Testing via Sequential Experiments

Best Practice and Tutorial

Authored by: James Simpson, PhD, Consultant to the STAT COE

The goal of the STAT T&E COE is to assist in developing rigorous, defensible test strategies to more effectively quantify and characterize system performance and provide information that reduces risk. This and other COE products are available at www.AFIT.edu/STAT.

STAT T&E Center of Excellence
2950 Hobson Way – Wright-Patterson AFB, OH 45433

Table of Contents

Executive Summary

A series of tests, referred to as sequential experimentation, is a recommended practice that can be best planned, at least in terms of a general strategy, prior to test. The information gained at each stage of experimentation is invaluable in considering how to continue the investigation. We can use the output of earlier 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. This paper is structured to provide compelling reasons to sequentially test. The discussion addresses the need for test-analyze-test, the use of blocking, along with the why's and how's of multi-stage testing from an experimental design perspective. Because the series of tests after the initial design are fairly well known from the outset, both in terms of objective and test strategy, it is not difficult to estimate the total number of tests, number of test stages, and even the types of design points (actual test conditions) prior to test in test planning. A tutorial of sequential testing is provided as an appendix to demonstrate some likely reasons for augmentation, show the benefits of leveraging current system knowledge to make best use of subsequent test resources, and in the end, showcase an efficient and effective way to maximize system understanding.

Introduction

Testers are routinely afforded more than a single test session to conduct their full complement of tests for a given project. Data acquisition and reduction systems are becoming more automated and sophisticated, enabling near real-time feedback with results capturing system performance. The primary purpose of testing is to gain knowledge regarding the system under test in order to make informed decisions, typically in terms of the next phase of the acquisition process. The best way to gain knowledge rapidly is through a series of hypotheses to be evaluated by scientific inquiries followed by data analysis, leading to new hypotheses, and so on, until all the important and relevant questions have been sufficiently answered. This process of induction – deduction – induction is often referred to as test-analyze-test, or sequential experimentation.

Fundamental to the practice of the design of experiments, is the best practice of planning for an initial set of tests that comprise perhaps only 25% of the resources available (Montgomery, 2009), followed by another set of tests, which under favorable circumstances leverages the results of the analysis (the knowledge gained) from the first set of tests. Often, yet another series is conducted after the second, usually with specific intent, and again until enough understanding is gained to make the decision to end test. Typically the last set of tests involves some test events useful for system model validation. The successive series of tests often require fewer runs in the latter stages than in the first run. Each stage of testing (including the first) should have specific experimental objectives for properly and effectively statistically modeling the factor-response relationships. Rationale and methods for satisfying these

objectives will be the primary thrust of the following discussions. There are many sound reasons for undertaking a sequential test strategy, and several will be discussed in this paper.

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. Typically, just a few of many factors positively affecting 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). 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.

A Need for Test-Analyze-Test

Box (1992) noted that Fisher once quipped that, "the best time to design an experiment is after it has already been conducted." Obviously, his reference to hindsight as illuminating is understandable, but this simple reminder also conveniently promotes sequential experimentation. Why not take advantage of what we learn early on, to shape our plans not for the next project, but for *this* test project? Sequential testing also answers the critic that says, "if we learn in the first few tests" (red light caution here on drawing any conclusions with insufficient information), "whether a factor matters, why do we need to complete the full factorial matrix?" A staged or phased process of testing allows for pauses, ideally with time for analysis to make any needed corrections in secondary phases. If learning takes place during test with data analysis confirmation, there may be cause to alter the plan.

Probably the strongest warning against the one-stage test then analyze strategy is that all their evidence gathering risk resides in one place. Consider a likely consequence of inadequate test planning, i.e. that a test is designed to answer the wrong question. What becomes of that test? Statisticians have run across this problem enough to assign it the error of the 3rd kind or Type III error – an elegant solution to the wrong problem (Kimball, 1957). Sequential testing by itself won't zero out Type III error rates, but opportunities are in place to recover when headed in the wrong direction.

In contrast to a single set of tests conducted in a single setting, a series of experiments or tests offers distinct advantages that permit the important process of discovery to learn from and modify successive test plans so that the latter tests are designed to answer refined hypotheses. The key contribution of phased testing toward system knowledge is that inferences proposed based on findings from the initial stage of testing can themselves be tested and even refined in subsequent test stages. Box, Hunter and Hunter (2005) discuss this in the first chapter of their experimental design text, and relate the importance of gaining knowledge through a series of reasoned arguments, supported by evidence. Many people distinguish between two basic kinds of arguments: inductive and deductive. Induction is usually described as moving from the specific to the general, while deduction begins with the general and ends with the specific. Arguments based on experience or observation are best expressed

inductively, while arguments based on laws, rules, or other widely accepted principles are best expressed deductively. The process typically starts with a hypothesized model, deduction is made for the scenario at hand, which guides the experiment design, which is executed. Data is then collected and analyzed, which is followed by induction and a possible modification to or confirmation of the model (Figure 1.)

