Using Design of Experiments to Improve Analyses, Simulations, and Cost

Steven Gordon, Ph.D.; Karen Dillard, Ph.D.
Georgia Tech Research Institute
Orlando, Florida; Atlanta, Georgia
Steven.Gordon@gtri.gatech.edu; Karen.Dillard@gtri.gatech.edu

ABSTRACT

The Department of Defense (DoD) is evaluating ways to accelerate acquisition and test and evaluation programs in order to field more effective weapon and training systems sooner. Nations that were near-peers are fast becoming peers, with capabilities in some areas outpacing those of the United States. We can regain our advantages in several ways, including improving weapon system's effectiveness and improving combat training. For weapon systems, we can accelerate the model-test-model process and provide early system prototypes for warfighter use in simulation and gaming environments. For combat training, we can improve validation of key training models.

An analysis and modeling method called Design of Experiments (DOE) has shown great promise to quickly model early prototype weapon systems, enable modeling for tradespace and requirements analyses, and to pinpoint designs that are precisely on target. DOE is, by policy, used wherever possible in DOD operational testing to assist in the evaluation of weapon systems and to improve the precision of weapons. Researchers in DoD and in the Department of Homeland Security are using DOE to augment Cyber Security teams evaluating system vulnerabilities in order to fully cover the threat landscape with fewer personnel. Also, for key models used in virtual and constructive training and in analysis, DOE is used to conduct validation of models compared to live test results.

This first known paper on DOE at I/ITSEC will discuss the methods used to develop DOE models and the simulations of those models. Several examples of DOE modeling will be presented. The value of using DOE will be discussed in an example of a Navy radar that was tested originally without DOE and then tested years later by the same Test Director, using DOE, but with 10% of the original resources – very high cost and time avoidance.

ABOUT THE AUTHORS

Steven Gordon is the Orlando Field Office Manager and a Principal Research Engineer for Georgia Tech Research Institute. He served 26 years in the United States Air Force with tours as an F-111 Weapons Systems Officer, Instructor, and Wing Electronic Warfare Officer; Air Staff Division Chief; 13th Air Force Director of Operations and Air Operations Center Director; and Air Force Academy Department of Mathematics Professor and Head. He also served as the first Technical Director for the Air Force Agency for Modeling and Simulation. Dr. Gordon has a Bachelor's Degree in Mathematics (Marymount); Master's Degrees in Education (Peabody/Vanderbilt), Industrial Engineering/Operations Research (Purdue), and in Business (Florida); and a PhD in Aero and Astro Engineering (Purdue). His research interests include return on investment for simulation-based training, tradespace tools for training systems, statistical techniques for test and evaluation, and decision support tools for military operations.

Karen Dillard is a Principal Research Scientist in the Advanced Concepts Laboratory at GTRI. Her interests include numerical analysis, optimization, experimental design, and test & evaluation. She retired from the United States Air Force after 22 years of service. As a program manager and scientific analyst, she has experience in undergraduate/graduate mathematics instruction and research, leadership development training, analysis of alternatives, and test & evaluation and tactics development of cyber/electronic warfare/surveillance/energy security technologies for operational use. She received her Ph.D. in Applied Mathematics from North Carolina State University, her M.S. degree in Applied Mathematics from University of Massachusetts – Lowell, and her B.S. degree in Mathematics from Rensselaer Polytechnic Institute.

Using Design of Experiments to Improve Analyses, Simulations, and Cost

Steven Gordon, Ph.D.; Karen Dillard, Ph.D.
Georgia Tech Research Institute
Orlando, Florida; Atlanta, Georgia
Steve.Gordon@gtri.gatech.edu; Karen.Dillard@gtri.gatech.edu

INTRODUCTION

Design of Experiments (DOE), sometimes called Experimental Design, is a mathematical/statistical method used to model complex systems, processes, or products. DOE was developed as a very efficient modeling method to capture the behavior of a system and to explain the system's behavior via equations and graphs. DOE derives models using a minimal set of experimental points to adequately capture the behavior of the system. The selected data points can be incrementally supplemented, if necessary, to produce a better model or confirm the model at additional points. In this way, DOE methods allow modelers to determine when sufficient data has been collected. DOE-developed equations relate the outputs (responses) of a system being modeled to the inputs (factors) of the system. By evaluating the equations derived from DOE modeling, characteristics of the system can be determined anywhere in the region modeled. Because DOE is a straightforward modeling method using a minimal set of data points, it can be repeated quickly anytime the system changes; the consequences of the changes will be evident in the derived equations and statistical measures.

The DOE equations can be used to determine the average outputs and the variability of the outputs anywhere within the region being modeled. Statistical measures provided in the DOE solution are indicators of how strong the relationships are and where more data may be needed. The DOE solutions also allow simplified sensitivity analysis to pinpoint which inputs and interactions most significantly affect the outputs. Modeling the variability of the outputs, when sample sizes are large enough, is a unique benefit of DOE. Variability in outputs, in the quality of consumer goods, and in military systems designed to hit targets is something that we would generally like to minimize. The derived polynomial equations are easily used to simulate and calculate the outputs and their variability across the modeled region in order to characterize how the system performs. More importantly, most DOE software can be directed to optimize the system of equations to find minimums and maximums or to hit targets (for quality or military purposes) with minimal standard deviation (minimal variability).

