Quality Engineering
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/lqen20

Guidelines for Planning and Evidence for Assessing a Well-Designed Experiment

James R. Simpson [a], Charles M. Listak [a] & Gregory T. Hutto [a]
[a] U.S. Air Force, Eglin AFB, Eglin, Florida

To cite this article: James R. Simpson, Charles M. Listak & Gregory T. Hutto (2013) Guidelines for Planning and Evidence for Assessing a Well-Designed Experiment, Quality Engineering, 25:4, 333-355

To link to this article: http://dx.doi.org/10.1080/08982112.2013.803574

Quality Engineering, 25:333–355, 2013
ISSN: 0898-2112 print/1532-4222 online
DOI: 10.1080/08982112.2013.803574

Taylor & Francis
Taylor & Francis Group

Guidelines for Planning and Evidence for Assessing a Well-Designed Experiment

James R. Simpson,
Charles M. Listak,
Gregory T. Hutto
U.S. Air Force, Eglin AFB, Eglin, Florida

ABSTRACT Since the design of experiments was first introduced by Fisher 90 years ago, this scientific and statistical approach to system interrogation for acquiring knowledge has enjoyed success across industries and among products, processes, and services. In the last decade, military test organizations have been promoting the use of design of experiments (DOE) as the preferred method of constructing and analyzing test programs. Increasingly, design of experiments is being used to greater effect and its impact is reaching groups less experienced in the method. Stories of successful application continue to have a common thread: detailed, effective planning. But not all organizations have members experienced in DOE test planning. And although planning papers and how-to case studies have appeared in the literature, the volume of these contribution types is dwarfed by theory and methods papers. If an experiment is planned and the planning process is documented, how would one go about assessing that plan? If the desire is to gauge the probability of experiment success as defined by a robust and truthful understanding of the system revealed upon analysis of the data, can the plan be assessed prior to test execution? This article proposes guidelines and evidence that span all phases of the experiment cycle, which can inform assessment of experiment planning soundness. The experiment cycle of plan, design, execute, and analyze (consistent with literature and texts) is used to structure the discussion, geared toward an audience somewhat familiar with the DOE method. Checklists are provided for each experiment phase coupled with descriptions of what would constitute fingerprints of successful implementation.

KEYWORDS design of experiments, design metrics, experiment checklists, statistical power, test phase, test planning

> To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of.—*Ronald A. Fisher, Presidential Address to the First Indian Statistical Congress, 1938*

Address correspondence to James R. Simpson, U.S. Air Force, Eglin AFB, Group OA, 53d Test Management, Eglin, FL 32542. E-mail: James.Simpson@eglin.af.mil

INTRODUCTION

Those in the business of developing and applying statistical methods for knowledge discovery understand that statistically designed experiments are a series of purposed, systematic tests designed, conducted, and analyzed to

learn about a process or system under study. The statistically based experiment discipline is used increasingly in product and process life cycle testing to mitigate program risk by revealing problems early in system design or operation. Clearly, testing informs and guides the management of an acquisition system so, fundamentally, testing must control the risk of erroneous conclusions based on test outcomes. Natural variability in outcomes and background lurking variables that corrupt the data contribute to these erroneous conclusions of declaring systems ready for the next programmatic stage. The focus on correct conclusions not withstanding, in the face of tighter budgets, testers should always maximize system knowledge gained from efficient testing. As such, knowledge only of whether to pass or fail a system without insight into system capability dynamics pales relative to drawing the right conclusion based on knowledge of the factors driving performance. Statistical design of experiments (DOE) is the method for managing random variation uncertainty while learning the most from limited resources in which factors influence performance (Box and Draper 1987; Box et al. 2005; Mason et al. 2003; Montgomery 2012; Wu and Hamada 2000). Kass (2006), in his book on war fighting experimentation, underscores the need for knowing factor influence in stating that the experiment must have the ability to not only detect change but isolate the reasons for change and relate the results to operations. Consider the following two diverse examples of testing: (1) an industrial experiment to develop a process to produce magnetically aligned carbon nanotubes and (2) the military qualification of a recent "strap-on" sensor and weapons kit for a special operations cargo aircraft.

Example 1

Research into the development of carbon nanotube composites has shown huge potential for composites superior in strength, weight, electrical properties, and heat conduction. Applications are widely diverse, including armor, sports, biotechnology, sensors, capacitors, and solar technology. A possible manufacturing process is to make nanotubes into buckypaper form. An experiment was conducted to study the contribution of fabricating parameters, including suspension concentration, sonication power and time, filtration vacuum pressure, and surfactant types on nanotube bundle quality as measured by surface quality, rope size, and pore size (Figure 1). Statistical modeling was also used to estimate the variability associated with manufacturing, the image taken, and the measurement processes (Yeh 2004).

Example 2

A cargo aircraft is equipped with roll-on pallets containing a gun system, computer control system, and operator control and display stations. The cargo door is equipped with canisters holding guided missiles, and additional rotatable sensors are attached to the exterior of the aircraft. The purpose of the kit is to convert a plain cargo aircraft into a gunship capable of engaging ground targets while conducting armed escort of ground parties. Such a system must demonstrate the capability to (1) communicate with, find, and orbit over friendly parties; (2) search for, correctly identify, track, and destroy adversaries; (3) defend oneself; (4) prepare the aircraft for flight (maintenance, fueling, arming), and much more. No single experiment can test all of these capabilities. Instead, with a multidisciplinary team of operators, maintenance workers, and engineers, the purpose is to create a campaign over a dozen different experiments, all with different design strategies, number of test events, and criteria for success. Indeed, few of these experiments require expending weapons;

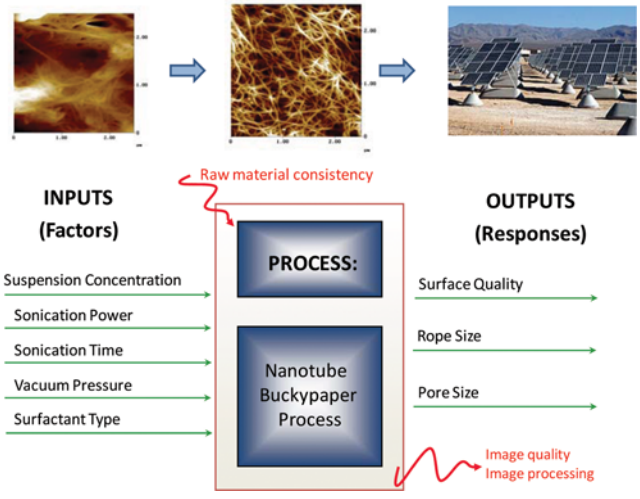

FIGURE 1 Nanotube buckypaper manufacturing process experiment to magnetically align nanotubes. (Color figure available online.)

they simply exercise communications or sensor systems. For example, one of the experiments examines the system's ability to automatically track a maneuvering ground target and another studies the system's capability to simultaneously target a moving target with both the gun and a missile. The checklists in this article are useful to grade the many experiments individually, as well as the collective set.

Though the designed experiments literature contains many case studies and some excellent planning sources (Atkinson and Cox 1974; Cox 1958; Shoemaker and Kacker 1988; Vanhatalo and Bergquist 2007; Viles et al. 2008), missing are what *characterizes* a well-planned experiment and how that excellence can be assessed and captured in indicators. Therefore, planning indicators associated with the four phases of an experiment were proposed. These criteria are particularly important in the beginning stages of any organization's conversion from "best efforts" to true designed experiments. The intent is to correctly explore a multidimensional test space and correctly distinguish the systems that work as desired from those that do not. The need for adequate testing in acquisition is clear: without a set of objective guidelines and indicators to assess a statistically designed experiment plan, we risk producing systems with an insufficient testing program to show that they are fit for their intended uses, which, in the military, has potential life or death implications. The detailed criteria presented herein reveal whether the design strategies contemplated are a well-considered series of investigation for increasingly expensive and technologically challenging operating environments; whether a thorough team approach process decomposition adequately specifies proper test conditions and measures the right success outcomes; whether the designs truly span the factor space (the region covering the range of factors) and have enough trials to arrive at the correct answer while minimizing risk; and, finally, whether the method of analysis is suitable to the design class chosen and is likely to correctly link changes in the factors to associated changes in performance.

The purpose of this article is to propose a rigorous experiment planning process useful for general application (industry and government examples provided) that will allow both testers and managers to improve and evaluate testing adequacy excellence. Fortunately, the probability of testing success can usually be well gauged immediately following the planning phase. Planning is critical because it not only requires a unified test team and a well-formulated series of objectives but the design of the test points, as well as test execution and expectations for analysis. The four phases of testing can be described as plan, design, execute, and analyze. The experimental design should be developed based on the needs of the analysis, and the analysis is only as capable as the experimental design allows. Proper execution will enable making sense of the data collected, and improper execution can completely invalidate an otherwise excellent design. As such, a test can be effectively vetted for planning, design, execution, and analysis goodness with indicators or evidence at the conclusion of the planning stage.

The article is organized according to the testing cycle phases and provides guidelines and indicators associated with each phase as evidence of success while stressing the interdependence of steps in one phase to elsewhere in the test cycle. For example, while one of the design (Phase II) guidelines addresses statistical power, higher power values also result in better empirical modeling precision in analysis (Phase IV). The guidelines and evidence can be effectively applied at the individual test entry level or to a series of tests required for complex systems with disparate objectives or even for single system evaluation across several phases of development and production. The four broad phases (Figure 2) with corresponding guidelines reflecting planning quality for testing are to (I) *plan* a series of experiments to

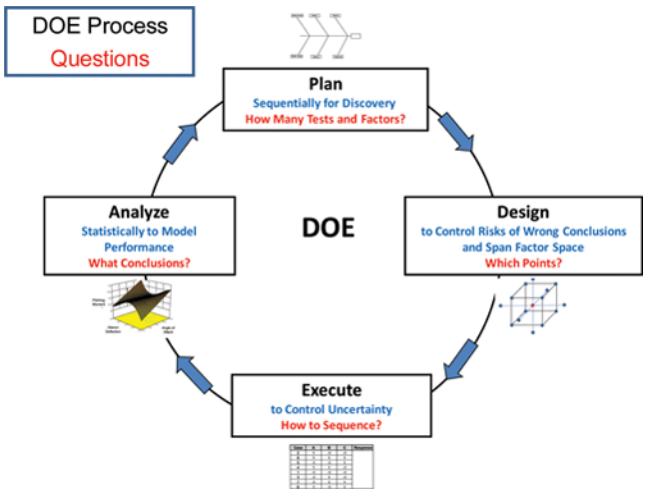

FIGURE 2 **Four phases of the test cycle using statistical design of experiments. (Color figure available online.)**

accelerate discovery, (II) *design* to control risks and to span the factor space, (III) *execute* to control uncertainty, and (IV) *analyze* statistically to model performance. Proposing these guidelines and evidence for assessment will hopefully encourage their use in DOE-based test plan development.

