INSTITUTE FOR DEFENSE ANALYSES

Designing Experiments for Model Validation – The Foundations for Uncertainty Quantification

Heather Wojton, Project Leader

Kelly Avery
Laura Freeman
Thomas Johnson

January 2019

IDA Document NS D-10456

Log: H 2019-000025

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

About This Publication
This work was conducted by the Institute for Defense Analyses (IDA) under contract HQ0034-14-D-0001, Task 229990, "Test Science Applications," 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.

For more information:
Heather Wojton, Project Leader
hwojton@ida.org • (703) 845-6811

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

INSTITUTE FOR DEFENSE ANALYSES

IDA Document NS D-10456

Designing Experiments for Model Validation – The Foundations for Uncertainty Quantification

Heather Wojton, Project Leader

Kelly Avery
Laura Freeman
Thomas Johnson

Executive Summary

Computer models and simulations are increasingly critical sources of information for developing, testing, and evaluating military systems. The validity of the information provided by models, and therefore the trustworthiness of the evaluations that use models, depends on a rigorous verification and validation process. A rigorous validation strategy requires not only a sound experimental design for the live data, but also a sound design strategy for covering the model domain. Design of Experiments (DOE) and statistical analyses are the foundational tools that support a comparison of simulation and live data, and quantification of statistical uncertainty in model validation.

Importance of Uncertainty Quantification

A quantitative validation analysis should accurately convey the uncertainty in any generalizations from models and data. Understanding the type of uncertainties that exist is a key consideration when developing the verification and validation plan and deciding how to allocate resources. DOE is critical for ensuring that statistical uncertainty is sufficiently small to support validation needs. A rigorous validation strategy should identify data collection requirements to achieve acceptable statistical uncertainty.

Experimental Design for Validation

DOE provides a defensible strategy for selecting the data needed from live testing and simulation experiments to support validation needs. Classical DOE (appropriate for physical experiments) and computer experiments provide the building blocks for conducting a validation. Even though they are often grouped into "DOE" as a whole, their design principles, algorithms for generating designs, and corresponding analysis techniques can be quite different.

A key feature of live testing is that response variables are stochastic. Thus, classical DOE emphasizes design robustness over optimality. Replication, randomization, and blocking are fundamental principles in physical experimentation. Typically only a small subset of the factors in the design will actually be statistically significant. This effect sparsity justifies the use of efficient screening experiments. The same reasons that make classical DOE appropriate for physical testing also apply to validation. The wide spectrum of validation techniques all require a robust estimate of the noise in the physical response variable. Classical DOE principles accurately estimate, making the comparison to simulation more powerful.

The literature on computer experiements assumes the outcome variable is deterministic. Thus, random variation, replication, or practical considerations about hard-to-change factors are not a concern. Design points in computer experiments are free to spread out across the design space, and emphasize optimality over robustness. These designs maximize the likelihood of identifying problems with the model or simulation by filling in the space, allowing for the discovery of local maxima/minima or non-linearity in the code.

Hybrid Approaches

In practice, a combination of classical and computer experiments has proven useful in the validation of defense models and simulations. Using hybrid approaches provides a strategy for two challenges with simulation experiments: 1) non-deterministic simulations and 2) matching points between live simulation experiments.

Despite the common assumption that computer models are deterministic, there are many instances in defense testing where the response variable of a simulation is non-deterministic. Human-in-the-loop or hardware-in-the-loop simulations introduce a physical component and thus random variation. While this random variation may lend itself to a classical DOE solution, there may still be a desire to use space-filling designs to cover more of the simulation space and gain a better understanding of the simulation response surface. In these cases, simulation designs paired with classical designs provide a straightforward strategy. The simulation design can cover the simulation input domain, and a classical design is used for selecting replicate points and/or matching points for live tests.

Direct matching of points between the live tests and simulation experiments provides the best validation strategy in all cases. If the subset of matched points are based on a classically designed experiment, the validation strategy can include regression analysis, which is the most powerful validation approach.

The best design for the computer experiment and live test ultimately depends on the analytical goal and the nature of the simulation and the data it produces. Statistical designs should support both a comparison with live data and exploration of the model space itself, including conducting sensitivity analyses and building emulators. DOE is the first step toward a sound methodology for data-driven validation.

