INSTITUTE FOR DEFENSE ANALYSES

Use of Design of Experiments in Survivability Testing

Mark A. Couch, Project Leader
John T. Haman
Thomas H. Johnson
Heather M. Wojton

March 2019

IDA Document NS D-10546

Log: H2019-000127

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
This work was conducted by the Institute for Defense Analyses (IDA) under contract HQ0034-14-D-0001, Task BD-9-1833(07), "JLF and Other Inventions," for the Office of the Director, Operational Test and Evaluation. The views, opinions, and findings should not be construed as representing the official position of either the Department of Defense or the sponsoring organization.

Acknowledgments
Review was conducted by Robert R. Soule, Director and Rebecca Medlin from the Operational Evaluation Division.

For more information:
Mark A. Couch, Project Leader
mcouch@ida.org • (703) 845-2530

Robert R. Soule, Director, Operational Evaluation Division
rsoule@ida.org • (703) 845-2482

INSTITUTE FOR DEFENSE ANALYSES

IDA Document NS D-10546

Use of Design of Experiments in Survivability Testing

Mark A. Couch, Project Leader
John T. Haman
Thomas H. Johnson
Heather M. Wojton

Executive Summary

Aircraft Survivability, a magazine published by the Defense Systems Information Analysis Center, has invited us to submit a short article about the use of Design of Experiments in survivability testing. The invitation comes in response to a presentation that an IDA Research Staff Member gave at an Aircraft Survivability conference in the summer of 2018. Mirroring that conference presentation, in this article we review the principles of Design of Experiments, and highlight its benefits in the context of aircraft survivability testing. This document is the article that we plan to submit to *Aircraft Survivability*.

Photo Courtesy of Darin Russell and U.S. Air Force

Use of Design of Experiments (DOE) in Survivability Testing

By Tom Johnson, John Haman, Heather Wojton, and Mark Couch

The purpose of survivability testing is to provide decision-makers with relevant, credible evidence, conveyed with some degree of certainty or inferential weight, about the survivability of an aircraft. In developing an experiment to accomplish this goal, a test planner faces numerous questions: What critical issues are being addressed? What data are needed to address them? What test conditions should be varied? What is the most economical way of varying those conditions? How many test articles are needed? Using Design of Experiments (DOE) provides an analytical basis for test planning tradeoffs when answering these questions.

DOE is the process of determining purposeful, systematic changes to factors (the independent variables) to observe corresponding changes in the response (the dependent variable). For instance, an experiment may investigate the survivability of various aircraft against engagements from different air defense systems, as shown in Figure 1. The selection of factors, response, and design points is called the experimental design. DOE provides a framework for this process that accommodates a statistical model fit, allowing an analyst to extract cause-and-effect relationships about the system under test.

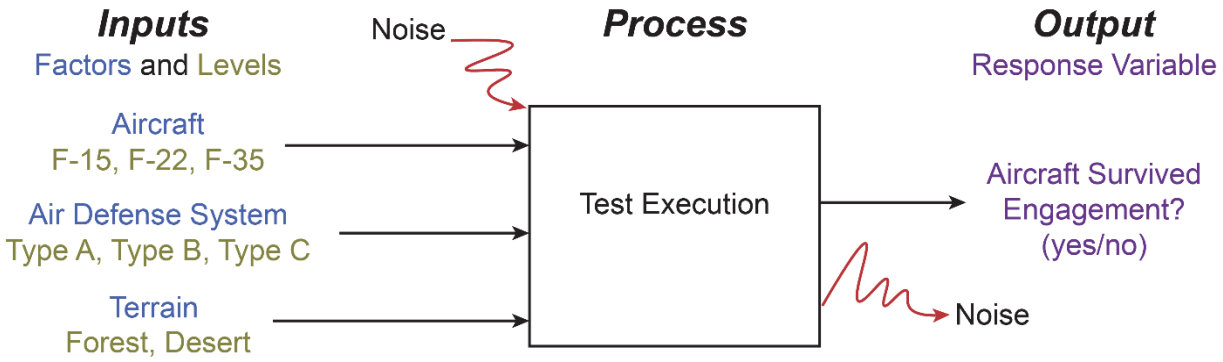

***Figure 1**: DOE test process*

Statistical measures of merit quantify the quality of a designed experiment. Power and confidence describe the risk in concluding that a factor does or does not have an effect on the response variable. A correlation matrix describes the degree to which the factor effects can be independently estimated. Optimality criteria metrics summarize the variance in the estimators. These measures of merit can be used in planning to characterize the quality of the prospective test length (sample size), run order, and design choice by considering the implications on precision, cost, and risk.