Phase I. Plan a Series of Experiments to Accelerate Discovery

The activities in the planning phase are straight forward and require no special statistical skills, just a collection of best practices and many hours of careful thought and effort by an assembled team of experts on the process under test. These planning and process decomposition activities usually occur over a period of several days, with our experience indicating 4–16 hours required for most tests. The team is usually led by a facilitator with experience in experimental design, with the most important contributions coming from the team of experts. Thorough process decomposition and understanding is the foundation of successful testing and is the most important aspect of systematic test planning (Barton 1997; Coleman and Montgomery 1993; Goh 2001; Hahn 1984). By examining the test planning documentation, it is usually obvious whether this step was done superficially or not at all.

A number of easy-to-use graphical tools may be employed to assist in the process decomposition, but the key is a positive attitude of commitment from all team members to participate fully in the process, foregoing criticism and pessimism during the activities. Tools used in other process-oriented strategies such as Lean and Six Sigma may contribute to this phase's success, including supplier–input–process–output–customer, process flowcharts, measures category charts, cause-and-effect (fishbone diagrams), and affinity diagrams. It should also be understood that, though the steps below are described sequentially, it is a rare test program that can complete this phase in one pass; two to three passes through the steps are more typical.

It cannot be overemphasized that successful tests are designed with all of the important stakeholders having a voice in the scope and outcome of the testing, as well as having contributions from design engineers, program office engineers, user representatives (operators), maintenance personnel, test facility experts, and anyone else with important information and views on the subject testing. The assembled team must balance sufficient expertise with a proper number of members to make progress. Groups of 5–10 have historically worked well, whereas groups of 20 or more are less effective. The team must commit the time to finish each stage of planning (Figure 3), discussed in detail here.

Problem Description

A concise description of the overall test problem, roles of each test organization, desired outcomes, and a draft of potential factors affecting performance are required. Significant budget constraints, physical limiting factors, overall comments on measurement limitations, and schedule expectations should be outlined for planning purposes in the problem description. Finally, the team should consider the acceptable level of risk for the current system under test; that is, what are the consequences of arriving at an incorrect conclusion as a result of the test program. The team should also identify the level of technical risk of the system. For instance, integrating a proven commercial jet engine on a large cargo aircraft carries much lower technical risk than an unproven sensor technology in a concept technology demonstration. Technically risky programs may suggest a more robust and powerful test design strategy than the test strategy for programs of much lower technical risk.

Relevant Background Research

The team should commission members to collect information that might shape the test program, such as previous testing of the system or analogous versions of the system. Previous data can be analyzed

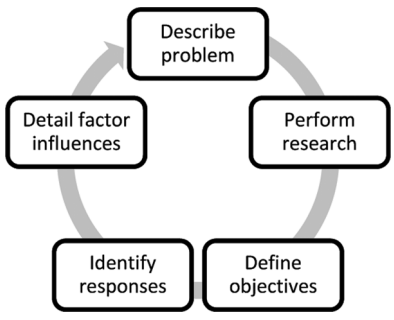

FIGURE 3 Stages of the planning phase.

to help predict probable performance or be used as a baseline for departure. History might suggest expected noise levels, suitable response variables (and ranges of values), active factors, and whether outlier runs might be expected. Historical research should be documented in the plan references and annotated in appropriate places of the test plan.

Objective(s) of the Test

Following the problem description and research, the team must determine the specific objectives of the test. Correctly specified test objectives, agreed to by all team members, are critical to progress. As a method of discovering objectives, it is often useful to explore questions along the following lines:

- How does one measure success (targets or standards) in the system under test?
- What constitutes failure in the system and how will we know the system has failed?
- Is the emphasis on making the process better or describing its current state?
- Is comparing performance to analogous systems or processes of interest?

Another useful approach is to ask what numerical answers are sought as a result of the test program—how fast, how far, how well? What knowledge is sought as a result of this test? Multiple objectives are routine and the following list describes commonly used objective categories for testing:

- *Screening* to find important factors affecting response performance and variability.
- *Comparing* results to a written *baseline*, standard, or goal (specification or requirement).
- Fitting a function of the factors to the response via modeling for *prediction* or interpretation as a concise explanation of the system in operation (*metamodeling*).
- Finding settings of controlled and uncontrolled variables that *optimize performance*.
- Finding *factors* that *affect* process *variation* in addition to the average response.
- *Characterizing* how the process works to reveal which and how factors matter.
- *Troubleshooting* the system to find conditions that lead to failure or poor performance.
- Explore factor variations in the presence of environmental noise that yield superior, stable performance (*robust product* or *process design*).
- Finding acceptable tradeoffs among conflicting objectives such as weight, cost, reliability, and performance (*multiple response optimization*).

Because discovering the appropriate outcomes of the test program is a vital step, this phase is best revisited, sometimes more than once, to clearly agree on the proper test program objectives.

Responses (Measures of Performance or Measures of Effectiveness)

Once the problem objectives are developed, the team must determine measures of system performance that constitutes "adequacy" while providing guidelines for measurement accuracy and precision. As an aid to discovering responses, a process flow diagram is often used by planning teams to specify each step in executing the system under test, including setup for the run, test event execution, and recovering from the run to nominal conditions. Figure 4 shows a simple process flow diagram depicting locating a ground party and establishing a protective orbit for the aforementioned strap-on sensor weapon kit.

The process flow diagram serves the twin purposes of defining the start (S) and end (E) of a test event while offering the opportunity to consider intermediate states of the process for *possible* response variables. In the current case, the test event consists of a searcher with specified initial conditions beginning a search pattern for a ground party. The event concludes when the ground party is identified and an orbit is established. Multiple intermediate system states can be measured as responses, including times, locations, and ranges to identify and locate the ground party; time to search; number of ground parties initially acquired then lost; false ground parties acquired; degree of correctness of the ground party identification; tracking errors on the ground party while it was being located; operator judgment of the ease of use of the human interface; operator workload; and the clarity/readability of the information displayed. In general, for any process it is useful to measure how long (time), how possible (success/failure), and how well (deviation from ideal).

To document the discussions on response measures, a table of response variables is useful for

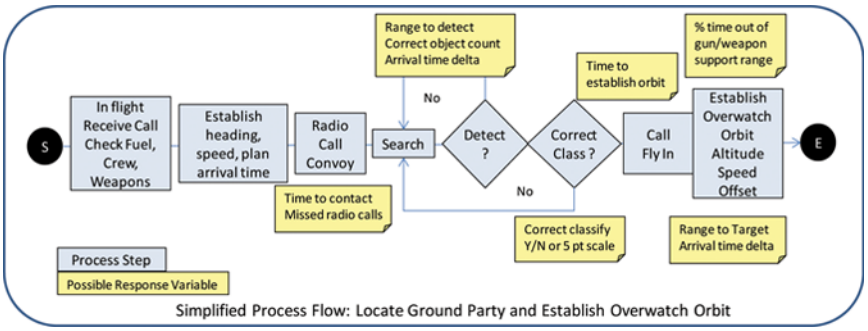

FIGURE 4 **Process flow diagram. (Color figure available online.)**

the team to compile (Table 1). Such a table forms a basis for team review of whether responses cover all objectives, whether the responses are largely objective and unbiased, and what data elements must be recovered (and how) in order to compute the response. Such a table can also serve to initiate creating detailed test procedures and a data collection plan.

A hallmark of a well-planned experiment is multiple, objective numeric responses listed for each set of test events and that the measurements have the precision and accuracy required of the experiment's objectives. Less useful outcomes are indicated by one or two responses, primarily consisting of "test team judgment," "operator opinion," "mission success," or other such subjective measures. Table 2 shows shaded example measures, with the most reliable being dark, less reliable measures colored medium, and the least informative light.

Well-designed, precise tests can answer their objectives most efficiently with fewer trials when the responses are repeatable, accurate, and measured on a *continuous* scale. As a rule of thumb, binary responses, though popular among defense testers, often require more than 10 times as many test events as objective measures for the same degree of statistical precision (Bisgaard and Fuller 1995; Montgomery 2009). Count or categorical responses fall somewhere between these extremes in terms of accuracy and reliability. Subjective judgments are an important part of both developmental and operational testing but should not be given undue influence on test decisions. The lack of information presented by binary responses has long been recognized in the statistical community (Bisgaard and Fuller 1995; Cook 1996, 1998). If possible, experimenters should try to develop alternative quality measures, including subjective grading if necessary. Experience has shown that many responses originally thought of as binary or categorical may be converted into continuous variables after thoughtful consideration of the physics behind the response.

Factors Considered and Chosen

With the responses documented for the test objectives, the team turns to considering the factors and conditions, both controlled and uncontrolled, that can affect the responses. In this step, the team brainstorms as many factors as possible and then considers how to treat them as part of the experiment. Considerable focus is important since undocumented factors with real effects on the responses can misguide an otherwise well-designed test program. Leaving a factor out of the factor list at this stage is equivalent to assuming that it has no effect on the response.

Brainstorming factors, typically using a cause-and-effect diagram, usually takes the longest clock time of all steps in this phase and should not be rushed by the team. In fact, it is good practice to revisit the list of factors after additional research into historical test findings. Once the cause-and-effect diagram has been

TABLE 1 **Table of Response Variables**

N	Variable	Units	Range	Priority	Accuracy	Data element	Source
1	Acquire time	Seconds	0–10	M	±1	Start, event time	Instrumentation
2	Acquire range	Meters	10–30 K	H	±50	Event time, position	Instrumentation
3	Track time	Seconds	10–15	M	±1	Start, event time	Instrumentation

TABLE 2 Color-coded Example Measures

	Objective (fact-based measures)	Subjective (opinion-based measures)
Quantitative (numeric)	Range to target (nm)	Likert rating scale choice (−2 to +2)
	Tracking errors (m)	Agree to disagree
	Target location error (m)	Better to worse than standard
	Weapon miss distance (m)	Cooper-Harper scale (1–10)
Qualitative (descriptive)	Correctly identify target (binary)	Open-end prose describing event
	False alarms per period (count)	Mission success–failure judgment (binary)
	Words correctly heard (count)	Opinion of fitness for use

populated, the results can be recorded in a summary table of potential factors and levels. Table 3 shows an example for an airborne searcher problem.

