A Tutorial on the Planning of Experiments

Laura J. Freeman, Anne G. Ryan, Jennifer L. K. Kensler, Rebecca M. Dickinson & G. Geoffrey Vining

Taylor & Francis
Taylor & Francis Group

A Tutorial on the Planning of Experiments

Laura J. Freeman[1],
Anne G. Ryan[2],
Jennifer L. K. Kensler[3],
Rebecca M. Dickinson[2],
G. Geoffrey Vining[2]

[1]Institute for Defense Analyses, Alexandria, Virginia
[2]Department of Statistics, Virginia Tech, Blacksburg, Virginia
[3]Scientific Test and Analysis Techniques Test and Evaluation Center of Excellence, Air Force Institute of Technology, Wright-Patterson Air Force Base, Ohio

ABSTRACT This tutorial outlines the basic procedures for planning experiments within the context of the scientific method. Too often quality practitioners fail to appreciate how subject-matter expertise must interact with statistical expertise to generate efficient and effective experimental programs. This tutorial guides the quality practitioner through the basic steps, demonstrated by extensive past experience, that consistently lead to successful results. This tutorial makes extensive use of flowcharts to illustrate the basic process. Two case studies summarize the applications of the methodology.

KEYWORDS design of experiment, experimental protocol, experimental strategy, scientific method, sequential experimentation

INTRODUCTION

The proper planning of an experiment requires the formation of the right team applying the right tools. A properly planned experiment requires people management, team facilitation, project management, and subject-matter expertise, both scientific/engineering and statistical. Successful experiments often require a great deal of thought, expertise, discussion, insight, and a basic understanding of the statistical analysis. This article outlines a general process for designing experiments proven to be successful in practice. The ultimate success of an experiment depends on the execution of every step in the planning process. There are some occasions where the experimenters may need to adapt some of the steps that we outline. However, such modifications require sound judgment and thought.

A common complaint about modern statistical texts on the planning of experiments is that they focus more on the analysis of the experimental results rather than the basic experimental protocol. However, adherence to the appropriate protocol determines the likelihood that the experiment successfully addresses the important questions at hand. All of the texts discuss basic principles such as randomization, replication, and local control of error; however, most texts pay mere lip service to the basic issues of selecting the response or responses of interest, the factors, and their levels. These texts discuss how certain experimental constraints lead to blocking or to split-plot structures. However, few go beyond these basics.

Coleman and Montgomery (1993) do an excellent job outlining the basic steps for planning experiments. The first chapter of Montgomery (2009) provides a summary of their paper. Unfortunately, this paper's primary

Address correspondence to Anne G. Ryan, Department of Statistics, Virginia Tech, 416-B Hutcheson Hall (0439), Blacksburg, VA 24061. E-mail: agryan@vt.edu

audience is industrial statisticians with some prior experience in planning experiments. Other good papers that discuss the planning of experiments include Shoemaker and Kacker (1988), Vanhatalo and Berquist (2007), and Viles et al. (2008). Unfortunately, these papers do not go far enough to address many basic practitioner needs for properly planning experiments. Johnson et al. (2012) covered design of experiments (DOE) as applied in the Department of Defense.

This tutorial starts by examining the role of experimentation within the application of the scientific method. Then it provides a high-level flowchart for the planning process. The tutorial continues by explaining in detail each stage in this process. An additional flowchart explaining steps for choosing designs is also given. Two case studies illustrate the total package.

THE SCIENTIFIC METHOD

The scientific method is a structured approach to problem solving and discovery. George Box (1997) referred to it as a tool for efficiently generating new knowledge. Box went on to note that humans have always been learners, but before the scientific method, progress depended on the chance meeting of an informative event and a perceptive observer. The scientific method, according to Box, has accelerated this learning process in the following four ways (Box 1997):

- Providing a better understanding of the interactive nature of learning
- Deducting the logical consequences of a group of facts, each individually known but not previously brought together
- Passively observing and analyzing systems already in operation and data coming from these systems
- Deliberately staging artificial experiences by experimentation

In perhaps its simplest form, the scientific method can be viewed as a linear progression through the following steps:

1. Define the problem
2. Propose an educated theory, idea, or model
3. Collect data to test the theory
4. Analyze the results
5. Interpret the data and draw conclusions

Steps 3–5 go to the core of well-planned and analyzed experiments, which involve collecting the relevant data, properly analyzing them, and then giving an appropriate interpretation of the results. In essence, the scientific method requires experimentation to support investigation.

Most often the conclusions and results themselves will prompt further investigation—sometimes even changing the original objectives as new knowledge is brought to light. This process then becomes iterative and the conclusions serve as a starting point for new ideas or models to be tested. This iterative learning process allows for the questions and problems at hand to be explored at deeper and deeper levels. And even when a problem is solved, the inquiries do not cease because new discoveries can lead to more questions.

The scientific method is certainly an inherent part of all experimental sciences. Donald Marquardt, in his 1986 Presidential Address at the American Statistical Association's Annual Meeting, stated:

> The unique characteristic is that statistics is the generic field of technology, among all the disciplines that use the scientific method. From this perspective, all other fields are special cases for specific subject matter, and the unifying concept is the scientific method.
>
> The scientific method is an inherent part of all experimental sciences. Statistics is the discipline responsible for studying the scientific method with the greatest intensity and for providing in-depth expertise to other disciplines. (Marquardt 1987, p. 4)

Taking a similar stance, Box views statistics as a catalyst to investigation and discovery. For example, response surface methodology (RSM) was originally introduced as an attempt to catalyze a process of investigation and iterative learning (Box 1999).

The scientific method is a sequential learning exercise, and proper experimentation must facilitate this learning process. Consequently, the experimentation supporting the scientific method is sequential—building on what was learned in a previous phase. In some cases, the investigation uses a formal sequential strategy such as RSM. In other cases, the experimental planning builds upon similar experiments conducted in the past. Rarely do people plan experiments in a complete vacuum with no prior experimental information. The key point is that proper methods of experimentation that build upon previous knowledge

and experimentation are extremely powerful and lead to successful investigation.

It is understood that every experiment brings to light its own set of characteristics and complexities; however, all experimentation has the same underlying process. In this article we provide a flowchart that takes the underlying process of experimentation and breaks it down into steps that should be followed for every experiment. The use of the flowchart is illustrated through a case study where the objective is to improve the yield of alcohol from an ethanol–water distillation column. This is followed by a case study using experimental design in a large United States Department of Defense (DoD) system test. The diversity in case studies illustrates the applicability of experimental design ranging from highly controlled engineering processes to large-scale, complex system testing, which are subject to large amounts of variability.

