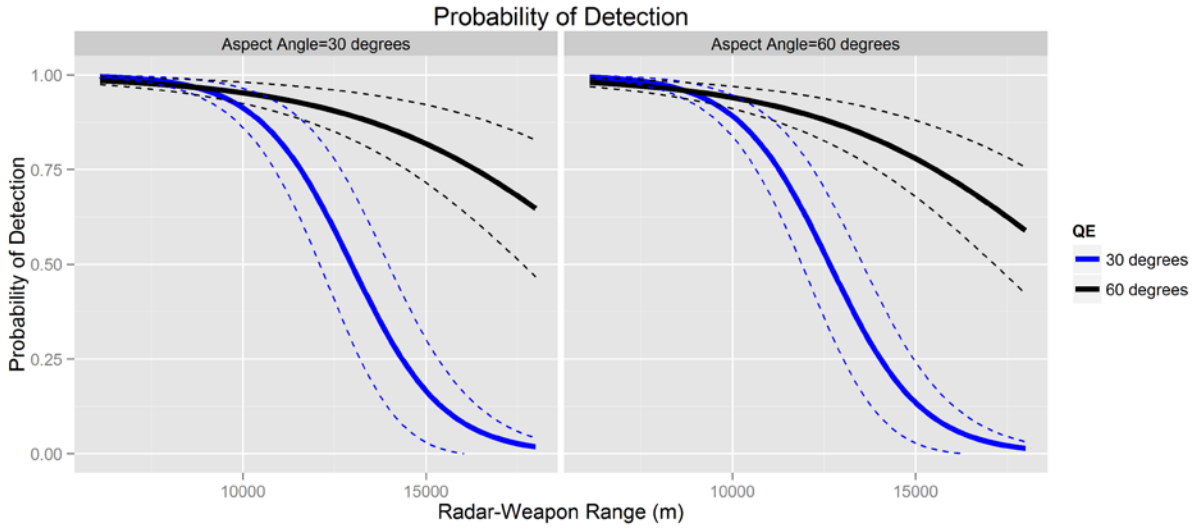
performance and to generate estimates of probability of detection for all combinations of factors. Most importantly, this approach identified the effect of each factor, after accounting for the others, helping determine which factors have the largest impact on performance. The general logistic regression equation is

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

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 naïve 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.

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 X.17 below shows how the probability of detection for artillery projectiles 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 had large impacts on 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 more shallow trajectories that stay closer to the ground (low quadrant elevation) and are more likely to be masked by terrain. Shots with trajectories exposing larger cross-sections of the projectile to the radar (smaller aspect angles) were easier for the Q-53 to detect, although the data showed this factor to be less important than radar weapon range and quadrant elevation.

The logistic regression approach also allows us to analyze the impacts of these factors simultaneously and observe how they interact. In Figure X.17, as the radar-weapon range increases, the probability of detection drops sharply around 12 km for shots with 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 longer ranges. While these factors have large effects, other factors such as aspect angle have relatively minor effects on probability to detect. Comparing the left and right panels of Figure X.17, we can see that a 30-degree change in aspect angle results in a small change in probability of detect.



QE – Quadrant Elevation

Figure X.17. The probability of detection for the Q-53 counterfire radar using the 360-degree operating mode against single-fired artillery.

In addition to detecting incoming projectiles, the Q-53 counterfire radar will also estimate 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 the true point of origin and the location estimated by the Q-53 is referred to as target location error (TLE). 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 analysis than the probability of detection analysis.

Figure X.18 below shows quantile plots of TLEs for the 360-degree operating mode, broken down by munition type. The lines represent what the distribution would look like if the data were normal. The chart on the left shows the raw data. The data do not fall along the straight lines, indicating that the underlying distribution does not conform to the normal distribution. The data are skewed to the right, with many large TLEs in excess of what would be expected for normal data. The figure on the right shows the same data on the log scale; the data fall much closer to the lines, indicating that the log scale is more appropriate.