In planning an experiment, DOE ensures adequate coverage of the operational envelope, determines the amount of testing required to answer the survivability questions under study, and provides an analytic basis for assessing test adequacy. After data are collected, analysis involves the application and interpretation of statistical models to make estimates and draw inferences. Statistical analysis methods aim to incorporate all relevant information to ensure that conclusions are objective and robust.

Alternatives to DOE include observational studies and "one-factor-at-a-time" tests. An observational study wields little or no control over factors; it simply observes and records their settings as they happen to occur. This approach often suffers from confounding between factors and a lack of randomization, making it difficult to attribute a change in the response to one particular factor. In many fields, such as social sciences, observational studies are unavoidable. Because survivability tests typically occur in controlled environments, observational studies should be avoided.

A one-factor-at-a-time test is an intuitive approach to designing an experiment that involves varying a factor in the experiment while holding all other factors constant. After a "sweep" of data points is collected, the next factor is varied and all other factors are held constant. This approach is repeated for each factor in the experiment. The problem with the one-factor-at-a-time approach, however, is twofold. First, the approach becomes expensive as the number of factors increases, as the total sample size is the number of factors multiplied by the number of data points collected in the sweep. Second, the approach precludes the estimation of interaction effects between factors. A properly designed experiment varies factors simultaneously to efficiently cover the design space and accommodate the estimation of interaction effects in the statistical model.

Fundamentals of DOE

At its core, DOE provides a strategic framework for controlling factors to form a statistical relationship with the response variable. This statistical relationship, or *model*, can then be used to provide

quantifiable, statistically defensible conclusions about aircraft survivability. The six fundamental steps of DOE are as follows [1]:

1. Define the objective of the experiment
2. Identify the response variable (dependent variable)
3. Identify the factors (independent variables)
4. Select the experimental design
5. Perform the test
6. Statistically analyze the data.

Defining the Objective of the Experiment

A clearly defined objective is vital in aircraft survivability testing. What is the survivability of an aircraft against small-arms fire? What is the probability that a ballistic projectile will penetrate aircraft armor? The test objective will answer these questions and drive the experimental design and analysis options. Establishing an objective is a collaborative process that occurs early in the test planning phase, and is the point at which requirements representatives, program managers, users, testers, and subject-matter experts come to agreement (ideally). The most general types of test objectives include screening, characterization, comparison, and optimization.

Screening tests are designed to help identify the most important factors in the test. They are especially useful at the onset of testing, when little is known about how the system responds to various factors. They also allow one to narrow the field of possible factors prior to more costly testing.

Testers conducting a screening test generate a list of all potential factors that are thought to affect the response variable, design and execute a simple test of these factors, and identify the factors that have the largest impact on the response. Further testing in a sequential test program would focus only on these previously identified influential factors.

Characterization tests describe performance at each level of the experiment's several factors. These tests have some of the most important and common goals for live fire testing. Characterization of a system helps one determine whether a system meets requirements across a variety of operational conditions. These tests accomplish this by accommodating a mathematical model that produces accurate predictions of the response variable.

Tests are often designed to compare two or more systems or variants across multiple conditions. These tests aim to quantify and test for statistically significant differences in performance among systems that are operating under similar conditions. For example, whether fuel type (biofuels vs. fossil fuels) impacts the self-sealing properties of fuel tanks is a comparison question. In this example, fuel tanks holding different mixtures of fuel might be fired upon and fuel leakage compared.

Optimization tests seek to find the combination of controllable factors or conditions that lead to optimal test outcomes (i.e., maximize desirable or minimize undesirable outcomes). For example, in a composite armor test, an optimization may inform the optimal number of plies and type of resin that achieves the lowest probability of projectile penetration at the lowest weight. These tests are extremely useful in system design and manufacturing, as well as in the development of military tactics, techniques, and procedures.

Identifying the Response Variable

The response variable is the measurable output from the experiment and is directly tied to the objective. It sounds obvious to take response measurements that map to the survivability question, but this is sometimes overlooked or ignored because it can be inconvenient. If the objective is to characterize helicopter susceptibility to infrared seekers, then the response variable may be whether or not the seeker could lock on to the helicopter, or it may be the time it takes for the seeker to acquire the helicopter.