The desire to explain how DOE may be useful to simulation professionals was the motivating factor for developing this paper. When researchers want to characterize or assess a system or evaluate or justify improvements that are most effective, DOE is a method that can be used, in a straightforward way, to justify the claims. In this paper, we discuss the history, basics, limitations, benefits, and selected use cases for DOE. We also discuss the use of DOE modeling in live or simulated environments. Then, we discuss ways DOE is used and will be used to optimize products, characterize systems, evaluate alternatives, and validate models. The number of use cases for DOE are increasing due to innovation. We have limited the discussions of the use cases to motivations, diagrams, and/or general results in order to include more numerical examples describing why and how DOE is used and the results of its use.

HISTORY OF DOE

The theory of DOE was developed in the early 1900s by Ronald Fischer[1] for use in developing better treatments in agriculture. He wanted to replace the cumbersome, time-consuming methods at the time with a more efficient and accurate way to discover the best agricultural methods (Encyclopedia Britannica, 2013). His work included the theory of randomization to reduce the biased effects of unmodeled inputs to a system (School of Mathematics and Statistics, 2013). His DOE theory leveraged regression or curve-fitting methods and the science of statistics. Sir Ronald Fischer wanted to determine the optimal treatments needed to produce the highest quality crops or other agricultural products

[1]The first textbook that specifically focused on DOE was written by Sir Ronald Aylmer Fischer (Knighted in 1952) and was published in 1935 as The Design of Experiments. His previous book, Statistical Methods for Research Workers (published in 1925), had established the basis of the design and analysis of experiments and randomization to minimize the impact of lurking (unmodeled) inputs.

in the most cost effective way. At that time, solutions did not permit any insight into treatment interactions, nor could the treatments be optimized. Through the discovery and use of DOE, these concerns were resolved, and DOE use in agriculture increased worldwide.

INNOVATIVE USES FOR DOE

The simplicity, flexibility, and ability of DOE to efficiently deliver a great deal of information about a system is now proving valuable to many other industries and use cases. The variety of applications of DOE has steadily increased because DOE provides a means to model inputs and outputs of a system (including product manufacturing) across the operational envelope and to pinpoint the inputs that most affect the variability in product quality. DOE provides equations that are used to adjust the inputs so the outputs are on target consistently (with minimal variability). In addition to agriculture, DOE is also used extensively in consumer product development and manufacturing, Department of Defense (DoD) test and evaluation (T&E), systems engineering, cyber security analyses to cover the threat envelope with fewer personnel, and validation of models in virtual and constructive simulations. These use cases are discussed briefly in the bullets listed below:

- Based on DOE use in agriculture, it was realized that DOE modeling of outputs and the variability of the outputs could be used to improve the production processes of consumer goods. Variability in product quality, even in automated factories, was causing a lack of consumer satisfaction and low brand loyalty. DOE modeling of the average and the standard deviation of the key attributes to quality allowed manipulation of the inputs in order to hit targeted average output, and DOE was used to minimize variability in the keys to quality. Modeling methods that pinpoint the causes of variability will always be useful in production of high quality products and, by the way, in fielding of high quality, on-target military systems.
- Systems engineering was known as a valuable process, but it was also noted for time-consuming planning, paper documents, many presentation charts, and lengthy decision processes. Computing helped, but tradespace meetings where attributes of various alternatives would be analyzed, depicted in viewgraphs, and deliberated over many sequential meetings were still too time consuming. The use of "surrogate models" developed using DOE allows government decision-makers to make tradespace decisions in one sitting. A unique attribute of DOE is that the process of modeling in DOE software also codes all the inputs to common units and thus scrubs much of the proprietary information from the alternatives, allowing the government to conduct analysis of alternatives in a fair, nonproprietary systems engineering setting. Engineers supporting programs that needed to conduct tradespace studies or evaluations of alternatives have often been frustrated by lack of available models or simulations that could be used to conduct those trades (Bruni, 2014). DOE provides a simple way to model a system frequently as the design changes or to model average outputs and variability of outputs from simulation.
- Methodologies like model-based systems engineering (MBSE), Digital Engineering (DE), and Virtual Prototyping (VP) are being used to improve systems engineering processes. These processes can accelerate acquisition and improve the functionality and quality of products. MBSE develops and executes a model in simulation as a means of gathering confirmatory information on the operation of the modeled system and the interrelationships with supporting systems. One part of the DE initiative is to promote model-centric interaction between industry and government, and this interaction includes data exchange of model details, functionality, and interaction with its environment. DE will enable the use of digital models throughout the system lifecycle to represent the system of interest and its dependencies. VP allows early operator use of a proposed product or system in a representative simulation environment. Warfighters and decision makers can decide if this system or a variant of the system can make a difference in operational use. MBSE, DE, or VP may use DOE models because the DOE models can be developed quickly from analysis of the gathered data from operational prototypes or simulated systems. DE may use coded and uncoded models of the system (discussed later in this paper) to carry information of the system to different audiences.
- By using DOE in T&E, government and industry T&E professionals can estimate sample sizes and data requirements prior to tests. The methods of previous tests (that tested a few times at many/hundreds of locations) can be replaced by using DOE designs that test more times at fewer points, based on intelligent selection of test locations and statistical power calculations. DOE provides more information, including equations for outputs and the variability of outputs, sensitivity analysis to determine the most important inputs, and, often, a reduction in

test costs. Measuring and knowing the variability in systems like autopilots, weapons, and electronic warfare countermeasures is also vitally important for success in combat.

