Quality Engineering, 24:60–79, 2012

ISSN: 0898-2112 print/1532-4222 online
DOI: 10.1080/08982112.2012.627288

Taylor & Francis
Taylor & Francis Group

Designed Experiments for the Defense Community

Rachel T. Johnson[1], Gregory T. Hutto[2], James R. Simpson[2], Douglas C. Montgomery[3]

[1]Naval Postgraduate School, Monterey, California
[2]Eglin Air Force Base, Eglin, Florida
[3]Arizona State University, Tempe, Arizona

ABSTRACT The areas of application for design of experiments principles have evolved, mimicking the growth of U.S. industries over the last century, from agriculture to manufacturing to chemical and process industries to the services and government sectors. In addition, statistically based quality programs adopted by businesses morphed from total quality management to Six Sigma and, most recently, statistical engineering (see Hoerl and Snee 2010). The good news about these transformations is that each evolution contains more technical substance, embedding the methodologies as core competencies, and is less of a "program." Design of experiments is fundamental to statistical engineering and is receiving increased attention within large government agencies such as the National Aeronautics and Space Administration (NASA) and the Department of Defense. Because test policy is intended to shape test programs, numerous test agencies have experimented with policy wording since about 2001. The Director of Operational Test & Evaluation has recently (2010) published guidelines to mold test programs into a sequence of well-designed and statistically defensible experiments. Specifically, the guidelines require, for the first time, that test programs report statistical power as one proof of sound test design. This article presents the underlying tenents of design of experiments, as applied in the Department of Defense, focusing on factorial, fractional factorial, and response surface design and analyses. The concepts of statistical modeling and sequential experimentation are also emphasized. Military applications are presented for testing and evaluation of weapon system acquisition, including force-on-force tactics, weapons employment and maritime search, identification, and intercept.

KEYWORDS factorial design, optimal design, power, response surface methodology, space filling design, test and evaluation

Address correspondence to Rachel T. Johnson, 1411 Cunningham Rd., Naval Postgraduate School, Glasgow Hall–Rm 253, Monterey, CA 93943. E-mail: rtjohnso@nps.edu

WHY DOES THE DEFENSE COMMUNITY NEED DESIGN OF EXPERIMENTS?

Any organization serious about testing should embrace methods and a general strategy that will cover the range of product employment, extract the most information in limited trials, and identify parameters affecting

performance. For example, the purpose of Air Force (AF) test and evaluation (T&E) is "mature system designs, manage risks, identify and help resolve deficiencies as early as possible, and ensure systems are operationally mission capable (i.e., effective and suitable)" (AF/TE 2009, p. 1). Similar instructions and regulations guide the other U.S. armed services. The fields of designed experiments and industrial statistics, with their rich histories spanning over a century, provide the framework for test science excellence. Large-scale efforts are underway in the Department of Defense (DoD) to replace current test strategies of budget-only-driven test events, combat scenarios, changing one factor at a time, and preserving traditional test programs with a scientific and statistically rigorous approach to test—design of experiments. Design of experiments improves DoD test rigor by objectively justifying the number of trials conducted based on decision risk, well apportioning test conditions in the battle space, guiding the execution order to control nuisance variation, and objectively separating the signal of true system responses from underlying noise.

Effectiveness and efficiency are essential to all testing but especially military test and evaluation. The footprint of the military T&E enterprise is substantial, whether measured in resources, people, or national defense capacity. The DoD spent nearly $75 billion in research, development, test, and evaluation in fiscal year 2008. To illustrate the scope in one service, AF T&E accounts for an estimated 25–30% of all 11,000 AF scientists and engineers; in expenditures, AF research, development, test, and evaluation was $26.7 billion—20% of the U.S. Air Force (USAF) budget (Secretary of the Air Force Financial Management Office [SAF/FM] 2007). Design of experiments (DOE) enables effectiveness of system discovery with detailed process decomposition tying test objectives to performance measures, together with test matrices that span the operating region and allow for faults to be traced to causes. Efficiencies are gained by combining highly efficient screening designs with initial analyses to learn early, followed by knowledge-based test augmentation for continued learning via statistical modeling, culminating in validation tests—all with the purpose of full system understanding using only the resources necessary. The DoD is moving toward the use of DOE as the primary method of test. As stated in the guidance document (2010), published by the Director of Operational Test and Evaluation, there is a specific request to "increase the use of both scientific and statistical methods to in developing rigorous, defensible test plans and in evaluating their results" (p. 1). These guidelines require test programs not to explicitly "do DOE" but to report the evidences of well-designed experiments including continuous response variables, how test factors are to be controlled during test, and the strategy (family of test designs) used to place individual points in the space to be explored. This article supports the reshaping of the DoD T&E policy by detailing basic experimental design tools and their application in military context.

Military T&E is serious business, because it dictates the future effectiveness of U.S. defense forces. Test programs designed using the principles of designed experiments stand to improve the cost-effectiveness of defense acquisition by ensuring that experimentation and failures occur during development and not in the field; that correct decisions are reached in fielding new combat capability; and that only the appropriate amount is expended during test in an era of declining defense budgets.

BACKGROUND AND HISTORY OF DESIGNED EXPERIMENTS

Statistically designed experiments are among the most useful, powerful, and widely applicable statistical methods. They are used extensively in many industrial and business settings, with applications ranging from medical/biopharmaceutical research and development to product design and development across virtually the entire industrial sector, agriculture, marketing, and e-commerce. In this section we present a brief overview of the methodology aimed at helping the members of the DoD test community who have had little exposure to designed experiments but understand some of the basic concepts and principles.

There have been four eras in the modern development of statistical experimental design. The first or agricultural era was led by the pioneering work of Sir Ronald A. Fisher in the 1920s and early 1930s. During that time, Fisher was responsible for statistics and data analysis at the Rothamsted Agricultural Experimental Station near London, England. Fisher recognized that flaws in the way the experiment that

generated the data had been performed often hampered the analysis of data from systems (in this case, agricultural systems). By interacting with scientists and researchers in many fields, he developed the insights that led to three basic principles of experimental design: randomization, replication, and blocking. By *randomization* we mean running the trials in an experiment in random order to minimize systematic variation from variables that are unknown to the experimenter but that vary during the experiment. *Replication* is repeating at least some of the trials in the experiment so that an estimate of the experimental error can be obtained. This allows the experimenter to evaluate the change observed in response when a factor is changed relative to the probability that the observed change is due to chance causes. This introduces scientific objectivity into the conclusions drawn from the experiment. *Blocking* is a technique to prevent the variability from known nuisance sources from increasing the experimental error. Typical sources of nuisance variability include operators or personnel, pieces of test equipment, weather conditions, and time.

Fisher systematically introduced statistical thinking and principles into designing experimental investigations, including the factorial design concept and the analysis of variance. His two books (Fisher 1958, 1966) had a profound influence on the use of statistics, particularly in agriculture and many of the related life sciences. For an excellent biography of Fisher, see J. F. Box (1978).

Though industrial applications of statistical design began in the 1930s, the second, or industrial, era was catalyzed by the development of response surface methodology (RSM) by G. E. P. Box and Wilson (1951). They recognized and exploited the fact that most industrial experiments are fundamentally different from their agricultural counterparts in two ways: (1) the response variable can usually be observed (nearly) immediately and (2) the experimenter can quickly learn crucial information from a small group of runs that can be used to plan the next experiment. G. E. P. Box (1999) called these two features of industrial experiments *immediacy* and *sequentiality*. Over the next 30 years, RSM and other design techniques spread throughout the chemical and process industries, mostly in research and development work. George Box was the intellectual leader of this movement. However, the application of statistical design at the plant or manufacturing process level even in the chemical industry and in most other industrial and business settings was not widespread. Some of the reasons for this include inadequate training in basic statistical concepts and experimental methods for engineers and other scientists and the lack of computing resources and user-friendly statistical software to support the application of statistically designed experiments.

The increasing interest of Western industry in quality improvement that began in the late 1970s ushered in the third era of statistical design. The work of Genichi Taguchi (Kackar 1985; Taguchi 1987, 1991; Taguchi and Wu 1980) also had a significant impact on expanding the interest in and use of designed experiments. Taguchi advocated using designed experiments for what he termed *robust parameter design*, or

1. Making processes insensitive to factors that are difficult to control (i.e., environmental factors).
2. Making products insensitive to variation transmitted from components.
3. Finding levels of the process variables that force the mean to a desired value while simultaneously reducing variability around this value.

Taguchi suggested highly fractionated factorial designs and other orthogonal arrays along with some novel statistical methods to solve these problems. The resulting methodology generated much discussion and controversy. Part of the controversy arose because Taguchi's methodology was advocated in the West initially (and primarily) by consultants, and the underlying statistical science had not been adequately peer reviewed. By the late 1980 s, the results of an extensive peer review indicated that although Taguchi's engineering concepts and objectives were well founded, there were substantial problems with his experimental strategy and methods of data analysis. For specific details of these issues, see G. E. P. Box (1988), G. E. P. Box et al. (1988), Hunter (1985, 1989), Pignatiello and Ramberg (1992), and Myers et al. (2009). Many of these concerns are also summarized in the extensive panel discussion in the May 1992 issue of *Technometrics* (see Nair 1992).