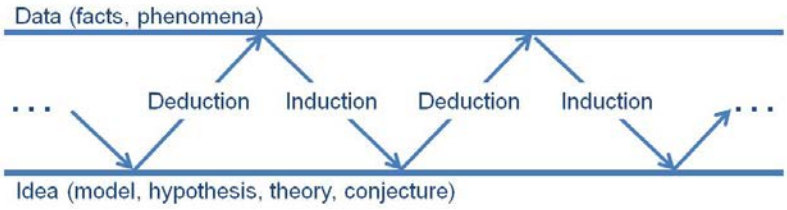


Figure 1. Iterative learning process via experimentation. (Box, Hunter and Hunter, 2005)

As applied to experimental design and analysis, suppose that an initial test planning process suggests the use of an experimental design to potentially capture main effects, 2-factor interaction effects and perhaps even pure quadratic effects. Suppose further that the resulting design developed contains no validation points, no degrees of freedom for lack of fit (LOF), few replicates and partial aliasing of model terms such that the model correlations could potentially inhibit uncovering the true model effects. It would be difficult in a single-shot test to determine the possible effects active in the presence of the experimental error. An iterative approach with augmentation runs would have a far superior chance of success in understanding true system behavior.

A couple of aspects of military testing actually work in our favor if sequential testing is utilized. As it is generally true that the test design is required long before execution is scheduled, it is also true that there is some flexibility to change the actual test point settings for a given test event. The vast majority of executed tests, when compared to the original test plan, bear little resemblance to each other. This reality can be used to argue for proposing a sound general sequential test plan (with all the test points specified), and then making modifications as necessary based on knowledge gained through test execution and analysis of the data. The complete set of tests required using a sequential strategy can be estimated reasonably well by knowing the capability of the initial design, the likelihood that factors will need to be added or modified, the need to estimate well more complex model effects, and the requirement to validate the statistical model with some additional tests. There are often breaks between the execution of test events, allowing for data to be processed and analyzed. Regardless of the access to data, simply the process of setting up, performing pilot or trial runs, making sure all data acquisition systems are operating, discovering new contributors to background variation, and executing initial test events often reveals an abundance of information relevant to and useful for test design. This insight can be successfully leveraged to make proper adjustments to the test design, working in concert with the analyst.

Blocking Invokes Sequential Testing

Often the stages of tests can't be conducted under like or similar conditions, so recognizing that the testing environment is changing during the duration of testing is beneficial. Fisher (1935) proposed the concept of blocking to handle such situations, with the intent of controlling nuisance variability locally, and thereby preventing the error term from incorporating known, yet unavoidable contamination of the response data. The procedure for blocking is to conduct each block separately and sequentially.

Blocking necessities are abundant in military testing; they only have to be recognized and included as a part of the plan. Blocking also represents a form of sequential experimentation. The full complement of test events is partitioned into subgroups or blocks. Often the intention is to simply enable the partitioning of the block variability apart from error or input factor contributions to variability. However, blocking can be used also as a means to logically stage testing, such that more basic models are fit in initial stages, followed by the capability to model more complex models in subsequent stages. Figure 2 shows an example test plan initially in 4 factors and 4 blocks.

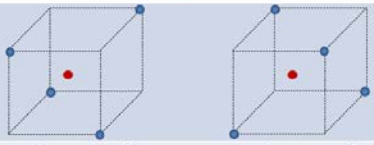
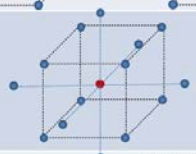
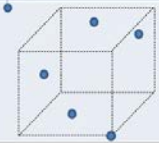
Block	Design	Model
1: Screen – linear model		$Y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \text{some} \sum_{i<j} \beta_{ij} x_i x_j + \varepsilon$
2: Augment – interaction model		$Y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i<j} \beta_{ij} x_i x_j + \varepsilon$
3: RSM – 2nd order model		$Y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i<j} \beta_{ij} x_i x_j + \sum_{i=1}^{k} \beta_{ii} x_i^2 + \varepsilon$
4: Validation runs		Actual: 0.315 / Predicted: (0.30 , 0.33) / Valid: ✓

Figure 2. Possible use of blocking to perform sequential testing to grow knowledge and increase the complexity of the statistical model, only as required.

The first block (or set of tests) is intended to screen the important factors, primarily focusing on main effects and some ability to estimate interactions. In this example, the first block consists of 10-12 runs, depending on the number of replicates used. The second block, in this example, would require about the same number of runs as the first, and by the end of Block 4, it would be anticipated that about 40-45 runs would be executed.

A sequential approach to test has been recommended by the pioneers and chief scholars in the field as a principle of profound importance. Fisher (1952), Box (1992), Montgomery (2009) and Vining (2011) all agree on the importance of leveraging knowledge from earlier tests to inform future test planning.