- DOE is now used in cyber vulnerability evaluations. Testing systems for cyber security is time consuming, requires a significant amount of expert labor, and must be periodically repeated. Consequently, some organizations have developed DOE methods to assist the experts in testing (or even continuous testing) for cyber security. Use of DOE for cyber security testing helps satisfy the increasing need for this testing with fewer cyber experts (Kim et al., 2018). The DOE designs facilitate the ability to expand test coverage, compensate for the shortage of cyber security experts, and increase the return on investment because more balanced, space-filling testing can be completed using DOE. DOE designs can be used to span the space using a mathematically-derived coverage plan to evaluate all, or most, known vulnerability approaches in one design. Using DOE, cyber experts can leverage, supervise, and review the semi-automated search-and-detect evaluations. Designs use inputs such as computer ports, operating systems, type system (server/storage/network), and other potential vulnerability types. Other cyber security experts are using DOE to test effectiveness (performance and interoperability), suitability (availability, throughput, and usability), and security (denial, degradation, manipulation, exfiltration, and pivoting) across mission areas and tasks (Hoover & Wells, 2018).
- Modeling and simulation (M&S) is a critical resource for training, rehearsal, experimentation, wargaming, and T&E. M&S can portray future systems and environments, and M&S will be increasingly used for T&E of advanced systems because M&S can portray the complexity of systems of systems and combat environments. These environments may only exist otherwise in combat, and the use of M&S facilitates multiple replays for trial and error, sufficient sampling, and repeated practice away from view of unwanted spectators. Simulations and simulators need to be validated for T&E, especially for advanced programs (Freeman, 2018). The Handbook on Statistical Design & Analysis Techniques for Modeling & Simulation Validation (IDA, 2019) was sponsored by the Director of Operational Test and Evaluation (DOT&E). The guidance in the Handbook includes use of DOE, comparison of live and M&S data, and hypothesis testing for accuracy determination. The differences between simulation and live data may not matter in some cases, and for some uses, and validation may stipulate areas where the simulation is good enough for locating trends or capturing key characteristics (Freeman, 2019). In other cases, M&S tools must be upgraded in accuracy to allow M&S use in testing, analysis, and advanced systems training.

This use case section summarized how DOE was repurposed from the original discovery and how DOE has been used for other tasks. In the next sections, we will briefly cover the basic DOE process, show some examples with numerical results, and then we end this paper with a more in-depth look at current research in validation of simulation.

BASIC DESIGN OF EXPERIMENTS PROCESS

The application of DOE begins with an upfront analysis, including decisions about outputs that need to be modeled and inputs that might affect the outputs. Prior knowledge of the system being modeled is a great benefit, but DOE procedures also allow for discovery of the significant inputs. Depending on the textbook used, the full DOE process has approximately 10 steps in a standard process for planning a designed experiment. The key steps for use of DOE are also covered on the DOT&E web site (DOT&E, 2019). The purpose of DOE is to characterize and better understand how the outputs and their standard deviations (variability) change when the inputs are varied. The design method of varying the input levels in a structured way provides a great deal of information to help gain as much knowledge as possible about the outputs and their variability relative to the changes in the inputs. We first describe a generic example of a use case with an input-output flow diagram as depicted in Figure 1.

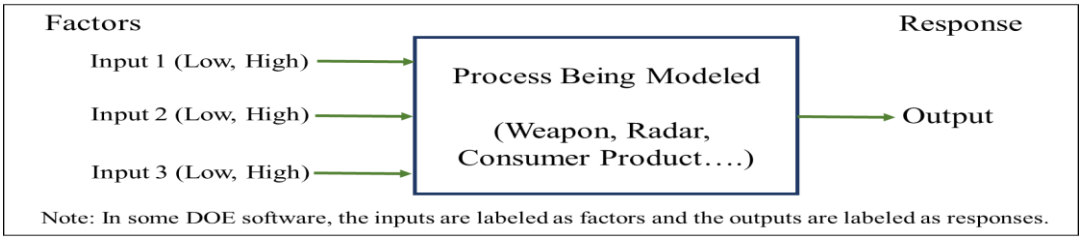

Figure 1. DOE Flow Diagram for Generic 3-Input, 1-Output Linear (Two-Level) Design

The number of inputs and the number of levels of the inputs affect the size and complexity of the design. The number of outputs does not affect the size of the design, but it does require reliable measurement of each output. We seek to model the key inputs that might affect the outputs. In the example in Figure 1, we would gather data to determine how the output and the variability of the output are affected by the inputs. Based on historical information or other evidence, one may decide that the inputs should be at 3 or more levels, and this would indicate that we are looking for quadratic (second-order) or higher order terms in the output equations. These higher order terms would likely cause the response surface graphs to show curvature, such as concave up or concave down regions. Using higher numbers of levels for the factors or adding more factors complicates the model, but there are alternative designs that moderate the complexity.

As a first numerical example in Figure 2, we use a classroom example where students model a Statapult[2] using 2 levels for each of the 3 factors: Tension Pin (TP), Stop Pin (SP), and Pull Back (PB). Because this experiment is at 2 levels, we are looking for a linear model with some interactions.