FLOWCHART FOR DESIGNING EXPERIMENTS

The flowchart in Figure 1 methodically presents the steps for both designing and analyzing experiments. Essential questions are highlighted in the flowchart that must be contemplated and answered by an interdisciplinary team. This team should consist of engineers, scientists, statisticians, and other subject-matter experts. Additionally, the team should involve key stakeholders in the outcome of the experiment. The composition of this team will be determined based on the experiment at hand, where each person in the group will have the opportunity to use his or her expertise to help design the experiment. As statistical software packages become more advanced and readily available, many practitioners are turning to software for both planning and analyzing experiments. These software packages allow practitioners to design experiments quickly, such as a second-order response surface design or an optimal design, but software lacks the ability to ensure that these designs will answer the pertinent research questions efficiently. Our flowchart allows for the use of statistical software in experimentation, but the use of software is only one step of many that practitioners should follow when designing and analyzing experiments. Each part of the flowchart is explained in detail in the sections that follow.
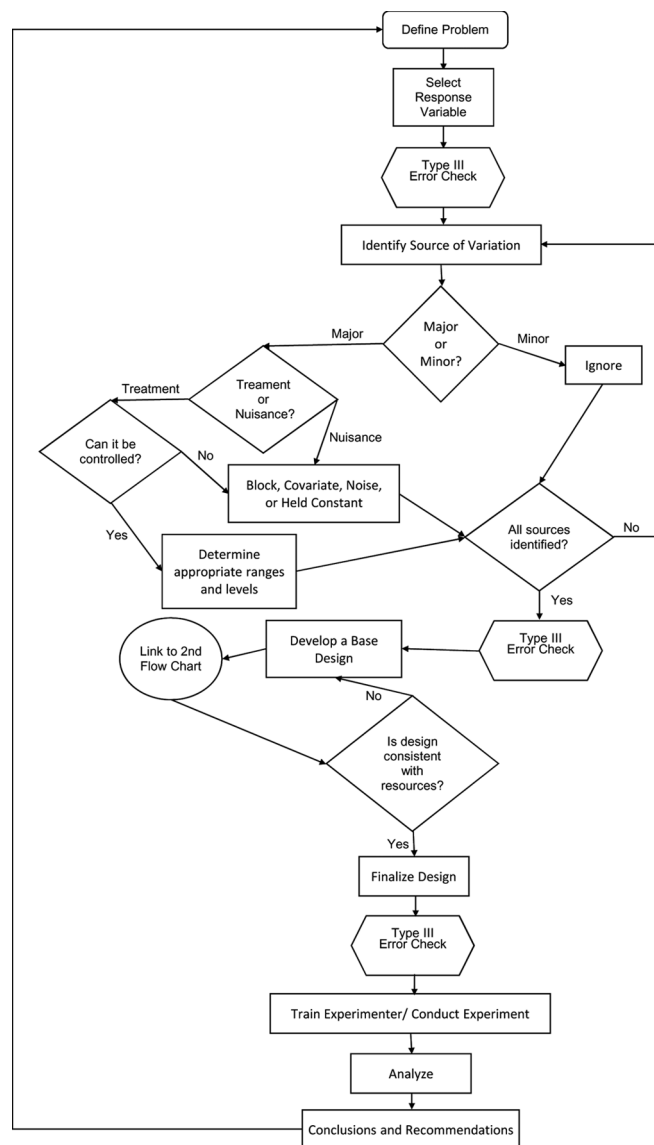

FIGURE 1   Flowchart for designing experiments.

Section 1: Define the Problem

The first step to designing experiments is to define the problem. This involves determining the overall research goals and the specific questions to be answered by the experiment. The goals of the experiment should express the value added to the scientific community from this research, which also gives a sense of purpose for experimentation. In many cases this is the hardest step of the checklist. It is also essential for ensuring that all stakeholders understand the scope and purpose of the experiment. Unfortunately, it is often overlooked by experimenters, who tend to begin the experimentation process with collecting data. A list of the overall research goals and specific scientific questions to

be answered by the experiment should be created and referred to throughout experimentation. Coleman and Montgomery (1993) advised that these objectives for experimentation should be unbiased, specific, measurable, and of practical consequence.

When developing the list of objectives for the experiment one should be careful to list only the scientific questions that relate to the overall research goals (Dean and Voss 1999). A common mistake made by experimenters is to try to answer too many questions with one experiment, which often increases cost and complexity. When determining the specific questions to be answered, it is important to decide whether the experiment will be "stand-alone" or sequential in nature. By *stand-alone* we mean that there is no current plan for immediate follow-up experimentation. By *sequential* we mean that the experimenter intends to perform follow-up experiments essentially immediately after analyzing the results of the current experiment. Box (1999) referred to this classification of experimentation as one-shot versus immediate sequential experimentation. In a one-shot experiment the model form is assumed to be known and fixed, and a single experiment will be conducted to estimate the model as best as possible; for example, alphabetic-optimal designs. See Box (1982, 1999) for more discussion on optimal designs and one-shot experimentation.

In contrast, a designed experiment that is part of immediate sequential experimentation formally builds upon the information from previous runs. In this case, the model and relevant factors evolve as more information is obtained. For example, the primary objective for the first of a set of experiments may be factor screening. After the important variables are identified, the next experiment determines whether the current settings for the factors result in a response that is near optimum (Myers et al. 2009). Response surface methods are a collection of statistical tools that allow for this type of iterative learning with the use of mainly classical designs.

Generally, one-shot experiments are also part of a sequential process that builds upon previous experimental knowledge. For example, consider a one-shot industrial experiment where during the experiment the process is adjusted because experimenters notice that the temperature settings are too high. The one-shot experiment is now more of a sequential experiment because experimenters are

using previous information about the process to make decisions on how to move forward. A more efficient way of conducting this industrial experiment may have been to consider sequential experimentation from the beginning.

Almost all one-shot experiments have a more long-term sequential component. Most agricultural studies seem to be one-shot experiments because, unlike many industrial experiments, the results are not immediate. It may take months or even years to carry out the experiment and obtain data. However, it would be senseless to ignore these results when designing a future study. The use of previous information in a designed experiment is a type of sequential experimentation. Therefore, it can be argued that all experimentation is sequential in nature. This statement was argued in Box (1992–1993) and illustrated by Box (1992–1993) as shown in Figure 2. Vining (2011) also provided technical advice on the relationship between DOE, RSM, and sequential experimentation.

When determining the overall objectives for a planned experiment, it is important to discuss whether the experiment is immediate sequential or one-shot. This classification should also be kept in mind at all stages of design.
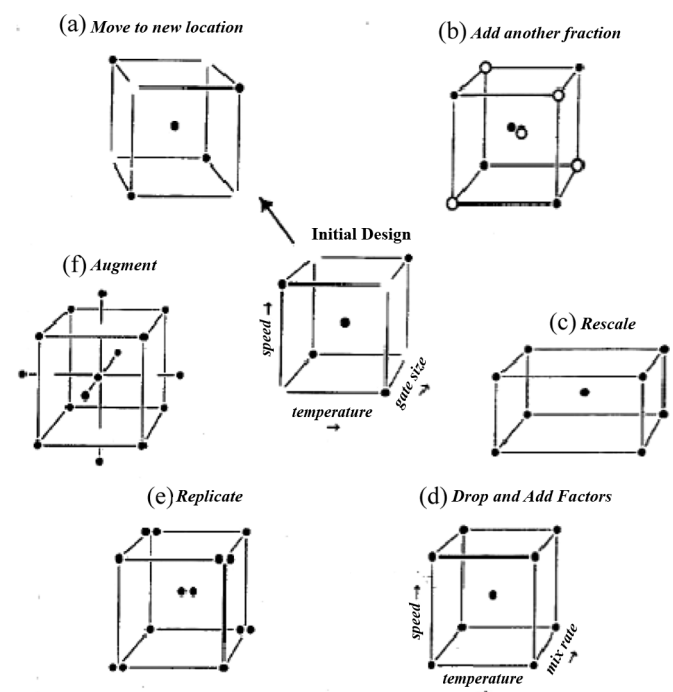

FIGURE 2 Alternatives for a second set of runs depending on the results from the initial set.

Section 2: Select Response Variables

The second step of the flowchart is to select response variables. When the measureable objectives have been stated, the response variables can be determined. It is important in this step to list the response variables along with the units of measurement for each response variable. It is also advisable to create a detailed protocol with directions on how measurements should be taken and how the data will be collected for analysis. The protocol reduces unnecessary variability in the data collection process and eases data analysis at later stages of the experiment.

When determining response variables it is important to obtain as much information-rich data as possible. In most cases this involves measuring on a continuous scale rather than an ordinal or binary scale, which can be considered information-poor data. Data with a higher scale can always be converted to lower scale data but not vice versa. For example, suppose that a local pub is holding a darts competition. In order to be in the competition, interested participants must go to the pub and throw 50 darts. The miss distance from bulls-eye for each of the 50 darts is recorded. These data are then used to classify each participant in either the advanced or intermediate dart throwing group. The miss distance measure is considered information-rich data. We have a much better idea about the abilities of the participants by looking at the miss distance data compared to hit/miss data. The hit/miss data are considered information-poor data. One participant may hit the bulls-eye 10 times but the rest of the darts miss the dartboard completely. Is this participant better than a person who hits the bulls-eye 9 times but the remaining 41 darts all hit the dart board? The miss distance data can also be easily converted to hit/miss data, but miss distance data cannot be reconstructed from the hit/miss data.