The "units" column contains information about the unit of measurement for the factor. Lengths in feet or miles or temperatures in degrees Celsius are continuous variables and variables containing scenarios or labels are "categorical." For the "range" column, one should record the physically operating range of the factor; the design range chosen for the experiment will be recorded in the last column. The column labeled "change" refers to how readily the factor level can be changed in test execution. Some control factors, termed *hard-to-change* (HTC) factors, cannot be easily or practically changed as often as a completely randomized scheme requires. So for efficiency's sake, HTC factors are changed less often, affecting the benefits of randomization. Though HTC factors are fairly common in practice, the team should understand that complete randomization is preferred and the design with HTC factors alters the nature of the method into what is referred to as a *split-plot* design and analysis, with attendant challenges that should not be lightly undertaken (more in Phase III). The "experimental control" column marks how each factor will be treated in the test program: varied as control factors (C); held constant (H); and nuisance (N). Some nuisance variables may be measured and recorded and used in the analysis phase as covariates. When in doubt, conscious effort should be made to include factors as experimental.

After a first pass through this table, the team should make every effort to discover and record the physical variable(s) underlying any variable labeled "categorical" in the units column. This practice is, again, not suggested lightly. Not only does this process of physical understanding increase insights into the process under test, but real-valued, physically based variables offer a much richer and more statistically powerful set of design and analysis options than those of categorical variables. With sufficient expertise, thought, and effort, many variables formerly labeled categorical can be relabeled in different units, at least approximating real-valued continuous or broad-ranging discrete factors.

Checklists are displayed at the end each section (Figure 5) to provide both the guidelines (first-order bullets) and the material solution evidences (second-order bullets). A collapsed first-order checklist could be used for planning, whereas the expanded version, including second-order, is intended for use by both the test team for internal assessment and for outside entities for evaluation of a test plan.

TABLE 3 Airborne Searcher Problem Factors/Levels

Number	Variable	Units	Range	Change:ETC, HTC	Exp. Control: C, H, N	Design range
1	Sensor	Categ.	Old–new	ETC	C	Old–new
2	Searcher Alt	Kft	0–45	HTC	C	15–30
3	Searcher Vel	kts	300–540	ETC	H	480
4	Tgt Loc Error	m^2	0–10,000	ETC	C	1,000–5,000
5	Tgt Backgnd	Ratio	0–100	ETC	H	25
6	Tgt Speed	m/s	0–30	ETC	C	0–20
7	Tgt Length	m	4–20	ETC	C	4–20
8	Visibility	m	100–10,000	HTC	N	Covariate

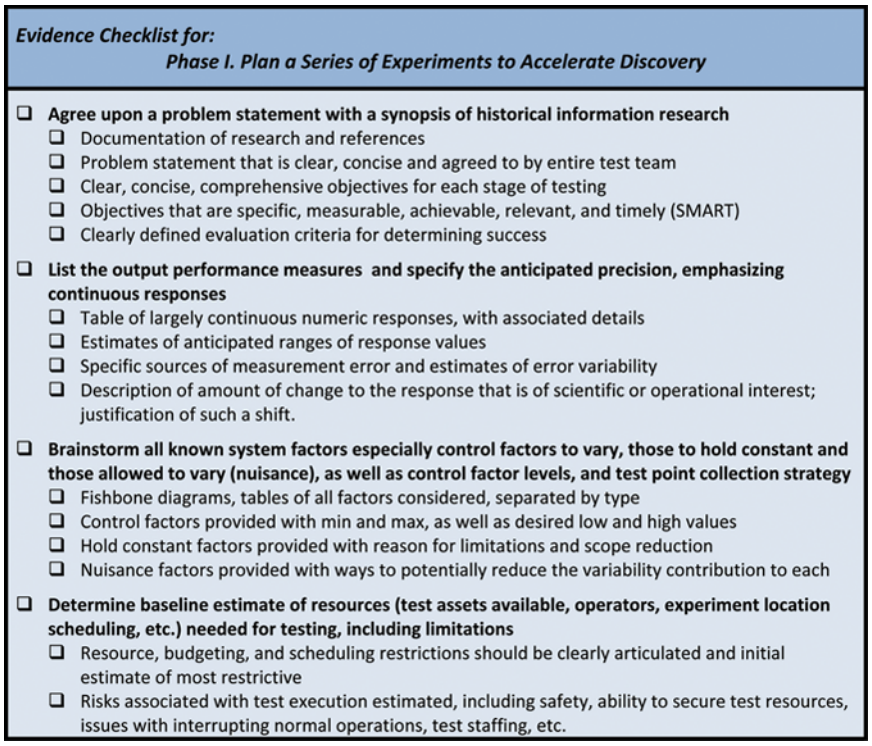

FIGURE 5 Phase I—Planning evidence checklist. (Color figure available online.)

Phase II. Design with Power and Confidence to Span the Factor Space

System planning is understandably only Phase I of test science, but that process requires complete team attention and commitment so that it lays the foundation for success in the three subsequent phases. The second phase, design, involves the task of devising the test point strategy to determine how many test points, what factor combinations, and in which experimental groupings. It cannot be overstated that the plan phase, combined with test science expertise, is the primary driver of success in the design phase. The work of the design phase must be done using sound statistical practices and knowledge of classical experimental design and involves comparing alternative testing schemes. Here, the success indicators are more abundant and reflect not only how well the test is designed but how well it has been planned. The abundant fruits of exemplary planning are on display in the evidence of this phase.

DOE IIa. Span the Factor Space

Determining the test events that will effectively cover the entire region of the factor space is critical to system understanding. For example, in investigating multipurpose military systems, a large factor space is the norm. That is, the system should work across a number of ship, vehicle, or aircraft types; in diverse environmental and weather conditions; against many types of enemy systems and deceptions; and from a variety of geometric engagement conditions. In practice using experimental design, 10 or more factors are commonly encountered (Johnson et al. 2012; Simpson and Wisnowski 2001). Clearly, testing *all* combinations of even eight variables is often prohibitive. The problem, nonetheless, remains: a high-dimension factor space that must be fully explored, but testing all possible combinations may not be feasible. Fortunately, alternative experimental designs for large factor spaces are available. The experimental design guidelines and evidence described here address evaluating the breadth of coverage of the design space and the selection of test points, including the number of factors, the levels chosen for each factor, and the experimental design strategy to purposefully spread test points across the factor space.

Brainstorming Potential Designs

Once the team prioritizes the list of factors using the best available knowledge, it is time to address test points. The experimentalist considers alternate

design strategies to span the desired space in a sequential campaign of test matrices. Many experimental design classes can be considered, suitable for a wide range of objectives and circumstances, including comparative two-level designs, general factorial designs, fractional-factorial designs, response surface designs for nonlinear processes, several variants of algorithmic single-criterion optimal designs, and extended capability classes such as split-plot, robust, space-filling, and mixture designs. Often there are several good choices that can be made for a project emphasizing different strategies. In the context of the design phase discussion, the focus will be on efficiently and effectively spanning the factor space, but the test matrices must also be designed for practical execution (considered in Phase III), sequential discovery, and, ultimately, insightful analysis and reporting (covered in Phase IV).

Experimental design is undertaken to uncover the true input–output relationship. Empirical statistical modeling using least squares estimation of some low-order polynomial of the inputs is routine. Consider two design strategies, both involving two input factors (scalable to higher dimension without loss of generality). One strategy (Figure 6a) involves a two-level full-factorial. Clearly, replication is required to estimate noise, but these four unique design points are capable of a linear plus interaction response surface, illustrated by the adjacent response surface. This response surface just graphically displays an underlying estimating equation, capable of predicting responses for any input conditions within the factor ranges tested. The second design (Figure 6b) is a

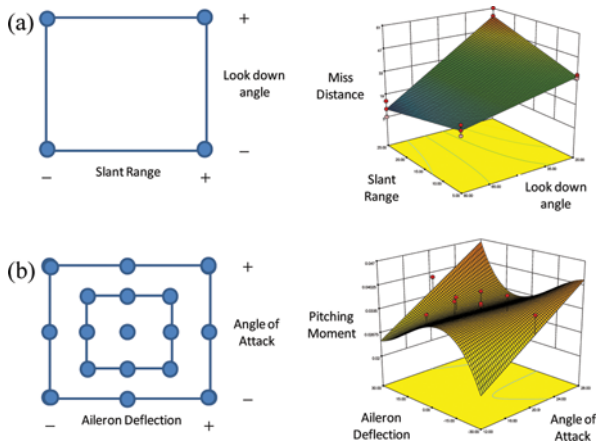

FIGURE 6 **Connection between experimental design selected and model complexity capability. (Color figure available online.)**

nested face-centered design (Landman et al. 2007), which calls for increased coverage of the test space, enabling a model of at least cubic order. The data collected in this example resulted in the nonlinear surface shown. An efficient approach for fitting this surface would usually entail several test–analyze iterations, perhaps starting with the two-level design.

The decision for determining test points in matrices depends on the assumed polynomial order of the statistical model to be fit. The design and analysis phases are intimately bound together such that the analysis and reporting capability is directly related to the location (factor settings) of test points chosen and executed. The desired polynomial order for the general model can be determined using historical test findings or can be reasoned by expert (engineer, scientist, or operator) knowledge of the underlying physics. Regardless of the assumed general model, it is always prudent to add (lack of fit) test points to ascertain whether a higher order model is warranted. If so, an additional set of points should be added to the original to support the higher order model. This sequential strategy avoids waste by leveraging new knowledge to choose the points needed to adequately model the response surface.

Classical Experimental Design

Among the choices for populating the factor space with test points are classical experimental design strategies including the general factorial design (all factor combinations), the two-level full- and fractional-factorial designs, and response surface designs for quadratic models (Box and Draper 2007; Myers et al. 2009). These choices have been used for over 50 years with astounding success and are commended for their ability to adequately and efficiently model, sequentially build via test–analyze–test, predict new test events, and provide robustness when problems arise in test execution. Figure 7 shows representative designs in three factors, displaying example locations of the points in the factor space. The red interior points are center point runs often replicated to estimate noise and test for possible curvature.