Factor:	A	B	C	Shot Distance (Inches)			
Row #	Tension Pin (TP)	Stop Pin (SP)	Pull Back (PB)	Y1	Y2	Average	Standard Deviation
1	2	4	160	55	53	54	1.41
2	2	4	180	80	80.5	80.25	0.35
3	2	5	160	29	30	29.5	0.71
4	2	5	180	60	60	60	0
5	3	4	160	65	64.5	64.75	0.35
6	3	4	180	100	98	99	1.41
7	3	5	160	34	33	33.5	0.71
8	3	5	180	80.5	80	80.25	0.35

Figure 2. Data Entered into DOE Design (3 Inputs at 2 Levels for an 8-Run Design)

We ask the software to analyze the data (it uses a form of least squares regression), and the software provides an equation for the average shot distance in inches and for the standard deviation of shot distance in inches. The DOE solution provides statistical metrics that show the importance of each coefficient and the strength of the regression relationship, but, due to page constraints, these statistical tables are not shown here. Even for very complex models, the regression solution is delivered quickly. For the software[3] used for this paper, the data is entered as depicted in Figure 2. However, in the analysis, the inputs are automatically "coded" (coordinates transformed) to -1 for the low setting and +1 for the high setting for the inputs. This coding keeps all the inputs on the same scale and permits more accurate statistical evaluations. We can determine the sensitivity analysis by noting the size of the coefficients (in absolute value terms) in each of the equations shown in Figure 3 below. Sensitivity analysis determines which inputs most or least affect the outputs. For the shot distance in Figure 3, the PB input has the most impact on shot distance. For the standard deviation equation, the interaction of Tension Pin and Pull Back has the most impact of the standard deviation (variability) in the system. These two equations, for the region modeled, allow us to target any distance between approximately 30 inches to 99 inches to hit a target; and, for that target distance, we would also know the calculated standard deviation for that shot. The "carrots" or hats above the Y and the S in the equation below indicate the equations are derived from least squares regression curve-fitting, and the equations are treated as averages. The first equations provided are coded, so the highest any input can be is +1 and lowest value for any input is -1.

$$\hat{Y} = \text{Shot Distance} = 62.56 + 6.72\,TP - 11.84\,SP + 17.22\,PB + 3.03\,TP{*}PB + 2.09\,SP{*}PB \text{ inches}$$

$$\hat{S} = \text{Standard Deviation of Shot Distance} = .66 + .04\,TP - .22\,SP - .13\,PB + .31\,TP{*}PB \text{ inches}$$

Figure 3. Equations for the Average Shot Distance and the Standard Deviation

Since there was one response in the design in Figure 3, there are only two equations derived – one equation for the average response (shot distance) and a second equation for the standard deviation (variability) of the shot distance. If we had measured two responses, there would have been four derived equations. Once the analysis of the derived equations is completed, the user can ask the software to uncode the equations to original units of measure for the inputs, and these uncoded equations can be used in publications and customer deliverables. One of the steps in the

[2]Statapult® Catapult is a registered trademark of Air Academy Associates. Statapults are available from many sources.

[3]Design of Experiments (DOE) Pro XL, www.SigmaZone.com

DOE process and for this example is to determine how much data to collect for each input, and this is based on a statistical power calculation. DOE software facilitates this power calculation, and the typical standard is to have a sample size that provides a statistical power of approximately .80 or 80% power (a probability of .8 for finding the desired shifts in the output). Knowing statistical power allows experimenters to right-size the experiment and not use too many or too few resources (Montgomery, 2013). Statistical power is dependent on several settings in the experiment, some able to be set by experimenters and one (estimated noise) is part of the outcome of the experiment.

For the classroom example in Figure 2, we calculated the statistical power before we conducted the experiment, and that power calculation helped us determine that 2 Statapult shots for each run/row in the design was enough. We defer the detailed discussion of statistical power to a later paper, but DOE software calculates estimated power based, roughly, on the statistical confidence you have in your model, the number of rows in the design, the number of times the experiment is conducted for each row, the shift in the output that is important to know, and the noise (uncertainty/variability) expected in the output. Before the Statapult shots were taken, we entered the answers to those questions, and the calculated statistical power was 79.2% -- which was judged close enough to 80% statistical power to be acceptable. We could increase statistical power by increasing the sample size (more than 2 shots per row or adding more rows), decreasing the noise or standard deviation in the Statapult shots (discipline in standard procedures), only caring about larger shifts in the output, or decreasing the statistical confidence in our results.

This academic example allowed a more complete explanation of the DOE process and results. Typical use cases for DOE are often sensitive due to classification or the proprietary nature of the purpose or results. Examples used in the remaining portion of this paper are truncated to protect sensitivities.

DOE DESIGN EXAMPLES AND SOLUTION STEPS

This first example is a DoD radar test where the test results are summarized. The radar was tested prior to fielding using traditional (not DOE) methods in the 1990s and then again in 2012 using DOE. The two tests were managed by the same Test Director (Ahner & Cortes, 2013). The original test was planned to have 96 combinations of inputs and to test each combination 30 times. The plan included 80 hours of testing, 18 hours of live aircraft raids, 110 electronic attack techniques, and 1900 simulated anti-ship cruise missile launches. The original test required 2880 test runs but many were modified for redirected tests, so 670 data points were eventually collected. The pass/fail criteria for the original test was based on checking if the median from the tested points at each location was greater than the requirements threshold, and if so, the system passed at that point. Some odd behaviors of the radar were noticed but not explained by the original testing. In 2012, the DOE solution helped explain the odd behaviors and provided equations for the outputs. In the table below, the original radar testing is compared to the subsequent testing using DOE. Clearly, DOE provided much more information with less than 10% of the resources.