Why Sequential? An Experimental Design Perspective

By the very nature of pausing after a stage of testing, there are inherent advantages that can be leveraged in subsequent stages. Data is often available soon after test, such that exploratory data analysis and more formal tests of hypotheses and statistical testing can be conducted in a matter of minutes. Regardless of data availability, just observing the conduct of the test often provides sufficient information to make substantial modifications to the next stage of testing.

Design and model link. Aside from the common sense and test risk management reasons for sequentially testing in subgroups, there are compelling reasons to test in stages based on the tenants of design of experiments (DOE). Well recognized in a statistically designed experiment approach is the tremendous value in developing an empirical statistical model that relates change in the response to changes in location in the design space (or changes in factors individually or together). This practice requires an accurate estimate of experimental error (contributions to response variability due to unknown or unaccounted for variables) to determine statistical significance, and a test design strategy that enables the development of that statistical model. The model is then used as a surrogate for the system to explain which factors generate response changes, to convey how factors working together can change the response, to estimate performance for conditions not tested, to predict future performance, and to optimize performance while identifying conditions that lead to that best or worst response value. Each design proposed is capable of building a model of a particular complexity for some percentage of the number of factors tested. Figure 3 illustrates this connection between the design and the statistical model. Here the model is represented graphically by a response surface with domain axes reflecting the factor space, while the vertical dimension captures how the response changes.

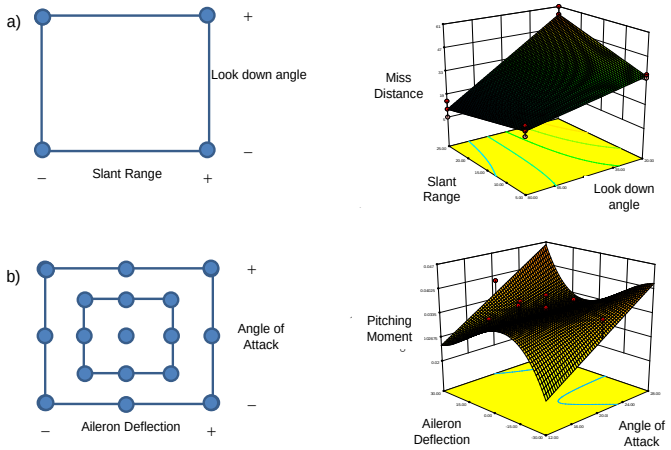


Figure 3. Each design is built with an intended purpose and underlying model capability.

Build the model in stages. As a matter of common practice, we screen the often extensive list of factors which possibly matter, using a fractional factorial design. Resource restrictions are nearly always (low cost computer simulation models being one exception) a reality for program management and the test team, so efficiency is a goal. Because the final model may have linear, interaction and second order effects, several test stages are generally expected to rightly characterize the factor-response

relationship. Starting with a design capable of screening *and* fitting the final model simultaneously is difficult if efficiency is required. Successively growing the model complexity in stages is not only efficient, but it is often the only way to arrive at a proper understanding of the system. Most often the initial design stage involves some screening design followed by additional tests for various purposes. A tutorial, together with some rationale for taking a particular approach, is given in the appendix.

Complex Model. In rare instances, it may be strongly anticipated that the factor-response relationship is complex and best approximated by a high order polynomial or maybe even by a piece-wise regression model due to possible response change discontinuities. In these circumstances, it makes sense to start with a high order design approach and to populate the input space fairly extensively (e.g. 4 or more levels of each factor) with design points. Even with this unusual model starting point, the chances of sufficiently populating the space to model well, while remaining efficient, is nearly unattainable. Some nonlinear model is assumed and additional points are purposely placed to test for lack of fit for higher order effects or to identify regions of the input space that require additional testing to better fit the response. Validation is also important.

Model refinement. Natural reasons from a test science perspective to perform testing sequentially are not only to efficiently grow the model complexity, but to add points set at particular factor levels (or in locations of the factor space) so as to refine the model. Points are then again added to confirm that the model proposed is indeed adequate. Stages of testing are often necessary from a modeling perspective to block out known sources of nuisance variability, also called blocks. Finally, stages of testing enables discovery of the factor space not initially explored, so the region of interest moves to a new location based on the knowledge we gain in earlier testing. This exploration can obviously lead to continued movement in the factor space or simply result in sequential testing to grow, refine and validate the model.

Factor changes. For example, it could be that a variable not previously considered is revealed during test setup, or even during the initial testing, that potentially plays a role in system performance. Factors thought to be systematically varied with ease turn out not to be so easy to change. Set point error can be discovered, making it difficult to set factor levels precisely, or perhaps a learning curve is present. Perhaps some of the known uncontrollable background variability (e.g. cloud cover) takes on certain conditions (no clouds, blue sky only) that would clearly bias conclusions and limit the scope of the possible inferences. Box (1992) provides an excellent depiction of many reasons for running a series of experiments based on an initial design.