Designing Experiments for Model Validation – The Foundations for Uncertainty Quantification

Dr. Laura Freeman

Kelly Avery

Thomas Johnson

Computer models and simulations are increasingly critical sources of information for developing, testing, and evaluating military systems. In the development of new systems, systems engineers and developmental testers may use engineering and engagement-level computer models to refine system design and evaluate design tradeoffs in meeting performance requirements. Once a system design is finalized, engagement and mission-level models can be used to design live tests to answer informed questions about operational effectiveness, suitability, or survivability. Computer models and simulations can benefit the Department of Defense (DoD) Test and Evaluation (T&E) by:

- Enabling tests to focus on critical information suggested by the model
- Identifying edges of the operational space
- Extrapolating performance into conditions where live testing is constrained by safety or environmental considerations
- Bolstering conclusions from live testing by filling in gaps from live tests
- Creating realistic test environments by using stimulators to augment live test conditions.

The Director, Operational Test and Evaluation (DOT&E) has noted the usefulness of models, simulations, and stimulators in planning, executing, and evaluating operational tests. In a June 2002 memo to the Operational Test Agencies (OTA), DOT&E discussed modeling and simulation as a data source supporting core T&E processes. In 2016 and 2017, DOT&E noted shortfalls in the analytical tools used to validate models. A 2016 guidance memo from DOT&E requires the use of a statistically-based quantitative method to compare "live data" to "simulation data" when models are used to support operational tests or evaluations. The guidance requires that test planning documents capture 1) the quantities on which comparisons will be made (response variables), 2) the range of conditions for comparison, 3) a plan for collecting data from live testing and simulations, 4) the analysis of statistical risk, and 5) the validation methodology.

In a follow-on clarification memo in 2017, DOT&E emphasized that validation not only required a sound experimental design for the live data (ideally matched with the simulation data), but also a sound design strategy for covering the model domain. This clarification is consistent with the National Research Council (2012) recommendation that validation activities can be separated into two general categories: 1) external validation (i.e., comparison to live test data), and 2) sensitivity analysis (i.e., investigate model outcomes across the model input domain).

Figure 1 conceptually contrasts two strategies for selecting points from models and live testing. A key element of this selection is what points to conduct in both live and model-based testing. Ideally, the live design should encompass the simulation design (left panel) so that comparisons made between the two are interpolations. However, this is often not feasible due to practical constraints that exist in live testing. In these cases, illustrated in the right panel, the domain of the live testing should span the maximum possible domain of the simulation experiment, and regions of extrapolation should be clearly identified in the validation limitations. Additionally, uncertainty should be propagated with the extrapolation to capture the state of knowledge in the extrapolation.

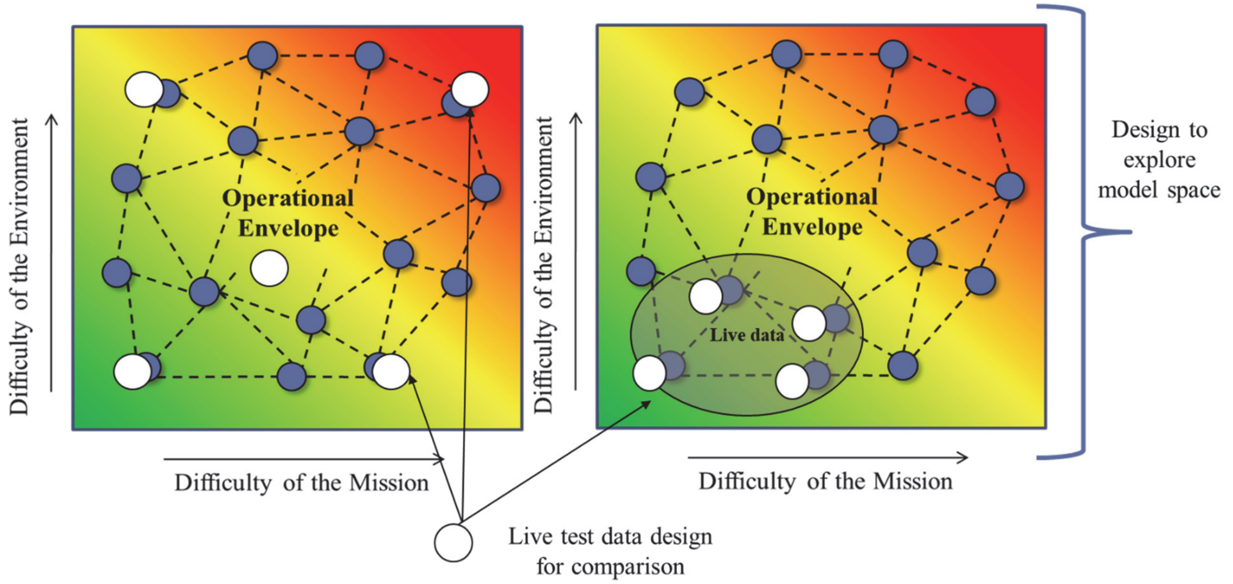
Difficulty of the Environment
Operational Envelope
Difficulty of the Mission
Difficulty of the Environment
Operational Envelope
Live data
Difficulty of the Mission
Design to explore model space
Live test data design for comparison

