A Brief Introduction to Design of Experiments

Jacqueline K. Telford

Design of experiments is a series of tests in which purposeful changes are made to the input variables of a system or process and the effects on response variables are measured. Design of experiments is applicable to both physical processes and computer simulation models. Experimental design is an effective tool for maximizing the amount of information gained from a study while minimizing the amount of data to be collected. Factorial experimental designs investigate the effects of many different factors by varying them simultaneously instead of changing only one factor at a time. Factorial designs allow estimation of the sensitivity to each factor and also to the combined effect of two or more factors. Experimental design methods have been successfully applied to several Ballistic Missile Defense sensitivity studies to maximize the amount of information with a minimum number of computer simulation runs. In a highly competitive world of testing and evaluation, an efficient method for testing many factors is needed.

BACKGROUND

Would you like to be sure that you will be able to draw valid and definitive conclusions from your data with the minimum use of resources? If so, you should be using design of experiments. Design of experiments, also called experimental design, is a structured and organized way of conducting and analyzing controlled tests to evaluate the factors that are affecting a response variable. The design of experiments specifies the particular setting levels of the combinations of factors at which the individual runs in the experiment are to be conducted. This multivariable testing method varies the factors simultaneously. Because the factors are varied independently of each other, a causal predictive model can be determined. Data obtained from observational studies or other data not collected in accordance with a design of experiments approach can only establish correlation, not causality. There are also problems with the traditional experimental method of changing one factor at a time, i.e., its inefficiency and its inability to determine effects that are caused by several factors acting in combination.

BRIEF HISTORY

Design of experiments was invented by Ronald A. Fisher in the 1920s and 1930s at Rothamsted Experimental Station, an agricultural research station 25 miles north of London. In Fisher's first book on design of experiments[1] he showed how valid conclusions could be drawn efficiently from experiments with natural fluctuations such as temperature, soil conditions, and rain fall, that is, in the presence of nuisance variables. The known nuisance variables usually cause systematic biases in groups of results (e.g., batch-to-batch variation). The unknown nuisance variables usually cause random variability in the results and are called inherent variability or noise. Although the experimental design method was first used in an agricultural context, the method has been applied successfully in the military and in industry since the 1940s. Besse Day, working at the U.S. Naval Experimentation Laboratory, used experimental design to solve problems such as finding the cause of bad welds at a naval shipyard during World War II. George Box, employed by Imperial Chemical Industries before coming to the United States, is a leading developer of experimental design procedures for optimizing chemical processes. W. Edwards Deming taught statistical methods, including experimental design, to Japanese scientists and engineers in the early 1950s[2] at a time when "Made in Japan" meant poor quality. Genichi Taguchi, the most well known of this group of Japanese scientists, is famous for his quality improvement methods. One of the companies where Taguchi first applied his methods was Toyota. Since the late 1970s, U.S. industry has become interested again in quality improvement initiatives, now known as "Total Quality" and "Six Sigma" programs. Design of experiments is considered an advanced method in the Six Sigma programs, which were pioneered at Motorola and GE.

FUNDAMENTAL PRINCIPLES

The fundamental principles in design of experiments are solutions to the problems in experimentation posed by the two types of nuisance factors and serve to improve the efficiency of experiments. Those fundamental principles are

- Randomization
- Replication
- Blocking
- Orthogonality
- Factorial experimentation

Randomization is a method that protects against an unknown bias distorting the results of the experiment. An example of a bias is instrument drift in an experiment comparing a baseline procedure to a new procedure. If all the tests using the baseline procedure are conducted first and then all the tests using the new procedure are conducted, the observed difference between the procedures might be entirely due to instrument drift. To guard against erroneous conclusions, the testing sequence of the baseline and new procedures should be in random order such as B, N, N, B, N, B, and so on. The instrument drift or any unknown bias should "average out."

Replication increases the sample size and is a method for increasing the precision of the experiment. Replication increases the signal-to-noise ratio when the noise originates from uncontrollable nuisance variables. A replicate is a complete repetition of the same experimental conditions, beginning with the initial setup. A special design called a Split Plot can be used if some of the factors are hard to vary.

Blocking is a method for increasing precision by removing the effect of known nuisance factors. An example of a known nuisance factor is batch-to-batch variability. In a blocked design, both the baseline and new procedures are applied to samples of material from one batch, then to samples from another batch, and so on. The difference between the new and baseline procedures is not influenced by the batch-to-batch differences. Blocking is a restriction of complete randomization, since both procedures are always applied to each batch. Blocking increases precision since the batch-to-batch variability is removed from the "experimental error."