It is important to determine the response variable or variables of interest early in the design stages because the type of data collected weighs heavily on the selection of the appropriate analyses. It is also important to relate each scientific question from the first checklist stage with a response variable to ensure that all scientific questions will be answered with the data collected. Coleman and Montgomery (1993) provide further guidelines for choosing response variables.

TABLE 1 Outcomes of a Hypothesis test

		Truth	
		Null hypothesis is true	Null hypothesis is false
Test decision	Null hypothesis is true	Confidence level$(1 - \alpha)$ (correct decision)	Type II error(β)
	Null hypothesis is false	Type I error(α)	Power$(1 - \beta)$ (correct decision)

After the response variable has been chosen, the next step on the flowchart is a hexagon labeled "Type III Error Check." Many are familiar with Type I and Type II errors, which measure the risk of reaching the wrong decision based on data collected. Table 1 shows the relationship between Type I error, Type II error, confidence, and power for a given hypothesis under test. A Type III error, on the other hand, is committed by providing an elegant answer to the wrong question. Type III errors often occur when the responses and/or the factors truly necessary to address the question of interest are inconvenient or expensive to measure or control. The experimenter then chooses to use surrogates that provide poor information about the actual problem. Because the responses and factors used in the experiment are now easy to change and to measure, a very elaborate experimental design and analysis can be developed. Ultimately, however, the experimental results do not address the actual problem. Instead, they address some other problem that happened to be convenient to solve. Kimball (1957) illustrated Type III errors through statistical consulting examples.

It is important to take a step back and ask yourself whether your conclusions at each step of the flowchart correspond to the overall questions you set out to answer in the beginning. The Type III hexagon appears throughout the flowchart, indicating that this question of whether a Type III error is being committed should be considered often. Specifically, at this Type III error check, the team should ask themselves, "Are the response variable(s) identified appropriate to address the goals of the experiment?"

319

Section 3: Identify Sources of Variation

A critical next step when planning for any experiment is to identify all sources of variation. This step can be done after, or in conjunction with, selection of the response variable. A source of variation is anything that might influence the performance of the process or system under consideration. In other words, it is anything that could cause the data collected in one observation to differ numerically from the data collected in another observation. Some sources of variation may be very minor, having relatively little impact on the response variable, and can be ignored, whereas other sources of variation will significantly impact the response variable. It is these major sources of variation that must be planned for in the experiment. We can further classify these major sources into one of two categories: treatment factors and nuisance factors.

This exercise, identifying and classifying all sources of variability, helps to facilitate important and necessary conversations between the experimenter and the statistician. For example, it is likely that the experimenter is able to easily identify the potential sources of variation; label how controllable, if at all, a source of variation may be; provide insights into how large the impact of a source may be; suggest appropriate and inappropriate treatment factor levels/settings; and recognize the practicality, speed to implement, and costs associated with the sources of variation. The statistician can assist the experimenter in correctly classifying the sources (treatment or nuisance) and how they should be dealt with in the context of the experiment. The statistician also can make sure that the selected range of factor levels/settings is wide enough to identify observable effects and can explain to the experimenter the impact that a choice may have on final conclusions.

Selecting Treatment Factors, Levels, and Ranges

Treatment factors are the major sources of variation that are of particular interest to the experimenter. They are measurable, controllable, and thought to be (very) influential. We define an influential factor to mean

1. in general, the factor does in fact impact the response of interest.
2. the range of proposed settings for the factor is in a region where they will impact the response of interest.

Example 6-2 in Montgomery (2009) outlines an experiment that illustrates this basic point. The response is the production rate for an industrial filter press. From chemical engineering theory, one knows with certainty that the biggest driver of the production rate is the pressure. However, the one factor that is not statistically significant is the pressure. Why? The actual relationship between pressure and the production rate is highly nonlinear. There reaches a point where the rate hits an asymptote once the pressure exceeds a certain threshold. The levels for pressure used in the experiment were in the region where the production rate was no longer dependent on pressure. Pressure met condition 1 above, but it did not meet condition 2.

Treatment factors can be continuous or categorical and narrowly defined or broadly defined. It is important to clearly specify how these factors will be measured and how they will be controlled at the desired values.

Coleman and Montgomery (1993) recommended recording the following information for the treatment factors: the normal control variable level at which the process is run along with the distribution or range under normal operation conditions, the possible range to which it can be set and precision to which it can be measured, proposed treatment factor level settings, and the predicted effect that a setting will have on the response variable. We also suggest recording any potential constraints to take into consideration (e.g., easy-to-change vs. hard-to-change, problematic treatment combinations with other treatment factors, etc.).

The experimenter plays a key role in determining the ranges over which the factors will be varied and the number of levels at which the design will be run. Two concepts are very important to this conversation: the region of operability and the experimental region. The region of operability represents all possible settings for the factors. In many engineering situations, the region of operability represents the ranges that the equipment and materials can operate. The experimental region is a subset of the region of

operability. The experimental region represents the range of settings for the factors being considered for *a specific experiment.*

The objective for the experiment being planned helps to determine how large to set the experimental region. Most experiments use low-order Taylor series approximations as the basis for the model. In fact, most experimenters assume either a strict first-order Taylor series (only main effects) or main effects plus some interactions. For optimization or for predictive capability, we often assume a second-order Taylor series. The choice of the experimental region must reflect the order of the model proposed for the specific experiment. Taylor series assure that the model is a reasonable approximation *over some neighborhood.* If we pick a neighborhood that is too small, then the important effects do not have the opportunity to show their influence. If we pick a neighborhood that is too large, we introduce bias due to an underspecified model. Myers et al. (2009) emphasized that the choice of factor levels (selecting the experimental region) is essential to the success of the experiment. For instance, selecting ranges that are too narrow may result in an important factor becoming insignificant in the analysis. This is particularly important for a screening/characterization experiment. Keeping in mind the sequential nature of experimentation, as one learns more about the process, choosing ranges will become easier.

Determining how many levels to use also depends on the objective of the experiment. For example, typically "screening" experiments, which seek to identify the most important factors, use only two levels for each factor. On the other hand, designs intended for formal product or process optimization or for characterization typically use either three or five levels for each factor. Designs involving categorical factors may have even more levels.

Proper selection of factors and their specific levels requires true subject-matter expertise. In that sense, all experiments require and build upon prior knowledge, which sounds very Bayesian. Well-planned experimentation does not occur in a vacuum. Almost always the engineer or scientist has at least some tentative theory to guide the selection of the factors and their appropriate levels. In the vast majority of experiments, the scientist/engineer has much more previous experimental results for the same or similar processes, products, or systems upon which to build.

One of the authors helped to facilitate a series of experiments to correct a problem with the nickel–hydrogen batteries intended for the International Space Station. An engineer from the National Aeronautics and Space Administration actually approached this author saying that he would be a "hero to the country" if this experiment proved successful. Of course, the author involved is purely a statistician, with no background in electrochemistry. He was in no position to postulate the important factors, much less their levels. That responsibility belonged to the chemists and chemical engineers assigned to the team. Their knowledge was essential for the success of the experiment. The statistician's contribution was to make the process more efficient and informative. The true heroes were the chemists and chemical engineers applying their subject matter expertise to select the correct factors and their levels.

It is very tempting to conclude from this discussion that experimenters should pursue formal Bayesian approaches to planning experiments in order to take full advantage of the prior information available. One must exercise extreme caution in this regard. When employing a Bayesian design methodology, the success of the experiment depends on the specification of the prior. In extreme cases, an overly precise prior can result in the experiment having little additional benefit beyond confirming prior beliefs. Classical designs, on the other hand, are robust to the accuracy of prior information (as long as factors are influential) and do not prohibit incorporating prior information into a Bayesian analysis if so desired.