Optimal Designs

Optimal designs can take on many forms (e.g., *A*-, *D*-, *G*-, or *I*-optimality) and are constructed via computer algorithms intended to optimize some scalar

representation of the model matrix (X) in quadratic form ($X'X$). A- and D-optimality are designed to minimize statistics associated with regression coefficient variances, whereas G- and I-optimal designs address prediction variance properties. Although the computer-generated optimal designs can be excellent for their intended criterion, they do not directly address many of the other design properties of interest (e.g., replication, number of factor levels, robustness to a misspecified general model, aliasing, etc.). However, optimal designs certainly have a role in situations when no classical design well suits the needs of the problem. In situations involving constraints on the input space, nonstandard polynomial models, unusual sample size requirements, mixed-level fractional factorials, or small run design augmentation schemes, optimal designs offer a flexible alternative (Goos and Jones 2012; Johnson et al. 2011).

Figure 8 provides a checklist to assess success in test design planning after completion of the process planning. The emphasis in this phase is on sound and efficient experimental designs that can consider all relevant influences, uncover factor effects truly influencing performance, and estimate system noise.

DOE IIb. Design Controls Risk of Wrong Conclusions

The practical limitations on time and resources together with the presence of system noise affects outcomes and the ability to effectively measure factor influence, presenting us with inescapable risks of making wrong inferences based on the events observed in testing. When shifts in performance are observed via data analysis, one must decide whether the shift is potentially due to a casual effect based on purposed changes in factor(s) or whether the shift instead is merely a product of the underlying noise present.

Type I and II Errors

Designed experiments are unique in statistically based studies in that it is possible to quantify and manage the probabilities of incorrect conclusions associated with hypotheses to test factor significance. Convention states that the probability of wrongly declaring a response shift due to factor changes when the factor is not influential is the α, or Type I, error. Conversely, the probability of failing to detect a shift due to a factor when it truly exists is the β, or Type II, error. Making correct decisions in the face of noise is a hallmark of a well-designed experiment: the key is setting a low α risk and then providing adequate test resources to achieve high statistical *power* $(1-\beta)$ (Lenth 2001).

Complementary hypotheses are established to quantify α and β risks. As an example, suppose a factor of interest in infrared air-to-air missile (AIM-9X) performance is an operating environment concern, background clutter (Figure 9). Suppose that clutter can be controlled in a high-fidelity missile fly-out simulation and one of the goals of the test is to determine whether clutter matters. A default hypothesis is set (clutter does not matter), together with an alternative (clutter matters). So α is the probability of declaring that clutter matters when it does not—it would occur if clutter truly does not matter but a relatively large change in performance (say miss distance) is observed when clutter is purposefully changed. β is determined by the probability, represented by an area associated with the alternate hypothesis world (clutter matters) of observing a small change in miss distance when the true average change is actually relatively large.

Relationship of α, β, σ, δ, and N

In order to compute β, the test team must agree upon a minimally acceptable difference in

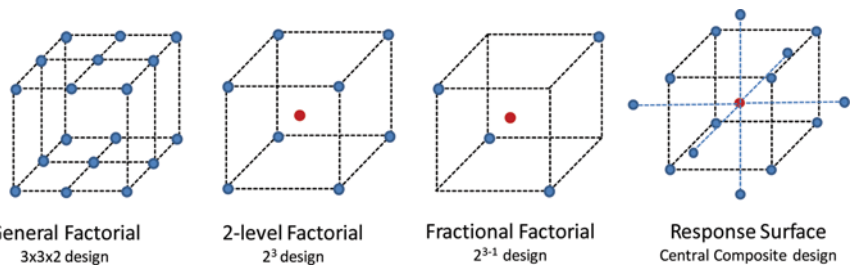

FIGURE 7 Classical experimental designs. (Color figure available online.)

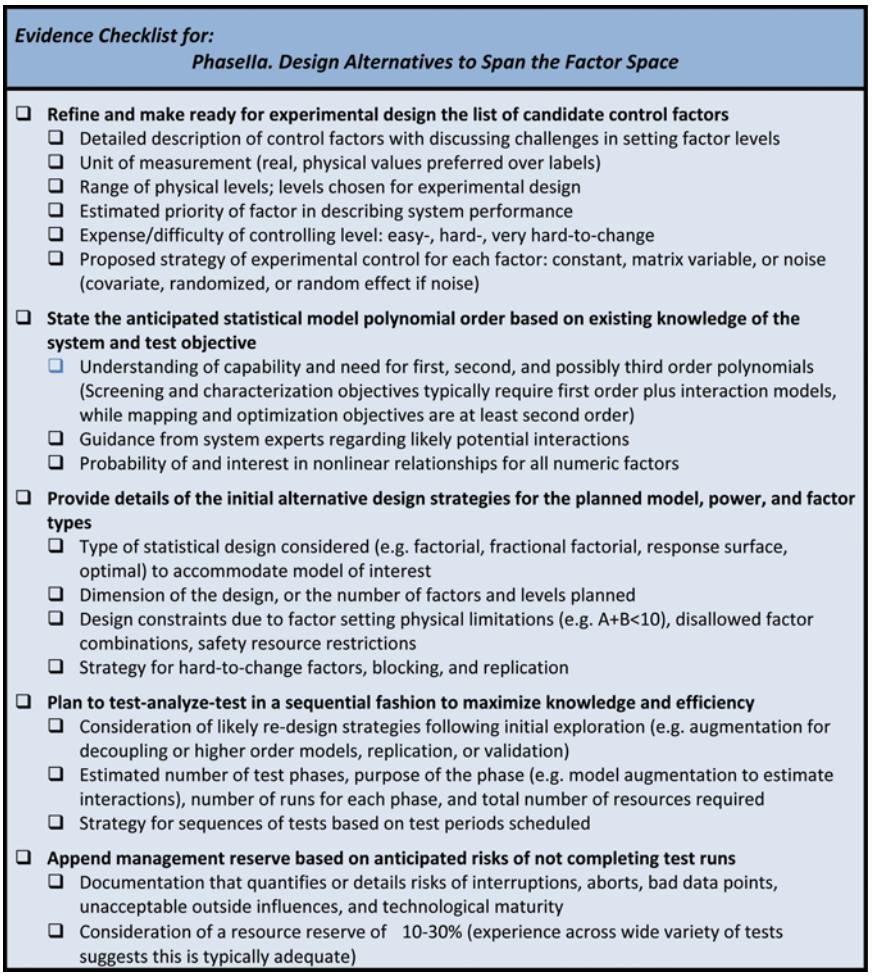

FIGURE 8 Phase IIa—Design evidence checklist. (Color figure available online.)

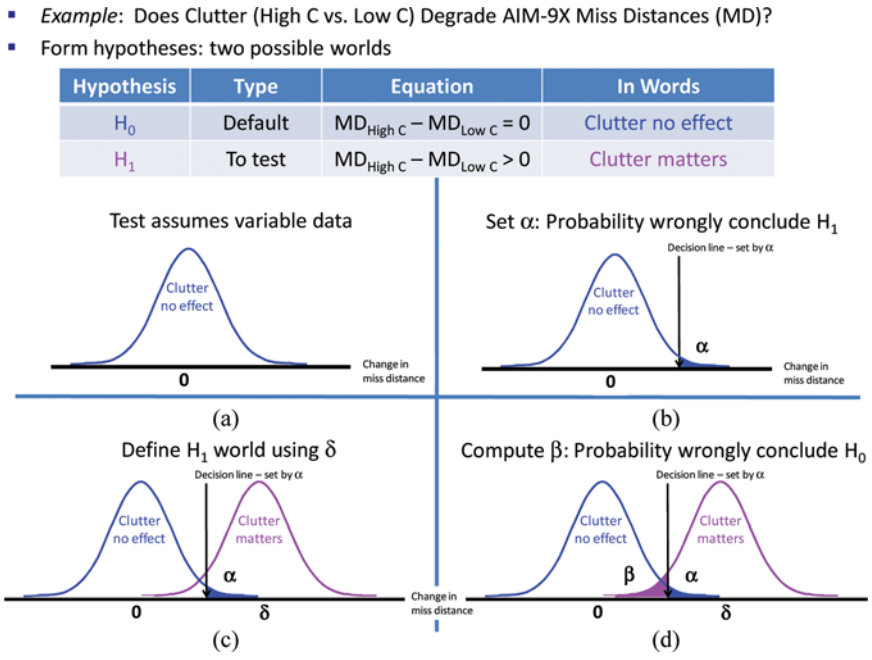

FIGURE 9 Hypothesis testing and statistical errors, α and β. (Color figure available online.)

performance to be detected (δ). In Figure 9, suppose that the test team determined that it was important to see a change in miss distance of at least 5 feet; β is the area under the alternative world (clutter matters, on average by $\delta = 5$), associated with average small changes in miss distance. The α and β risks will always be nonzero and are influenced not only by δ but by a number of other parameters.

Certainly the most important other parameter influencing test risk associated with incorrect conclusions is the total number of runs (N) contemplated for a design. As a general rule, increasing well-placed trials leads to higher power. More trials can result in simultaneous reductions of α and β. Clearly δ, the change in performance that matters, also affects risk. In that situation, α is usually set so power can be calculated for increasing values of δ (Lenth 2001). The relationship between power and sample size is one of marginal decreasing returns. Power grows drastically initially, but as the number of trials continues to increase, power improvement slows. Assuming a stable system under test and little chance of missing data, trials to obtain power values above 95% are usually not necessary.

Power Analysis

Determining the risk-controlled test size per objective is often referred to as a *power analysis*. *P*ower analysis involves mainly information gathering, a series of decisions, and an iterative evaluation of alternative experimental designs, primarily involving test size and power, until a final test matrix is developed (Lenth 2001). Power analysis requires input from all four phases of the science of test: the number of factors from planning; estimates of σ, α, and δ from design; restrictions on randomization from execution; and the complexity of the assumed model from analysis. Power analysis begins in process decomposition with identifying the factors of interest, the levels of each factor, and the anticipated order of the model. Then system noise level must be estimated, the α risk should be set, and the sensitivity to changes in the response magnitude (δ) should be decided. Multiple sequential experimental designs can then be developed and assessed for statistical power. The final experimental design evaluation will assess the evidence of a successful design, including power. A summary description of the parameters involved in a power analysis is provided in Table 4.

The checklist shown in Figure 10 provides the checklist for developing the final, multicriteria test design along with the evidence table to use in assessing the quality and level of effort of this phase of the process.