Bad or Missing Data. Clearly, test execution is often challenged by tight or insufficient schedules, particularly in allocating time for set up and conducting dry runs. The repercussions of shortening the pre-test timeline are numerous: 1) inadequate assurance that learning curves are minimized, 2) test input conditions wrongly set, 3) data acquisition and response measurement systems malfunctioning or not calibrated, and 4) test execution procedures are not clearly understood nor followed. All these contributors tend to not only inflate system noise or generate run order dependent data, but can cause bad or even missing observations. Unfortunately, bad or missing data are more common than rare. To increase the probability of successful test, either the size of the test planned must be inflated to

compensate or conversely, a sequential strategy can be used. Testing in stages allows for corrections to the execution procedures and can target those test events that require re-test, thereby minimizing the requirement for additional testing.

Moving outside the original test space. The initial settings for factors are determined through careful planning based on the knowledge of the system at that time. First, minimum and maximum values are estimated, followed by the low and high (or the multiple categorical choices) settings for test. Often we discover that more interesting performance is obtained outside the originally planned values. Discovery of these new settings often takes place during test, so having the flexibility to test outside the region of initial experimentation is welcome later in testing. A sequential approach well suits this need and is illustrated in the appendix tutorial.

Leveraging initial findings. In general, we learn more in test setup than in all of planning, and obviously even more in initial testing. Accordingly, by scheduling pauses in test to enable data acquisition and reduction, as well as preliminary analyses, sizeable portions of the subsequent testing can either be saved or modified to make best use of those initial findings. The data analysis using commercial statistical software takes very little time, often less than an hour, and can provide a great deal of insight. For example, some factors previously thought to be inconsequential become major system performance drivers, previous noise estimates may have been overly conservative so the test duration and resources can be significantly scaled back, or test conditions previously thought practical and relevant turn out to be infeasible. All these discoveries are not only beneficial in the end, but are also exceedingly better to know about early on in test in order to remedy with corrective action.

Sequential How – Initial Stages

The fact that many practitioners think of an initial screening experiment as the only design needed is arguably a hurdle to get over when thinking about fundamentally instituting sequential testing as standard practice. There is little defensible rationale for planning only a single small-run screening experiment if the objective is to know more than the difference between just two, hardly differentiable conclusions: which factors might matter and which factors most likely matter. Questions that come to mind are:

- How do they matter, in what manner, and by how much
- What to expect for conditions not tested
- Which areas of the factor space give better or worse performance than requirements
- What conditions give the best or worst performance, or
- What are the performance trade-offs given competing objectives

These thoughts are all perfectly suitable test and evaluation knowledge discovery kinds of inquiries which are worthy of suitable answers. It takes more than a single screening design to provide these answers. Other reasons for subsequent tests are given throughout this paper, but it is important to recognize that a major shortcoming of novice DOE practitioners is planning a simple screening design as the complete test solution.

Regarding the design choice for the initial test stage, the classical two-level fractional factorial (2^{k-p}) with center point replication, if practical, is preferred. Two levels for each factor stresses the importance of initially choosing those levels that would result in most different response outcomes, and would be very efficient in terms of number of runs. This class of designs was fully introduced by Box and Hunter (1961a, 1961b), as a means to efficiently test and discern the underlying factor-response relationships. The papers come with guidance on picking the right initial fraction via resolution, the concept of aliasing, augmentation via fold-over designs, and blocking strategies. Fractional factorial designs were developed with the intent that sequential experimentation is used and the fraction would be the first stage in a series of test stages.

Some of the higher priority reasons for performing tests after an initial 2^{k-p} fractional factorial design are to decouple alias chains of interest and/or to add design points (axial points) to fit a second order model. Regarding the goal of decoupling aliased model effects (could be main effects or two-factor interactions), there are three primary fold-over choices: the full fold-over, the semi-fold and the factorial optimal (a specified number usually less than the semi-fold). For more details, see Daniel 1962, John 1966, Montgomery and Runger 1996, Emanual and Palanisamy 2000, Mee and Peralta 2000, Li and Lin 2003, Misra et al. 2013, and Rios et al. 2011. Each option typically trades off the number of runs required for less capability to break alias chains. A distinct advantage of the test-analyze-test approach is that the analysis of the stage one testing can be used to target specific alias chains to decouple, typically in a very modest (2-6) number of runs. The analysis of the combined data from the initial and augmentation stages will enable modeling all of the effects of interest. Additional stages can be added to increase the precision of estimation in certain test space locations and to collect additional data for validation. Running successive stages in blocks to separate the stage-to-stage differences from error and the factor effects, is also generally recommended.

Clearly realities of test conduct often result in aberrations from the intended design and schedule, which supports even more emphatically the need to test in stages. As such, assessments of earlier stages will help decide which points need to be re-accomplished, whether factor combinations are infeasible or whether factor levels need to be modified. Replication from early stages can be used to adequately estimate error and determine the amount of additional testing necessary for sufficient modeling. Each stage's data and analysis informs the test conditions and test objectives in successive stages, plus ultimately determines when enough testing has been done.