Orthogonality in an experiment results in the factor effects being uncorrelated and therefore more easily interpreted. The factors in an orthogonal experiment design are varied independently of each other. The main results of data collected using this design can often be summarized by taking differences of averages and can be shown graphically by using simple plots of suitably chosen sets of averages. In these days of powerful computers and software, orthogonality is no longer a necessity, but it is still a desirable property because of the ease of explaining results.

Factorial experimentation is a method in which the effects due to each factor and to combinations of factors are estimated. Factorial designs are geometrically constructed and vary all the factors simultaneously and orthogonally. Factorial designs collect data at the vertices of a cube in p-dimensions (p is the number of factors being studied). If data are collected from all of the vertices, the design is a full factorial, requiring 2^p runs. Since the total number of combinations increases exponentially with the number of factors studied, fractions of the full factorial design can be constructed. As the number of factors increases, the fractions become smaller and smaller ($^1/_2$, $^1/_4$, $^1/_8$, $^1/_{16}$, …). Fractional factorial designs collect data from a specific subset of all possible vertices and require 2^{p-q} runs, with 2^{-q} being the fractional size of the design. If there are only three factors in the experiment, the geometry of the experimental design for a full factorial experiment requires eight runs, and a one-half fractional factorial experiment (an inscribed tetrahedron) requires four runs (Fig. 1).

Factorial designs, including fractional factorials, have increased precision over other types of designs because they have built-in internal replication. Factor effects are essentially the difference between the average of all runs at the two levels for a factor, such as "high" and "low." Replicates of the same points are not needed in a factorial design, which seems like a violation of the replication principle in design of experiments. However, half of all the data points are taken at the high level and the other half are taken at the low level of each factor, resulting in a very large number of replicates. Replication is also provided by the factors included in the design that turn out to have nonsignificant effects. Because each factor is varied with respect to all of the factors, information on all factors is collected by each run. In fact, every data point is used in the analysis many times as well as in the estimation of every effect and interaction. Additional efficiency of the two-level factorial design comes from the fact that it spans the factor space, that is, puts half of the design points at each end of the range, which is the most powerful way of determining whether a factor has a significant effect.

USES

The main uses of design of experiments are

- Discovering interactions among factors
- Screening many factors
- Establishing and maintaining quality control
- Optimizing a process, including evolutionary operations (EVOP)
- Designing robust products

Interaction occurs when the effect on the response of a change in the level of one factor from low to high depends on the level of another factor. In other words, when an interaction is present between two factors, the combined effect of those two factors on the response variable cannot be predicted from the separate effects. The effect of two factors acting in combination can either be greater (synergy) or less (interference) than would be expected from each factor separately.

Frequently there is a need to evaluate a process with many input variables and with measured output variables. This process could be a complex computer simulation model or a manufacturing process with raw materials, temperature, and pressure as the inputs. A screening experiment tells us which input variables (factors) are causing the majority of the variability in the output (responses), i.e., which factors are the "drivers." A screening experiment usually involves only two levels of each factor and can also be called characterization testing or sensitivity analysis.

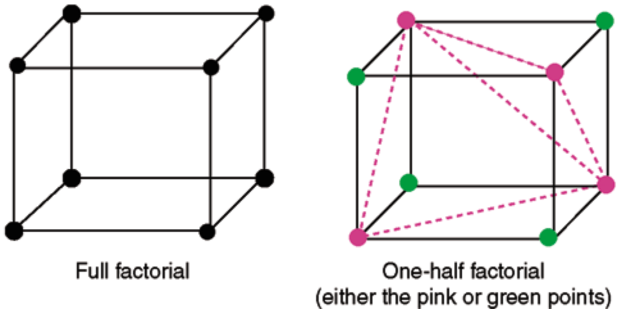

Figure 1. Full factorial and one-half factorial in three dimensions.

A process is "out of statistical control" when either the mean or the variability is outside its specifications. When this happens, the cause must be found and corrected. The cause is found efficiently using an experimental design similar to the screening design, except that the number of levels for the factors need not be two for all the factors.