There are many options for response variables, with some better than others. Some tests will measure more than one response, as survivability can be measured in numerous ways. The key is to select the most important and informative responses, as these will be used to determine the scale and cost of the test and enable the testers to draw definitive conclusions from the test data.

A continuous response variable, such as an elapsed time measured in seconds, is more "information-rich" than a binary response variable, such as whether an aircraft was shot down (yes/no). A sample size determination calculation, called a power analysis, generally shows that fewer continuous response variables are needed to achieve a desired level of risk.

A seemingly obvious yet important practice is to make sure that the response variable is measurable. That is, the response of interest can be observed and recorded at a reasonable cost and within a reasonable schedule, and it can be measured with precision and consistency. Additionally, the response should be sufficiently informative to address the objectives.

Identifying the Factors

Factors are the independent variables that are expected to influence a change in the response variable. Factors have varying levels, which are specific values or conditions they take on. To illustrate, a factor could be projectile velocity that has three levels—250, 500, and 750 ft/s. Another factor could be aircraft type with levels F-15, F-22, and F-35. A well-designed experiment strategically and simultaneously varies the levels of all factors in the experiment to construct an efficient statistical relationship between the factors and the response variable.

Selecting which factors to account for often begins with generating a large, exhaustive list. This list can be narrowed by prioritizing the factors that are central to the test objective. For instance, if the objective is to assess the survivability of an aircraft under nominal flight conditions, the factors and levels should span those nominal conditions.

Factors that have a large impact on the response variable should also be emphasized. If factors that have a large impact are neglected, their effect will be perceived as random variation on the response variable, possibly leading to a poorly fitting model. Attempting to control many factors that have a negligible impact on the response variable can become needlessly expensive.

The factors that are selected will affect the design, analysis, and conclusions that can be drawn from the data. It is a good idea to involve statistical expertise even at the earliest stages of test design. Evaluators are often concerned with how the inclusion and management of factors in a test will affect test size and cost. Contrary to common assumptions, adding or removing factors often does not change the required test size, as changes can be made to the experimental design to accommodate the new number of factors.

Once factors are selected, the next question is how many levels each factor should have. The test objective, resources available, and subject-matter expertise can help determine the number of factor levels. Characterization experiments may require several factor levels to cover the operational envelope, but comparison or screening experiments may require only two appropriately selected levels.

Selecting the Experimental Design

The experimental design specifies the factor settings for each run in the experiment. Once the objective, response variable, and factors have been identified, the experimental design is selected. The following are viable experimental designs for aircraft survivability tests.

Factorial designs include at least two factors and include a run for each unique combination of factor settings. This combination 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 informative designs, though potentially overly expensive when many factors are involved. For example, a factorial experiment with 5 three-level factors has 243 unique runs. In destructive live fire testing, procuring 243 coupons could be cost prohibitive.

A special class of factorial experiment is called the 2^k full-factorial experiment. Here, the experiment has k factors, and each factor has two levels. These are often used to screen for important factors, but they also become more costly as the number of factors increases. For instance, if k equals 8, the 2^8 factorial experiment has 256 runs. A full-factorial design provides information on the main effects (i.e., each individual factor) and how the factors interact (i.e., two-factor interactions, three-factor interactions, etc.). Replication can be used to provide estimates of error; however, if the test is not designed well, confounding can occur, in which two factors vary in the same pattern in the test matrix.

A more efficient class of factorial experiment is the fractional factorial. These variations of full-factorial designs do not require all combinations of factor levels to be included in the test, and they include only a fraction of the test points that would be in a full factorial, as shown in Figure 1. They provide information on the main effects of each factor, as well as some information about how the factors interact.