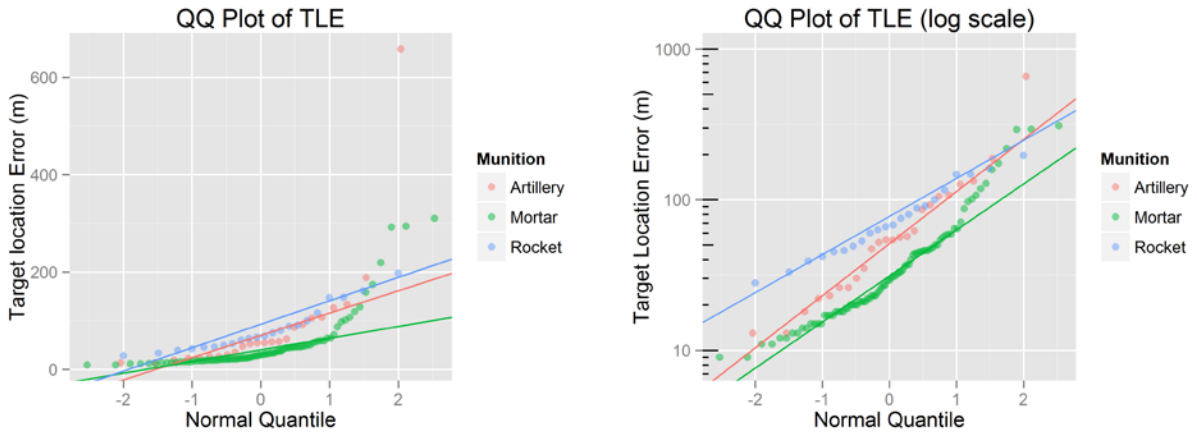


Figure X.18. Quantile-quantile (QQ) plots are used to visually assess normality.

As a result, the TLE data were analyzed 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 X.19 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, 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.

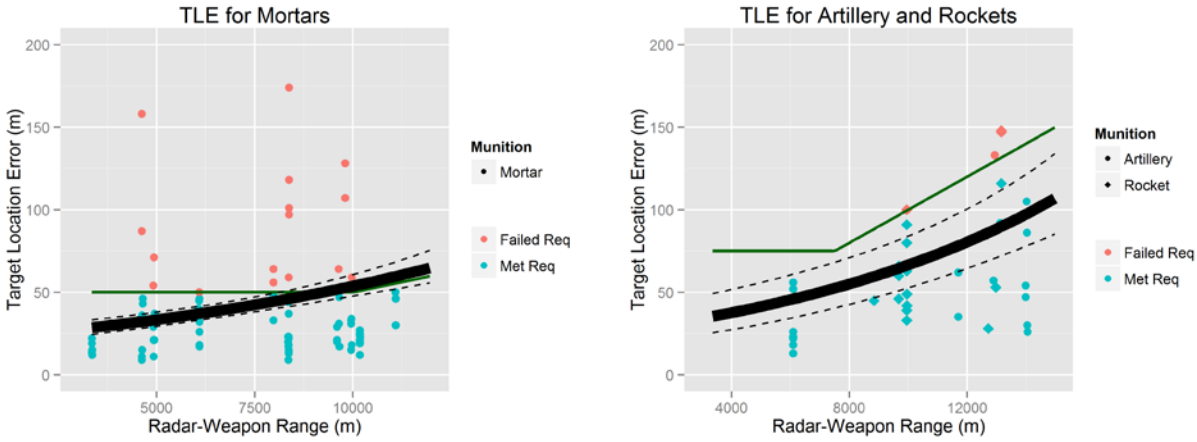


Figure 6. Q-53 target location error for estimated weapon locations.

Physical factors related to the shot's geometry as well as threat and operating mode, impact Q-53 performance. Understanding the effects of these factors helps commanders in the field chose 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. The statistical regression

techniques used to analyze the data identified which factors affected system performance and quantified their impact and practical significance for soldiers employing this system.

X.5 Summary

The methodologies of design of experiments and the corresponding statistical analyses provide a framework for testing a variety of complex military systems. However, examples in existing literature that show how these methods apply to military systems are far from widespread. This chapter provided an overview of the process of designing experiments for military systems with operational users in an operational environment. An important distinction of operational testing is that often the data collected is not exactly the same as the planned data. The analysis model should reflect the observed data and not the planning process. The sonar testing case study provided an example of how the data collected can deviate from the design. In other cases the operating environment results in completely uncontrolled data collection. Kenett and Nguyen (2017) discuss the process for evaluating the information quality of unplanned experiments or experiments that change during execution. The analysis should reflect the data collected, and that data must also be reviewed to ensure it reflects the original goals of the experiment and caveated appropriately if limitations exist. In the case of the new sonar software testing the impact of the new software ultimately was larger than expected, resulting in a larger detectable difference. Therefore the lower replication than planned with imbalance across conditions was ultimately acceptable.

This chapter also summarized some essential parametric statistical analyses used in the analysis of defense systems. Because full system tests in the operational environment are often very expensive, parametric analysis provides the most information for limited sample sizes. However, it is important to ensure that the parametric model reflects the qualities of the data observed (e.g., skewness).

Three case studies illustrated different design and analysis approaches for a bomb, a sonar software upgrade, and a radar.

References – See attachment

Terms for Index

Performance

Operational Testing

Developmental Testing

Design of Experiments

Linear Models

Logistic Regression

Lognormal Distribution

Lognormal Transformation
Power
Confidence
Statistical Model
Statistical Measures of Merit
Defense
Complex Systems