Optimizing a process involves determining the shape of the response variable. Usually a screening design is performed first to find the relatively few important factors. A response surface design has several (usually three or four) levels on each of the factors. This produces a more detailed picture of the surface, especially providing information on which factors have curvature and on areas in the response where peaks and plateaus occur. The EVOP method is an optimization procedure used when only small changes in the factors can be tolerated in order for normal operations to continue. Examples of EVOP are optimizing the cracking process on crude oil while still running the oil refinery or tuning the welding power of a welding robot in a car manufacturing assembly line.

Product robustness, pioneered by Taguchi, uses experimental design to study the response surfaces associated with both the product means and variances to choose appropriate factor settings so that variance and bias are both small simultaneously. Designing a robust product means learning how to make the response variable insensitive to uncontrollable manufacturing process variability or to the use conditions of the product by the customer.

MATHEMATICAL FORMULATION AND TERMINOLOGY

The input variables on the experiment are called factors. The performance measures resulting from the experiment are called responses. Polynomial equations are Taylor series approximations to the unknown true functional form of the response variable. An often quoted insight of George Box is, “All models are wrong. Some are useful.”[3] The trick is to have the simplest model that captures the main features of the data or process. The polynomial equation, shown to the third order in Eq. 1, used to model the response variable Y as a function of the input factors X’s is

$$Y = \beta_0 + \sum_{i=1}^{p} \beta_i X_i + \sum_{\substack{i=1 \\ i \neq j}}^{p} \sum_{j=1}^{p} \beta_{ij} X_i X_j + \sum_{\substack{i=1 \\ i \neq j \neq k}}^{p} \sum_{j=1}^{p} \sum_{k=1}^{p} \beta_{ijk} X_i X_j X_k + \cdots, \quad (1)$$

where

β_0 = the overall mean response,
β_i = the main effect for factor (i = 1, 2, … , p),
β_{ij} = the two-way interaction between the ith and jth factors, and
β_{ijk} = the three-way interaction between the ith, jth, and kth factors.

Usually, two values (called levels) of the X’s are used in the experiment for each factor, denoted by high and low and coded as +1 and −1, respectively. A general recommendation for setting the factor ranges is to set the levels far enough apart so that one would expect to see a difference in the response but not so far apart as to be out of the likely operating range. The use of only two levels seems to imply that the effects must be linear, but the assumption of monotonicity (or nearly so) on the response variable is sufficient. At least three levels of the factors would be required to detect curvature.

Interaction is present when the effect of a factor on the response variable depends on the setting level of another factor. Graphically, this can be seen as two nonparallel lines when plotting the averages from the four combinations of high and low levels of the two factors. The β_{ij} terms in Eq. 1 account for the two-way interactions. Two-way interactions can be thought of as the corrections to a model of simple additivity of the factor effects, the model with only the β_i terms in Eq. 1. The use of the simple additive model assumes that the factors act separately and independently on the response variable, which is not a very reasonable assumption.

Experimental designs can be categorized by their resolution level. A design with a higher resolution level can fit higher-order terms in Eq. 1 than a design with a lower resolution level. If a high enough resolution level design is not used, only the linear combination of several terms can be estimated, not the terms separately. The word “resolution” was borrowed from the term used in optics. Resolution levels are usually denoted by Roman numerals, with III, IV, and V being the most commonly used. To resolve all of the two-way interactions, the resolution level must be at least V. Four resolution levels and their meanings are given in Table 1.

Table 1. Resolution levels and their meanings.

Resolution level	Meaning
II	Main effects are linearly combined with each other ($\beta_i + \beta_j$).
III	Main effects are linearly combined with two-way interactions ($\beta_i + \beta_{jk}$).
IV	Main effects are linearly combined with three-way interactions ($\beta_i + \beta_{jkl}$) and two-way interactions with each other ($\beta_{ij} + \beta_{kl}$).
V	Main effects and two-way interactions are not linearly combined except with higher-order interactions ($\beta_i + \beta_{jklm}$ and $\beta_{ij} + \beta_{klm}$).

IMPLEMENTATION

The main steps to implement an experimental design are as follows. Note that the subject matter experts are the main contributors to the most important steps, i.e., 1–4, 10, and 12.