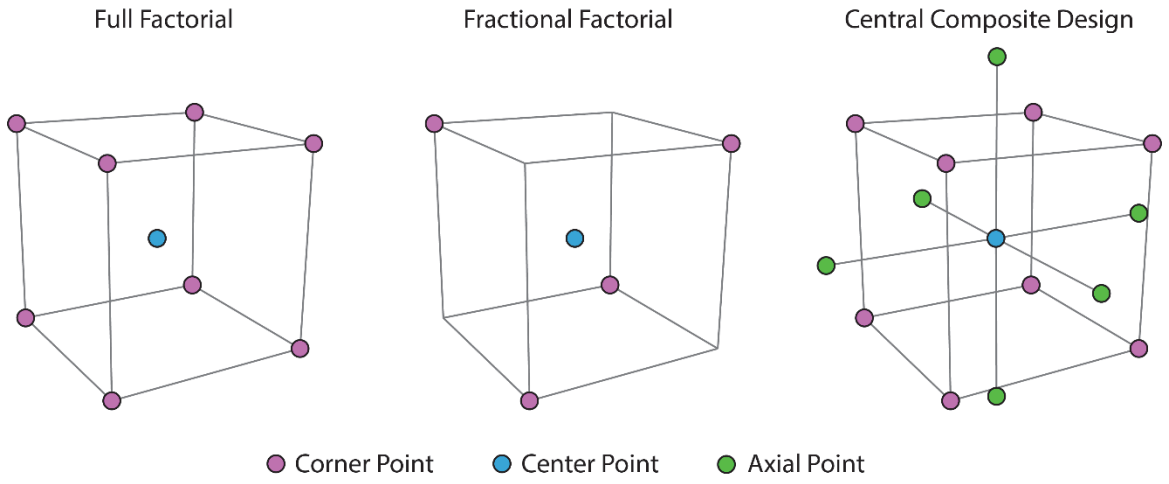

***Figure 2**: Common Experimental Designs*

Response Surface Designs, such as a central composite design, are 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 approach often involves the inclusion of higher-order model terms 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 2^k factorial or fractional factorial base designs.

Optimal designs are constructed using algorithms that modify test point placement until some criterion is maximized or until the design is optimal in terms of that specific criterion (i.e., optimality criteria). That is, the evaluators may want to generate a model that produces minimal average prediction variance over the operational envelope or has minimal variance in model parameter estimates. Optimal designs are often used when constraints prevent the use of standard designs (e.g., cannot test one or more factors at the extremes).

In developing an optimal design, evaluators input what model parameters they would like to estimate and how many runs are available, and the algorithm generates a design specifying the ideal placement of these test points. For example, consider an armor panel test in which the test program has exactly 20 panels at its disposal. The response variable is the penetration depth of the projectile, and the factors are the projectile's velocity, panel thickness, and obliquity angle. The evaluator believes that the model should include main effects and two-factor interactions, in addition to a quadratic effect for velocity. The evaluator can input this model and the desire for 20 runs, and the d-optimal algorithm will output the experimental design that minimizes the uncertainty on the model coefficient estimates.

Performing the Test

Equally important as the selection of response variables, factors, and an experimental design is the management of the test execution. Each run in the experiment should be executable. The factors should be configurable, and the response variables should be measurable with some degree of precision and consistency.

When it is known in advance that a preplanned run order cannot be executed, grouping schemes might mitigate this problem. A split-plot design executes runs in groups and holds a hard-to-change factor constant within a group, while varying the other easy-to-change factors. For example, in an armor panel test that includes two panels of different sizes, it may be determined that the panel size is hard-to-change, and each panel size should be shot 10 times. Thus, a split-plot design executes the test in groups and holds the panel size constant, while firing 10 shots under easy-to-change factor settings that vary. The size of the panel is then changed, and this process is repeated.

Split-plot designs are incredibly useful experimental designs that allow analysts to maximize the information obtained from an experiment and minimize the amount of time they spend collecting the data.

Another issue involving run schedule is encountered when humans are used in the testing. If aircraft survivability is evaluated using actual pilots, learning effects (improved performance after continued operation/testing) and fatigue (decreased performance after continued operation/testing) can confound the results. Furthermore, if multiple pilots are used in a test and they are not equivalent in skill,

training, or other relevant characteristics, this disparity can also adversely affect results. These issues can be mitigated, however, by pulling test pilots from a large pool of pilots who have homogeneous characteristics, planning sufficient rest times, and randomly assigning these pilots to different missions.

Analyzing the Data

Analysis occurs after the test is executed and the data are tabulated. The analysis of the data is often one of the most neglected aspects of DOE, but if the tests are to provide meaning to decision-makers, the data must be analyzed to get the full value out of DOE. The selections of the factors, experimental design, and response variable determine the general form of the statistical model that will be fitted. It is ideally known prior to test execution which general type of model will be used and what the particular analysis of that model will be.