Key Strengths	Original Radar Evaluation	Radar Evaluation Using DOE
Number of Total Runs	670	64
Statistical Power Calculated?	No	Yes
Equations Developed?	No	Yes
Pass/Fail Criteria	Median > Requirement at Tested Points	Equations for Operational Region to Determine if > Requirement
Anomalies Identified?	No	Yes
Reused at Other Points?	No	Yes, if in Region Modeled

Figure 4. Original Radar Evaluation Compared to Evaluation Using DOE

A more recent DOE example from 2019 (Kiemele, 2019) illustrates a similar trend to that depicted in the example above in Figure 4. A company wanted to test a new system and conducted 7105 flight tests at many input settings. Because these tests were not conducted in a specific pattern of input setting levels, the equations relating the inputs to the outputs of the 7105 flight tests were highly correlated. Regression analysis or DOE software could not determine what inputs most affected the outputs. Correlation tends to hide the impact of any particular input on the output. In a subsequent gathering of data using selected arrangements of input settings in a 16-row DOE design, with 3 flight tests for every row (48 total flight tests), the company was able to derive equations, conduct sensitivity analysis, and determine the uncorrelated effects for each of 5 inputs on the output and the standard deviation of the output. This

example from 2018 and 2019 illustrates that use of the DOE process can provide more information from just 48 test flights compared to the strongly correlated data from 7105 test flights conducted by the same team – illustrating acceleration of the analysis and a healthy return on investment. DOE encourages careful selection of test points and efficient and effective selection of sample sizes. The DOE designs are developed to deliver uncorrelated coefficients for the inputs and interactions where possible. Well planned selection of test points could have saved 7057 flight tests.

The next example in Figure 5 is an evaluation of how two inputs (electronic countermeasures and target speed) can affect the detection range of a sensor (DOT&E, 2013). The output measured is detection range of the sensor in miles. The two inputs are speed of the target (slow or high) and countermeasure use (off or on). For this example, the software asked (during data entry) what coding we wanted to use; in this case, we declared that Target Speed Slow was -1 and Target Speed High was +1. We also declared Countermeasures Off (None) was -1 and Countermeasures On was +1. Absent sensitivities, the factors would be expressed in actual quantitative lows and highs (vice just "Slow" and "Fast" or "None" and "On") to enable improved optimization.

Factor	A	B	Detection Range				
Row #	Target Speed	Countermeasures	Y1	Y2	Y3	Y bar	S
1	Slow	None	4.9	6.4	7.5	6.27	1.31
2	Slow	On	0.2	1.7	2.2	1.37	1.04
3	Fast	None	3.2	3.8	5.1	4.01	0.92
4	Fast	On	2.1	3.4	4.1	3.21	1.01

Figure 5. Electronic Warfare Countermeasures and Target Speed Effects on Detection Range

The analysis of this data using DOE software provides equations for the average detection range and standard deviation of detection range in Figure 6 below. In the first analysis from DOE software, the inputs are coded from -1 to +1. So, the largest value for any input in these equations is +1. Once the DOE solution is provided, we can evaluate the coefficient sizes and the statistical metrics to see how we can simplify equations, determine sensitivity analysis, and calculate detection range and standard deviation of detection range anywhere in the region modeled. Typically, users of DOE select designs that are "orthogonal" so that the derived coefficients for the inputs in the coded output are independent relative to the other coefficients. Due to this mathematical independence, the relative importance of the inputs can be determined by their absolute value size relative to the other coefficients (Loper, 2015). For this example, we will not ask the computer to uncode our data, and we will work with the coded equations. Note the equations have coefficients for Target Speed, Countermeasures, and the 2-way interaction of Target Speed and Countermeasures.

$$\hat{Y} = \text{Detection Range} = 3.71 - .11\ \text{Target Speed} - 1.43\ \text{Countermeasures} + 1.03\ \text{Target Speed-Countermeasures}$$

$$\hat{S} = \text{Standard Deviation of Detection Range} = 1.07 - .10\ \text{Target Speed} - .04\ \text{Countermeasures} + .09\ \text{Target Speed-Countermeasures}$$

Figure 6. DOE-Derived Coded Equations for Detection Range in Terms of Countermeasures Use and Target Speed

Looking at the coded equations in Figure 6 (or asking the software to optimize), we can determine that the maximum detection range is when the Speed of the Target is low (-1) and the Countermeasures are off (-1). At that point, the Detection Range would be 6.28 miles, and the standard deviation of the distance to detect at those same settings would be 1.30 miles. Because these equations are coded, it would also be easy to locate the inputs that would provide the lowest standard deviation (variability) in detection range. That setting would be high (+1) for Target Speed and off (-1) for Countermeasures, where the standard deviation is .92 miles. For very complex models with many equations, the DOE software can be used in this same way to rapidly find maximums and minimums and find locations where a target can be engaged with the lowest standard deviation (tightest shot package).

Some designs use fewer experimental points because they do not intend to model all interactions (Box, et al., 1978). These designs could be selected to not model specific interactions. For instance, the examples used so far in this paper have not included quadratic interactions that capture concave up or concave down trends in the data. Models that are intended to locate quadratic effects are used when those quadratic effects are expected because of the nature of the system they are modeling or because of prior information. The quadratic effects can be captured by modeling a system at greater than 2 levels in the design. If researchers do not know what to expect, they can model the system as if it is linear, and then check the model with a row of data gathered at the middle of the region or at the middle of an edge of

the region. If strong quadratic effects are noticed, the linear design can be fully augmented with a center row of data and all center edge rows. The simplest adjustment that can be made to a linear model with 4 rows is to add a 5th row of center point shots in order to investigate or capture some of the quadratic (concave up or concave down) effects. This augmented design for a ground-launched munition is depicted in Figure 7 with the added center point row highlighted in yellow.