There were several positive outcomes of the Taguchi controversy. First, designed experiments

became more widely used in the discrete parts industries, including automotive and aerospace manufacturing, electronics and semiconductors, and many other application areas that had previously made little use of the techniques. Second, the fourth era of statistical design began. This era has included a renewed general interest in statistical design by both researchers and practitioners and the development of many new and useful approaches to experimental problems in the industrial and business world, including alternatives to Taguchi's technical methods that allow his engineering concepts to be carried into practice efficiently and effectively (e.g., see Myers et al. 2009). Third, formal education in statistical experimental design is becoming part of many engineering programs in universities at both the undergraduate and graduate levels. The successful integration of good experimental design practice into engineering and science is a key factor in future industrial competitiveness and effective design, development, and deployment of systems for the U.S. military.

Applications of designed experiments have grown far beyond their agricultural origins. There is not a single area of science and engineering that has not successfully employed statistically designed experiments. In recent years, there has been a considerable utilization of designed experiments in many other areas, including the service sector of business, financial services, government operations, and many nonprofit business sectors. An article appeared in *Forbes* magazine on March 11, 1996, entitled "The New Mantra: MVT," where MVT stands for *multivariable testing*, a term some authors use to describe factorial designs (Koselka 1996). The article described many successes that a diverse group of companies have had through their use of statistically designed experiments. The panel discussion edited by Steinberg (2008) is also useful reading. The increasingly widespread deployments of Six Sigma, Lean Six Sigma, and Design for Six Sigma as business improvement strategies have further driven the increase in application of designed experiments (e.g., see Hahn et al. 2000; Montgomery and Woodall 2008). The Define–Measure–Analyze–Improve–Control (DMAIC) framework that is the basis of most deployments utilizes designed experiments in the Improve phase, leading to designed experiments being considered the most important of the DMAIC tools.

FACTORIAL EXPERIMENTS

Most experiments involve the study of the effects of two or more factors. In general, factorial designs are most efficient for this type of experiment. By a *factorial design* we mean that in each complete trial or replication of the experiment all possible combinations of the levels of the factors are investigated. For example, if there are two factors, say, A and B, and there are a levels of factor A and b levels of factor B, each replicate of the experiment contains all ab combinations of the factor levels. When there are several factors to be investigated, factorial experiments are usually the best strategy because they allow the experimenter to investigate not only the effect of each individual factor but also the interactions between these factors.

Figure 1 illustrates the concept of interaction. Suppose that there are two factors, A and B, each with two levels. Symbolically we will represent the two levels as A^- and A^+ for factor A and B^- and B^+ for factor B. The factorial experiment has four runs: A^-B^-, A^-B^+, A^+B^-, and A^+B^+. In Figure 1a we have plotted the average response observed at the design points as a function of the two levels of factor A and connected the points that were observed at the two levels of B for each level of A. This produces two line segments. The slope of the lines represents a graphical display of the

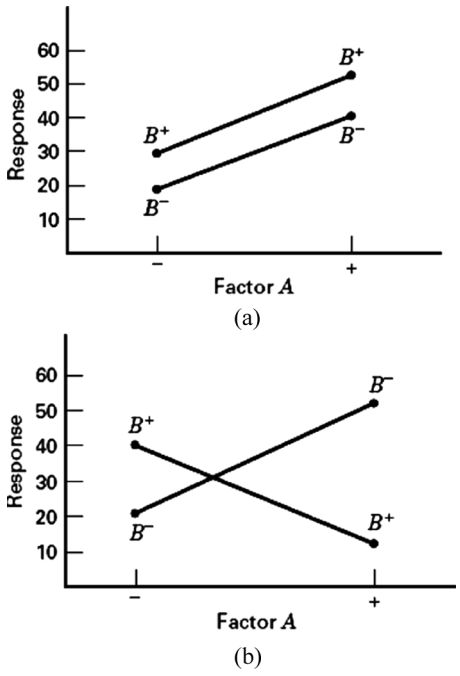


FIGURE 1 **Illustration of interaction: (a) no interaction and (b) a two-factor interaction.**

effect of factor A. In this figure, both line segments have the same slope. This means that there is no interaction between factors A and B. In other words, any conclusion that the experimenter draws about factor A is completely independent of the level of factor B. Now consider Figure 1b. Notice that the two line segments have different slopes. The slope of the lines still represents the effect of factor A, but now the effect of A depends on the level for B. If B is at the minus level, A has a positive effect (positive slope), whereas if B is at the plus level, A has a negative effect (negative slope). This implies that there is a two-factor interaction between A and B. An interaction is the failure of one factor to have the same effect at different levels of another factor. An interaction means that the decisions that are made about one factor depend on the levels for the other factor.

Interactions are not unusual. Both practical experience and study of the experimental engineering literature (see Li et al. 2006) suggest that interactions occur in between one third and one half of all multifactor experiments. Often discovering the interaction is the key to solving the research questions that motivate the experiment. For example, consider the simple situation in Figure 1b. If the objective is to find the setting for factor A that maximizes the response, knowledge of the two-factor or AB interaction would be essential to answer even this simple question. Sometimes experimenters use a one-factor-at-a-time strategy, in which all factors are held at a baseline level and then each factor is varied in turn over some range or set of levels while all other factors are held constant at the baseline. This strategy of experimentation is not only inefficient in that it requires more runs than a well-designed factorial but it yields no information on interactions between the factors.

It is usually desirable to summarize the information from the experiment in terms of a mathematical model. This is an *empirical* model, built using the data from the actual experiment, and it summarizes the results of the experiment in a way that can be manipulated by engineering and operational personnel in the same way that mechanistic models (such as Ohm's law) can be manipulated. For an experiment with two factors, a factorial experiment model such as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \varepsilon \qquad [1]$$

could be fit to the experimental data, where x_1 and x_2 represent the main effects of the two experimental factors A and B, the cross-product term $x_1 x_2$ represents the interaction between A and B, the βs are unknown parameters that are estimated from the data by the method of least squares, and ϵ represents the experimental error plus the effects of factors not consider in the experiment. Figure 2 shows a graphical representation from the model

$$\hat{y} = 35.5 + 10.5x_1 + 5.5x_2 + 8.0x_1 x_2 + \varepsilon$$

Figure 2a is a response surface plot presenting a three-dimensional view of how the response variable is changing as a result of changes to the two design factors. Figure 2b is a contour plot, which shows lines of constant elevation on the response surface at different combinations of the design factors. Notice that the lines in the contour plot are curved, illustrating that the interaction is a form of curvature in the underlying response function. These types of graphical representations of experimental results are important tools for decision makers.

Two-level factorial designs are probably the most widely used class of factorial experiment used in the industrial research and development environment (see Montgomery 2009). These are designs

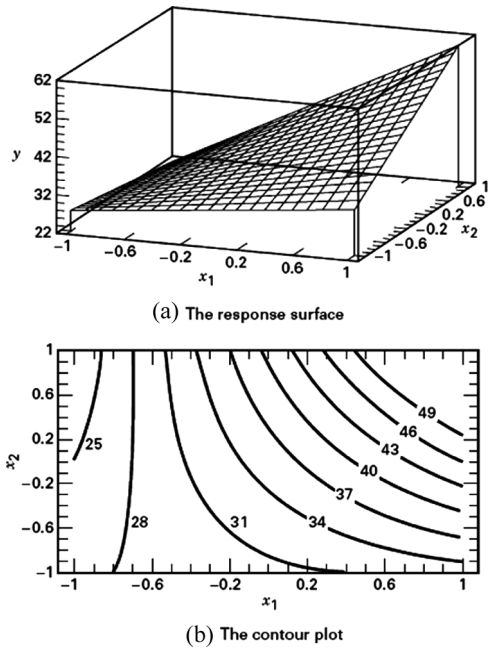
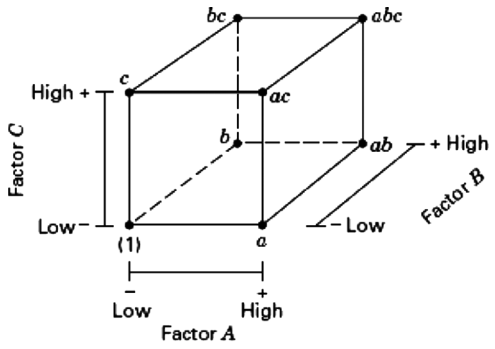


FIGURE 2 Graphical displays of the model $\hat{y} = 35.5 + 10.5x_1 + 5.5x_2 + 8.0x_1 x_2 + \varepsilon$: (a) response surface plot and (b) contour plot.



(a) Geometric view

Run	Factor A	B	C
1	–	–	–
2	+	–	–
3	–	+	–
4	+	+	–
5	–	–	+
6	+	–	+
7	–	+	+
8	+	+	+

(b) Design matrix

FIGURE 3 **The 2^3 factorial design: (a) geometric view and (b) design matrix.**

where all factors (say k) have two levels, usually called *low* and *high* and denoted symbolically by -1 and $+1$. In these designs, the number of runs required is $N = 2^k$ before any replication. Consequently, these designs are usually called 2^k designs.