Phase III. Execute Sequentially with Randomization and Blocking to Control Uncertainty

An often neglected aspect of planning for designed experiments is managing the fashion in which the design will be executed. Execution order and what happens between tests can greatly influence the purity of the data, the level of noise contributing to experimental error, and the average time per test event. Other concerns associated with execution include replicating identical factor conditions, controlling measurement error, and ensuring independence of data collected sequentially. Important execution principles including randomization, blocking, covariates, split-plot experiments, replication, and sequential experimentation are explained in this section.

Randomization

Background variability causes the responses to have different outcomes for identical input settings. Some known sources of variability are controlled via planning by choosing to set them held constant for the entire experiment, such as system operator. Typically unbeknownst to the experimenter, though, are nuisance or lurking variables changing while the test is being conducted. These influences can include lot-to-lot variations in raw material and experimental units, changes in the design or operation of the system, learning, warm-up, wear-out, fatigue, boredom, changes in environmental conditions, etc. Within a homogeneous unit of experimental trials, the solution to background changes is *randomization*. That is, the order of the trials is randomized to ensure that factor setting changes do not line up with, or become influenced by, the background changes (Fisher 1935).

Blocking

Another important principle of experimentation from Fisher (1935) is the concept of *blocking*, also

TABLE 4 Power Analysis Parameters Description

Parameter	Description	How obtained	Relevance in planning
***k*: factors**	Number of factors in the experiment	Determined in planning	Key finding from planning
df_{error}: model error	Amount of data reserved for estimating system noise	Desired model order (e.g., linear, quadratic)	Estimate of complexity of input–output relation
α: alpha	Probability of declaring factor matters when it doesnot	Set by test team	Fix and leave alone
δ: delta	Size of response change expert wants to detect	Experts and management determine	Some ability to vary
σ: sigma	System noise—run-to-run variability or repeatability	Historical data; pilot tests; expert judgment	System driven but can be reduced by scripting
$1-\beta$: power	Probability of declaring a factor matters when it does	Lower bound set by test team	Primary goal is to set *N* to achieve high power
***N*: test size**	Number of samples or runs	Based on other parameters	Direct, should modify to satisfy power

known as local control of error. Sometimes the nuisance sources of variability are not only known but controllable for the purposes of testing. A standard approach to execution is to break the total test design into sets of trials that can be accommodated by the experimental setup: number of tests permission or field exercise, number of formulations that batches of raw materials can accommodate, the number of troop evolutions that can be run during an exercise shift, etc. It has been noted that factors serving as blocks are not usually factors of interest but blocks also restrict complete randomization of the design, a characteristic shared by HTC factors of interest to be introduced momentarily.

Covariates

More complex patterns of experimental execution can be implemented as well. Nuisance factors thought to influence the response that are uncontrollable but can be measured are treated as *covariates*.

Evidence Checklist for:
PhaseIIb. Deciding on a Design Strategy to Control the Risk of Wrong Conclusions

- ❑ **Report statistical power and type I error probability of incorrectly declaring a factor significant for proposed model effects and key performance measures (e.g. distance, accuracy, task time)**
  - ❑ Type I error level (α) is appropriate for risk mitigation and justified for given testing
  - ❑ Values for delta (δ) clearly derived and justified from expert and decision maker input
  - ❑ Estimates of sigma (σ) provided from relevant historical data, pilot test, or expert judgment
  - ❑ Power values reported by factor type if mixed-level design
- ❑ **Report metrics per statistical design, weight metrics, and report final design scores**
  - ❑ Design approaches specified and justified
  - ❑ Power (designs for screening) or prediction variance estimation (response surface designs)
  - ❑ Alignment between model order and design capability (i.e. adequately estimate terms plus additional points to guard against model misspecification or lack of fit)
  - ❑ Sufficient replication to estimate pure error
  - ❑ Validation strategy
  - ❑ Flexibility in testing sequentially
  - ❑ Ability to predict well new observations
- ❑ **Decide on final set of designs to accomplish all test objectives**
  - ❑ Confirmation of limitations and restrictions, including hard-to-change factors
  - ❑ Comparison of measurement metrics across designs, with metric weighting if appropriate
  - ❑ Designs graded the highest from multiple criteria compared to alternative designs chosen
- ❑ **Decide final sequential strategy for experimentation all design phases and test schedule**
  - ❑ Estimates of test entries and time per event
  - ❑ Priorities set to obtain essential test points (e.g. complete fractional factorial design)
  - ❑ Replicates as blocks, or blocking effects separate from model effects
  - ❑ Addition of test events only as necessary to build on existing knowledge
  - ❑ Strategy for validation points and model checking

FIGURE 10 Phase IIb—Power analysis and design strategy evidence checklist. (Color figure available online.)

Examples include environmental factors (wind, temperature, humidity, sea state), system operator characteristics (training, experience, skills), or experimental unit condition (age, wear). Statistical modeling involves the analysis of covariance that essentially extracts the variation due to the covariates prior to modeling the relationship between the factors of interest and the response. Like blocking, analysis of covariance correctly deals with known sources of nuisance variability separating them from the model and from experimental error.

Split Plot Experiments

Hard-to-change factors that would otherwise argue against complete randomization can be changed according to an experimental design called a split-plot design (see Jones and Nachtsheim 2009; Kowalski et al. 2007; Simpson et al. 2004). Split-plot designs are not without drawbacks because the lack of complete randomization weakens the cause-and-effect relationship between the HTC factor and the response, but such designs can maintain a well-defined relationship between the easy-to-change (ETC) factors as well as the HTC factor by ETC factor interactions. Examples of HTC factors are hardware configurations, software loads, product design settings, altitude, and oven temperature. A key distinction between blocked experiments and split-plot experiments is that, though both restrict the randomization of runs, blocks are nuisance sources of variation, whereas HTC factors are factors whose effects are of interest.

Replication

As stressed in the design phase discussion, some *replication* during the experiment is essential to estimate pure error, the better form of experimental error. Replicating design points requires that the factors be reinitiated or reset and typically replicates are separated substantially in execution order under a complete randomization scheme. By contrast, *repetitions* consisting of collecting multiple data observations *without resetting* factor levels are not normally recommended unless measurement error can be mitigated by their use (e.g., surveys). If repetitions are used incorrectly as replicates, the observations are often correlated in time and experimental error estimation from repetitions is biased downward.

Plan Using Sequential Experimentation

A series of tests conducted according to well-understood principles is the best way to limit risk, manage chaos, and maximize the likelihood of correct conclusions. The factor space for many systems undergoing testing and evaluation is vast—dozens of possible variables resulting in many thousands or millions of possible unique test conditions. Testers seldom know which of the variables will matter most in driving system performance, though they may have suspicions. Military testing shares this environment with many other domains, including product design, basic research, and manufacturing or processing. Consequently, an experimental best practice is to structure the overall test program in such a way as to test in stages with appropriate objectives and experimental designs for each stage, thereby providing periods for analysis, understanding, and redesign. It is seldom wise to devote more than 25% of total test resources to one experiment (Box et al. 2005; Montgomery 2012). The information gained at each stage of experimentation is invaluable in considering how to continue the investigation. We can use the output of early stages to accelerate our learning about the process, validate (or improve) our various simulations, and refine the active factor space for later stages of testing. At the outset of the test, there is limited knowledge of which factors are important, the appropriate factor level ranges, the degree of repeatability or noise in the process, and many other facets. Sequential experimentation helps build that knowledge in stages so that the experimentation is increasingly beneficial and in the end much more effective than one-stage test.

Within each experimental environment, it is appropriate early on to outline a sequence of experiments. Initially, the experiment objective is often to *screen* many factors and identify the few that drive the process outcome. Typically, just a few of many factors considered to affect performance actually do so. The testers may then wish to reduce the size of the factor space explored in subsequent tests. Just as important, we may discover unexpected features of system performance such as nonlinear behavior, unanticipated noise levels, isolated unusual runs that do not conform to similar conditions, or aspects of the factor space that are not well represented in this early stage of experimentation (target backgrounds,

natural environments, human reactions, etc.). An example of sequential testing is given in Figure 11. In the case of each of these test outcomes, we can learn valuable lessons about process performance that should be used to redesign and execute the next stage of testing.

Some principles of sequential assembly include the following:

- Screen with many factors, fewer levels, and simple models—leverage efficiency
- Re-assess or improve test execution procedures to better manage noise
- Revise factors (drop due to negligible effect or add due to new area of interest), levels, or range based on newfound knowledge
- Pause to assess noise, right-size testing to estimate uncertainty, and adjust for identified lurking variables
- Place next points where needed to better model factors to responses or capture nonlinear performance
- Validate or confirm predicted or unusual performance with one or more confirmation runs
- Move in factor space to improve performance
- Allow for system repair and retest
- Experiment in natural groupings of test points and employ blocking
- Make engineering or logic changes to simulations because predicted performance was not validated in live testing

A typical sequence of testing that should be outlined at each stage may include screening, confirmation of previous stage results, investigation, exploration, and, lastly, prediction and confirmation. Each stage should be appropriately budgeted and scheduled in the test strategy up front. The checklist shown in Figure 12 provides an assessment tool for the general considerations of test execution.

Phase IV. Analyze Statistically to Model Performance

Analysis of test data represents the fourth and final step of the iterative cycle associated with design of experiments. Most, if not all, of the analysis discussion is relevant to the process decomposition and planning stages of testing. The steps for analyzing test data are provided here so that it is clear not only what should be emphasized post-data collection during empirical modeling but also which seeds to plant in the minds of the test team as the test is planned and scoped. Analysis intent must be integrated early in planninng or we suffer the consequences of useless data and limited findings.