Factor	A	B	Payload Distance					
Row #	Engine Size	Angle	Y1	Y2	Y3	Y4	Y bar	S
1	10	55	59.25	59.13	61.38	59.88	59.91	1.03
2	10	75	86	86.06	81.5	82.19	83.94	2.43
3	20	55	34.75	39.13	39.25	34.06	36.81	2.77
4	20	75	75.38	75.81	77.31	75.25	75.94	0.95
5	15	65	73.2	78.7	76.25	77.1	76.31	2.31

Figure 7. Sample Data with an Added Center Point 5th Row

Once the 5th row is added to the design to capture a portion of the quadratic (concave up or concave down) effect, the DOE software is asked to model the new data and look for possible second-order (quadratic) effects. We asked for that analysis, and one additional **bolded** term in each equation in Figure 8 is discovered.

$\hat{Y}$ = Payload Distance = 75.84 – 7.78 Engine Size + 15.79 Angle **- 11.695 Engine Size^2** + 3.78 Engine Size-Angle

$\hat{S}$ = Standard Deviation of Payload Distance = 2.29 + .06 Engine Size – .11 Angle **- .49 Engine Size^2** – .81 Engine Size-Angle

Figure 8. DOE-Derived Equations for Delivery of a Payload Downrange with Added Center Point 5th Row

For the equations in Figure 8, DOE software used data from all 5 rows and calculated two equations that include the bolded second order (quadratic, concave down) terms. Quadratic effects often indicate the possibility of maximums and/or minimums for the output(s) inside the region modeled, and these can be valuable discoveries. For the 5-row design in Figure 7, with equations in Figure 8, the graphs for the output average and standard deviation are depicted below in Figure 9.

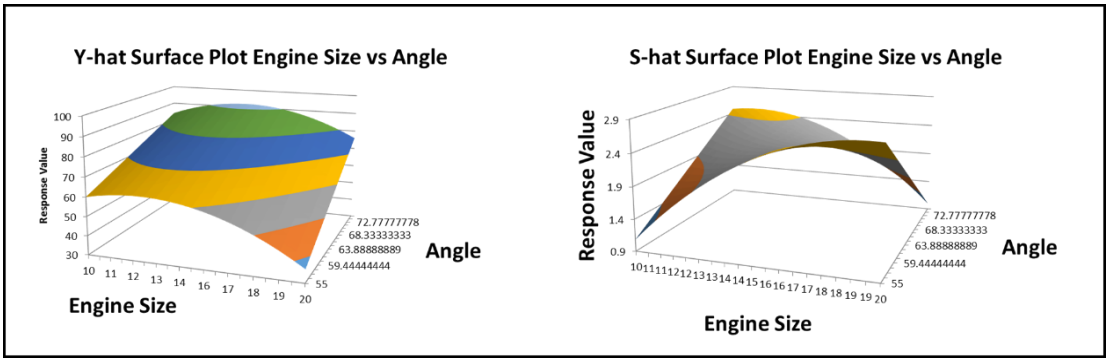

Figure 9. Graphs of DOE-Derived Equations for Average Response (Y-hat) and Standard Deviation (S-hat)

These two graphs indicate a significant concave down surface for both, and the experimenter in this case would likely also add the 4 center-edge rows to the design to capture the full quadratic (concave) characteristics at the edges of the region. If we did add all 5 center and center edge rows (the one already added plus 4 more rows at the center edges), this would more fully capture the quadratic effects near the edges. The original linear model was a 2-level, 2 factor design (2^2); the 9-run design would be a 3 level, 2 factor (3^2) design.

USING DOE TO VALIDATE A SIMULATION OF A COMPLEX MILITARY SYSTEM

The cost, time, and representative operational landscape required to operationally test and evaluate a complex military system in a live environment has become prohibitive. The use of M&S to support T&E can significantly contribute to the success of tests, if the community has confidence in that M&S. Validation of the M&S, a non-trivial process,

builds the community's confidence. A complex military system essentially fuses data from different systems and presents results to the operator in a meaningful way. The systems and connections represented in M&S must perform as the live military system performs in the known operational environment, so it will be trusted when applied to an operational scenario. The community requires a methodical, rigorous approach to validate the M&S necessary for T&E (Freeman, et al., 2018; Freeman, et al., 2019; IDA, 2019). One example of this approach is described below.

A five-step methodology is developed to compare simulation data to live test data of a complex military system with many system components, factors, and responses: 1) Use DOE to identify and scope the factors, responses, and runs; 2) Use historical live test data to characterize the military system; 3) Adjust the DOE based on the characterization and other considerations; 4) Compare live test data and simulation data; 5) Consider the operational significance of the detected lack of agreement to a mission in order to identify acceptable simulation performance.

Step 1: Build DOE design using operational and technical experts. Including technical input, modified by operational awareness, produces a DOE with specific factors/levels and responses deemed necessary for the M&S validation. Otherwise, the DOE could become intractable if all the potential factors/levels and responses were included.

Step 2: Characterize the system with live test data. During development of the military system, operators executed live tests, researchers analyzed data, and engineers improved the system. This live test data serves another purpose: to characterize the performance of the military system for comparison to the simulation. The simulation must perform as characterized from live test data (or "close" to it) to be validated.