As an illustration, Figure 3 shows a 2^3 factorial design in the factors A, B, and C. There are $N = 8$ runs (before any replication). Figure 3a is the geometric view of the design, showing that the eight runs are arranged at the corners of a cube. Figure 3b is a tabular representation of the design. This is an 8×3 design matrix, where each row in the matrix is one run in the design and each column is one of the three design factors. This design will support the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3 + \beta_{123} x_1 x_2 x_3 + \varepsilon \quad [2]$$

where x_1, x_2 and x_3 are the main effects of the three design factors, x_1x_2, x_1x_3 and x_2x_3 are the two-factor interactions, and $x_1x_2x_3$ is the three-factor interaction. Methods for the statistical analysis of these experimental designs, estimating the model parameters, and interpretation of results are described in Montgomery (2009).

FRACTIONAL FACTORIAL DESIGNS

As the number of factors in a factorial design increases, the number of runs required for the experiment rapidly outgrows the resources of most experimenters. For example, suppose that we have six factors and all factors have two levels. A complete replicate of the 2^6 design requires 64 runs. In this experiment there are six main effects and 15 two-factor interactions. These effects account for 21 of the 63 available degrees of freedom (DOF) between the 64 runs. The remaining 42 DOF are allocated to higher order interactions. If there are eight factors, the 2^8 factorial design has 256 runs. There are only eight main effects and 28 two-factor interactions. Only 36 of the 255 DOF are used to estimate the main effects and two-factor interactions. In many experimental settings, interest focuses on the main effects of the factors and some of the low-order interactions, usually two-factor interactions. The occurrence of three-factor and higher order interactions is relatively rare, usually occurring in less than about 5% of typical engineering and scientific experiments. In the experimental design literature, this is called the *sparsity of effects principle*. Consequently, it is often safe to assume that these higher order interactions can be ignored. This is particularly true in the early stages of experimentation with a system where system characterization (determining the most important factors and interactions) is important, and we suspect that not all of the original experimental factors have large effects.

If the experimenter can reasonably assume that most of the high-order interactions are negligible, information on the main effects and low-order interactions may be obtained by running only a fraction of the complete factorial experiment. These fractional factorial designs are among the most widely used types of experimental designs for industrial research and development. The 2^k factorial designs are the most widely used factorial as the basis for fractional designs. The 2^k factorial design can be run in fractional sizes that are reciprocal powers of 2; that is, $\frac{1}{2}$ fractions, $\frac{1}{4}$ fractions, $\frac{1}{8}$ fractions, and so on. As examples, the $\frac{1}{2}$ fraction of the 2^5 design has only 16 runs in contrast to the full factorial, which has 32 runs, and the $\frac{1}{16}$ fraction of the 2^8 has only 16 runs in contrast to the 256 runs in the full factorial. There are simple algorithmic methods for constructing these designs

(see Box et al. 2004; Montgomery 2009). These designs also lend themselves to sequential experimentation, where runs can be added to a fractional factorial to either increase the precision of the information obtained from the original experiment or resolve ambiguities in interpretation that can arise if there really are higher order interactions that are potentially important. These techniques are implemented in standard software packages that are easy for experimenters to use.

RESPONSE SURFACES AND OPTIMIZATION

The previous two sections introduced the concepts of factorial and fractional factorial designs, respectively, which are typically used for *screening*—determining what factors or combinations of factors impact a response variable of choice. Once the important factors are identified, a logical extension is to determine the levels of these factors that produce the best or most desirable results. One way this is accomplished is through the use of RSM. RSM, which was developed in the second era of statistical experimental design, is a collection of statistical and mathematical techniques that are used for improving and/or optimizing processes. These techniques can be generalized to their use for the development of mathematical models that describe the response variable as a function of factors of interest. For example, suppose that you have a set of predictor variables $x_1, \ldots, x_k$ and a response variable y. The response can be modeled as a function of the input (predictor) variables. RSM can aid in the development of this function (or mathematical model). For example, consider the function

$$y = f(x_1, \ldots, x_k) + \varepsilon$$

where $f(x_1, \ldots, x_k)$ represents a function consisting of the predictor variables and ϵ represents the error in the system. This model can be used in any capacity of interest to the researcher (such as visualization of the response variable(s) or optimization of the response). Equations [1] and [2] show polynomial functions in two and three variables, respectively, with main effects and interactions.

The development of a function that translates the input variables into an output response plays a key role in the three main objectives of RSM, which are (1) mapping a response surface over a particular region of interest, (2) optimization of the response, and (3) selecting operating conditions to achieve a particular specification or customer requirement. Though these objectives are often described in the context of industrial problems, they are also prevalent in the defense community.

Factorial and fractional factorial designs are sometimes used in RSM as an initial design intended to provide insight such as what factors are most important in the experiment. Recall that G. E. P. Box (1999) stressed the use of a sequential experimental design strategy. This means that after the initial experiment is conducted and analyzed to identify the important factors, more sophisticated experimental techniques can be used to describe and model the complexities in the response surface. A classic response surface design that is both efficient and highly effective in fitting second-order models is the central composite design (CCD; see Box and Wilson 1951). This design consists of factorial corner points (either a full factorial or appropriate fraction), center points, and axial points. The distance from the center of the design space to the axial points is often based on the shape of the region of interest. A spherical region would call for axial points at a distance of ± 1.732 in coded units. Alternatively, a CCD with axial distances set to ± 1 fits into a cubical region as shown in Figure 4. The addition of these center and axial points in the CCD allows the experimenter to fit higher order terms, such as squared terms in the inputs.

The use of higher order models provides valuable insights and allows the objectives of RSM (mapping the response surface, optimization, and selecting

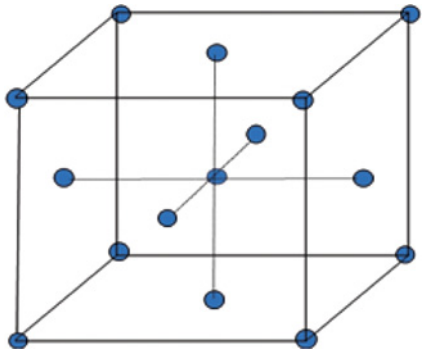


FIGURE 4 Test point geometry of a face-centered CCD in three factors. (Color figure available online.)

operating regions based on specifications) to be met. An application of RSM in the defense community is presented in the next section.

EXAMPLE DOE APPLICATIONS

Two example applications of DOE are presented in this section. First an example of a military force-level encounter is given. In this example, a fractional factorial is used to study the relationship between the input factors and the output response. Next, an example of an air-to-air missile simulation model using RSM to study seven factors of interest is illustrated.

Force-Level Encounter Assessment

Frequently, military testers encounter the problem of engaging in simulated combat operations against an "aggressor" adversary to determine methods of employing some new system or capability—tactics development. In the Air Force, force sizes range from one versus one to 50–75 aircraft encounters ("many vs. many") in the periodic Red Flag exercises outside Las Vegas, Nevada. Valiant Shield, a June 2006 exercise, involved 22,000 personnel, 280 aircraft, and more than 30 ships (including three aircraft carriers and their strike groups) in the Pacific Ocean and surrounding lands.

Such large-scale force encounters offer appropriate scale to realistically exercise military systems against an unpredictable thinking adversary. In this sense, exercises are the best simulation of combat short of war. On the other hand, large-scale encounters are unwieldy, noisy, and offer fewer battles as experimental units than smaller force exercises. Experimental controls may restrict tactical free-play, thus hindering fighting force training. Nevertheless, force exercises are an important opportunity to test our military systems and tactics in an environment far too expensive for any single military test activity to afford on its own. This case illustrates effective experimentation in the midst of large force exercises. The case was adapted from McAllister's dissertation research (2003) concerning tactical employment of fighters. Air Force doctrine calls for rapidly establishing air supremacy—the unrestricted use of air and space—while denying it to the adversary. For the case study, eight friendly (traditionally "Blue") fighters with modern sensors, weapons, and communications contest the airspace with eight adversary ("Red") fighters. Engagements of this size are typical of air combat exercises such as Red Flag.

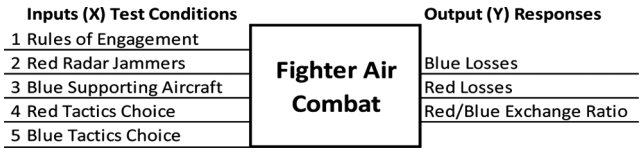


FIGURE 5 Notional Blue–Red force engagement of eight fighters per side.

Figure 5 illustrates some possible input and output conditions for the engagement. The Appendix contains more complete lists. "SA" refers to the gold standard of air combat: *situational awareness*—accurately knowing where friends and enemies are. Lack of (or loss of) SA is frequently a terminal condition in air combat.

The tables in the Appendix further show inputs and outputs measured on as rich a measurement scale as possible. Real-valued variables (when possible) are a hallmark of a well-designed experiment (Coleman and Montgomery 1993). The output measures count the losses on both sides and the exchange ratio. Combat exchange ratios have a long history and useful interpretations but are uninformative if the losses are zero on either side. McAllister (2003) considered three adjustments to the exchange ratios to deal with these problems.