1. State the objective of the study and the hypotheses to be tested.
2. Determine the response variable(s) of interest that can be measured.
3. Determine the controllable factors of interest that might affect the response variables and the levels of each factor to be used in the experiment. It is better to include more factors in the design than to exclude factors, that is, prejudging them to be nonsignificant.
4. Determine the uncontrollable variables that might affect the response variables, blocking the known nuisance variables and randomizing the runs to protect against unknown nuisance variables.
5. Determine the total number of runs in the experiment, ideally using estimates of variability, precision required, size of effects expected, etc., but more likely based on available time and resources. Reserve some resources for unforeseen contingencies and follow-up runs. Some practitioners recommend using only 25% of the resources in the first experiment.
6. Design the experiment, remembering to randomize the runs.
7. Perform a pro forma analysis with response variables as random variables to check for estimability of the factor effects and precision of the experiment.
8. Perform the experiment strictly according to the experimental design, including the initial setup for each run in a physical experiment. Do not swap the run order to make the job easier.

9. Analyze the data from the experiment using the analysis of variance method developed by Fisher.
10. Interpret the results and state the conclusions in terms of the subject matter.
11. Consider performing a second, confirmatory experiment if the conclusions are very important or are likely to be controversial.
12. Document and summarize the results and conclusions, in tabular and graphical form, for the report or presentation on the study.

NUMBER OF RUNS NEEDED FOR FACTORIAL EXPERIMENTAL DESIGNS

Many factors can be used in a screening experiment for a sensitivity analysis to determine which factors are the main drivers of the response variable. However, as noted earlier, as the number of factors increases, the total number of combinations increases exponentially. Thus, screening studies often use a fractional factorial design, which produces high confidence in the sensitivity results using a feasible number of runs.

Fractional factorial designs yield polynomial equations approximating the true response function, with better approximations from higher resolution level designs. The minimum number of runs needed for Resolution IV and V designs is shown in Table 2 as a function of the number of factors in the experiment.

There is a simple relationship for the minimum number of runs needed for a Resolution IV design: round up the number of factors to a power of two and then multiply by two. The usefulness of Table 2 is to show that often there is no penalty for including more factors in the experiment. For example, if 33 factors are going to be studied already, then up to 64 factors can be studied for the same number of runs, namely, 128. It is more desirable to conduct a Resolution V experiment to be able to estimate separately all the two-way interactions. However, for a large number of factors, it may not be feasible to perform the Resolution V design. Because the significant two-way interactions are most likely to be combinations of the significant main effects, a Resolution IV design can be used first, especially if it is known that the factors have monotonic effects on the response variable. Then a follow-up Resolution V design can be performed to determine if there are any significant two-way interactions using only the factors found to have significant effects from the Resolution IV experiment. If a factorial design is used as the screening experiment on many factors, the same combinations of factors need not be replicated, even if the simulation is stochastic. Different design points are preferable to replicating the same points since more effects can be estimated, possibly up to the next higher resolution level.

Table 2. Two-level designs: minimum number of runs as a function of number of factors.

Factors	Runs
Resolution IV	
1	2
2	4 = 2^2
3–4	8 = 2^3
5–8	16 = 2^4
9–16	32 = 2^5
17–32	64 = 2^6
33–64	128 = 2^7
65–128	256 = 2^8
129–256	512 = 2^9
Resolution V	
1	2
2	4 = 2^2
3	8 = 2^3
4–5	16 = 2^4
6	32 = 2^5
7–8	64 = 2^6
9–11	128 = 2^7
12–17	256 = 2^8
18–22	512 = 2^9
23–31	1,024 = 2^{10}
32–40	2,048 = 2^{11}
41–54	4,096 = 2^{12}
55–70	8,192 = 2^{13}
71–93	16,394 = 2^{14}
94–119	32,768 = 2^{15}

APPLICATION TO A SIMULATION MODEL

Screening Design

Design of experiments was used as the method for identifying Ballistic Missile Defense (BMD) system-of-systems needs using the Extended Air Defense Simulation (EADSIM) model. The sensitivity analysis proceeded in two steps:

1. A screening experiment to determine the main drivers
2. A response surface experiment to determine the shape of the effects (linear or curved)

The primary response variable for the study was protection effectiveness, i.e., the number of threats negated divided by the total number of incoming threats over the course of a scenario, and the secondary response variables were inventory use for each of the defensive weapon systems.