Statistical analyses can summarize and characterize information better than simple bin averages or "roll-up" measures. Figure 3 illustrates this by displaying DOE inferences (red) next to roll-up inferences (blue and black). The DOE analysis results in smaller uncertainty intervals, and we learn that notional factor levels B and C have a significantly different effect on the response variable. The difference does not come out in the non-DOE approach because the blue uncertainty intervals overlap. If the difference between levels B and C is true, the non-DOE approach would require significantly more data (time and resources) to find it.

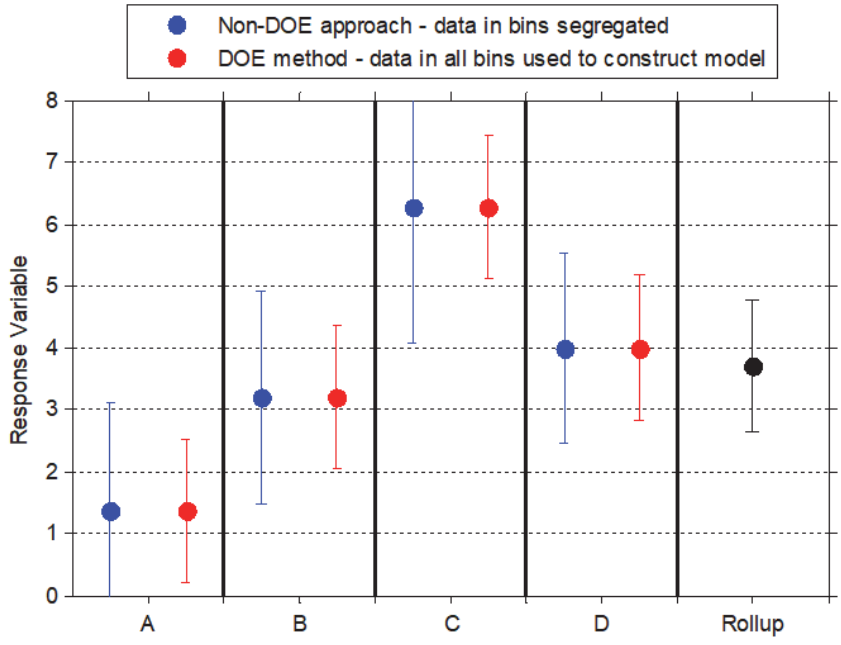

Figure 3: *The DOE Advantage. Differences in response between B and C are apparent when data are modeled with DOE.*

Numerous types of statistical models can be used for aircraft survivability data. A continuous response variable, such as the residual velocity of a projectile penetrating an armor coupon, can be fit with a linear model. A binary response, such as whether or not a projectile penetrates an armor coupon, can be fit with a logistic regression model. A response variable depicting time can be fit with an exponential, lognormal, or gamma model.

Statistical models provide numerous services to analysts, but, broadly speaking, models allow analysts to do two things: make inferences and make predictions. Analysts can estimate the effects (coefficients) of the model and determine which factors have large or small, and positive or negative, effects on the response.

Statistical models have methods that allow analysts to quantify the uncertainty in the estimates obtained from the model, informing them of the extent to which they should believe an effect is large or small. This is the inference part of model analysis.

Statistical models are also prediction devices. Analysts can use statistical models to make informed predictions about the range of plausible responses over the operational envelope, including at factor conditions that were not even tested!

Model-checking is an important part in the analysis stage. Typically, statistical models have some assumptions that allow analysts to make the leap from raw data to inference. And different models may have different assumptions. Generally, in real data analysis, the assumptions of the model are not met, but the model should still be investigated so that the degree to which the assumptions are violated is understood. Many statistical models are robust to deviations from the model assumptions, meaning that statistical models can still be useful when assumptions are not grossly violated.

The statistical model provides the vocabulary to talk about how our test data relate to the research objective at the beginning of the DOE process. This vocabulary includes the magnitude of effects, strength and types of relationships, and degree of uncertainty in a result. The statistical model can be used to make statements about changes in performance across the operational space, predict system performance, and assess requirements.

A variety of statistical software is available to assist with analysis and test design. Design Expert by StatEase is great for learning DOE and has many powerful analysis capabilities. It is set up in an intuitive layout that walks the user through planning, execution, and analysis. In addition, it has an easy-to-use graphical user interface (GUI) that is useful for exploring different types of experimental designs and focuses primarily on simple linear models. JMP by SAS also has a GUI, but it has more functionality and complexity and is thus a bit more challenging to use. Finally, the R computing language, which is free, offers even more functionality and complexity, but it is more challenging to use because it requires coding skills.