Justification for Statistical Modeling

A unique feature of tests constructed with statistically designed experiments methods is the ability of the tester to link suspected causes to the observed effects. That is, the test matrix is designed in such a way as to effectively link changes in system

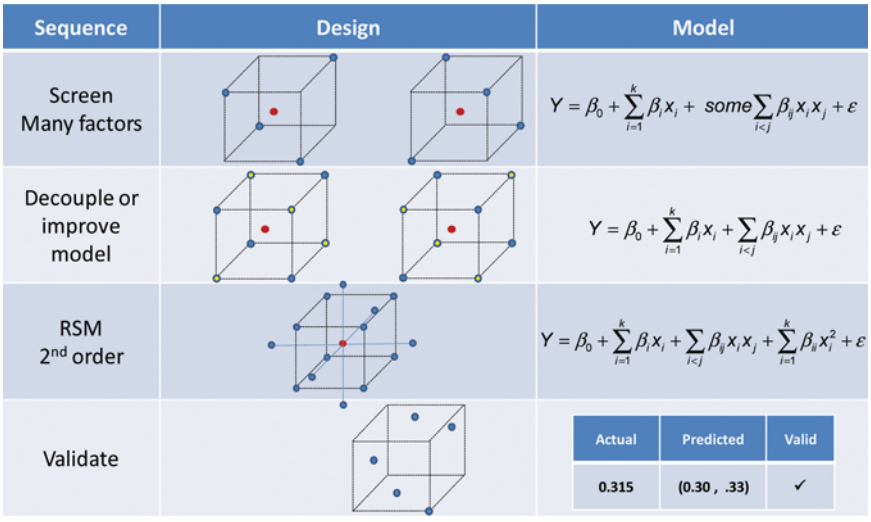

FIGURE 11 Sequential experimentation from screening, to decoupling, to second order and validation. (Color figure available online.)

performance to the purposed changes in the input factors. It is too strong to claim "cause-and-effect" relationships because statistical sampling precludes such a statement, but verifiable association or correlation can certainly be detected (Kutner et al. 2004). Nonetheless, being able to link changes in performance to factor effects is a powerful feature of designed experiments and should be exercised routinely to maximize knowledge gained from testing. The mechanism for characterizing and assessing this relationship is the empirical statistical model. It turns out that desirable properties of a statistical model such as low uncertainty, the ability to correctly identify a factor as significant, and the ability to predict performance within the factor space are heavily dependent on the test points prescribed. So it is essential to understand that the quality of the experimental design and the strength of the analyses are directly connected. As planning commences for a test strategy, the design and model are properly presented in tandem to weigh strengths and weakness of alternative designs and models (Figure 13).

The analysis process involves making sense of data typically collected on a complex system in a foggy environment. A fundamental challenge is to ascertain the contributors to either dispersion or average of a response in the presence of random noise. Analytical tools provide a mechanism for determining likely influential inputs, as determined by statistical significance. System experts are relied upon to then decide whether statistically significant factors have a correspondingly sufficient practical significance. The steps outlined below describe the analysis procedure for assessing performance in the presence of purposed or observed changes in inputs. Each step with the exception of diagnostic verification (outside the scope of this article) will be briefly discussed as they relate to evidence of a well-designed test.

Analysis Steps

1. Exploratory data analysis
2. Iterative empirical modeling
3. Model diagnostic verification
4. Model prediction
5. Model validation

Exploratory Data Analysis.

Prior to formally modeling the potential input–output relationship from a statistically designed

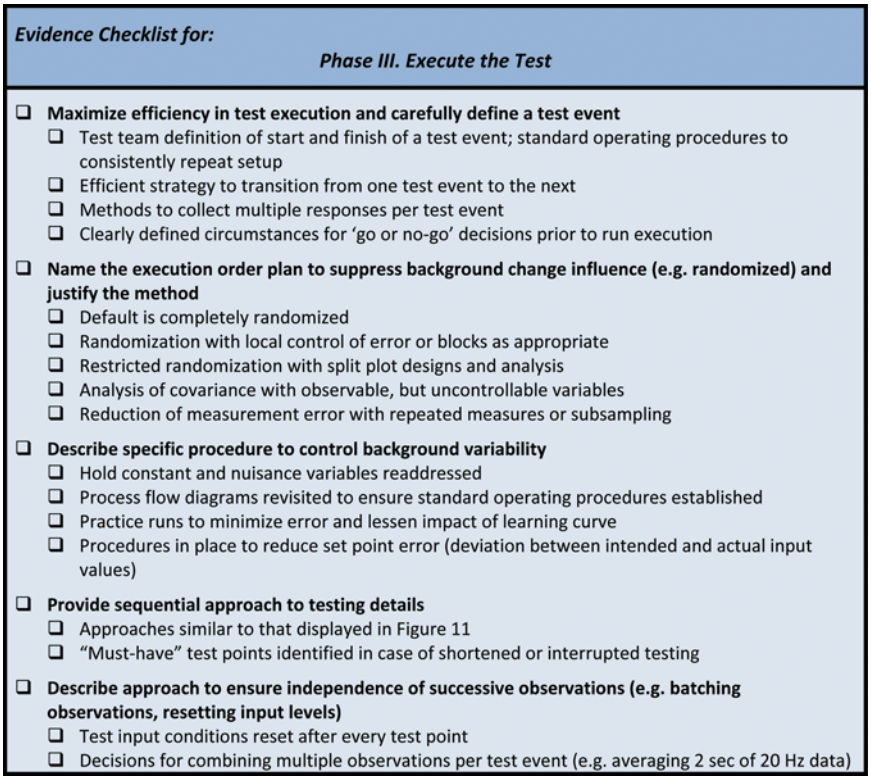

FIGURE 12 **Phase III—Execution evidence checklist. (Color figure available online.)**

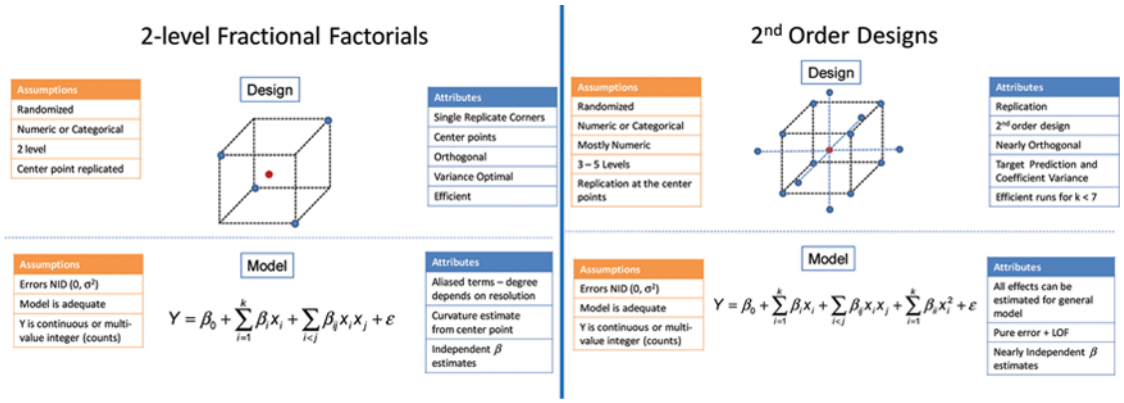

FIGURE 13 **Relationship between example experimental designs and supported models. (Color figure available online.)**

experiment, the data should be assessed using graphics and summary statistics, a method called *exploratory data analysis* (Tukey 1977). Unusual observations are typically present and often difficult to identify without careful examination of the data. Helpful graphs include scatter plots in small multiples of inputs versus outputs, histograms, and box-whisker plots. Summary statistics including the mean, standard deviation, percentiles, as well as the median and median absolute deviation are useful. An important first step in data analysis, exploratory data analysis outcomes include suspected influential inputs or even interaction effects, outlier identification, measures of response central tendency, as well as a sense of the distributional characteristics of the parent population.

Empirical Modeling.

The analysis of data from a designed experiment is relatively straightforward compared to analysis of a retrospective or even an observational study. Nevertheless, reality tends to invade and alter sound test designs via nuisance lurking variation, unaccounted for factors, outliers, missing observations, and measurement error. Fortunately, the design strategies together with the resulting analytical tools are quite robust to these routine outside influences. The standard approach to analysis for a designed experiment is to build an empirical statistical model. Least squares regression is the default modeling method due to its ability to effectively capture the factor effects and interactions while remaining insensitive to modest violations of model error assumptions (normality, independence, constant variance). Alternative modeling techniques (discussed later) can be employed as needed.

To illustrate the value of statistical modeling, consider the DOE conducted to characterize the terrain following/terrain avoidance capability of a tilt-rotor CV-22 aircraft (Figure 14).

Among the factors are ride mode, airspeed, and turn rate on the response altitude deviations, where the deviations are an absolute percentage, as too high or too low are equally detrimental. Based on execution of a two-level, replicated factorial design, an empirical regression model is built consisting of the statistically significant model terms, where x_i represent the coded unit ($-1 =$ low and $+1 =$ high) factor settings. The regression model can be depicted directly or graphically as a response surface over the range of input settings. Figure 15 shows the influence of airspeed and turn rate on altitude deviation.

Model Prediction.

The factor main effect and interaction plots can provide a simple means for communicating the essence of the relation between inputs and responses.

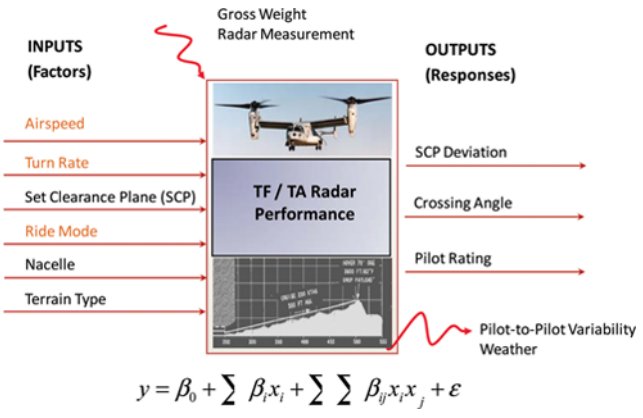

FIGURE 14 **CV-22 flight test input–process–output diagram. (Color figure available online.)**

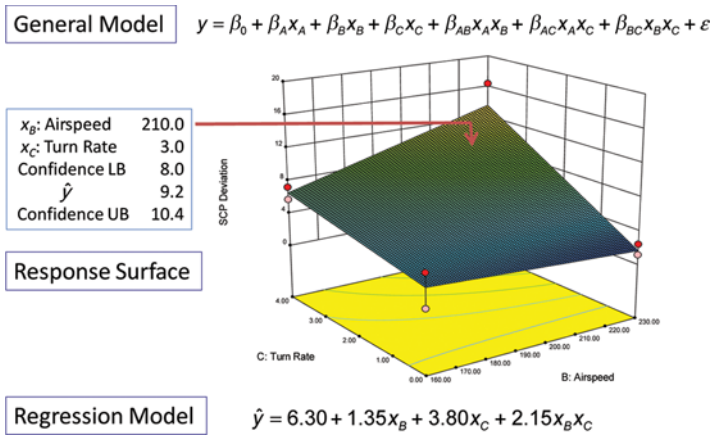

FIGURE 15 **CV-22 example general model, response surface, regression model, and prediction. (Color figure available online.)**

The model is also useful to estimate or predict outcomes for test point conditions anywhere across the response surface, enabling the user to determine input regions of superior and inferior performance relative to an expectation or specification. More important than point predictions are uncertainty intervals, either confidence intervals for the average response or prediction intervals for a future observation. Tolerance intervals can also be constructed if the goal is to capture some high percentage (e.g., 90%) of future observations with some high confidence level. For the CV-22 example, suppose that an aircrew has planned a mission involving an airspeed of 210 knots in airplane mode, using hard ride mode, and a turn rate of 3.0 degrees/second. The predicted deviation from set clearance plane is 9.2% with a confidence interval lower bound of 8% and upper bound of 10.4% (Figure 15).