Figure 1. Notional picture of data selection across M&S and live testing

The validity of the information provided by models, and therefore the trustworthiness of the evaluations that use models, depends on a rigorous verification and validation process. In recent years, evaluations of operational effectiveness, suitability, and survivability have increasingly relied on models and simulations (M&S) to supplement live testing. In the 2017 DOT&E Annual report, Director Behler noted that we need to "improve upon current M&S capabilities, including the verification, validation, and accreditation of M&S assets." In order for the information extracted from models to be valuable, we must understand how well the model represents the system or process being simulated. In the past, validation processes have lacked sufficient statistical rigor to support the quantification of uncertainty in accreditation decisions. Design of Experiments (DOE) and statistical analysis are the foundational tools that support a statistical comparison of simulation and live data.

Importance of Uncertainty Quantification

Historical verification and validation processes have not formally acknowledged uncertainty quantification as a critical aspect of using models for evaluations. Recent research and access to better mathematical tools for quantifying uncertainty make it possible to quantify uncertainty from models as well as live data. A quantitative validation should aim to accurately convey the uncertainty in any generalizations from models and data.

The two major categories of uncertainty are aleatoric (statistical) uncertainty and epistemic uncertainty (uncertainty due to lack of knowledge). Understanding the type of uncertainties that exist is a key consideration when developing the verification and validation plan and deciding how to allocate resources. Epistemic uncertainty reflects data inaccuracy that is independent of sampling; in other words, collecting more of the same data does not reduce epistemic uncertainty. Epistemic uncertainty can only be reduced by improving our knowledge of the conceptual model (e.g., improved intelligence on threats) or by incorporating more proven theories into the models (e.g., incorporating of the tactical code from systems).

Statistical uncertainty captures information on the quantity of data and the variability of the data that was collected under a certain set of conditions. It cannot be reduced or eliminated through improvements in models, but can be reduced by collecting more data. DOE is critical for ensuring that aleatoric uncertainty is sufficiently small to support validation needs. A rigorous validation strategy should identify data collection requirements to achieve acceptable statistical uncertainty.

Experimental Design for Validation

DOE provides a defensible strategy for selecting what data is needed from live testing and simulation experiments to support validation needs. The most powerful validation analysis techniques (e.g., match pairs in a regression analyses) require some degree of coordination between the designs of the physical and simulation experiments. DOE for physical experiments, referred to here as "classical DOE," was developed in the 1920s by Ronald Fisher and is typically taught in introductory DOE courses. DOE for simulation experiments, referred to here as "computer experiments," was developed in the late 1970s and early 1980s and is still researched today. Classical DOE and computer experiments provide the building blocks for conducting a validation and, even though they are often grouped into "DOE" as a whole, their design principles, algorithms for generating designs, and corresponding analysis techniques can be quite different.

A challenge for analysts is that popular textbooks for Classical DOE and computer experiments do not explicitly lay out the necessary steps to tackle validation. Classical experimental design has been used extensively in development and operational tests over the past several years. See Johnson, Hutto, Simpson, and Montgomery (2012) for a review of DOE in Defense testing. A feature of live testing is that response variables are stochastic. That is, re-running a trial with the same factor settings gives different observations.

This non-deterministic nature of physical tests is why Classical DOE emphasizes design robustness over optimality. Robustness, here, refers to the capability of the experimental design and its associated analysis to withstand complications that arise in physical testing. These include outliers, missing data, nuisance errors, and violations to the statistical assumptions.

Replication, randomization, and blocking are fundamental principles in physical experimentation. Replication is repeating at least some of the trials in the experiment in order to estimate of the experimental error. Randomization refers to the practice of running the trials in the experiment in random order to minimize systematic variation and to provide a valid estimate of uncertainty. Blocking is a technique to prevent the variability from known nuisance sources from increasing the experimental error.