Summary

Because aircraft survivability testing is complex and resource-intensive, it is important that data be collected and evaluated in the most efficient and meaningful way possible. DOE provides an analytic work space in which testers and analysts can balance the test objectives with the test resources to find the optimal way to collect data.

Statistical methods for test design and analysis provide an effective means for doing so. In addition, as aircraft systems become even more complex, the test community should continue to evolve applications of state-of-the-art statistical methodologies to confront these new challenges.

ABOUT THE AUTHORS

Dr. Tom Johnson is research staff member with the Institute for Defense Analyses (IDA), where he supports Live Fire Test and Evaluation (LFT&E), providing expertise in statistics, sample size determination, sensitivity experiments, and acceptance sampling plans. His particular focus is on operational tests for personal protective equipment, as well as tests for Army helicopters. Dr. Johnson received his B.S. degree from Boston University and his M.S. and Ph.D. degrees from Old Dominion University, all in aerospace engineering.

Dr. John Haman is a research staff member of IDA's Test Science Team. His focus is on statistical methodologies for operational evaluation, and he is currently supporting electronic warfare systems for the Navy and Air Force. Dr. Haman received a B.S. in mathematics from Truman State University and a Ph.D. in statistics from Bowling Green State University.

Dr. Heather Wojton is a Research Staff Member at the Institute for Defense Analyses. She provides expertise in the evaluation of human-system interactions. She currently aids in the test and evaluation of a broad range of major defense systems, including both training and operational aircraft, and information systems. Prior to taking a position at IDA, she obtained her PhD in Experimental Psychology from the University of Toledo. Dr. Wojton is currently interested in measuring how trust affects people's behavior toward complex technologies and improving methods for measuring common human-system constructs, including workload and usability.

Dr. Mark Couch is the Warfare Area Lead for LFT&E in IDA's Operational Evaluation Division. He supports a variety of LFT&E projects, most notably fixed- and rotary-wing aircraft programs and Joint Live Fire programs. He has analyzed aircraft combat and safety data from recent conflicts in Iraq and Afghanistan, where his analyses have been used to implement improvements in aircraft survivability and crew protection. Previously, he enjoyed a 23-year Navy career flying the MH-53E helicopter. Dr. Couch has a Ph.D. in aeronautical and astronautical engineering from the Naval Postgraduate School.

References

[1] Montgomery, Douglas C. *Design and Analysis of Experiments.* John Wiley & Sons, 2017.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
XX-03/2019	OED Draft	March 1 - March 15, 2019

4. TITLE AND SUBTITLE

Use of Design of Experiments in Survivability Testing

5a. CONTRACT NUMBER

HQ0034-14-D-0001

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Couch, Mark; Haman, John; Johnson, Thomas, Wojton, Heather

5d. PROJECT NUMBER

BD-9-1833

5e. TASK NUMBER

1833(07)

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Institute for Defense Analyses
4850 Mark Center Drive
Alexandria, Virginia 22311-1882

8. PERFORMING ORGANIZATION REPORT NUMBER

D-10546-NS

H 2019-000127

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Director, Operational Test and Evaluation
1700 Defense Pentagon
Washington, DC 20301

10. SPONSOR/MONITOR'S ACRONYM(S)

DOT&E

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release. Distribution is unlimited.

13. SUPPLEMENTARY NOTES

Project Leader: Couch, Mark A.

14. ABSTRACT

The purpose of survivability testing is to provide decision makers with relevant, credible evidence about the survivability of an aircraft that is conveyed with some degree of certainty or inferential weight. In developing an experiment to accomplish this goal, a test planner faces numerous questions: What critical issue or issues are being address? What data are needed to answer the critical issues? What test conditions should be varied? What is the most economical way of varying those conditions? How many test articles are needed? Design of Experiments provides an analytical basis for test planning tradeoffs when answering these questions.

15. SUBJECT TERMS

Design of Experiments (DOE), aircraft survivability

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Mark A. Couch
Unclassified	Unclassified	Unclassified	Unlimited	13	19b. TELEPHONE NUMBER *(Include area code)* 703-845-2530