Step 3: Adjust DOE after review of the characterization. Some factors/levels may not impact performance as expected or collection/calculation of some responses may not describe the performance as needed. Also, the DOE should be adjusted to factors/levels of the live test data actually collected. Since live test data was collected for a development objective, some runs were prioritized to answer specific development questions or to just demonstrate a particular capability, rather than meet DOE full-factorial cases. Cost and time are other considerations for adjusting the DOE.

Step 4: Apply statistical methods to compare live test data and simulation data. In this example, selected methods provide insight into the distribution, rank, mean/median, and variance differences between the datasets, as well as the practical impact of those differences. The latter leads directly to the crucial step of evaluating operational significance of the result.

Step 5: Consider operational significance, after results of statistical significance (i.e., difference between live and simulation data meets certain mathematical conditions). The term operational significance denotes that the difference between live and simulation data has practical or real consequences in the employment of the military system.

Additional discussion of statistical methods should be included here too. A variety of hypothesis tests determine statistical significance of differences in distribution, rank, mean/median, and variance between the datasets. Since the result is a probability, there is a chance, a risk, that the conclusion does not represent reality. In addition to these hypothesis tests, there are statistical measures which indicate the size of the effect. Descriptions of these effect size measurements can be found in the literature (Fritz, et al., 2012).

Because of complexity, many comparisons will be required in the data analysis. The use of simple techniques helps provide an initial screening through the process of many comparisons. Applying a distance measurement between the two sets of data has been shown to be a simple way to view "closeness" of the data. Initially, responses are compared one-by-one (univariate analysis). Multivariate analysis (comparison of correlated responses) may be extended for appropriate situations (Rebba & Mahadevan, 2006). The Mahalanobis distance, which is a known clustering metric, can be used in both univariate and multivariate analyses (Mahalanobis, 1936).

The squared Mahalanobis distance is distributed as the well-known Chi-squared distribution with the number of degrees of freedom equal to the dimension of the data, such as the number of different responses (Mardia, et al., 1979). As such, the distance calculated in validation may be tested against a Chi-squared probability. Fortunately, this has a real physical meaning: the resulting percentage from the Chi-squared test is the same as the percentage number of points in the cluster that are farther away from the mean of the cluster than the test point in question.

For example, an initial threshold for acceptance is taken to be a distance which has a Chi-squared probability of 0.5. This is interpreted to indicate that the simulation data point of interest is as close to the live data mean as 50% of the live data. Simulation data within this threshold can be interpreted to be close enough to the performance of the actual system under test. Results outside this threshold will undergo further review by an operational expert. If outside the threshold, but the difference is insignificant to the mission, the expert may document that simulation performance can be considered acceptable. If outside of the threshold and the difference could have a tactical impact on the mission, it will be noted that, for this measure of performance, the simulation differs from the live system in an operationally significant manner.

Lessons learned from challenges and solutions to issues advance the evolution of this validation process. Most lessons learned/issues stem from the need to post-process/analyze/compare data categorized by the DOE case. Test directors should prepare comprehensive test instructions prior to the test and ensure accurate note-taking of actual events during the test. Before the test, data analysts can plan data organization, in terms of data fields, naming conventions, etc. to identify data for each DOE case and assist in automated coding scripts. After the test, data analysts need to handle issues with time synch, coordinate systems, units, truth sources, and data retrieval.

The simulation can be improved with multiple iterations of this process. The investment in this process over time not only builds a crucial simulation for the operational test and evaluation phase of the military system, but also builds a future simulator for training the next generation of military operators. Training will continue to cost time and money; however, improving models for T&E can also improve training realism. DOE could be used to shape a more focused training plan with a balance of effective simulation and live training, perhaps even tailored on an individual basis.

CONCLUSION

This paper discusses the history and advantages of using DOE, describes applications of DOE, and demonstrates parts of the DOE process. DOE is an efficient and powerful method to model a system or process by developing equations that relate the inputs to the outputs across the region of interest. The derived equations can be used to calculate the output values for any set of input levels within the region used to develop the equations. This DOE capability can then be used to calculate maximum levels of the outputs and/or the minimum standard deviations (variability) of the outputs. The DOE process also gives researchers a way to justify the statement that a system, process, or method is optimum, most effective, or best in some other way. Some of the early use cases for DOE, such as agricultural treatment selections and improving the quality (and success) of consumer products, were briefly discussed. For systems engineering, DOE can be used to develop models of new and emerging systems by using small sample sizes, and these easily-modified models are used in tradespace and design evaluations. In DoD testing, DOE has enabled the test community to develop estimates for economical sample sizes needed in tests before the tests are conducted. It answers the question of how much testing is enough, correcting the historical trend of often testing too much and too long. And, the DOE solutions characterize the system throughout the tested region. Now, new uses for DOE are being developed. Currently, new efforts are finding innovative ways to use DOE to facilitate cyber testing of the entire space using fewer cyber security team members. DOE may also become an embedded continuous cyber testing tool. In order to use a mix of live and simulation results in tests, research is ongoing to use DOE to help to correct the differences between the live and simulation data. For live and virtual simulations for training, DOE use may indicate ways to improve simulations for training realism, thus improving the simulation to better represent the actual combat environment. Perhaps some readers of this paper may encounter tasks where DOE could be used to model (characterize) the systems of interest. If so, we hope this article has helped when that circumstance arises.