On the input side, some discussion is in order. Rules of engagement (ROE) specify the conditions under which a fighter is authorized to engage and destroy another aircraft. Rules of engagement may range from loose—allowing the destruction of any aircraft not positively identified to be friendly (a relatively quick process)—to tight ROE calling for closing the target for positive visual identification. Looser ROE allow sensors and missiles to be employed at maximum range (usually to Blue's advantage), whereas tighter ROE delay missile firings considerably. Radar jammers are employed to mask own-side aircraft from the enemy. This condition counts the number of dedicated stand-off jamming aircraft available to the Red forces. Blue supporting assets refers to the number of airborne early warning, command and control, and intelligence aircraft available to the Blue side. Finally, the Red and Blue tactics options are inserted in the experiment in an attempt to answer whether one Blue tactic is

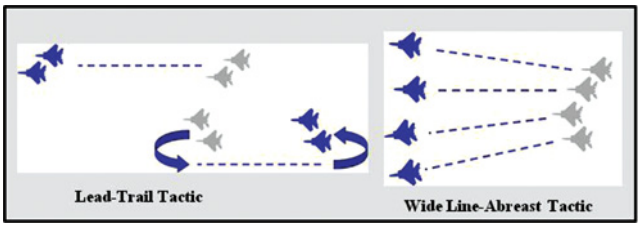


FIGURE 6 Notional Blue tactical employment choices. (Color figure available online.)

universally superior to the other and whether Red's choice of tactics should influence Blue's tactical choices. As an illustration of such tactics, consider Figure 6 and the two notional tactics developed for the Blue forces.

A prime tenant of modern air warfare is to avoid closing (merging) with the adversary and engaging in what is popularly known as a *dogfight*. Such turning engagements nullify superior U.S. weapons and sensors, putting even relatively unsophisticated opponents in a position from which they may be able to destroy Blue aircraft. With the Lead-Trail tactic, one pair of fighters is always positioned to engage the adversary while the other turns away to maintain stand-off distance from the adversary. With the Line-Abreast tactic, all four shooters are available for the initial salvo, maximizing the number of first-shot missiles in the air. The drawback to line abreast is that all four fighters turn away simultaneously, increasing the risk of a dogfight when Blue fighters turn back into the engagement.

Choice of Experimental Designs and Data Generation

As originally stated, the objective is to determine whether any tactical choices are superior for the Blue forces across an array of typical combat encounters. In line with G. E. P. Box's (1999) advice on sequential experimentation referenced earlier, the experiment begins with a fractional factorial screening design[1] with five factors, each at two levels: a $\frac{1}{2}$ fraction requiring 16 trials and yielding excellent information on the five main effects and 10 two-factor interactions.[2]

[1]In reality, because the 16 trials might take 8-10 days to complete, the design might be further blocked in groups of 4-8. Additionally, it would be a good practice to replicate one or more points to objectively estimate pure error.

[2]In DOE terminology, this is a Resolution V design. One can estimate main effects clear of all but a single four-way interaction, and each two factor interaction is aliased with a single three-factor interaction. Sparsity of effects has empirically found these higher order interactions to be rare.

TABLE 1 Design factors and levels

Factor	Name	Units	Type	Design Values
A	ROE_t_ID	seconds	Numeric	10,60
B	Red_Jammers	count	Numeric	0, 2
C	Blue_Spt_AC	count	Numeric	2, 8
D	Red_Tactic	nm	Numeric	0, 5
E	Blue_Tactic	nm	Numeric	0, 5

The design table and constructive response data are provided in Tables 1 and 2. The ROE values represent the number of seconds typically required for a positive identification under the two rule sets; both Red and Blue supporting aircraft are represented by numeric counts, and the Red/Blue tactics choices are designated by the closest approach of the two adversary forces, with "0" representing a possible merge and resulting dogfight between Red and Blue fighters.

The simulated data shown in Table 2 were generated by an Excel Monte Carlo simulation created some years ago. The simulation has been used to produce sample data for classroom instruction, tactics development planning discussions, and a variety of technical papers (McAllister 2003 is an example). The Excel simulation emulates up to four missile exchanges between Red and Blue forces. It ends when the simulated missiles are exhausted or one force loses 50% of their aircraft.

Discussion of Results

Table 3 shows that 8 of the potential 32 terms in the regression model appear to have an effect on the exchange ratio. The main effect of variable D, the Red Tactic, was included for hierarchy, because the interaction BD between Red Tactic and Red Jammers was highly significant. We shall focus on the model terms involving the factor E—Blue tactical choice. Plots of the AE and CE interactions are shown in Figures 7a and 7b. In both interaction plots it is clear that the tactical choice maintaining larger separation distances between the Blue and Red Forces (E at + 5 level, red lines) exploits the benefits from both looser ROE and additional supporting aircraft. With the other tactical choice (E at 0 level, black lines), neither looser ROE nor additional supporting aircraft lead to increased kills of Red aircraft. Examination of residuals shows no apparent violations of assumptions.

TABLE 2 Simulated tactics—Development design and exchange ratios

Std	A:ROE_t_ID	B:Red_ Jammers	C:Blue_ Spt_AC	D:Red_ Tactic	E:Blue_ Tactic	Red/Blue_KRatio
Units>>	sec	count	count	nm	nm	ratio
1	60	0	2	0	0	0.3
2	10	2	2	0	0	1.3
3	10	0	8	0	0	1.0
4	60	2	8	0	0	1.3
5	10	0	2	5	0	2.0
6	60	2	2	5	0	0.3
7	60	0	8	5	0	3.0
8	10	2	8	5	0	0.0
9	10	0	2	0	5	1.0
10	60	2	2	0	5	0.3
11	60	0	8	0	5	1.0
12	10	2	8	0	5	9.0
13	60	0	2	5	5	3.0
14	10	2	2	5	5	0.5
15	10	0	8	5	5	9.0
16	60	2	8	5	5	0.0

In a noisy exercise, the experimenter should have reasonable expectations for what sorts of effects can be detected. Pilot learning, daily weather changes, aborted sorties due to aircraft malfunctions, and the "fog of war" can lead to substantial swings in outcomes from day to day. In such a noisy environment, tactics and equipment that double or triple the effectiveness of a given force should be readily detectable; conversely, tactics that lead to modest improvements of 20–30% may be masked by exercise noise. To illustrate, in this tabletop

TABLE 3 ANOVA table for simulated tactics—Development design

Analysis of variance table [Partial sum of squares - Type III]

Source	Sum of Squares	df	Mean Square	F Value	p-value Prob>F
Model	120.2	9	13.4	32.9	0.0002
A-ROE_t_ID	13.3	1	13.3	32.7	0.0012
B-Red_Jammers	3.4	1	3.4	8.5	0.0271
C-Blue_Spt_AC	15.2	1	15.2	37.3	0.0009
D-Red_Tactic	0.4	1	0.4	1.0	0.3503
E-Blue_Tactic	13.3	1	13.3	32.7	0.0012
AC	10.2	1	10.2	25.0	0.0025
AE	15.5	1	15.5	38.1	0.0008
BD	38.8	1	38.8	95.4	<0.0001
CE	10.2	1	10.2	25.0	0.0025
Residual	2.4	6	0.4		
Cor Total	122.7	15			

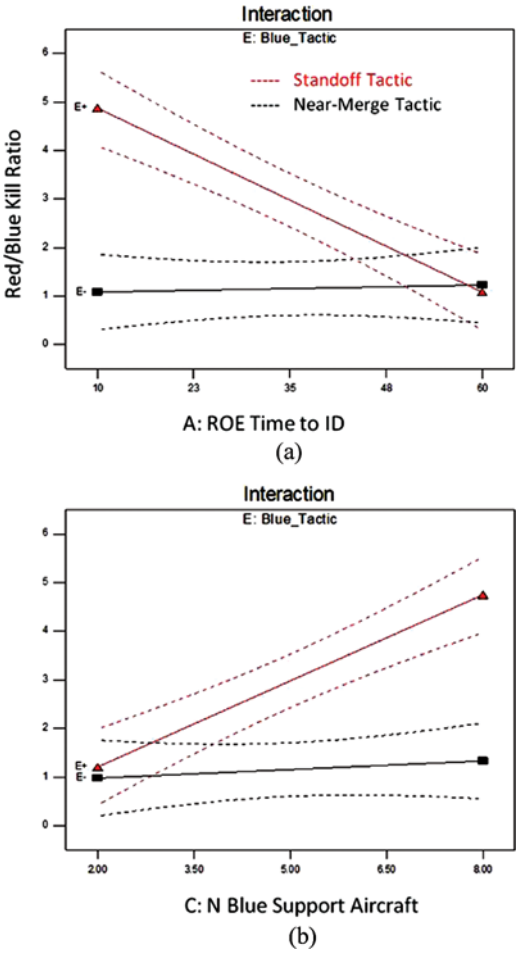


FIGURE 7 Interaction between Blue tactical choice and (a) Blue rules of engagement and (b) Blue support aircraft. (Color figure available online.)