Optimal Designs in Stage One

The previous discussion and associated appendix provide some best practices for initial fractional factorial designs, including mixed-level fractions. Certainly there are times when an optimal design is the right choice as an initial stage design. Example justifications for this decision include designs with many categorical factors, disallowed factor combinations, constraints in the design space, or unusual statistical model term needs. Optimal designs differ from classical fractions in that they are computer generated using a single optimality objective (usually the D- or I-optimality criterion). Benefits of an optimal design are that the design can be built and tailored to a specified number of runs and to a

specific (and perhaps special case) model. Although these designs can be augmented for sequential experimentation, it is often thought that the optimal design is a one-stage test solution.

Setting aside the benefits of test-analyze-test, it is often impossible to know the correct model order before the test. Optimal designs will not add center points or replicates, unless the person interfacing with the software knows better. Validation and lack of fit points are also not factored into these designs. Two suggestions for stage one optimal designs are to make sure to add points for replication and to remember that center points are always informative, assuming the factors are mostly numeric. Another thought when considering optimal designs in a stage one screening mode (where Stage 2 is intended to increase model order fitting capability), is to include lack of fit points in stage one. JMP statistical software (JMP, 2013) has a nice feature which allows the user to specify which terms must be estimated (necessary) and which are desirable (if possible). This flexibility could be used in a screening design context so that design points, for example pure quadratic terms, could be added if possible. Another optimal design two-stage strategy would be to use a second order definitive screening design (Jones and Nachtsheim, 2011) as the initial stage and augment that design to better estimate the 2-factor interactions in Stage 2. Always take care to add replicates and lack of fit points to any optimal design.

Sequential How – Higher Order Polynomial Models

Although sometimes it is known from past experience or first-principles knowledge that a design is needed to fit a higher order model, more often the more complex models are discovered sequentially. Design types that contain test points intended to fit pure quadratic (A^2) effects, nonlinear interactions (AB^2), or even pure cubic (A^3) effects are most advantageous if they can be built up sequentially starting with an initial first-order design. The reason we know to run augmented designs after the first order design is because “lack of fit” points are included in the earlier stages with the purpose of answering the question, “are higher order model points needed in subsequent testing?” Probably the best example of a LOF point is the center point location (in coded units the point at the design origin or centroid of (0, 0, …, 0)). Center points are useful for a number of reasons, but are well recognized to provide evidence to test for curvature. Here, a statistical test is performed to compare the response average from the center points to the average of the corner points, in order to determine whether pure quadratic effects (essentially a higher order model) are needed to rightly fit the factor-response model.

Some higher order model designs are displayed (for 3 factors) in Figure 4. A central composite (either the classical rotatable or the face-centered version) design (CCD or FCD) consists of factorial or fractional factorial corners to estimate the main effects and two-factor interactions, the center point to provide the LOF test and subsequently aid in second order estimation, plus the axial points which are used to estimate the pure quadratics. Typically the factorial portion plus replicated centers comprise the first stage, with axials plus additional replicated centers for the second stage. A third stage could consist of additional points (usually selected by an optimal design algorithm) to estimate any additional effects needed (e.g. pure cubics). A final fourth stage can be used for validation. Other central composite designs that can be built sequentially have been proposed (see Nguyen and Lin, 2011).

Suppose it is suspected that third order or higher polynomials are needed or that the test space region is vast in factor level ranges (e.g. many range bins of interest, from a few hundred feet to say, fifty miles). Highly nonlinear regions, or even discontinuities, are possible. A reasonable strategy would be to use a nested central composite design approach, where one CCD is embedded within another. One possible choice would be to set the low/high factorial values for the outer CCD at -1/+1, while placing the inner CCD at -0.5/+0.5. Several sequential stage alternatives are possible to build the nested CCD. Depending on the locations of the axial points (both inner and outer), this design could have 5-9 levels for each factor, providing the points necessary to build upwards from a quartic model.

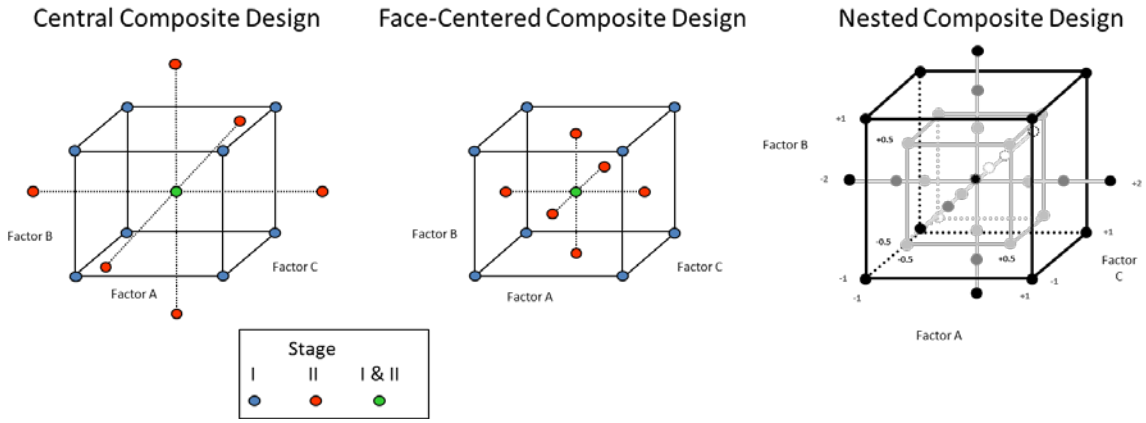


Figure 4. Second and higher-order designs which are typically executed sequentially. The stages for the CCD and FCD are shown by coloring the points.