Another distinguishing feature of physical experiments is that, in general, only a small subset of the factors that are part of the design will actually be statistically significant. The random variation in the response variable means that the impact of many factors, controlled or uncontrolled, cannot be distinguished from the background noise. Referred to as effect sparsity, this concept justifies the use of efficient screening experiments. Figure 2 shows examples of classical experimental designs. Full factorial and fractional factorial are some of the most common classical design techniques. They push design points to high/low levels in the region and add center points to check for curvature in the response surface. The central composite design shown is a class of response surface designs that enable more flexible modeling. Another broad class of classical designs are optimal designs, which are defined in terms of an optimality criteria. They are algorithmically generated based on a researcher-specified model and a fixed sample size (the fixed sample size must exceed the number of model terms). Common optimization approaches include minimizing parameter estimate variance, minimizing average prediction variance, and minimizing the maximum prediction variance.

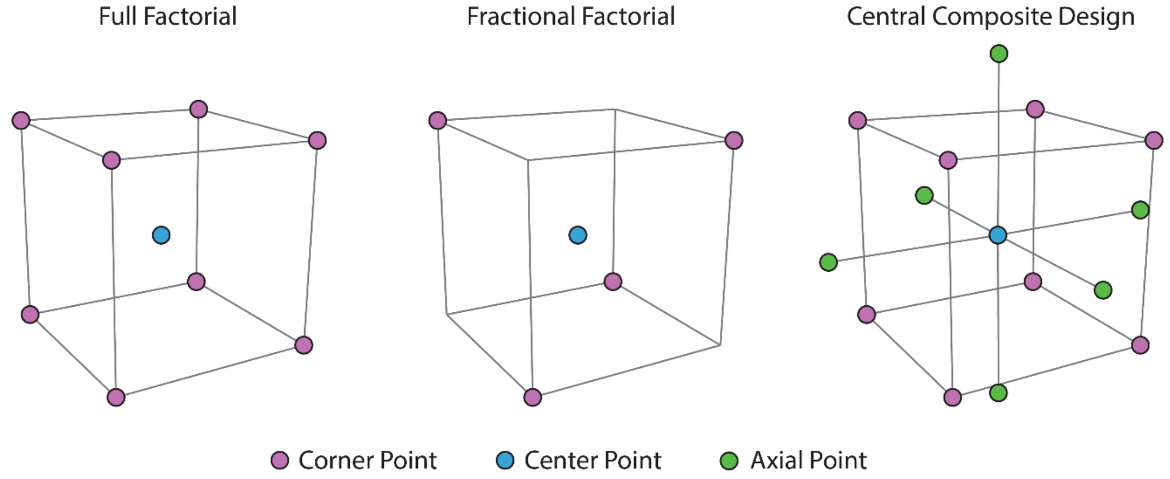

Figure 2. Examples of classical experimental designs used for live testing

The same reasons that make Classical DOE appropriate for physical testing also apply to validation. The wide spectrum of validation techniques all require a robust estimate of the noise in the physical response variable. Classical DOE principles accurately estimate, making the comparison to simulation more powerful.

Computer experiments, on the other hand, are not as well known. The literature on computer experiments assume the outcome to be deterministic. Therefore, they do not worry about random variation, replication, or practical considerations about hard-to-change factors. As a result design points in simulation experiments are free to spread out across the design space. Space filling designs are the general class of experimental designs for computer experiments. As the name suggests, the purpose of these experiments is to fill the space. Space filling designs maximize the likelihood of identifying problems with the model or simulation by filling in the space, allowing for the discovery of local maxima/minima or non-linearity in the code.

Figure 3 shows examples of common space filling designs. The simplest way to fill the design space is to create a uniform grid. Though these grids achieve good spacing, a disadvantage is that the number of design points increases exponentially as the number of factors and levels increases. Minimax and maximin designs offer a solution to this problem by creating an algorithm that optimizes design point spacing for any given number of design points, and for any given number of factors. A disadvantage of the minimax and maximin designs is that their projections onto subspaces are not good. For example, if we project the 16 design points in Figure 3 on the horizontal axis, then we will only get 11 distinct points for the maximin design. This means that if the vertical-axis variable does not affect the output, then five runs (or points) of the design would be wasted because they do not provide any additional information over the other 11 runs.