The boxed insert shows the 47 factors screened in the study. These factors were selected by doing a functional

FORTY-SEVEN FACTORS TO BE SCREENED TO IDENTIFY BMD SYSTEM-OF-SYSTEMS NEEDS

- Threat radar cross section
- Satellite cueing system probability of detection
- Satellite cueing system network delay
- Satellite cueing system accuracy
- Satellite cueing system time to form track
- GB upper tier time to acquire track
- GB upper tier time to discriminate
- GB upper tier time to commit
- GB upper tier time to kill assessment
- GB upper tier probability of correct discrimination
- GB upper tier probability of kill (P_{k}) assessment
- GB upper tier launch reliability
- GB upper tier reaction time
- GB upper tier P_{k}
- GB upper tier burnout velocity (Vbo)
- GB lower tier time to acquire track
- GB lower tier time to discriminate
- GB lower tier time to commit
- GB lower tier probability of correct discrimination
- GB lower tier 1 launch reliability
- GB lower tier 1 reaction time
- GB lower tier 1 P_{k}
- GB lower tier 1 Vbo
- GB lower tier 2 launch reliability
- GB lower tier 2 reaction time
- GB lower tier 2 P_{k}
- GB lower tier 2 Vbo
- SB lower tier time to acquire track
- SB lower tier time to discriminate
- SB lower tier time to commit
- SB lower tier time to kill assessment
- SB lower tier probability of correct discrimination
- SB lower tier P_{k} assessment
- SB lower tier launch reliability
- SB lower tier reaction time
- SB lower tier P_{k}
- SB lower tier Vbo
- Network delay
- Lower tier minimum intercept altitude
- Upper tier minimum intercept altitude
- ABL reaction time
- ABL beam spread
- ABL atmospheric attenuation
- ABL downtime
- GB upper tier downtime
- GB lower tier downtime
- SB lower tier downtime

decomposition of the engagement process for each defensive weapon system, that is, a radar must detect, track, discriminate, and assess the success of intercept attempts and the accuracy, reliability, and timeline factors associated with each of those functions.

A fractional factorial experimental design and EADSIM were used to screen the 47 factors above for their relative importance in far-term Northeast Asia (NEA) and Southwest Asia (SWA) scenarios over the first 10 days of a war. A three-tiered defense system was employed for both scenarios, including an airborne laser (ABL), a ground-based (GB) upper tier, and a lower tier comprising both ground-based and sea-based (SB) systems.

We initially conducted 512 EADSIM runs to screen the sensitivities of the 47 factors in the NEA scenario. This is a Resolution IV design and resolves all of the 47 main factors but cannot identify which of the 1081 possible two-way interactions are significant.

After analyzing results from the initial 512 runs, 17 additional, separate experimental designs were needed (for a total of 352 additional EADSIM runs) to identify the significant two-way interactions for protection effectiveness. We learned from the NEA screening study that more runs were warranted in the initial experiment to eliminate the number of additional experiments needed to disentangle all the two-way interactions. For the SWA screening study, we conducted 4096 EADSIM runs to find the 47 main factors and all 1081 two-way interactions for the 47 factors. This was a Resolution V design. An added benefit of conducting more experiments is that SWA error estimates are approximately one-third the size of NEA error estimates, i.e., the relative importance of the performance drivers can be identified with higher certainty in SWA compared to NEA, which can be seen in Fig. 2. Note that only a very small fraction of the total number of possible combinations was run, 1 in 275 billion since it is a 2^{47-38} fractional factorial, even for the Resolution V design.

Figure 2 illustrates the main factor sensitivities to the 47 factors for both the NEA and SWA scenarios, labeled F1 to F47. The colored dots represent the change of protection effectiveness to each factor, and the error bars are 95% confidence bounds. The *y*-axis is the difference in the average protection effectiveness for a factor between the "good" and "bad" values. Factors are determined to be performance drivers if the 95% confidence