The key point is that subject-matter expertise is necessary as a starting point to plan experiments, but one should put very serious bounds on the extent to which they apply to the new experimental situation. Each new experiment, even within a formal sequence of experiments, deals with a new situation. Past information is relevant as a starting point, but it is not wise to let this prior information dominate. Classical experimentation uses prior information only to suggest the proper region, factors, levels, and a tentative model. Formal Bayesian optimal designs allow the prior information to play a larger role; in some cases, a much larger role. This larger role is beneficial only to the extent that the prior information truly applies to the new situation. In many

cases, too much dependence on the prior information leads to misleading and potentially poor results. Ultimately, within the scientific method, the data must stand alone, independent of any specific theory that claims to explain them.

Nuisance Factors: Blocking, Covariates, and Noise

Nuisance factors are those sources of variation that are not of particular interest for the experiment or cannot be controlled due to constraints in the experimental setup but have too much of an influence on the response variable to simply be ignored. We want to design the experiment in such a way as to enable the effects of the nuisance factors to be distinguished from those of the treatment factors.

Determining how to deal with a particular nuisance factor and reduce its impact depends on its nature (controllable, uncontrollable, or noise). A few common options include fixing the level of the nuisance factor (which may limit the scope of the experiment), treating it as a blocking factor, or, if measurable, treating it as a covariate. Keep in mind, though, that the analysis will change depending on the approach.

If the experimenter is interested in the variability of the response as the experimental conditions are varied, then the nuisance factors are deliberately included in the experiment. Such nuisance factors are termed *noise factors*, and this type of experiment is referred to as a *robust design*. An example of this would be designing systems that are insensitive to weather conditions.

Coleman and Montgomery (1993) recommended recording each nuisance factor, its measurement precision, the strategy (blocking, covariate, etc.), and any anticipated effects. The process of identifying sources of variation is continued until all known sources of variation have been identified and dealt with.

After all sources of variation have been identified and classified, the next step on the flowchart is a hexagon labeled "Type III Error Check." Again, a Type III error is committed by providing an elegant answer to the wrong question. Specifically, at this Type III error check, the team should ask themselves, "Are all of the sources of variation identified, are they correctly classified, are they expected to affect the response variable, and, most important, are they appropriate to address the goals of the experiment?"

Section 4: Choose the Experimental Design

The fourth section of the flowchart for planning an experiment is to choose the experimental design. Choosing the experimental design involves developing a base design for the experiment, which will then be evaluated on whether or not the design is consistent with resources. If this proves to be the case, then the base design becomes the final design used for experimentation. If not, the process is repeated until a base design is developed that is consistent with available resources or resources are expanded to accommodate a larger test design. Often, documenting Type I and Type II error rates as a function of test size provides a useful tool for justifying additional test resources. The choice of experimental design is often an iterative process. The statistician contributes heavily to this process, reflecting the focus of many existing experimental design textbooks. However, it is common for a statistician, using standard designs or statistical software, to generate a design, share it with the team, and learn of a new constraint or challenge that was not discovered in the planning process up to this point.

Figure 3 illustrates the process for determining the base design. It elaborates on the steps outlined in the main flowchart given in Figure 1. The first step in the design flowchart is to consider whether any of the controllable treatment factors are hard to change. If there are hard-to-change treatment factors, split-plot designs should be used to structure the experimental design from this point on. The next step of the design flowchart is to determine the number of levels for the factors. The resulting base design will be dependent on whether all factors have two levels or some have more than two levels.

Key considerations in choosing a design include the shape of the experimental region, restrictions on randomization, and cost. The simpler the design, the more likely it is to be carried out correctly. Furthermore, a complex design will have a complicated analysis. Therefore, choose the simplest, most straightforward design that the situation allows.
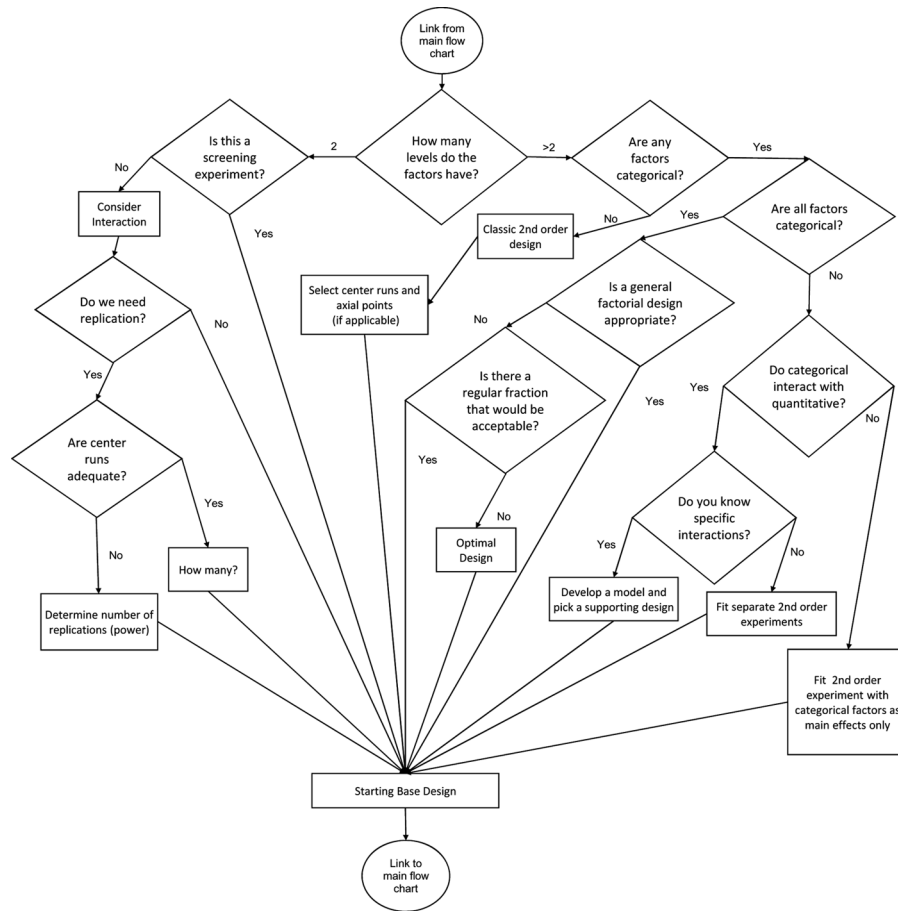
FIGURE 3 Flowchart for choosing designs.

The experimental region greatly impacts the choice of design; thus, we must characterize the shape of the experimental region. Is the experimental region cuboidal, spherical, or irregular? The experimental region is cuboidal if all combinations of factor highs and lows make sense, which is illustrated in Figure 4. On the other hand, suppose some combinations of factor highs and lows lead to problems, but all high and low values can be run at some
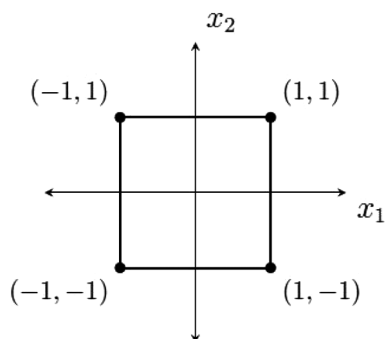

FIGURE 4 Factor levels for cuboidal experimental regions.

moderate combination of the other factors. In this case the experimental region is spherical, as seen in Figure 5. We call the experimental region irregular if it is neither cuboidal nor spherical. Classical designs can be used when the experimental region is cuboidal or spherical. Common classical designs include fractional-factorial designs, central composite designs, and Box-Behnken designs. In many cases modifications can be made to transform irregular experimental regions into spherical or cuboidal experimental regions. Though transformations of the experimental region are case dependent, one possible option is to reduce the ranges of the factors to settings where the new design region can accommodate classical designs. This idea is illustrated in the distillation column case study section.