TABLE 4 Representations of weapon systems used in the product acquisition life cycle

	Simulation of Reality					
Acquisition Phase	Modeling & Simulation		Hardware		System/Flight Test	
Requirements Development Material Solution Analysis Technology Development Engineering, Manufacturing & Development	Warfare	Physics	Hardware-in-the-Loop / Systems Integration Lab	Captive	Subsystem	Prototype
Production & Deployment						Production Representative
Operations and Support						Production

simulation, a noise standard deviation of 0.64 implies that day-to-day swings of ± 1 unit in force exchange ratios would not be remarkable (or found to be statistically significant).

Tactics and equipment development are analogous to robust product design in that Blue tactics are design parameters under USAF control, whereas environmental conditions and adversary equipment and tactics are uncontrollable noise variables. In this particular example, happily, Blue tactic effectiveness does not depend on Red equipment or tactics, making the Blue tactics choice robust to anything Red chooses to do.

Air-To-Air Missile Capability Assessment

The military is engaged in the continual development and acquisition of highly complex, sophisticated and technologically superior warfighting systems, from helmet-mounted information systems to aircraft carriers. Among these capabilities requiring enhancement are aircraft-launched weapons for attack against ground and air targets—a key capability for all services in close air support, destruction of air defenses, or counter air operations. The weapons must perform as required and function reliably under diverse operating conditions. In this example we consider just one of the services' weapon variants from the classes of air-to-air or air-to-surface missiles. Examples of such munitions include AIM-120 Slammer, AIM-9X Sidewinder, AGM-65 K Maverick, and AGM-114 Hellfire.

These weapon systems undergo product development in phases based on their levels of acquisition maturity, and test and evaluation is used to assess readiness for the next phase. Various computer simulation and flight test capabilities are utilized for weapon system performance evaluation, depending on the available fidelity level and resources required per test point (Table 4). For missile design, development, and evaluation, the tools typically involve computational fluid dynamics aero simulations; physics-based 6-DOF kinematic models; integrated constructive, or hardware-in-the-loop (HWIL), simulations; captive carry flight test; and delivery of inert or live weapons.

Of the test entries for a next-generation air-to-air missile acquisition, three primary tests include (1) early developmental testing to perform product design initial assessments using digital simulation, (2) later stage developmental test capability assessments using a validated integrated flight simulation or HWIL simulation, and, finally, (3) operational test for weapons effectiveness using captive carry and weapon releases. Figure 8 shows how various simulation forms can be used for test affordably to support system assessment along the various stages of the product life cycle. The tests performed earlier in development feed the experiment designs for future phases, whereas the more realistic complex hardware-based simulations in turn serve to validate physics-based purely constructive simulations. More tests are required earlier and these experiments are typically more affordable. The factor and run numbers are only notional to provide a rough sense of

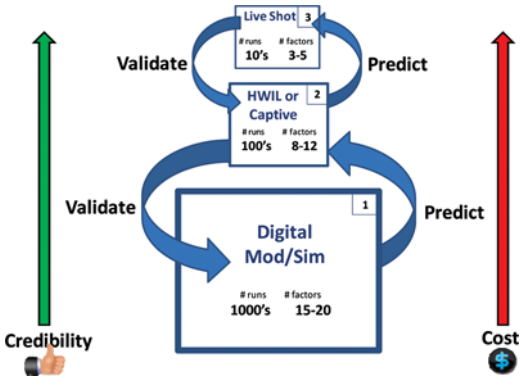


FIGURE 8 **Integrated testing across stages of developmental and operational test using simulation and open air testing. (Color figure available online.)**

the relative magnitudes of the experimental designs. Experimental design is an elegant solution to the complex challenge of comparing simulations of reality as to which factors affect performance and which ones do not. Empirical statistical models of proper polynomial degree (e.g., Eqs. [1] and [2]) serve to directly compare the predictions from each succeeding level of simulation.

This example details the testing of an air-to-air missile during an advanced stage of product development using a high-fidelity, stochastic, multiple-component missile fly-out simulation passing end-game fuzing and fragmentation to a terminal engagement model. It is assumed that the target has been tracked and correctly identified.

A designed experiment approach to building the test strategy and analyzing the data will be illustrated. The key relevant factor categories include weapon deployment geometries bounded by limitations on the missile kinematics, target characteristics, guidance challenges, environmental influences, and terminal flight condition variables. Regardless of the test scenario, careful planning using all the relevant test team representatives (program management, aircrew operators, engineers, and analysts) must jointly develop the test program specific objectives, the influential factors, the responses to be measured, and the appropriate test matrices (i.e., experimental design).

Choice of Experimental Designs and Data Generation

A sequence of test matrices should be planned to leverage knowledge gained from each test phase, feeding the findings of the previous test into the scope of the one succeeding. As such, a reasonable strategy in the developmental phase is to conduct a screening experiment followed by augmentation experiments to discern the true influential interactions and/or nonlinear effects. Often a response surface design capable of mapping the underlying input space is the end objective. Conducting several separate, sequential experiments, each building on knowledge gained from the previous experiment (see Montgomery 2009), is encouraged. Table 5 shows some of the factors typically considered for air-to-air missile capability assessment.

These factors are generated during a rigorous planning session in which the full test team decomposes the process. The team decides on objectives

TABLE 5 **Partial list of typical variables for an air-to-air missile engagement test**

Number	Variable	Variable range
1	Angle off the nose (boresight) of the shooter	0–90
2	Range to target (in % of max range for that set of conditions)	20–90
3	Target type	A, B, C, ...
4	Shooter aircraft type	A, B, C, ...
5	Target aspect angle	0–180
6	Target maneuver	0–90° of turn
7	Shooter altitude	15–25
8	Target altitude	5–30
9	Shooter velocity	300–500
10	Target velocity	300–500
11	Infrared (IR) detector resolution	1–4
12	Target countermeasure (CM) type	A, B, C, ...

and performance measures (parameters measured during flight and at the target) key to answering the objective and then well defines all of the relevant factors associated with the shooter, target, and engagement scenario. For this example, the objective of the test is to assess the lethality performance of an improved air-to-air missile against a known threat aircraft using a previously validated integrated flight simulation. A reduced set of factors and responses used for this example is provided in Figure 9; from an analysis perspective the purpose is to fully characterize the lethality of this missile across the spectrum of its kinematic envelope. Factors include those associated with the relative location, direction, speed, and tactics of the target, as well as a missile design change ultimately increasing the resolution of the infrared (IR) detection. Air-to-air missiles guide using either radio frequency or infrared tracking. Essentially two IR missile variants are tested here, one with traditional resolution (IR detector resolution = 1) and one with enhanced resolution (IR detector resolution = 4).

Suppose initially that the team is primarily interested in modeling miss distance across this seven-variable input region (some variables fixed, others combined from Table 5). Factors with quantitative levels, if applicable, are always preferred because the experiments and subsequent analysis provide insight across the entire region of factor space between the low and high settings. It turns out that each of the seven inputs can be defined such that numeric continuous values are appropriate. Based on engineering knowledge and historical performance of related missiles, it is suspected that at least a second-order polynomial relationship will exist between some inputs and outputs. Because third-order polynomial terms are anticipated to well model miss distance, it makes sense to span the input space such that both the interior and perimeter of the region are reasonably populated with design or test points.

As mentioned, the classic CCD (G. E. P. Box and Wilson 1951) is useful for experiments where the anticipated model is second order. In this case, a cubical region is a natural fit, so an axial distance = +1 in coded units is selected. From our previous discussion, we recognize this design (Figure 10a) with axial distances set to ±1 as a face-centered design (FCD).

Because there is also sufficient rationale for highly nonlinear relations between inputs and the response, and because runs are relatively inexpensive, a second FCD design will be embedded or nested in the interior of the first FCD canvassing the perimeter of the input space (Landman et al. 2007; Tucker et al. 2010). The interior design would place the corner and axial points at ±0.5 in coded units. This nested FCD design (Figure 10b) structure well populates the interior of the input space, has nice symmetry and low correlation among input variables, and is quite efficient when alternate, small-run fractions are used for the corner point designs (Yang 2008).