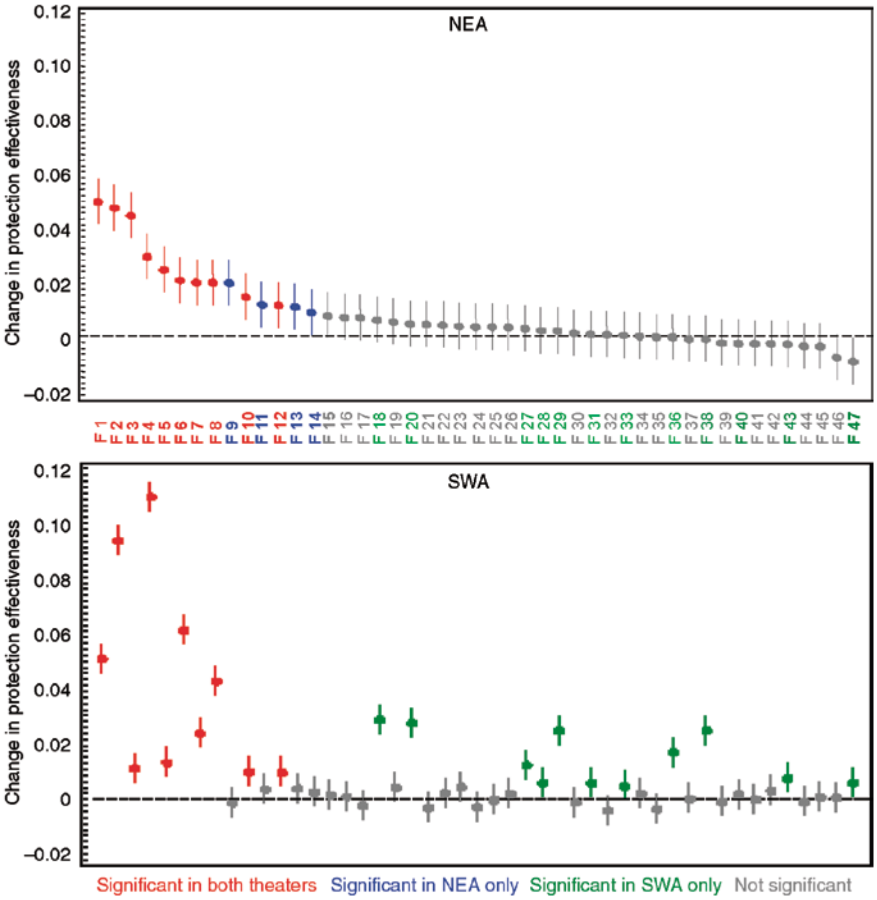

Figure 2. Change in protection effectiveness: 47 main effects and 95% confidence limits.

bounds do not include zero as a probable result. Factors shown in red in Fig. 2 were found to be performance drivers in both scenarios. Factors in blue were found to be drivers in NEA only, and factors in green were found to be drivers in SWA only. Factors that were not found to be drivers in either scenario are shown in gray. (The factors in Fig. 2 are not listed in the same order as they appear in the boxed insert.)

The factors in Fig. 2 are sorted in numerical order of their effects in the NEA scenario. The red factors all appear in the left quarter of the SWA graph, indicating that many of the same factors that are most important in the NEA scenario are also the most important in the SWA scenario. The important factors that differ between the two scenarios, coded in blue and green, result from the geographic (geometric, laydown, and terrain) differences in those two theaters.

The two-way interactions (1081) are too numerous to show in a figure similar to the one for the main effects. However, the vast majority of the two-way interactions are quite small. An example of a significant interaction effect can be seen in Fig. 3, shown graphically by the two lines not being parallel. The increase in protection effectiveness from improving Factor 6 is large if Factor 9 is at the low level, but essentially zero if Factor 9 is at its high level. (Factors 6 and 9 are not the sixth and ninth values listed in the boxed insert.) Data would not have been collected at the +1 level for Factors 6 and 9 in the traditional change-one-factor-at-time experiment, starting at the −1 level for both factors. The protection effectiveness value at +1 for both factors would probably be overestimated from a change-one-factor-at-time experiment. Only by varying both factors at the same time (the Factorial principle) can the actual effect of two factors acting together be known.

Response Surface Design

Once a screening experiment has been performed and the important factors determined, the next step is often to perform a response surface experiment to produce a prediction model to determine curvature, detect interactions among the factors, and optimize the process. The model that is frequently used to estimate the response surface is the quadratic model in Eq. 2:

$$Y = \beta_0 + \sum_{i=1}^{p} \beta_i X_i + \sum_{\substack{i=1 \\ i \neq j}}^{p} \sum_{j=1}^{p} \beta_{ij} X_i X_j + \sum_{i=1}^{p} \beta_{ii} X_i^2 \;, \quad (2)$$

where

β_0 = the overall mean response,
β_i = the main effect for each factor (i = 1, 2, ... , p),
β_{ij} = the two-way interaction between the ith and jth factors, and
β_{ii} = the quadratic effect for the ith factor.

To fit the quadratic terms in Eq. 2, at least three levels for the input X variables are needed, that is, high, medium, and low levels, usually coded as +1, 0, and −1. A total of 3^p computer simulations are needed to take observations at all the possible combinations of the three levels of the p factors. If 2^p computer simulations represent a large number, then 3^p computer simulations represent a huge number. The value of conducting the initial screening study is to reduce p to a smaller number, say k. Even so, 3^k computer simulations may still be prohibitively large.