Sequential How – Validation Stage

Typically the most ignored set of points are those needed to confirm that the proposed understanding of the system is reflective of the true system. This set of points is referred to as validation points, and their label conveys their purpose. Not many points are typically required (4-8 runs), but again it is the lack of their consideration in planning that leaves them off the final test design.

Some meritorious suggestions have been posed for the types of points to consider when deciding on the exact points to run. Possible choices include:

- Points not in the original design
- Points that would aid in further developing the statistical model
- Points which improve some of the design optimality statistics
- Points of particular interest, either known before testing or after initial stages
- Points in the interior of the test space (as most designs place points on the perimeter)
- Points for lack of fit for higher order models
- Points set at particular factor levels

The luxury of so many ways to decide the settings for these points makes it a relatively easy undertaking. In fact, one could subdivide the validation set to accommodate multiple criteria or apply

multiple criteria directly to the complete set. The validation set can also be viewed as the check for whether to proceed with the next stage of testing. Scuillo (2007) considered sequential experimentation stage beyond the second order design (to fit some cubic or higher order polynomial) and proposed a two-stage validation augmentation strategy to first decide a validation set consisting of LOF points and then used the analysis to determine the best set of points to improve the higher order model fit.

Validation points can be used to confirm a proposed statistical model simply by comparing the actual result to the prediction interval (or confidence interval if multiple observations are collected and a mean value is computed). If the actual falls within the prediction, that point is considered valid. If all the points fall within their respective prediction intervals, the model is essentially validated. Degrees of success in validation can be achieved based on the percentage of the points that validate and then by the number of points outside the interval that can be explained or are close to the bounds. Clearly the model summary statistics factor into final model assessment, even if the points are validated. For instance, suppose the response can only take on values over the range of 1-10. If the model from the test analysis had a large noise ratio, it is conceivable that the prediction intervals for each validation point nearly span the entire range of feasible values, such that any actual value would validate. In this case, the validation exercise fails to add much value to the process and the excessive noise present, or weakness of the statistical model, dominates the discussion of test findings.

Validation points are obviously important to validate the proposed model, but shouldn't be ignored after validation is conducted. It is always advisable to then include the validation points in the final model fit, as these points should always benefit the model resulting in better prediction capability in terms of tighter parameter estimate intervals and tighter prediction or confidence intervals. So, another way to look at these points is that they serve important functions. They serve to certify the viability and reliability of a statistical model developed from test data for use in general inference, prediction and optimization. These points also provide improved fit capability and reduced uncertainty bounds as they are incorporated into the model dataset.

Power for Sequential

Right sizing any test program is at or near the top of most everyone's list and if it isn't, it should be. In right sizing a test endeavor, one obviously considers the programmatic side of budget, resources (manpower, test ranges, assets, instrumentation, etc.), but may not readily consider two very vital additional aspects. These two other aspects come directly from planning using a design of experiments' approach: incorrect finding/conclusion risks (probabilities), and failure to cover the intended use space of the system. The planners should address this second aspect adequately, assuming they identify all the major control factors (say *k* of them) to be systematically varied and use a design strategy that sufficiently populates that *k*-dimensional factor space.

The first topic though, that of controlling risks of incorrect conclusions, is becoming more standard and is supported by available statistical software. The two types of risk are: the Type I or α error – a probability of incorrectly stating a factor influences the system performance (measured by the response), and Type II or β error – the probability of incorrectly stating a factor does not affect

performance, when it actually does matter. Statistical power is the complement of the β error (power = $1 - \beta$), which is the probability that the data will support concluding a factor matters given that it really does matter. For a given test design with N runs, an assumed underlying statistical model order, estimated noise standard deviation, a desired magnitude of response change to detect, and a set α error, power can be directly computed. Understand that: 1) not all the factors and or associated effects will have the same power for a given design and 2) power changes for each response, so reporting a table of power values is appropriate. With several stages of testing though, the question becomes, "Which part of the design is used to calculate statistical power?" The quick answer is simple, "As much of the known design as possible, because adding tests nearly always increases power." What about the desire to not fully commit to the later stage test points ahead of time? That's not a problem, because a notional set of points can be used that will give very close approximations of power.