Some people today advocate for the universal use of optimal designs in experimentation. These people rightfully point out that many classical designs are optimal, and the computer algorithms used to generate optimal designs can produce classical designs
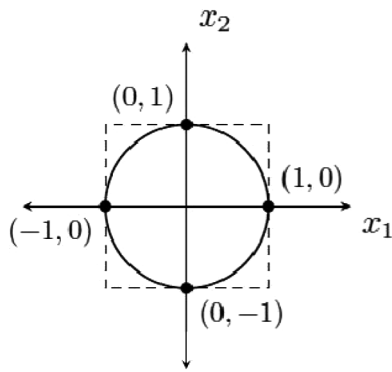
FIGURE 5 Factor levels for spherical experimental regions.

under certain conditions. As a result, they argue that one should use one's favorite optimal design software to plan all experiments.

Box (1982) forcefully argued against the universal use of optimal designs. He pointed out that these optimal designs are "optimal" only with respect to a very narrow mathematical criterion. An optimal design in terms of a very specific mathematical criterion can be very non-optimal in terms of other very important statistical criteria. Box argued that the classical designs possess many highly desirable statistical characteristics above and beyond the narrow mathematical criterion used to generate optimal designs. Optimal design software typically recommends non-classical designs even when there are highly competitive classical designs that would be appropriate for the specific experimental context. We suggest extreme caution in the use of optimal designs. We do recognize that optimal designs are well-suited for certain experimental design situations, including designs where a large number of categorical factors are necessary or when the experimental design region is highly constrained and cannot be transformed to a regular region. However, we do not recommend their universal use. The use of classical designs is a robust test design strategy. When optimal designs are employed, experimenters should always consider multiple design optimality and robustness criteria.

Any restrictions on randomization must be accounted for when selecting a design. Restrictions on randomization may necessitate the use of blocks and/or split-plots. Scheduling issues could lead to blocking. For example, perhaps the experimenters only have access to the production line for a limited amount of time each day; thus, the experiment will need to be performed over several days. Another example that leads to blocking occurs when materials used in the experiment are produced in batches. In some cases one batch does not contain enough materials for the entire experiment and multiple batches must be used. Having some factors whose levels are difficult to change could lead to a split-plot design. In a split-plot design the hard-to-change factors are changed less frequently than the easy-to-change factors. For example, in industrial experiments oven temperature is often time consuming to change. Thus, an experimenter may want to change the oven temperature less often than the other factors.

Cost must also be considered when choosing a design. How many runs can we afford? Keep in mind that we do not want to spend all of our resources on one experiment. Consider the budget for this experiment within the context of the budget for the overall project. Box (1992–1993) recommended using no more than 25% of resources in the first experiment.

Selecting a design involves using all of the information gathered in the first three sections of the flowchart. Subject-matter experts provide essential knowledge regarding any constraints on the experimental region and any restrictions on randomization. Once a potential design is identified, determine how the data will be analyzed after the experiment is performed. Make sure you know how to analyze the experiment before the data are collected. If several designs are under consideration, compare the trade-offs of the various designs.

After selecting a design and ensuring that it is consistent with resources, the next step on the flowchart is a hexagon labeled "Type III Error Check." This is a very important Type III error check. There are many questions the team should ask themselves at this point, including

- "Is the design executable?" Often elegant statistical designs fail to be executable in application, which results in a different experiment being run than the one planned. It is hard to know whether a design is executable until the full run matrix is viewed by the whole team. As a result, it is essential at this point to make sure that all stakeholders have reviewed the full design to ensure that all treatment conditions are possible and the execution order is reasonable.

- "Will the analysis that results from this design address the goal of the experiment?" It often helps to simulate data under the given design and run a mock analysis to illustrate to the test team the types of conclusion that can be drawn from the proposed experiment.
- "Does the design provide adequate power to address the question of interest?" Power analysis tells us about the probability of detecting significant test outcomes and is therefore essential for assessing test adequacy. Power calculations based on the proposed sample size, assumed size of the factor effect, and assumed variance in the process should be computed before conducting the experiment to determine whether the analysis for the proposed design will have the ability to detect meaningful scientific effects.

Section 5: Train Experimenter/ Conduct the Experiment

The fifth section of the flowchart for designing experiments is to train the experimenters and conduct the experiment. When conducting the experiment, make sure that the experimental protocol is clear to everyone and that everyone understands why the protocol is important. This protocol should include detailed steps for both collecting and recording data. It is vital that any deviations from the protocol be recorded because these deviations may lead to adjustments in the analysis. Run sheets are useful for recording information during an experiment. Run sheets also provide the experimenter with the run order for the experiment. It is important that this run order is strictly followed to preserve the validity of the experimental design. Even when the run order is randomized, situations arise when the factor levels remain the same from one run to the next. It is important for the experimenter to be trained about resetting factor levels after each experimental run. Webb et al. (2004) provided a discussion of the impact of not resetting factor levels. Even when the experimental protocol is strictly followed, unforeseen complications may arise during any experiment. It is useful for the experimenters to have contingency plans outlined prior to conducting the experiment in case adjustments to the experiment must occur. It can also be beneficial to perform trial runs before the actual experiment is conducted to determine potential problems that may arise.

Section 6: Analyze the Data

The sixth step is to analyze the data. Do not analyze the experiment you planned; analyze the experiment that you actually ran. If there were deviations from the experimental protocol, then these deviations need to be taken into consideration in the analysis. Graphical representations of data are often an invaluable aid to the understanding of the data. Exploring the data using graphical methods often provides useful insights that cannot be obtained by a formal test. As part of the analysis, fit a model relating the explanatory variables and the response variable. The analysis may include hypothesis testing, confidence intervals, and prediction intervals.

Bayesian approaches to statistical analyses are very popular now. Once again, experimenters need to exercise caution as to their use. The scientific method requires the data to stand alone. The use of prior information via Bayesian methods leads to an analysis that is biased to the specific explanation for the data upon which the prior information resides. This bias is why analysts in the past have emphasized "diffuse" and "non-informative" priors when doing Bayesian analyses. Bayesian analyses can provide a methodology for incorporating prior information. However, this information is beneficial only to the extent that the prior information truly applies to the new situation.

Deming (1975) categorized studies as either enumerative, which deal with static populations, or analytic, which deal with dynamic situations. Bayesian approaches to analysis have their greatest potential for success within an enumerative study of a static population. In such a situation, the prior information has the greatest probability of being relevant. The vast majority of experimentation involves a "new" set of conditions, which then implies that prior information may be of limited value. As a result, it is typically useful to put severe bounds on its use. Classical, frequentist approaches only use the prior information to suggest potential factors, levels, and a tentative "full" model.

It is true that Bayesian approaches allow the data to dominate the analysis once the sample size becomes sufficiently large. However, most experimental design scenarios require much smaller sample sizes to be cost and time effective. One way to view the impact of prior information (through the prior distribution) is

"inertia." Such inertia often is a good thing for enumerative studies with large sample sizes. The large sample sizes allow the data to overcome the inertia provided by the prior. The small sample sizes often dictated by practical experimental situations cannot overcome this inertia.

A real example helps to illustrate the inertia in Bayesian analysis with small sample sizes. One of this article's authors consults with the National Aeronautics and Space Administration. One project involved an experiment intended as a follow-up to a previous experiment that suffered from several issues. A Bayesian analysis was used to analyze the results from the new experiment with a prior distribution based on the results of the original experiment. The analysis provided a full distribution of the posterior distribution and not just the mean and variance. However, the posterior distribution was bimodal, with one peak clearly reflecting the prior distribution and the other peak reflecting the data from the new experiment. As with many real experiments, the sample size of the new experiment was not large enough to overcome the inertia of the prior distribution. From a frequentist perspective, the prior information significantly biased the experimental results.