From the CV-22 example, it is evident that the process of design and analysis of experiments can satisfy several fundamental objectives in testing. Ultimately the test provides valuable information for postprocessing. It is helpful to know during planning the general purpose or objective for the analysis. Obviously the analysis objective comes directly from the test objectives laid out in Phase I. For example, early in testing the CV-22 terrain following/terrain avoidance system, the goal might be to determine among the many possible factors, which ones are largely influential—a test objective called *screening*. The plan, design, execute, analyze cycle can also have as its primary analysis objective to *characterize* the nature of the relationship between the significant factors and performance. Here the intent is typically to understand the magnitude and direction (or sign) of the relationship, whether interactions play a role, and whether the input–output function is possibly nonlinear. Once nonlinear response surfaces are developed, the objectives often lead to a desire to *map* or *optimize* the system. The guidelines used and evidence to assess the quality of a test plan are different depending on the analysis objective.

Model Validation.

Models constructed from even a well-designed test are of no use if they fail to capture the system studied, so a critical analysis step is to validate the model to ensure that it reflects true system performance. In all situations, a small number (three to eight) of validation runs at factor conditions not previously tested is highly recommended. The validation runs are compared to predictions from the statistical model to assess prediction performance. The executed points should fall within the limits of the prediction intervals or there is reason to believe that there is more to learn. Validation runs should be explicitly budgeted and scheduled and can be used to improve model fit (Box et al. 2005).

Modeling for Success—Multiple Analysis Techniques

Empirical modeling success is dependent on every phase of the designed experiment process. All of the relevant factors and appropriate levels must be identified in the planning phase. For the design phase, the statistical model capability depends largely on the test point allocation and adequate coverage of the factor space via the experimental design. In the third phase, execution, steps taken to mitigate noise and control

lurking variable contamination via randomization allow the statistical model to clearly reveal potential causal factors. Lastly, in the analysis phase, a host of analytical tools for modeling are at the ready to apply and then compare relative to one another to determine the best model to validate and report in fully capturing system behavior.

In many tests, standard least squares regression modeling is the better choice. Usually least squares works when the test was well planned, a subset of the factors and their interactions are statistically significant, and the regression model assumptions are not grossly violated. Sometimes, though, it is not clear which modeling technique is best, so comparing different model forms can be valuable. The analyst has modeling alternatives (Table 5) to handle almost any peculiarity of a test.

Other Considerations: Modeling Diversity Due to Design, Factor, and Data Type Variations

Although the modeling details are beyond the scope of this article, it is important to plan for the proper error structure of the model. Consider the factor types themselves. Often fixed effects (e.g., material type, chemical ingredient, hardware variant) are accompanied by random effects (e.g., operator, hardware serial number), plus covariates, or variables that can be observed but not controlled (e.g., temperature, pressure, humidity). Designs with fixed effects, random effects, and covariates require a model form that allows for the proper analysis (mixed models). In addition, factor levels may be difficult to change (e.g., chamber temperature), so restricting the randomization in execution is required. The proper design is called a split plot and its analysis requires estimating two error terms. Factors can also be nested within each other. For example, airspeed is nested within the nacelle setting for the CV-22 (low airspeed in helicopter mode is not the same as low airspeed in airplane mode). Finally, constraints on the input space can occur because some factor combinations are not feasible. If the restrictions are prevalent, significant correlation among factors can exist, requiring careful analysis.

Responses can also take on various alternative forms. Although we desire all responses to be continuous variables that vary over a wide range and can be collected with perfect accuracy, this is not always the case. Sometimes a single response is actually a sequence of measurements over time or space (e.g., radar cross section measurement of a target rotated in aspect 360 degrees), so the performance is now a function or profile (Chicken et al. 2009). Other times, surveys such as aircrew feedback on a helmet display are the primary response instrument. Consider responses known to behave such that extreme values occur up to 20% of the time, resulting in valid outliers. In this case, perhaps an exponential generalized linear model is a better modeling construct. A final class of testing involves software-intensive systems such as

TABLE 5 Alternative Statistical Modeling Techniques for Unusual Factor or Response Conditions

Factor and response conditions	Method alternative	Strengths
Fixed and random factors	General linear model	For random effects and split plots
Outliers, heavy-tail responses	Generalized linear models, robust regression	Alternatives to normal errors and can fit multiple outliers
Correlated factors	Ridge or partial least squares regression	Helps identify the correct model factor influences
Noisy responses, many categorical factors	Classification/regression trees (CART), bagging	Provides insight when analysis of variance fails, groups multiple levels
Highly nonlinear response surfaces, many factors	Multivariate adaptive regression splines (MARS)	Complex model without over fit so often predicts well
Deterministic responses	Kriging	For computer experiments
Functional (profile) responses	Wavelets	Excellent fit, few parameters, variable selection
Very few runs with informative historical data	Bayesian methods	Leverages prior knowledge to make inference from sparse data

communication or automated mission planning systems. If the purpose is to assess software functionality, the response is deterministic and has no noise component; therefore, running replications adds no value to the test. Alternative designs such as space filling and factor covering arrays are suggested.

For complex nonlinear factor–response relationships, a host of statistical learning methods are available (Hastie et al. 2001), both parametric and nonparametric. As the number of influential factors grows ($k > 4$), and assuming that the experimental design is capable of estimating second- or higher order polynomials, consider using multivariate adaptive regression splines. For deterministic response computer simulations, a common modeling strategy called *kriging* offers the flexibility to fit closely to the data. In each of the above examples, though, some alternative modeling techniques are worth considering. As the control over the points in the test space is diminished and the study becomes more observational in nature, consider additional modeling techniques including robust or ridge regression (Montgomery et al. 2012; Ryan 1997), and generalized linear models (Myers et al. 2012).

Description of Advanced Model-Based Measurements

The purpose of this phase of designing a test is to not only try multiple modeling strategies on the available data but, more important, to plan the test to maximize the chances of success in the test data analysis. It is worthwhile to understand that the capability to properly analyze data from a design is best accomplished during planning. Ultimately the user benefits the most by a design strategy that controls risk and supports the ability to build a viable and capable model of performance throughout the region of interest. There are several model-based options to be considered in this analysis phase of the experimental design.

A relatively new development in DOE is the *fraction of the design space* (FDS) plot (Zahran et al. 2003). Such a plot shows the scaled variance of the predicted response value for a given model type and noise level as a function of the volume of the design space. An ideal FDS plot is low and flat across the design volume. The FDS plot shows how a proposed design performs against this standard and is particularly useful to compare competing designs for the same factor space.

Of interest regardless of the analysis objective category is the ability to uniquely estimate the regression coefficients, enabled by orthogonal experimental designs. As the factor effects, represented by the columns of the X matrix, become more dependent or correlated, the regression coefficient estimates become more unstable, increasing their variance. The correlation of the factor effects is easily measured using the variance inflation factor. A variance

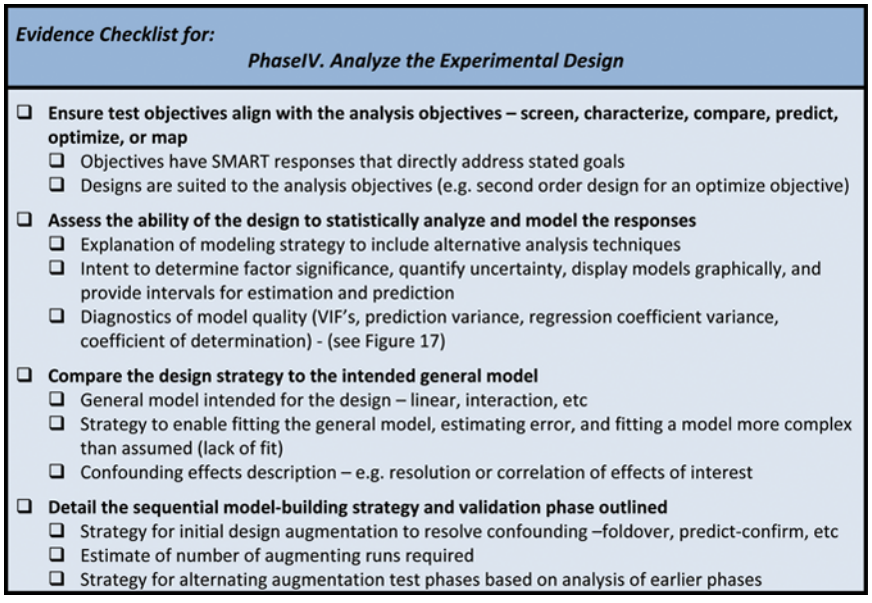

FIGURE 16 Phase IV—Analysis evidence checklist. (Color figure available online.)

inflation factor of 1 is ideal, reflecting orthogonal columns (Montgomery et al. 2012).

Benefits of Modeling

The analysis phase is often the least time consuming and often easiest of the four test phases. Worry-free analyses reflect the culmination of careful, deliberate, and diligent efforts that transpired in planning, design, and execution. Experienced analysts focus on a myriad of indicators that ultimately ensure success in testing, success measured by the amount of knowledge gained about the system under test. The statistical model is extremely powerful, so what are the benefits? The terms in the model best characterize system performance while performance is robust or insensitive to terms or factors not in the model. The model provides information about conditions when the system works well and when it may not work as well. Figure 16 provides a checklist of the primary analysis guidelines that should be addressed to ensure effective statistical analysis of experimentally designed tests.