The recommended strategy for performing a power analysis on a sequential test plan is to put together the most likely set of points to be executed across the multiple test stages. This set of points could also be used as the notional test plan, assuming all goes according to plan and no major departure from the plan is required based on findings during earlier stage execution and analyses. The complete set of points could also be collated into a single design matrix, probably safely partitioned by blocks, assuming that test conditions will noticeably change from stage to stage. Power can then be computed using this combined design and model degrees of freedom somewhere between *k* and 2*k*. The model degrees of freedom attempt to estimate the total number of model terms that will be significant. For screening designs, *k* is typically used. As such, a conservative estimate of 2*k* for a higher order model is reasonable.

One-stage Testing

Suppose there is no alternative but to run all the tests in a single test session[1]. There are still some principles of sequential assembly that apply to this situation. In fact, some of these suggestions, such as prioritizing the arrangement of test subgroups within the full set of test, are even more important to consider here. The basic reality is that only one shot is available to execute all your tests. It makes sense to consider the likelihood that certain tests, types of tests, or a certain percentage of tests may not be executed. As such, having a test design strategy that is robust to some missing data is beneficial. Some important sequential test practices to employ in one-stage testing are to subdivide the complete test set into groups that each form a collective whole from a modeling perspective, but also build upon each other in model complexity. One-stage tests can fall prey to time and resource constraints during

[1] Box, Hunter and Hunter (2005, pp 251-2) remind us that "the one-shot philosophy of experimentation described in much statistical teaching and many textbooks would be appropriate for situations where irrevocable decisions must be made based on data from an individual experiment that cannot be augmented…. It is the goal of this book [and this paper] to emphasize the great value of experimental design as a catalyst to the *sequential* process of scientific learning. It must be remembered that the framework for an experimental design, fractional or not, and indeed for any investigation is a complex of informed *guesses* that profoundly influence its course. The need for these guesses has nothing to do with the use of statistical experimental design. They must be made whatever the experimental method. These *guesses* include initially what factors to include, what responses to measure, where to locate the experimental region, by how much to vary the factors, and once the data are available, *how to proceed*. All these guesses are treated as "givens" in a one-shot philosophy… An opportunity for second guessing provides the best chance of understanding what is going on. The subject matter specialist can build on this understanding.

execution, such that later tests fail to be completed. One example of a multi-group arrangement for one-stage tests is to have 3 groups to: 1) screen and model first order plus interaction 2) fit higher order polynomials, and 3) demonstrate certain capabilities and validate. Arranging the subgroups is important, and if there is flexibility consider any necessary adjustments based on time constraints. Consider the above 3 groups as those planned. If it looks like not all of the tests will be conducted, and that maybe first order plus interaction modeling is sufficient, perhaps Group 2 could be bypassed to ensure validation and certain input combinations of special interest are executed.

The above grouping example includes a subgroup for validation, which of course is important in any testing. One-stage testing is no exception. In fact there might be circumstances in one-stage testing when the validation runs should be the demonstration tests (to save resources), and included in the randomization scheme for all the tests. Randomizing the validation runs within the larger tests aids in averaging out the lurking variability for that set of runs and helps ensure validation points are collected should testing be cut short.

When considering a safe approach for a one-stage design, there are advantages to using a classical design with points for replication and validation. With proper expertise and sufficient additional points of the right type, an optimal design could also be considered, especially if factor constraints exist or multiple factor levels are required. The better choice for the initial portion of one-stage designs are classical screening or second order designs because they tend to be robust to missing data, and lend themselves nicely to a blocked strategy within that one stage. If an optimal design is used, ensure that a conservative model is selected (higher order than anticipated), together with replicates, points to improve parameter estimate precision, and lastly validation points. Regardless of the choice for the initial design portion, care must be taken to ensure that all the right types of points are included.

One final suggestion for those planning one-stage tests is to take some extra time in planning to potentially identify the larger contributors to noise variability. These sources of noise (environmental, human, test conditions, set up, test articles, etc.) will not only tend to drive the ability to ascertain statistical significance, but can shape the types of inferences that can be made. For example, suppose for testing the effectiveness of a new round of 20mm ammunition it is decided to launch a single sortie with an inexperienced (at least not recently with guns) pilot. If pilot experience is not a factor of interest, it will be a part of noise. However, for this test, we only have results from an inexperienced operator, so the findings would really only be valid for an inexperienced crew. In certain testing, this outcome may be valid because the test objectives may state that the new ammunition round should provide as good or better performance for inexperienced aircrew. The bottom line though is to ensure the contributors to noise variability are known to the best extent possible and acceptable for the objectives of the test.

Sequentially throughout Acquisition – Passing from DT to DT/OT to OT

Clearly, in military testing a new weapon system undergoes dozens of test periods over the lifetime of its research, development and acquisition. It should then follow that the defense contractor, program office and test organizations would naturally collaborate in passing test findings and knowledge gained

from one phase to the next. Unfortunately, even in the fairly close-knit community of test and evaluation, those charged with developmental test (DT) don't always have a similar view on the objectives of test as the operational test (OT) folks. Those given the decree to do integrated test often find it hard to meld DT with OT.