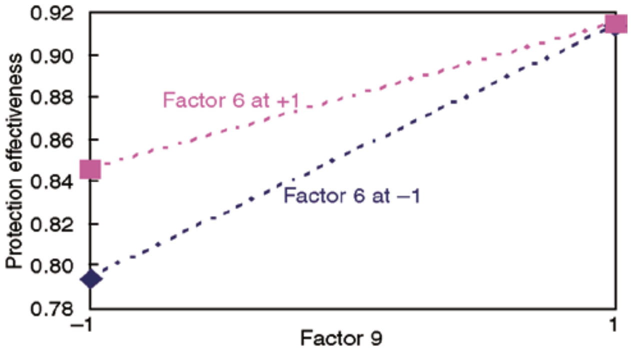

Figure 3. Protection effectiveness: two-way interaction between Factors 6 and 9 from the screening experiment.

The minimum number of runs needed for a three-level Resolution V design as a function of the number of factors is shown in Table 3. From the two-level screening designs, 11 main effects were statistically significant and have at least a 1% effect on protection effectiveness. Table 3 shows that for 11 factors, a minimum number of 243 runs are needed. Notice that 36 factors out of the original 47 have been deemed nonsignificant and will be dropped from further experimentation.

An example of a significant quadratic main effect (Factor 9) and a significant two-way interaction between Factors 6 and 9 for the three-level fractional factorial response surface experiment is shown in Fig. 4. There are different values in protection effectiveness when Factor 9 is at the low level (−1), depending on whether the level of Factor 6 is a the low, medium, or high level, but very little difference if Factor 9 is at the high level (+1). The shape of the lines in Fig. 4 is curved, indicating that a quadratic term is needed for Factor 9 in the polynomial equation. (Factors 6 and 9 are not the sixth and ninth factors listed as in the boxed insert.)

The polynomial equation for protection effectiveness with quadratic and cross-product terms resulting from the 3^{11-6} fractional factorial response surface experiment is shown in Eq. 3. The size of a factor effect on protection effectiveness is actually twice as large as the coefficients on the X terms since the coefficients are actually slopes and X has a range of 2 (from −1 to +1).

$$\begin{aligned} PE = {} & 0.938 + 0.035X_9 + 0.026X_{11} + 0.017X_5 \\ & + 0.016X_2 + 0.015X_6 + 0.014X_1 + 0.012X_7 \\ & + 0.011X_4 + 0.007X_3 + 0.006X_8 - 0.011X_6X_9 \\ & - 0.007X_8X_9 - 0.007X_2X_5 - 0.006X_5X_7 \\ & - 0.005X_3X_9 - 0.005X_5X_6 - 0.005X_1X_5 \\ & - 0.019X_9^2 - 0.011X_5^2 - 0.009X_{11}^2 - 0.008X_4^2 \\ & - 0.006X_3^2 - 0.006X_2^2 \,. \end{aligned} \tag{3}$$

Table 3. Three-level Resolution V designs: minimum number of runs as a function of number of factors.

Factors	Runs
1	3
2	$9 = 3^2$
3	$27 = 3^3$
4–5	$81 = 3^4$
6–11	$243 = 3^5$
12–14	$729 = 3^6$
15–21	$2187 = 3^7$
22–32	$6561 = 3^8$

The full study comprised not only an examination of two theaters (NEA and SWA) but also four force levels in each theater. All of the analyses shown previously were conducted at Force Level 4, which is comparable to a "Desert Storm"–level of logistics support before the operation. Force Level 1 is a rapid response with no prior warning and limited weapons available. Force Levels 2 and 3 are intermediate between Levels 1 and 4. The response surfaces for the four force levels in the NEA scenario are shown in Fig. 5. The individual graphs are the response surfaces for Factors 9 and 11, the two largest main effects for Force Level 4. There is a very noticeable curvature for Factor 9, especially at Force Levels 1 and 2. As the force level increases, protection effectiveness increases. The different color bands are 5% increments in protection effectiveness: red is between 65% and 70% and orange is between 90% and 95%. The response surfaces flatten out and rise up as the force level increases, that is, protection effectiveness improves and is less sensitive to changes in the factors. As the force level increases, there are more assets available, so the reliance on the performance of any individual asset diminishes. (Factors 9 and 11 are not the ninth and eleventh values listed in the boxed insert.)