Section 7: Conclusions and Recommendations

After completing all of the necessary steps in the flowchart, we finally come to the seventh section, which is to draw conclusions and make recommendations based on the analysis that has been performed in section 6. The important questions that one must now address are what did the data tell us and how are we going to respond to it? Recall that, like the scientific method, experimentation is an iterative process. There is always something to be learned and acted upon, whether this is done immediately with another experiment or perhaps saved for a later time if the overall objectives and specific questions have been satisfactorily addressed. Properly documenting the conclusions and recommendations will only benefit the experimental progression.

It is still important to consider whether we have committed a Type III error: Did we come up with an elegant solution to the wrong problem? Clearly, finding out that we have made a Type III error at this point is a serious matter. What should we do if we did? First, what question did we really address by this experiment? How far off is this question from the original one? Does what we learn still have value? What should we have done differently? What recommendations do we have to prevent making a Type III error in future experimentation? The key points are to (1) summarize whatever valuable information was learned from the experiment, even if it really did not address the original questions; (2) identify where things went awry; and (3) make concrete recommendations to minimize the chances that future experimentation falls into the same trap.

It is crucial that the results be made interpretable to everyone who is involved and not just the statistician. Presenting the results in a meaningful way, possibly through graphs and figures, is absolutely essential when trying to explain the conclusions and when one has the difficult task of trying to convince others of the results and make them believers of DOE and the importance of a properly planned experiment.

After drawing conclusions from the experiment, the scientists must decide how to proceed in the experimental process. Box (1992–1993) included a wonderful diagram showing possibilities for following up a fractional factorial experiment, which is illustrated in Figure 2. Recommendations for follow-up include moving to a new experimental region, dropping or adding factors, augmenting the current design, or verifying conclusions using confirmation runs.

Case Study: Distillation Column

As our motivating case study, we consider improving the yield of alcohol from an ethanol–water distillation column. Our improvement project has a budget of 64 runs. In this process, the response of interest is the concentration of ethanol in the distillate. Figure 6 illustrates the distillation process.

In operating the distillation column, the experimenter controls several knobs. These knobs, along with their ranges of operability and recommended settings from a pilot study, are given in the Table 2. Note that the ranges in Table 2 are the ranges at which the column can operate; they do not necessarily reflect the ranges at which ethanol–water distillation works.

The feed consists of a solution of 6% by volume ethanol and 94% water. The feed enters the column
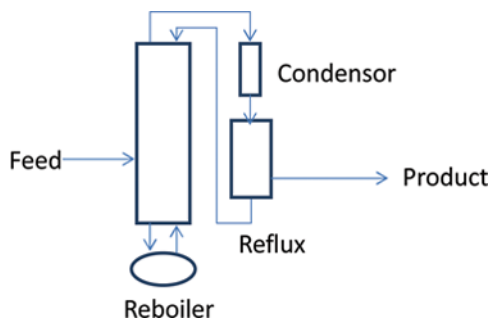
FIGURE 6 Distillation process. (Color figure available online.)

and goes into the reboiler, which can withstand a maximum pressure of 6 atm. The reboiler pressure is the pressure at which pure water boils at the set reboiler temperature. The solution then passes from the reboiler back into the column and then to the condenser. If the condensate temperature is too high, then the solution will not condense. The distillate immediately passes from the condenser into a holding tank with the same temperature as the condensate. The maximum pressure that the holding tank can withstand is 4 atm. The holding tank pressure is the pressure at which the distillate vapor is in equilibrium with the distillate liquid for the set condensate temperature. Chemical engineering theory and practice indicate that the condensate temperature must be lower than the feed temperature and that the feed temperature must be lower than the reboiler temperature. The reflux ratio is the amount of reflux divided by the amount of distillate. The reflux returns to the reboiler from the holding tank and the distillate is removed from the column. Note that the column cannot turn water into alcohol (i.e., the distillate rate cannot be greater than 6% of

the feed rate, because the feed contains only 6% alcohol).

Using the distillation column as an example, we will go through the steps for designing an experiment. In step 1 we must define the problem. The overall goal of our research is to improve the yield of alcohol in the distillation process. However, our goal for this particular experiment is to determine what factors influence the yield. In step 2 we select the response variable(s). In this case, the response of interest is the yield of alcohol. This response is measured as the concentration of ethanol in the final product.

In step 3 we identify the sources of variation in the process. We know that the knobs the operator can turn on the distillation column are feed rate, feed temperature, product rate, condensate temperature, reboiler temperature, and reflux ratio. An additional source of variation is the quality of the feed. Although the feed nominally contains 6% ethanol, this concentration may vary. Because the variation in feed quality is minor, we decide to ignore it for now. Recall that one restriction we have is that $T_r > T_f > T_c$, where T_r is the reboiler temperature, T_f is the feed temperature, and T_c is the condensate temperature. This constraint greatly reduces the combination of temperatures we can run. Thus, we define the temperature factors in terms of gradients: T_r, $T_{G1} = T_r - T_f$, and $T_{G2} = T_r - T_c$. An additional restriction is that the product rate must be less than the percentage of alcohol in the feed. Thus, we consider as a factor the product rate as a fraction of the feed rate: $R_{pf} = \frac{R_p}{R_f}$. Consultation with a chemical engineer led to the factors and ranges for the factors for the initial experiment given in Table 3.

In step 4 we choose a design. Because we are interested in determining which factors influence yield, we decided to use a screening design, with two

TABLE 2 Ranges of Operability and Recommended Setting for the Distillation Knobs

Input	Lower limit	Upper limit	Pilot settings
Feed rate (R_f)	75 GPM	125 GPM	100 GPM
Feed temperature (T_f)	70°F	300°F	270°F
Product rate (R_p)	2 GPM	20 GPM	6 GPM
Condensate temperature (T_c)	100°F	300°F	220°F
Reboiler temperature (T_r)	250°F	500°F	275°F
Reflux ratio (RR)	10	100	25

TABLE 3 Factors and Ranges for the Initial Distillation Experiment

Factor	Lower limit	Upper limit
R_f	80 GPM	120 GPM
T_r	275°F	295°F
$T_{G1} = T_r - T_f$	5	15
$T_{G2} = T_r - T_c$	55	65
$R_{pf} = R_p / R_f$	0.056	0.058
RR	15	35

TABLE 4 Screening Design and Results for Initial Screening Experiment

Run	R_f	T_r	T_{G1}	T_{G2}	R_{pf}	RR	Yield
1	−1	−1	−1	1	1	1	0.712
2	1	−1	−1	−1	−1	1	0.712
3	−1	1	−1	−1	1	−1	0.317
4	1	1	−1	1	−1	−1	0.320
5	−1	−1	1	1	−1	−1	0.325
6	1	−1	1	−1	1	−1	0.332
7	−1	1	1	−1	−1	1	*
8	1	1	1	1	1	1	0.680

*The distillation column flooded during run 7 and no product was produced.

levels for each factor. We used a 2^{6-3} fractional-factorial design with eight runs. We chose a small design to conserve our resources. However, if necessary we could follow up with a fold-over design. For more information on fold-over designs, see Myers et al. (2009). In step 5 we conduct the experiment. The design and results of this experiment are shown in Table 4.

In step 6 we analyze the data. In order to determine which factors impact the yield, we examined the half-normal plot in Figure 7. We do note that one observation is missing. Box and Draper (1975) discussed the robustness of classical factorial designs to missing observations and outliers. The classical design structure ensures that the assumptions for

the half-normal plot are still met even though one observation is missing. Box and Draper (1975) used this fact as one of the justifications for preferring classical designs over optimal designs.