CONCLUSIONS

The focus of this article was stochastic outcome testing planned using statistically based DOE, but the vast majority of the indicators can be easily adapted for deterministic systems such as software functionality test. A well-designed and appropriately analyzed test or experiment has desirable characteristics that distinguish it from most any other test strategy. Planning the test effort with a multidisciplinary team of experts cannot be overemphasized. In total, the recommended designed experiment should be crafted by seasoned analysts in tandem with experts in operations and science, conserve resources, accurately measure all of the relevant performance quantities, and culminate in an empirical statistical model that effectively resolves the existing relationships between the purposefully changed factors and the measures of performance. Data control should be rigorous, databases should be retrieved securely and rapidly, and data reduction should limit transcription error rates. Associated criteria presented in the evidence checklists for each of the planning stages should be established, understood, and practiced for organizations to master testing.

Once analysis, the final phase of planning, is complete, a thorough evaluation of design alternatives is recommended. Assuming that a design strategy has been developed that carefully considers the coverage of the factor space, execution order details, the statistical model form together with the analysis method(s), the alternatives should all be viable. Worry-free analyses reflect the culmination of careful, deliberate, and diligent efforts that transpired in planning, design, and execution. Therefore, the team can develop a comparative table to evaluate the design alternatives. Figure 17 provides design- and analysis-related criteria for consideration to compare strategies with

Characterize

Objectives	Criteria
Characterize	Sample Size
Compare	Power
Screen	ME
	2FI
Assume	Replicates - Pure Error
Same factors	Orthognonality
General Model = ME + 2FI	Terms aliased
Type I Error = 0.05	Word length count
	VIF
Response type: Numeric	Categoric balance
	Interaction balance (GBM)
	Partial aliasing
	Model misspecification (lack of fit)
	3FI
	Curvature
	Quadratic
	Range of Inputs
	Robustness to outliers/missing data
	Points total for rep/LOF
	Functionality
	Levels per factor - intended
	Levels per factor - design
	Ease of augmentation
	foldover strategy
	additional levels ease

Optimize

Objectives	Criteria
Estimate	Sample Size
Predict	Prediction Variance
Optimize	50% FDS
Map	90% FDS
Assume	95% FDS
Same factors	G-eff (min max prediction)
General Model = ME + 2FI+PQ	I-eff (avg pred variance)
Type I Error = 0.05	Replicates - Pure Error
	Orthognonality
Response type: Numeric	Condition number
	VIF
	Prediction Uniformity
	Rotatability
	Uniform precision
	Model misspecification (lack of fit)
	3FI
	Pure Cubic
	Range of Inputs
	Sensitivity to outliers/missing data
	Influence / Leverage
	Points total for rep/LOF
	Space Fill Properties
	entropy
	minimize Euclidean distance among pts
	Design Functionality
	Levels per factor
	Number of evenly spaced levels

Abbreviations
ME: Main effect
2FI: 2-factor interaction
3FI: 3-factor interaction
LOF: lack of fit
GBM: general balance metric

FIGURE 17 **Expanded list of design and model-based measurements. (Color figure available online.)**

similar objectives. Aspects of design, execution, and analysis are addressed, including design optimality measures, factor coverage, precise noise estimation, and practical execution issues.

The evidence checklists form a foundation for careful, consistent planning deliberation, leveraging the precepts of the scientific method for the successful application of design of experiments. One way the checklists could be employed is to use the primary bolded bullets for planning and the sub bullets for assessing. As such, evidence provided for each planning phase can serve not only test teams up front as a roadmap or procedure but as indicators to assess plans developed for science of test excellence. Due to the scope and limitations of this article, only the essentials are addressed that should apply to the vast majority of testing situations. Recognizing that all experiments are different in the challenges they present, adaptations, exceptions, and emphases are expected. Future papers are encouraged to provide more specifics or address specialty areas of experiment planning such as split plots, robust design, mixtures, or deterministic (e.g., software) systems.

ABOUT THE AUTHORS

James R. Simpson is Chief Operations Analyst for the Air Force's 53rd Test Management Group at Eglin Air Force Base, Florida. The wing analysts are responsible for the implementation, training, and mentoring of test planning and analysis methods and activities for Air Combat Command Operational Test. He is also an adjunct professor of industrial and systems engineering at the University of Florida and was formerly full-time faculty at Florida State University and the United States Air Force Academy. His research has focused on applied statistical methods, statistical process control, RSM, and experimental design. He is a past editor of *Quality Engineering* and currently chairs the ASQ Publication Management Board.

Dr. Charles Listak is the Director of Test Training for the Air Force's 53rd Test Management Group. He oversees the curriculum development and instruction of all test training programs for Air Combat Command's largest test and evaluation organization. He is a former B-1B and T-38A instructor pilot with 4,000 flying hours with extensive experience in bomber and air-to-ground operational testing. He is also an adjunct professor of instructional systems design and educational research at the University of Phoenix. Dr. Listak earned a Bachelor's in Mechanical Engineering from the University of South Carolina, an MBA from the University of South Dakota, and a Doctorate in Curriculum and Instruction from the University of West Florida. His research interests include human performance technology, educational program evaluation, and experimental design.

Mr. Gregory Hutto is Wing Operations Analyst for the 96th Test Wing, is responsible for embedding designed experiments as the principal test method for several hundred tests each year. He teaches an extensive series of short courses in test methods to all testers in the Wing from the Wing Commander to our 520 scientists and engineers. As a LtCol in the USAF Reserves, he served as senior military advisor to AF Operational Test & Evaluation Center Test Support Director and as special advisor for test design to the AF Flight Test Center commander at Edwards AFB, California. Mr. Hutto is a distinguished graduate of the US Naval Academy in Operations Research and holds a Master's in the same field from Stanford University.

REFERENCES

Atkinson, C., Cox, D. R. (1974). Planning experiments for discriminating between models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(3):321–348.

Barton, R. R. (1997). Pre-experiment planning for designed experiments: Graphical methods. *Journal of Quality Technology*, 29(3):307–316.

Bisgaard, S., Fuller, H. T. (1995). Sample size estimates for 2*k*-*p* designs with binary responses. *Journal of Quality Technology*, 27(4):344–354.

Box, G. E. P., Draper, N. R. (2007). *Empirical Model-Building and Response Surfaces*, 2nd ed. New York: Wiley & Sons.

Box, G. E. P., Hunter, J. S., Hunter, W. G. (2005). *Statistics for Experimenters: Design, Innovation and Discovery*, 2nd ed. New York: Wiley.

Chicken, E., Pignatiello, J. J., Simpson, J. R. (2009). Statistical process monitoring of nonlinear profiles using wavelets. *Journal of Quality Technology*, 41(2):198–212.

Cook, D. R. (1996). Graphics with regressions with a binary response. *Journal of the American Statistical Association*, 91(435):983–992.

Cook, D. R. (1998). *Regression Graphics: Ideas for Studying Regression through Graphics*. New York: Wiley.

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

Cox, D. R. (1958). *Planning of Experiments*. Oxford, England: Wiley.

Fisher, R. A. (1935). *The Design of Experiments*. Edinburgh, Scotland: Oliver and Boyd.

Goh, T. N. (2001). A pragmatic approach to experimental design in industry. *Journal of Applied Statistics*, 28(4):391–398.

Goos, P., Jones, B. A. (2012). *Optimal Design of Experiments: A Case Study Approach*. New York: Wiley.
Hahn, G. J. (1984). Experimental design in a complex world. *Technometrics*, 26(1):19–31.
Hastie, T., Tibshirani, R., Friedman, J. (2001). *Elements of Statistical Learning, Data Mining, Inference and Prediction*. New York: Springer.
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.
Johnson, R. T., Montgomery, D. C., Jones, B. A. (2011). An expository paper on optimal design. *Quality Engineering*, 23(3):287–301.
Jones, B., Nachtsheim, C. J. (2009). Split-plot designs: What, why and how. *Journal of Quality Technology*, 41(4):340–361.
Kass, R. A. (2006). *The Logic of Warfighting Experiments*. Washington, DC: CCRP Publications.
Kowalski, S. M., Parker, P. A., Vining, G. G. (2007). Tutorial on split-plot experiments. *Quality Engineering*, 19(1):1–15.
Kutner, M., Nachtsheim, C., Neter, J., Li, W. (2004). *Applied Linear Statistical Models*, 5th ed. New York: McGraw-Hill.
Landman, D., Simpson, J. R., Mariani, R., Ortiz, F., Britcher, C. (2007). Hybrid design for aircraft wind-tunnel testing using response surface methodologies. *Journal of Aircraft*, 44(4):1214–1221.
Lenth, R. V. (2001). Some practical guidelines for effective sample size determination. *The American Statistician*, 55(3):187–193.
Mason, R. L., Gunst, R. F., Hess, J. L. (2003). *Statistical Design and Analysis of Experiments: With Applications to Engineering and Science,* 2nd ed. Hoboken, NJ: Wiley-Interscience.
Montgomery, D. C. (2009). *Introduction to Statistical Quality Control*, 6th ed. New York: Wiley.
Montgomery, D. C. (2012). *Design and Analysis of Experiments*, 8th ed. New York: Wiley.
Montgomery, D. C., Peck, E. A., Vining, G. G. (2012). *Introduction to Linear Regression Analysis*, 5th ed. New York: Wiley.
Myers, R. H., Montgomery, D. C., Anderson-Cook, C. M. (2009). *Response Surface Methodology*, 3rd ed. New York: Wiley.
Myers, R. H., Montgomery, D. C., Vining, G. G., Robinson, T. J. (2012). *Generalized Linear Models: With Applications in Engineering and the Sciences*, 3rd ed. New York: Wiley.
Ryan, T. P. (1997). *Modern Regression Methods*. 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(2):95–103.
Simpson, J. R., Kowalski, S. M., Landman, D. (2004). Experimentation with randomization restrictions: Targeting practical implementation. *Quality and Reliability Engineering International*, 20(5):481–495.
Simpson, J. R., Wisnowski, J. W. (2001). Streamlining flight test using design and analysis of experiments. *Journal of Aircraft*, 38(6): 1110–1116.
Tukey, J. W. (1977). *Exploratory Data Analysis*. Boston: Addison-Wesley.
Vanhatalo, E., Bergquist, B. (2007). Special considerations when planning experiments in a continuous process. *Quality Engineering*, 19(3):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(1):44–51.
Wu, C. F. J., Hamada, M. (2000). *Experiments: Planning, Analysis, and Parameter Design Optimization*. New York: Wiley.
Yeh, I. (2004). Characterization of nanotube buckypaper manufacturing process. M.S. thesis, Tallahassee: Florida State University.
Zahran, A., Anderson-Cook, C. M., Myers, R. H. (2003). Fraction of design space to assess prediction capability of response surface designs. *Journal of Quality Technology*, 35(4):377–386.