The factors and settings are provided in Table 6. For proprietary reasons generic descriptions and coding of the input levels will be used to display the findings. Simulated data are used for the same reasons to illustrate potential influences due to the factors on the primary response, miss distance. The experiment used consists of a nested FCD, with complementary fractional factorial designs used for

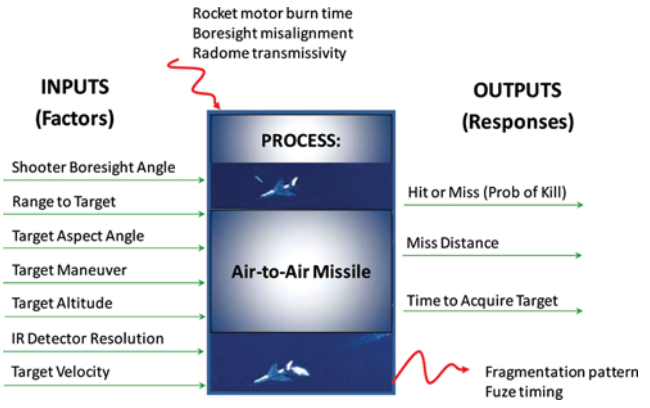


FIGURE 9 **Diagram showing the final input control factors and responses for a capability assessment of an air-to-air missile. (Color figure available online.)**

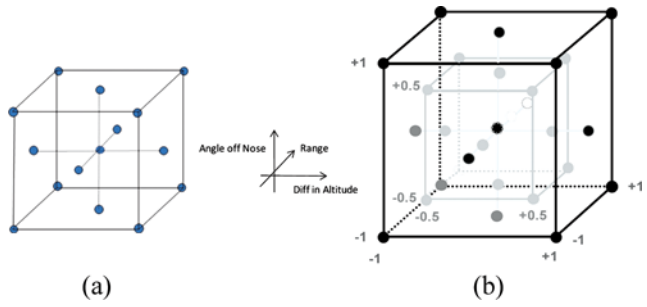


FIGURE 10 **Test point geometry of a (a) face-centered CCD in three factors: difference between shooter and target altitude, range to target, and the angle off the shooter aircraft nose; and (b) nested face-centered design. (Color figure available online.)**

TABLE 6 List of factors and settings for the capability assessment test

Factor	Name	Low actual	High actual
A	Shooter boresight angle	Nose	Beam
B	Range	Low	High
C	Target aspect	Inbound	Outbound
D	Target maneuver	None	90° turn
E	Target altitude	Co-altitude	Look down
F	IR detector focal plane array resolution (pixels)	1 (200 × 200)	4 (400 × 400)
G	Target velocity	Low	High

the corner points. Because each test point resulted from a simulated fly-out from an integrated flight model, the 100 points associated with this nested FCD were easily affordable.

Of note in simulation experimentation is that the fundamental principles of randomization and sequential experimentation play a less important role. The execution order of simulation experiments matters little as long as the noise component is accurately modeled. The sequential nature becomes relevant as simulation run time grows. So if runs are expensive or time consuming, we suggest a sequential strategy of a fractional factorial plus center points, followed by axial points to complete the FCD, followed by (if needed) the nested FCD.

Discussion of Results

The air-to-air missile experimental test points are typically conducted in batch mode using the integrated flight simulation over a weekend, causing little disruption in the acquisition program. The stochastic nature of the simulation allows for analysis using conventional empirical modeling techniques such as least squares regression. The simulated data are generated via Monte Carlo simulation based on behavior typical of traditional air-to-air engagements. Statistical modeling diagnostics are performed during analysis to check for possible violations of the model underlying assumptions. The residual errors from this investigation are well behaved such that the model assumptions are satisfied.

The nominal seven-factor second-order model contains the linear terms, two-factor interactions, and pure quadratics. The experiment design is capable of estimating all 35 model effects of this general second-order model plus higher order interaction and cubic terms. The analysis shows that a second-order model is sufficient. Only three of the seven factors influence the miss distance response and just six model terms of the 35 possible are significant (Table 7).

Because the statistical model is displayed for coded factor levels, the coefficients can be compared directly to determine which model terms are most influential. In this case, both the interaction between shooter angle and the target aspect (AC) and the pure quadratic for target aspect (C^2) have large explanatory power (see Figure 11).

Figure 12 conveys both the interaction and nonlinear relationship that A and C have with miss distance. There are several ways to interpret this response surface. One is that worse performance (higher miss distances) is achieved when the shooter

TABLE 7 Results of model for miss distance with three linear terms, two two-factor interactions, and a quadratic term

Factor	Coefficient estimate	Standard error	95% CI[a] Low	95% CI High
Intercept	14.68	0.41	13.86	15.49
A: Shooter angle	2.84	0.48	1.90	3.79
C: Target aspect	3.46	0.49	2.49	4.43
F: IR detector resolution	6.33	0.49	5.36	7.29
AC	−7.14	0.55	−8.23	−6.05
AF	−2.80	0.55	−3.89	−1.72
C^2	5.18	0.71	3.78	6.59

[a]CI = confidence interval.

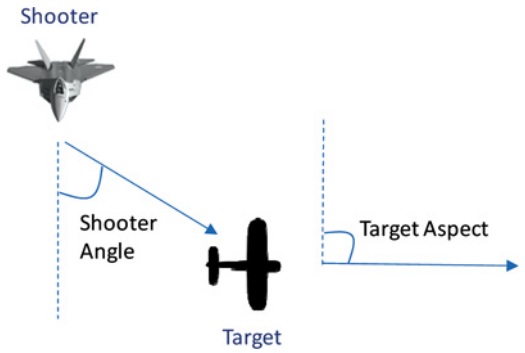


FIGURE 11 **Illustration of shooter angle off the nose and target aspect factor geometries. (Color figure available online.)**

angle is off the beam (A = 90°) and the target is moving away from the shooter (C = 180°) and lower miss distances are obtained if the target is approaching (C = 0°). By contrast, for shots off the shooter nose (A = 0°), miss distances are generally reasonable.

Another major finding involves the engineering design choice of IR array (control) and the shooter angle (noise). Figure 13 displays an interaction plot, indicating that the new IR detector resolution (red line) has the intended effect of reduced miss distances. The lower resolution IR detector performs worse except for shots directly off the shooter nose, and for this resolution the shooter boresight angle largely impacts performance. Conversely, for the improved resolution IR detector, lower miss distances are achieved and performance is insensitive to shooter angle. This result is an example of a meaningful finding in a robust design study. Robust

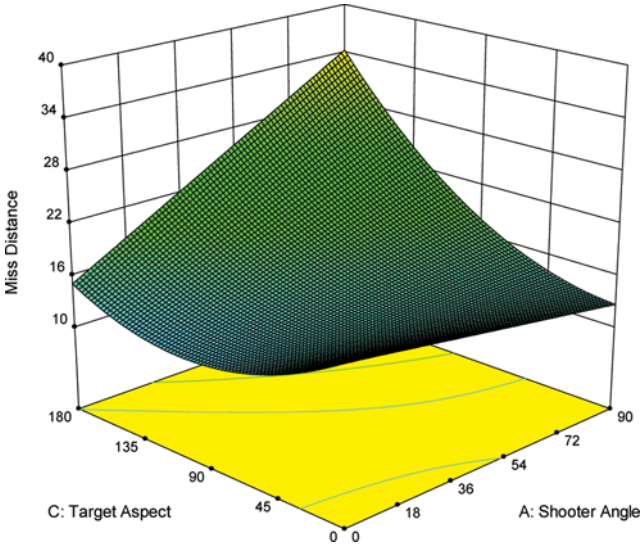


FIGURE 12 **Response surface characterizing the target aspect and shooter angle influences on air-to-air missile miss distance performance. (Color figure available online.)**

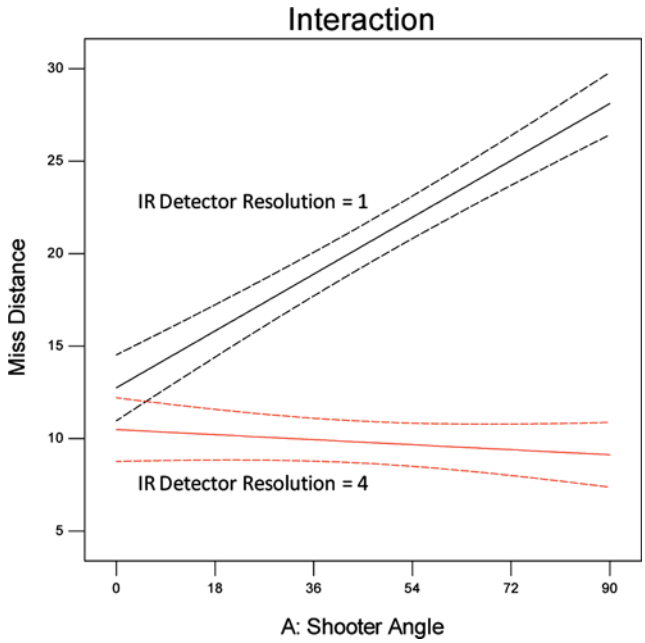


FIGURE 13 **Two-factor interaction plot showing the combined effects of two factors (CM difficulty and shooter angle) on air-to-air missile miss distance performance. (Color figure available online.)**

designs involve control factors that are set during employment (e.g., missile IR detector) and noise factors that vary during employment (e.g., shooter angle). A robust design problem is one that has a significant interaction between the control and noise factors. It is desired to determine control factor settings that provide acceptable overall average performance, as well as reduced response variability in the presence of noise variables. The enhanced resolution IR detector provides better average miss distance as well as resistance to the shooter angle setting.

ADVANCED DOE

Traditional experimental design tools are extremely powerful and provide great insight with the use of as few resources as possible. The advancement of technology and computing power has also expanded the ability of experimental design and analysis to solve more complex problems and tackle issues that previously could not be addressed. Areas considered advanced DOE include (but are not limited to) experiments with computer models, experiments with hard-to-change and easy-to-change factors, experiments in which there are constraints on the experimental region, and experiments where the response surface is expected to be a complex or

nonlinear model. Experiments that need to account for uncontrollable sources of variability, such as the impact of weather or other environmental forces, are not unusual in operational testing. Similarly, experiments that involve human operators are also relatively common, and these operators are random factors that have different levels of skill and/or experience that must be accounted for in both design of the experiment and analysis of the resulting data. Design of experiments for software testing or for testing of complex embedded software systems is also of growing importance. Another important topic is combining data from different sources, such as wind tunnel, computer model, and flight tests or from earlier stage development tests and current operational tests. Some of these topics are relatively well studied in the literature, and other topics are just emerging as areas of research.