This examination showed that the reflux ratio was the only significant factor and was the factor driving the yield. Next we fit the model with reflux ratio as the only factor. The first-order model in coded units is

$$\widehat{Yield} = 0.5124 + 0.1899RR.$$

In step 7 we provide conclusions and recommendations. Our results indicated that in our current experimental region the reflux ratio was driving the yield. In order to move our process into a region with a higher yield, we recommend following up this experiment with steepest ascent. After steepest ascent, we recommend performing a new screening experiment to determine which factors influence yield in the new region of experimentation.

Case Study: Mine Detection Experiment

In this case study, as a contrast to the previous case study, we illustrate the applicability of the experimental design process to a highly variable, open ocean mine avoidance test. In recent years, the Director of Operational Test and Evaluation has
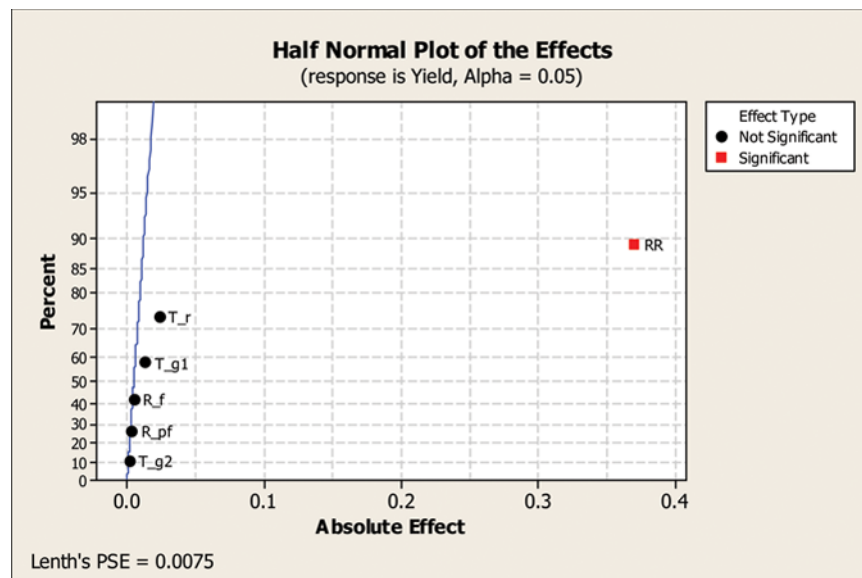

FIGURE 7 Half-normal plot showing that reflux ratio is the only significant factor impacting yield of alcohol. (Color figure available online.)

made it a priority for the DoD operational testing community to use experimental design in planning operational tests of major military acquisition systems. Often these systems are large and complex and the test environments contain many sources of uncontrollable variability. Using DOE provides a structured process for testing of these complex systems.

Forming an interdisciplinary team is necessary for the success of an experiment involving these large complex systems. In this case, the team must not only consist of engineers, scientists, and statisticians but also key stakeholders in the acquisition process, including members of the requirements community, developmental and operational testers, and the program manager. In the case study that follows the test team only consists of members of the oversight test organization. Therefore, even though we present a sound experimental design selection, when testing realities that were not considered in the design process were discovered, the team in charge of executing the test made real-time modifications to the test design that did not reflect planned experimental design.

It is common in testing major military systems, due to the high cost of the test and diverse group of stakeholders, to attempt to answer too many questions with one experiment. The result is often vaguely described experimental goals; for example, "Determine the survivability of a surface cargo ship." This goal, though providing a good starting point, does not focus the experiment to address a specific question. One important aspect of survivability is the ability to avoid underwater mines. In this example, we consider the effectiveness of a new degaussing system on a surface ship designed to reduce the ship's standoff range from a known mine danger area. This goal is more specific than the general goal of determining the ship's survivability and aids in refining the goal of the experiment. It is now clear that we want to characterize the ranges at which the cargo ship can safely transit mine danger areas with or without the degaussing system.

The next step is determining the response variables that measure the ship's survivability against different mine types. An obvious response variable is whether or not a mine detonates near the ship. However, this would be very expensive to test for two reasons. First and foremost it would result in loss

or damage of the surface ship (and present safety challenges) and, second, it is a binary response that is expensive to test. Fortunately, the Advanced Mine Simulation System (AMISS) provides the ability to simulate several mine types and provides information about whether or not the mine detonates based on the acoustic signatures of the ship. Figure 8 provides a notional picture of the AMISS.

The AMISS determines whether or not the mine detonates based on the acoustic signature. We can use the magnitude of noise in decibels, which is a key component of the acoustic signature and corresponds to whether or not the mine detonates, as a more informative response variable. Additionally, we can always translate this more informative response back to the less informative detonate/non-detonate response variable.

Now we must identify all sources of variation that might influence the outcome of the test. Clearly, the degaussing system status (on–off) is an important factor to consider. Other sources of variation identified by a team of testers and subject-matter experts include ship range from detection unit, ship movement direction, ship speed, machine line-up of the ship, depth of the ocean at the location of the mine simulator, and sea state. Not all of these factors can be controlled in the experimental design. For example, sea state is impossible to control but can be recorded. The ocean depth is another possible treatment factor, but this factor would be very expensive to study due to the fact that one would have to move the AMISS unit from one location to another. The AMISS unit is set to a fixed position on the ocean floor, so this factor is controlled by
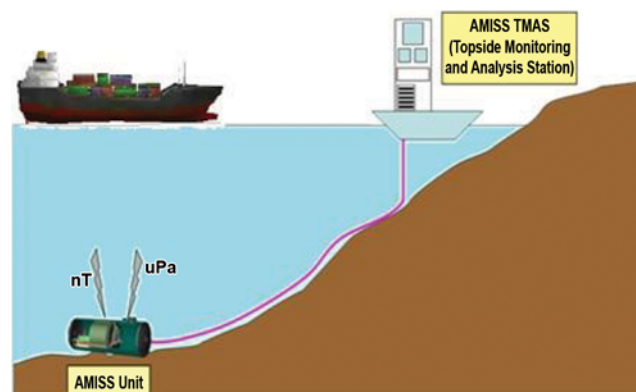

FIGURE 8 Notional picture of the AMISS system. (Color figure available online.)

TABLE 5 Sources of Variation for the AMISS Experiment Along with the Selected Management Strategy

Factor	Management strategy
Deguassing system (on–off)	Factor in experimental design
Ship speed	Factor in experimental design
Ship standoff range	Factor in experimental design
Ship direction	Fixed to run north–south legs; determined not to be a major factor
Machine line-up	Highly dependent on ship's speed, eliminated as a factor
Ocean depth	Fix at a constant level
Sea state	Uncontrollable, all runs will be conducted in one day to minimize variation; conditions will be recorded for reference

being held at a constant level. Table 5 summarizes the sources of variation for the experiment and the management strategy selected.

Determining the number of levels was not a straightforward process in this experiment. Subject-matter experts were not sure how much fidelity they required in the results and therefore did not know whether a first-order model (main effects only) or a second-order model was more appropriate. Furthermore, they were concerned that adding more levels to the test design would increase the size of the experiment beyond the available test resources. One day of testing on the ship was available, which allowed for somewhere between 20 and 30 runs.

As a result, several experimental designs were considered and presented to the test team. Table 6 provides seven common statistical designs for the three primary factors considered in this comparison study (speed, range, and degaussing status). These designs have been shown in the statistical literature to be a subset of the best designs available for three-factor tests.