Latin hyper cube designs provide a solution to this problem by ensuring good projection to smaller subspaces. For example, projecting the 16 design points on the right side of Figure 3 on the x-axis results in 16 unique levels. Similarly, projecting the design points on the y-axis results in 16 unique levels. Latin hypercube designs maintain good spacing and flexibility as the minimax and maximin, and gain good projection, making them the most popular space filling design for simulation experiments with continuous inputs.

Fast Flexible Filling (FFF) designs are the most versatile of space filling designs. They are the only space filling design that allows for categorical variables and therefore, extremely useful in defense applications. Lekivetz and Jones (2015) show how they can be used to fill non-regular design regions, which is another useful property. They are generated by a simple algorithmic process: 1) randomly generate n>>N data points, where N is the final desired design size, 2) cluster the n data points into N clusters using your favorite clustering technique, and 3) Use each cluster to form a design point using a summary statistic (for example the centroid of the cluster). Variations on the exact design and speed at which the designs are generated vary based on the clustering method and summary statistic selected.

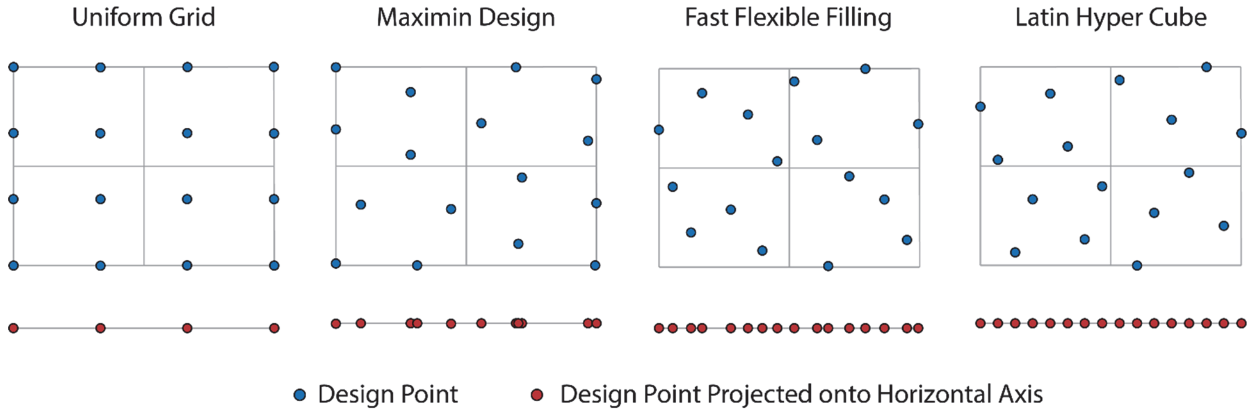

Figure 3. Common space filling designs used for computer experiments

Hybrid Approaches

In practice, a combination of classical and computer experiments has proven useful in the validation of defense models and simulations. Using hybrid approaches provides a strategy for two challenges with simulation experiments: 1) non-deterministic simulations and 2) matching points between live simulation experiments.

The computer experiments literature assumes that computer models are deterministic. Despite the prevalence of this assumption, there are many instances in defense testing where the response variable of a simulation is non-deterministic. Human-in-the-loop or hardware-in-the-loop simulations introduce a physical component and thus random variation. Examples include cockpit simulators, or simulations that incorporate some of the system's physical electronics.

The presence of random variation in the simulation response may lend itself to a classical DOE solution. However, in some cases there may still be a desire to use space-filling designs to cover more of the simulation space and gain a better understanding of the simulation response surface. In these cases, simulation designs paired with classical designs provide a straightforward strategy. The simulation design can cover the simulation input domain, and a classical design is used for selecting replicate points and/or matching points for live tests.

Direct matching of points between the live tests and simulation experiments provides the best validation strategy in all cases. If the subset of matched points are based on a classically designed experiment, the validation strategy can include regression analysis, which is the most powerful validation approach. If the matched data do not come from a designed experiment, statistical comparison is possible by aggregating the comparison across conditions.

The best design for the simulation experiment and live tests depends on the analytical goal and the nature of the simulation and the data it produces. Statistical designs should support both comparison with live data and exploration of the model space itself, including conducting sensitivity analyses and building emulators. Statistical measures of merit can help determine an adequate design and sample size, but the decision will typically be driven by the chosen effect size and amount of uncertainty that is acceptable. Power and confidence are directly related to the amount of uncertainty; as type I and type II errors are reduced, confidence interval widths also get smaller.