Some of the unsolved problems motivate the need for joint collaborative research between DoD partners and the DOE academic and practitioner community. To illustrate the application of one such advanced DOE topic, we will use a maritime domain awareness (MDA) application.

Chung et al. (2009) have developed a decision support tool for the task of search, identification, and interception (SII) of multiple objects in a maritime setting. This is a broad area of persistent surveillance vision with a limited number of assets, which requires an understanding of asset platforms and sensor characteristics. The SII tool is a simulation-based tool that is used to generate optimal routing of assets over time to most effectively search the area for hostile contacts. Typical assets include direct support unmanned aerial vehicles (UAVs), which provide situational updates to surface vessels.

The objectives of DOE and specifically RSM can help enhance the information provided by the decision support tool. The first objective in RSM, mapping a response surface over a particular region of interest, is particularly useful for visualizing a response or studying the effect of factors of interest based on the mathematical model created. Using the SII example, consider the sensor characteristics of the UAVs and how they influence the time to find a hostile object in an area of interest. Two sensor characteristics are α and β, which are the false-positive rates and false-negative rates, respectively, of detection.

In this example, factorial design and CCD could be used to map these input factors, α and β, to the output response (time to find hostile objects); however, there are special considerations. The first consideration is that the response, based on previous information, is expected to be highly nonlinear and may require the use of a nonlinear polynomial model or a special type of spatial correlation model, such as the kriging model, which is a special form of the Gaussian process model (see Jones and Johnson 2009; Santner et al. 2003). The use of these more complicated empirical models potentially warrants the use of an experimental design that has more levels than the factorial or CCDs. A good choice in this case might be a space-filling design, such as a sphere packing design (Johnson et al. 1990), a uniform design (Fang 1980), or a Latin hypercube design (McKay et al. 1979). For a review on empirical modeling and experimental design in computer simulation models, see Chen et al. (2006).

A space-filling experimental design was used to study the relationship between α and β, where the response of interest was measured in the number of cells (a cell 2D area on the surface) traversed by the UAV before the threat was found. The fewer cells traversed, the faster the hostile was intercepted. A response surface plot, created by using a Gaussian process model, is shown in Figure 14.

Figure 14 illustrates that as α and β approach zero (i.e., a perfect sensor) the number of cells traversed

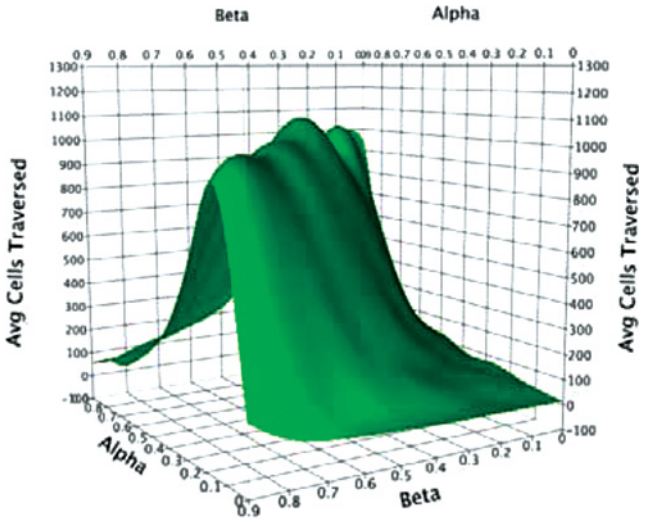


FIGURE 14 **Example response surface of average cells traversed by a UAV before the surface team intercepts the hostile entity as a function of α and β. (Color figure available online.)**

decreases dramatically. Now, imagine pairing this plot with information on the cost to obtain such sensor characteristics. This information could greatly influence the decision as to how good the sensors should be. For example, notice that in Figure 14 the response surface is relatively flat (unchanging) when α and β are below 0.35, but the surface increases exponentially from 0.35 to 0.5.

In addition to mapping the response surface, the Gaussian process model fit can aid in tasks such as selection of operating conditions. In a military environment or setting, there are always many factors that are uncontrollable and/or unpredictable. Given these uncontrollable factors, it is of utmost importance to provide adequate recommendations and draw accurate conclusions in the presence of these uncertain conditions. RSM can play a key role in these decisions. Maritime settings are often influenced heavily by weather conditions. Simulation models used to study the SII strategies take into account modeling these uncontrollable factors such as weather, location of hostile and neutral objects, and movement of hostile and neutral objects. It would be extremely desirable for decision makers to have the opportunity to select levels of controllable factors, such as number of assets, movement of assets, payload of assets, and speed of assets that provide things such as consistent performance and/or high probability of interdiction.

The air-to-air missile example and the SII example illustrate the use of experimental design and analysis techniques and emphasize the enormous potential for solving problems encountered in the defense community. This information is extremely important and there are situations (e.g., Nigerian river delta region, Horn of Africa, and Straight of Malacca) in which the benefit of decision support is greatly amplified by conducting these types of analysis techniques.

CONCLUSIONS

Statistically designed experiments have a long history of successful application in science, engineering, and business. As we moved from the agricultural era into the first industrial era new technical challenges had to be overcome and new methodology had to be developed so that designed experiments could be successfully employed. This led to the development and growth of response surface methodology throughout the 1950s, 1960, and 1970s. The second industrial era saw new methodology developed so that designed experiments could be successfully employed to make products and processes robust to uncontrollable sources of variability and to make the RSM framework more broadly applicable to product design and process development. The current era has seen designed experiments applied to new problems involving computer models, software development and testing, market research, e-commerce, and many other areas. The problems faced by the test community in the DoD are challenging and have many novel characteristics. Solving these problems and integrating statistically designed experiments into the DoD testing philosophy will require (1) broad education of current and future practitioners, (2) development of strong statistical expertise within the test community with high-level capabilities in designed experiments, and (3) research activities involving the test community and DOE researchers focused on specific problem areas vital to the DoD.

ABOUT THE AUTHORS

Dr. Rachel T. Silvestrini (née Johnson) is an Assistant Professor in the Operations Research Department at the Naval Postgraduate School. She received her B.S. in Industrial Engineering from Northwestern University and her M.S. and Ph.D. from Arizona State University. Her research and teaching interests are in statistics and operations research with focus in design of experiments.

Gregory T. Hutto is the Wing Operations Analyst for the Air Force's 46 Test Wing at Eglin AFB. He is a past Director and member of the Military Operations Research Society. Mr. Hutto has more than 21 years experience applying the principles of experimental design to military test and evaluation projects ranging from basic laboratory science efforts to large scale military exercises.

Dr. James R. Simpson is Chief Operations Analyst for the Air Force's 53rd Test Management Group at Eglin AFB, FL. He is Adjunct Professor at the University of Florida, served formerly as Associate Professor at Florida State University and Associate Professor at the Air Force Academy. He is Chair of the ASQ

Journal Editors' Committee, and serves on the ASQ Publication Management Board. He earned a B.S. in Operations Research from the Air Force Academy, an M.S. in OR from the Air Force Institute of Technology, and a Ph.D. in IE from Arizona State University.

Dr. Douglas C. Montgomery is Regents' Professor of Industrial Engineering and Statistics, ASU Foundation Professor of Engineering, and Co-Director of the Graduate Program in Statistics at Arizona State University. He received a Ph.D. in engineering from Virginia Tech. His professional interests are in statistical methodology for problems in engineering and science. He is a recipient of the Shewhart Medal, the George Box Medal, the Brumbaugh Award, the Lloyd S. Nelson award, the William G. Hunter award, and the Ellis Ott Award. He is one of the current chief editors of *Quality & Reliability Engineering International*.

REFERENCES

Air Force Test and Evaluation. (2009). *AFI 99–103 Capabilities-Based Test & Evaluation*. Air Force Test & Evaluation Executive.

Box, G. E. P. (1988). Signal-to-noise ratios, performance criteria, and transformation. *Technometrics*, 30:1–40.

Box, G. E. P. (1999). Statistics as a catalyst to learning by scientific method part II—A discussion [with discussion]. *Journal of Quality Technology*, 31:16–29.

Box, G. E. P., Bisgaard, S., Fung, C. A. (1988). An explanation and critique of Taguchi's contributions to quality engineering. *Quality and Reliability Engineering International*, 4:123–131.

Box, G. E. P., Hunter, J. S., Hunter, W. G. (2004). *Statistics for Experimenters*, 2nd ed. New York: Wiley.

Box, G. E. P., Wilson, K. B. (1951). On the experimental attainment of optimum conditions. *Journal of the Royal Statistical Society, Series B*, 13:1–45.

Box, J. F. (1978). *R. A. Fisher: The Life of a Scientist*. New York: Wiley.

Chen, V., Tsui, K.-L., Barton, R., Meckensheime, M. (2006). A review on design, modeling and applications of computer experiments. *IEE Transactions*, 38:273–291.

Chung, T. H., Kress, M., Royset, J. O. (2009). Probabilistic search optimization and mission assignment for heterogeneous autonomous agents. Paper read at 2009 IEEE International Conference on Robotics and Automation, May 12–17, Kobe, Japan.