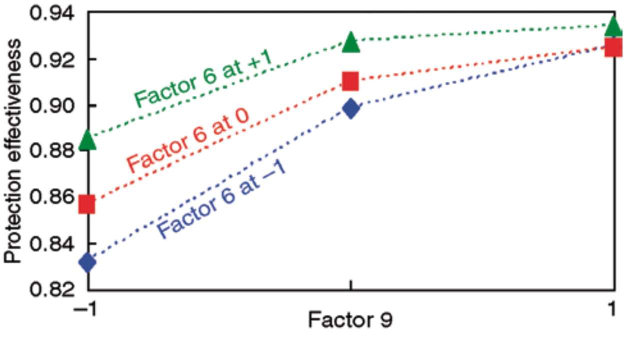

Figure 4. Protection effectiveness: quadratic main effect and two-way interaction between Factors 6 and 9 from the Response Surface Experiment.

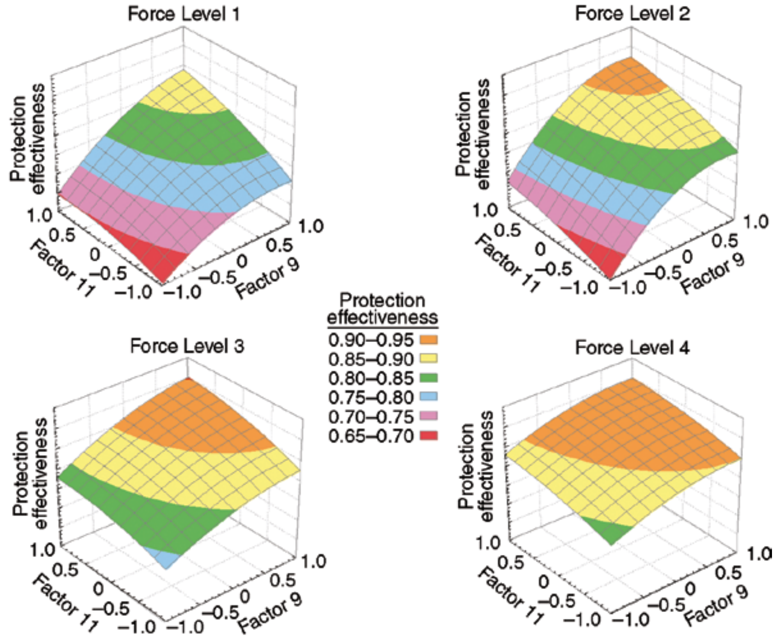

Figure 5. Protection effectiveness response surfaces for Factors 9 and 11 at four force levels in the NEA theater.

CONCLUSION

Design of experiments has been applied successfully in diverse fields such as agriculture (improved crop yields have created grain surpluses), the petrochemical industry (for highly efficient oil refineries), and Japanese automobile manufacturing (giving them a large market share for their vehicles). These developments are due in part to the successful implementation of design of experiments. The reason to use design of experiments is to implement valid and efficient experiments that will produce quantitative results and support sound decision making.

ACKNOWLEDGMENTS. The author would like to thank Roger L. West, Barry L. Mitchell, and Thomas P. Spriesterbach for their participation in the study that was used as the example for this article.

REFERENCES

[1] Fisher, R. A., *The Design of Experiments*, Oliver & Boyd, Edinburgh, Scotland (1935).
[2] Deming, W. E., *Out of the Crisis*, MIT Center for Advanced Engineering Study, Cambridge, MA (1982).
[3] Box, G. E. P., and Draper, N. R., *Empirical Model Building and Response Surfaces*, Wiley, Hoboken, NJ (1987).

The Author

Jacqueline K. Telford is a statistician and a Principal Professional Staff member at APL. She obtained a B.S. in mathematics from Miami University and master's and Ph.D. degrees in statistics from North Carolina State University. Dr. Telford was employed at the U.S. Nuclear Regulatory Commission in Bethesda, Maryland, from 1979 to 1983. Since joining APL in 1983, she has been in the Operational Systems Analysis Group of the Global Engagement Department working primarily on reliability analysis and testing, test sizing, and planning for Trident programs, and more recently for the Missile Defense Agency. She has successfully applied statistical analysis methods to numerous other projects, including evaluation of sensors, models of the hydrodynamics of radioisotopes, and lethality effects. She has taught statistics courses in the JHU graduate engineering program, mostly recently on the topic of design of experiments. Her e-mail address is jacqueline.telford@jhuapl.edu.

Jacqueline K. Telford