For completely deterministic simulations, such as finite element models, space filling designs are the recommended approach for both comparison and model exploration. On the other end of the spectrum, highly stochastic models, such as effects-based models, operator-in-the-loop simulations, or system of system models, classical designs are the recommended approach for both goals. For simulations somewhere in the middle in terms of randomness, such as a physics-based model with some built in Monte Carlo (random draw) input variables, a hybrid approach may be appropriate.

Summary

Advances in computational power have allowed both greater fidelity and more extensive use of such models. Numerous complex military systems have a corresponding models that simulate its performance in the field. In response, the DoD needs defensible practices for validating these models. DOE and statistical analysis techniques are the foundational building blocks for validating the use of computer models and quantifying uncertainty in that validation. Recent developments in uncertainty quantification have the potential to benefit the DoD in using modeling and simulation to inform operational evaluations. See Smith 2013 for a comprehensive review of uncertainty quantification. DOE is the first step toward a sound methodology for data-driven validation.

The recommendations included in this article come from a handbook on statistical methods for validation that the Institute for Defense Analyses has developed in support of DOT&E's initiatives for defensible model validation. We welcome feedback on the recommendations. Please contact Kelly Avery (kavery@ida.org) for more information on the handbook.

References

National Academy of Sciences Report (ISBN 0-309-06551-8), "Statistics, Testing, and Defense Acquisition, New Approaches and Methodological Improvements," 2012.

Guidance on the Validation of Models and Simulation used in Operational Test and Live Fire Assessments, March 2016, http://www.dote.osd.mil/guidance.html

Clarifications on Guidance on the Validation of Models and Simulation used in Operational Test and Live Fire Assessments, January 2017, http://www.dote.osd.mil/guidance.html

Modeling and Simulations, June 2002, http://www.dote.osd.mil/guidance.html

Johnson, R.T., Hutto, G., Simpson, J.R., Montgomery, Designed experiments for the defense community, Quality Engineering, 24 (2012) 60-79.

Lekivetz, R., & Jones, B. (2015). Fast flexible space-filling designs for nonrectangular regions. Quality and Reliability Engineering International, 31(5), 829-837.

Smith, R. C. (2013). Uncertainty quantification: theory, implementation, and applications (Vol. 12). Siam.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)* 01-2019	**2. REPORT TYPE** IDA Publication	**3. DATES COVERED** *(From - To)* 1/2/2019 – 1/24/2019

4. TITLE ANDSUBTITLE
Designing Experiments for Model Validation – The Foundations for Uncertainty Quantification

5a. CONTRACT NUMBER
HQ0034-14-D-0001

5b. GRANT NUMBER
____ ____ ____

5c. PROGRAM ELEMENT NUMBER
____ ____ ____

6. AUTHOR(S)
Laura J. Freeman (OED); Kelly M. Avery (OED); Thomas H. Johnson (OED);

5d. PROJECT NUMBER
BD-09-2299

5e. TASK NUMBER
229990

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-10456-NS
H 2019-000025

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

Director, Operational Test and Evaluation, 3E1088
The Pentagon
Washington, DC 20301

10. SPONSOR/MONITOR'S ACRONYM(S)
DOT&E

11. SPONSOR/MONITOR'S REPORT NUMBER
____ ____ ____

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
____ ____ ____

14. ABSTRACT
Advances in computational power have allowed both greater fidelity and more extensive use of such models. Almost all complex military systems have a corresponding model that simulates its performance in the field. In response, the DoD needs defensible practices for validating these models. Design of experiments and statistical analysis techniques are the foundational building blocks for validating the use of computer models for test and evaluation. Additionally, recent developments in uncertainty quantification have the potential to benefit the DoD in using modeling and simulation to inform operational evaluations. DOE is the first step forward towards a sound methodology for data-driven validation.

15. SUBJECT TERMS
Computer Experiments, Design of Experiments, Model Assessment and Validation, Uncertainty Quantification

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** Heather Wojton (OED)
a. REPORT Unclassified	**b. ABSTRACT** Unclassified	**c. THIS PAGE** Unclassified	Unlimited	15	**19b. TELEPHONE NUMBER** *(include area code)* (703) 845-6811