REFERENCES

Ahner, D., & Cortes, L. A. (2013). *Assessment of Radar Detection Performance Using Design of Experiments; A Case Study*, Retrieved December 7, 2017 via electronic mail from author

Box, G.E.P., & Hunter, W.G., & Hunter, J.S, (1978). *Statistics of Experimenters– An Introduction to Design, Data Analysis, and Model Building*, New York: John Wiley & Sons

Brockett, C. H., & Scott-Nash, S., & Pharmer, J. A., (2001). *Verifying and Validating the AEGIS Air Defense Warfare Human Performance Model*, retrieved from the I/ITSEC archives web site April 16, 2018

Bruni, S., & Riddle, K., & Ortiz, A., & Marshall H., & Gaughan, C., & Saffold, J., (2014). A Decision Aid for Optimizing Experimental Design Involving LVC Environments, paper 14139, retrieved from the I/ITSEC archives web site April 16, 2018

Curry, T. F., & Weese, D. L., & Senseny, P. E., (2007). *Using Experimental Design to Minimize Runs for Verification and Validation of Large-Scale Simulations*, copy received from lead offer

Director of Operational Test and Evaluation (DOT&E), (2019). Scientific Test and Analysis Techniques-Guidance, Retrieved May 1, 2019, https://www.dote.osd.mil/docs/TempGuide3/STAT_Guidance_3.0.pdf

DOT&E, (2017-DOE). *Design of Experiments – Guidance*. Retrieved April 16, 2018, from http://www.dote.osd.mil/docs/TempGuide3/DOE_0_Guidance_3.0.pdf

DOT&E, (2017-STAT). *STAT (Statistical Test and Analysis Techniques) - Guidance*. Retrieved April 16, 2018 from http://www.dote.osd.mil/docs/TempGuide3/STAT_Guidance_3.0.pdf

DOT&E, (2013). Test Sciences Roadmap, accessed 4/28/2019, Appendix 2-3, https://www.dote.osd.mil/pub/reports/20130711Appdxes2theTestScienceRoadmapReport.pdf

Elser, T., (2017). *Factorial Design*, North Charleston, South Carolina: CreateSpace

Encyclopedia Britannica, (2013). *Sir Ronald Aylmer Fisher, R*etrieved September 29, 2013 at www.britannica.com/EBchecked/topic/208658/Sir-Ronald-Alymer-Fisher

Freeman. L. (2018). Method (& Thoughts) Used to Validate/Accredit and Produce Accurate Data, April 10, 2018, Retrieved April, 17, 2018 from the 2018 M&S in T&E Technical Exchange Meeting web site

Freeman, Laura, Avery, Kelly, and Johnson, Thomas, (2019). Designing Experiments for Model Validation Test and Evaluation. *The ITEA Journal of Test and Evaluation*, 40-1, 10-15

Fritz, Catherine O., Morris, Peter E., & Richler, Jennifer J., (2012). Effect Size Estimates: Current Use, Calculations, and Interpretation. *Journal of Experimental Psychology: General*, 141, No. 1, 2-18

Hoover, A., & Wells, J. (2018). Mission Threat Analysis and Cyber Security, Retrieved April 22, 2018 from *http://www.itea.org/share/workshops/35-share/conferences/466-6th-cybersecurity-workshop-proccedings-2018.html*

IDA: Avery, Kelly M., Freeman, Laura J., Parry, Samuel H., Whittier, Gregory S., Johnson, Thomas H., and Flack, Andrew C. (2019). Handbook on Statistical Design & Analysis Techniques for Modeling and Simulation Validation, retrieved April 11, 2019 from IDA

Kim, M., & Kraus, P., & Mackertich, N., & Rogers, T. (2018). Statistical Test Optimization for Cyber Test, Retrieved April 22, 2018 from http://www.itea.org/share/workshops/35-share/conferences/466-6th-cybersecurity-workshop-proccedings-2018.html

Kiemele, M. (2019). Accelerating Knowledge Gain for Test Data, presented May 15, 2019 at the 23RD Test Instrumentation Workshop for the International Test and Evaluation Association (ITEA). Retrieved May 16, 2019 from http://www.itea.org/share/workshops.html

Mahalanobis, P. (1936). On the generalised distance in statistics. *Proceedings of the National Institute of Science of India, 12,* 49-55

Mardia, K., Kent, J., & Bibby, J. (1979). *Multivariate Analysis.* Academic Press

Montgomery, D. C. (2013). *Design and Analysis of Experiments, Eighth Edition*, Hoboken, New Jersey: John Wiley & Sons

Loper, M. L., *Editor* (2015). *Modeling and Simulation in the Systems Engineering Life Cycle*, Chapter 16, pages 187-200, Design of Experiments, London: Springer-Verlag

Rebba, R., & Mahadevan, S. (2006). Validation of models with multivariate output. *Reliability Engineering & System Safety*, 861-871

Schmidt, S.R., & Launsby, R.G., & Kiemele, M.J. (2004). *Understanding Industrial Designed Experiments, Fourth Edition*, Colorado Springs, Colorado: Air Academy Press

Schmorrow, D., & Cohn, J., & Bolton, A., (2006). *Inserting Science and Technology Into the Systems Acquisition Process*, paper 2937, retrieved April 16, 2018 from the I/ITSEC Paper Archives

School of Mathematics and Statistics (2013). *Sir Ronald Aylmer Fisher Biography*, University of St Andrews, Scotland, retrieved September 29, 2013 at www.history.mcs.st-andrews.ac.uk/Biographies/Fisher.html