TABLE 6 Designs Evaluated in Comparison Study

|	Design type	Number of runs	Estimable model terms	Design properties
1	Full-factorial (two-level)	8	6	Smallest possible design to investigate three factors and their interactions. Low power for detecting factor effects
2	Full-factorial (two-level) replicated	16	7	Increased power over unreplicated two-level factorial design. Adds the ability to estimate a three-way interaction over the un replicated design
3	General factorial ($3 \times 3 \times 2$), also referred to as a face-centered cubic design	18	9	Three-level designs for the continuous factors allow for the estimation of quadratic model terms
4	Central composite design (with one center point)	18	9	This design looks at five levels of speed and range. Balances variance and increases power across the design space
5	Central composite design (replicated center point)	20	9	Center point replication allows for an estimate of pure error (variability between runs under the same conditions) in addition to all other design benefits listed under design 4
6	Central composite design with replicated factorial points (large central composite design)	28	9	Large design has great power and the ability to estimate all desired model terms
7	Replicated general factorial	36	9	Large design with good power but not as powerful as the large central composite design
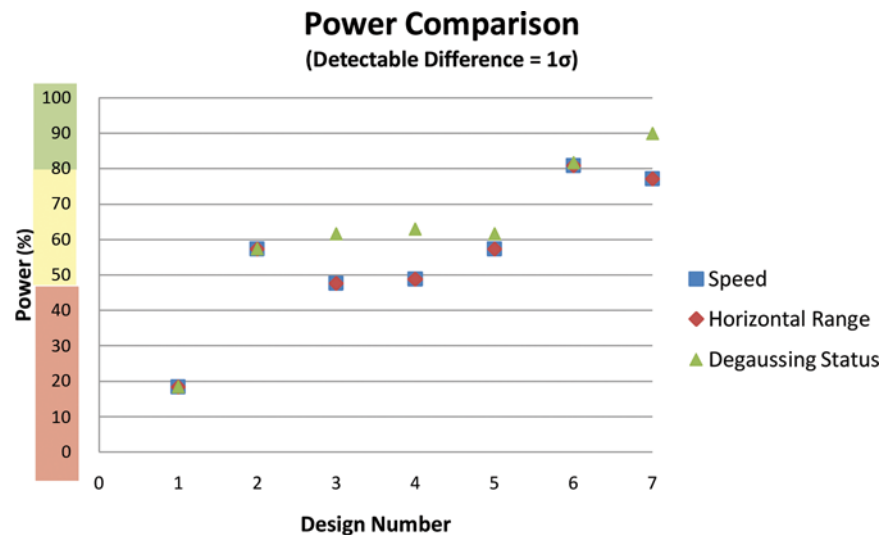
FIGURE 9 Power comparison for the main effects at the 90% confidence level. (Color figure available online.)

Figure 9 examines the trade space between design type and therefore sample size and power. The power was calculated using Design Expert 8 (Stat-Ease, Inc., Minneapolis, MN) for a change on the response of one standard deviation. Whitcomb and Oehlert (2001) discussed the mathematical details underlying the power calculations used in Design Expert 8. Notice only the smallest design (design 1) provides extremely low power, meaning that this test is at high risk for failing to detect the impact of the degaussing system (or any other factor).

After seeing the increase in power for degaussing status and the reasonable sample size associated with the central composite design (with replicated factorial points) the test team selected design 6. Additionally, the test team decided that the five levels of the central composite design in both speed and range were appropriate due to sensitivity of these factors on the level of acoustic noise required to detonate the mines.

A common problem with DoD experiments is that they generate classified data, as in this case. As a result, we regret that we cannot discuss the specific analyses and conclusions drawn in this case. However, it is worth while to note that despite the fact that the experiment was not executed as designed we were still able to draw meaningful conclusions about the effectiveness of the degaussing system from the data using regression analysis. This case study is an excellent example of the steps necessary to plan complex experiments.

CONCLUSIONS

Experimental design is a key aspect of the scientific method. The success of any experiment is highly dependent on the planning process and the experimental protocol established. In this article we illustrate important considerations that have led to success in practice. A few keys to remember:

1. Form the right team, involving all subject-matter experts and stakeholders.
2. Allow lots of time for the iterative planning process.
3. Clearly establish and document the goal of the experiment.
4. The selection of response variables, sources of variation, and their classification is important to the success of the experiment. Always do a Type III error check to make sure you have the right responses and factors selected.

Software can be useful in generating designs once the planning process is complete but, again, make sure to check that design is executable and addresses the goals of the experiments.

ACKNOWLEDGMENTS

funded by the Department of Defense under the advisement of Dr. Catherine Warner, Science Advisor DOT&E, and George Mumford, TRMC Test and Evaluation/Science and Technology Program Manager.

ABOUT THE AUTHORS

Laura Freeman is a Research Staff Member at the Institute for Defense Analyses. She provides statistical support to the Director, Operational Test and Evaluation on Department of Defense testing. Her areas of statistical expertise include designed experiments, reliability analysis, and industrial statistics. She has a B.S. in Aerospace Engineering, a M.S. in Statistics and a Ph.D. in Statistics, all from Virginia Tech. She is a member of ASQ and ASA

Anne Ryan is an Assistant Professor of Practice in the Statistics department at Virginia Tech. She received her Ph.D. in May 2011 from Virginia Tech. Her research interests include quality control, design of experiments, and statistics education. She is a member of ASQ.

Jennifer Kensler is a Research Staff Member at Riverside Research supporting the Scientific Test and Analysis Techniques in Test and Evaluation Center of Excellence at the Air Force Institute of Technology. She is an Adjunct Assistant Professor of Operations Research in the Department of Operational Sciences at the Air Force Institute of Technology. She received her Ph.D. in August 2012 from Virginia Tech. Her research interests include design of experiments and reliability.

Rebecca Dickinson is currently a PhD student in the Department of Statistics at Virginia Tech. Her research interests included reliability, design of experiments and quality control.

Geoff Vining is a Professor of Statistics at Virginia Tech. He is a Fellow of the ASQ.

REFERENCES

Box, G. E. P. (1982). Choice of response surface design and alphabetic optimality. *Utilitas Mathematica*, 21B:11–55.

Box, G. E. P. (1992–1993). George's column. *Quality Engineering*, 5: 321–330.

Box, G. E. P. (1997). Scientific method: The generation of knowledge and quality. *Quality Progress*, 30:47–50.

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

Box, G. E.P., Draper, N. R. (1975). Robust designs. *Biometrika*, 62: 347–352.

Coleman, D. D., Montgomery, D. C. (1993). A systematic approach to planning for a designed industrial experiment. *Technometrics*, 35:1–12.

Dean, A., Voss, D. (1999). *Design and Analysis of Experiments*. New York: Springer.

Deming, E. W. (1975). On probability as a basis for action. *The American Statistician*, 29:146–152.

Johnson, R. T., Hutto, G. T., Simpson, J. R., Montgomery, D. C. (2012). Designed experiments for the defense community. *Quality Engineering*, 24(1):60–79.

Kimball, A. W. (1957). Errors of the third kind in statistical consulting. *Journal of the American Statistical Association*, 52:133–142.

Marquardt, D. W. (1987). The importance of statisticians. *Journal of the American Statistical Association*, 82:1–7.

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

Myers, R. H., Montgomery, D. C., Anderson-Cook, C. M. (2009). *Response Surface Methodology*, 3rd ed. New York: Wiley.

Shoemaker, A. C., Kacker, R. N. (1988). A methodology for planning experiments in robust product and process design. *Quality and Reliability Engineering International*, 4:95–103.

Vanhatalo, E., Bergquist, B. (2007). Special considerations when planning experiments in a continuous process. *Quality Engineering*, 19: 155–169.

Viles, E., Tanco, M., Ilzarbe, L., Alvarez, M. J. (2008). Planning experiments, the first real task in reaching a goal. *Quality Engineering*, 21:44–51.

Vining, G. (2011). Technical advice: Design of experiments, response surface methodology, and sequential experimentation. *Quality Engineering*, 23:217–220.

Webb, D. F., Lucas, J. M., Borkowski, J. J. (2004). Factorial experiments when factor levels are not necessarily reset. *Journal of Quality Technology*, 36:1–11.

Whitcomb, P., Oehlert, G. W. (2001). Sizing fixed effects for computing power in experimental designs. *Quality and Reliability Engineering International*, 17:291–306.