Coleman, D. E., Montgomery, D. C. (1993). A systematic approach for planning a designed industrial experiment [with discussion]. *Technometrics*, 35:1–27.

Fang, K. T. (1980). The uniform design: Application of number-theoretic methods in experimental design. *Acta Mathematicae Applicatae Sinica*, 3:363–372.

Fisher, R. A. (1958). *Statistical Methods for Research Workers*, 13th ed. Edinburgh, Scotland: Oliver & Boyd.

Fisher, R. A. (1966). *The Design of Experiments*, 8th ed. New York: Hafner.

Gilmore, J. M. (2010). *Guidance on the use of Design of Experiments (DOE) in Operational Test and Evaluation*. Washington, DC: Operational Test and Evaluation Command.

Hahn, G. J., Doganaksoy, N., Hoerl, R. W. (2000). The evolution of Six Sigma. *Quality Engineering*, 12(3):317–326.

Hoerl, R. W., Snee, R. D. (2010). Statistical thinking and methods in quality improvement: A look to the future. *Quality Engineering*, 22(3): 119–129.

Hunter, J. S. (1985). Statistical design applied to product design. *Journal of Quality Technology*, 17:210–221.

Hunter, J. S. (1989). Let's all beware the Latin square. *Quality Engineering*, 1:453–465.

Johnson, M. E., Moore, L. M., Ylvisaker, D. (1990). Minimax and maxmin distance design. *Journal of Statistical Planning and Inference*, 26:131–148.

Jones, B., Johnson, R. T. (2009). The design and analysis of the Gaussian process model. *Quality and Reliability Engineering International*, 25:515–524.

Kackar, R. N. (1985). Off-line quality control, parameter design, and the Taguchi method. *Journal of Quality Technology*, 17:176–188.

Koselka, R. (1996). The new mantra: MVT. *Forbes*, March 11.

Landman, D., Simpson, J. R., Mariani, R., Ortiz, F., Britcher, C. (2007). Hybrid design for aircraft wind-tunnel testing using response surface methodologies. *Journal of Aircraft*, 44(4):1214–1221.

Li, X., Sudarsanam, N., Frey, D. D. (2006). Regularities in data from factorial experiments. *Complexity*, 11(5):32–45.

McAllister, B. (2003). Measures of effectiveness for testing and training. Barchi Prize Paper presented to Working Group 22, presented at the 71st MORS Symposium, June 10–12, Quantico, VA.

McAllister, B., Zessin, C. (2002). The use of design of experiments during tactics development. Barchi Prize Paper presented to Working Group 25, presented at the 70th MORS Symposium, June 18–20, Fort Leavenworth, KS.

McKay, N. D., Conover, W. J., Beckman, R. J. (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21:239–245.

Montgomery, D. C. (2009). *Design and Analysis of Experiments*, 7th ed. New York: Wiley.

Montgomery, D. C., Woodall, W. H. (2008). An overview of Six Sigma. *International Statistical Review*, 76(3):329–346.

Myers, R. H., Montgomery, D. C., Anderson-Cook, C. M. (2009). *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*, 3rd ed. Hoboken, NJ: Wiley.

Nair, V. N., Ed. (1992). Taguchi's parameter design: A panel discussion. *Technometrics*, 34:127–161.

Pignatiello, J. J., Jr., Ramberg, J. S. (1992). Top ten triumphs and tragedies of Genichi Taguchi. *Quality Engineering*, 4:211–225.

Santner, T. J., Williams, B. J., Notz, W. I. (2003). *The Design and Analysis of Computer Experiments*. Springer Series in Statistics. New York: Springer-Verlag.

Secretary of the Air Force Financial Management Office. (2007). *Air Force Research, Development Test and Evaluation Program Budget*. Secretary of the Air Force Financial Management Office.

Steinberg, D. M., Ed. (2008). The future of industrial statistics: A panel discussion. *Technometrics*, 50(2):103–127.

Taguchi, G. (1987). *System of Experimental Design: Engineering Methods to Optimize Quality and Minimize Cost*. White Plains, NY: UniPub.

Taguchi, G. (1991). *Introduction to Quality Engineering*. White Plains, NY: Asian Productivity Organization.

Taguchi, G., Wu, Y. (1980). *Introduction to Off-Line Quality Control*. Nagoya: Central Japan Quality Control Association.

Tucker, A. A., Hutto, G. T., Dagli, C. H. (2010). Application of design of experiments to flight test: A case study. *Journal of Aircraft*, 42(2):458–463.

Yang, Y. (2008). Multiple criteria third order response surface design and comparison. Master's thesis, Florida State University.

APPENDIX—MORE COMPLEX MEASURES AND TEST CONDITIONS FOR FORCE ENGAGEMENTS

TABLE A1 **Candidate real-valued response variables for fighter force engagements**

N	Measure of Performance - Responses - Y	Units	Rationale for Measuring
1	Red Killed	count	Tradition
2	Blue Killed	count	Tradition
3	Red Survive	count	Tradition
4	Blue Survive	count	Tradition
5	Red/Blue Exchange Ratio	ratio	Tradition
6	Blue Bombers Survive	count	The reason for fighter escort - bombers survive
7	Blue Bombs on time	seconds	bombers delayed/disrupted?
8	Blue Bombs on target	meters	bomber targeting disrupted?
9	Number of Red fighters unobserved	count	Superior Situational Awareness (SA) = no leakers
10	Number of Blue in Red rear unobserved	count	Superior SA = sneak into rear areas
11	Percent Blue Fighters that Merge	percent	Superior SA - hold at beyond visual range - no ''dog fight"
12	Percent Time Offensive/Defensive/Neutral	percent	Goal – Blue 100% offensive
13	Wasted/Denied Shots - Red	percent	Red fighters waste shots
14	Wasted/Denied Shots - Blue	percent	Blue fighter shots count
15	Number unobserved shots on Red	count	maximize unobserved shots on Red
16	Number unobserved shots on Blue	count	Superior SA – no unobserved shots Count or percent?
17	Time advantage to manuever	minutes	Blue sort/target/ react earlier and farther from Red formation
18	Number of saves by Wingman	count	Superior SA means no lack of mutual support after merge
19	Time to re-est mutual support after lost wingman	seconds	Red – maximize Blue – minimize
20	Time w/o mutual support	seconds	Alternate is expert judgment on a rating scale of adequacy
21	Picture accuracy (red and blue) - who/where/when	rating scale	1-10 scale? Worse/same or accurate/inaccurate
22	Num asymmetric engagements (2v1 or 4v2)	count	No fair fights - Blue gang up on Red
23	Time advantage in detect-shoot loop	seconds	direct measure of what we get from SA
24	Range advantage in detect-shoot loop	nm	direct measure of what we get from SA
25	Accuracy of data link positions – all players	meters	Compare to instrumentation measurements

TABLE A2 **Candidate categorical and physically based test conditions for fighter force engagements**

N		Potential test conditions - X's	Simple categoric levels	Physically-based levels
1	**Blue Control Variables**	Radar Support	Hawkeye, AWACS, none	detection range of fighter-target
2		Weapons	AIM-120 C3/C7/D, AIM-9X Blk I/II	missile launch range/lethality
3		Electronic Intel Support	RC-135 or none	1-10 scale of intel awareness
4		Radar Jamming Support	EA-6, B-52, EF-18 G, or none	watts/cm2 at radar, detection range
5		Blue Long Range ID Rules	loose, medium, tight	time required or distance required to ID
6		Blue Tactic Choices	1,2,3 . . .	intent of tactic: first shot, stealth, etc.
7		Comm Jamming Support	EC-130 Compass Call or none	percent comms degraded or allowed
8		Blue Fighter Force Size	Small, Medium, Large	Aircraft count: 2, 4, 8
9		Blue Fighter Mix	F-15/16, F-22/35, mixed force	1-10 scale of capability, detect range
10	**Red Control Variables**	Fighter Radar Control/Support	Ground radar or none	detection range of fighter-target
11		Weapons	AA-XX, AA-yy	missile launch range/lethality
12		Radar Jamming Support	Red jammers type 1,2,3	watts/cm2 at radar, detection range
13		Red Long Range ID Rules	loose, medium, tight	time required or distance required to ID
14		Red Tactic Choices	1,2,3. . .	intent of tactic: first shot, stealth, etc.
15		Red Fighter Mix	3rd/4th/5th generation	1-10 scale of capability, detect range
16		Red Fighter Numbers	Small, Medium, Large	Aircraft count: 2, 4, 8
17		Red Ground Defense Quality	Light, Medium, Heavy	Numbers, shot range, loss rates
18		Red Ground Defense Numbers	Few, Medium, Many	SAM count: 2,4,6, etc.
19	**Env. Variables**	Temperature	Cold, Ambient, Hot	Temperature degress C/F
20		Lighting	Day, Night	lumens available
21		Visibility (visual)	low, medium, high	visibility in nautical miles
22		Visibility infrared	low/medium/high humidity	water in gm/cm3 or IR absorbed (db/km)
23		Clutter Background (spectrum)	low/medium/high	1-10 clutter scale for spectrum
24		Weather (Percipitation)	clear, misty, raining	inches per hour or water density