Fortunately, for designed experiments and sequential testing, the maturity of the system under test has no bearing on the applicability or efficacy of the methods employed. To be clear, probably the single largest advantage to incorporating design of experiments across all aspects of military T&E is that earlier (e.g. modeling and simulation testing, or component level DT) designed experiments not only set the foundation by establishing what is known about the system, but are an excellent opportunity to build on that knowledge with sequential testing.

One of the best applications is associated with software operating systems existing on legacy weapon systems. For example, an aircraft operating system undergoes periodic upgrades every 18 months. Each time the new variant is ready for initial testing, the government (typically in concert with the contractor) devises a test schedule for "regression" testing. The purpose of regression testing is to, prior to testing any of the new features, make sure all the capabilities present in the previous version are still present. A natural use of sequential testing here would be to take the OT test plan (intended to fully exercise the system under the full complement of operational conditions) from the previous software version and either augment it or repeat a reduced subset of it and compare results to the OT statistical model.

Another logical use of sequential testing on a macro scale would be to take an early DT design and augment it to test the newly added capabilities. One could easily see taking late stage DT (system nearly ready for OT certification) test design and augmenting it to more comprehensively assess effectiveness and suitability. Probably a better way to plan isn't to throw the previous test plan over the fence, but to discuss early, when building the TEMP, what would be the best combined design for DT, for DT/OT, and for OT. Then decisions could be made as to which points DT will accomplish and which will be left for OT. Figure 5 provides a graphical depiction of the intent to use design of experiments throughout the acquisition life cycle and to pass the understanding of a system at a particular level of maturity to the next test period, such that the proceeding tests and analysis form the foundation for the test design in subsequent tests.

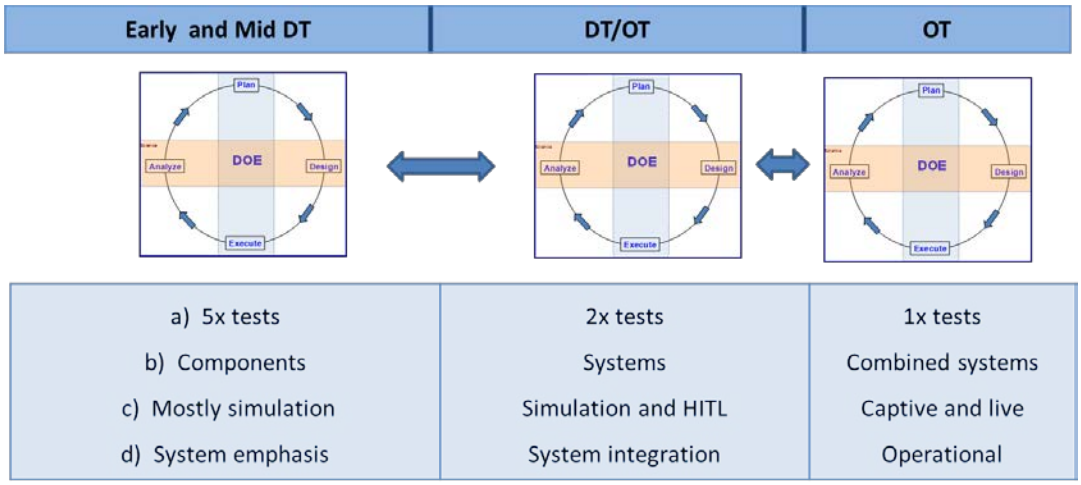


Figure 5. Sequential testing across acquisition phases. The table below distinguishes the emphasis in test phases in terms of the a) number of test events, b) type of system involved, c) system abstraction used as the test venue, and d) degree of system interface

Summary

Clearly the concept of sequential testing is worthy of consideration as we plan our tests. Some summary points to consider when considering the employment of test phases as part of your next test plan are listed below. This list is not intended to be exhaustive, but it is sufficient to start seriously thinking about the prospects and benefits of a multi-stage approach to test execution. Remember that even one-shot tests can be subdivided to take advantage of multiple test priorities and to minimize the risk of having insufficient data should the test be cut short.

Some rationale for separate phases of tests would be that you have more than one goal or objective for testing or that you might want to screen and optimize. Consider the following possible list of test needs all arguing for sequential testing, all from the same test project:

- Screen with many factors, less levels and simple models – leverage efficiency, then
- Re-assess or improve test execution procedures to better manage noise, or
- 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 the next stage of points where needed to better model factors to responses or capture nonlinear performance
- Augment to decouple interactions aliased together, either partially or perfectly
- Validate or confirm predicted or unusual performance with one or more confirmation runs
- Move in factor space to improve performance so there is a need to change the factor levels
- Allow for system repair, recalibration and retest
- Always think to 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

If for no other reason than to identify problems or successes early, every tester should be motivated to adopt sequential testing or the practice of test-analyze-test. A recent test practice encountered may serve to highlight the benefits to such a test philosophy. Traditional test methods for a large weapon delivery platform integration with weapons consist of enumerating thousands of test conditions to be executed by the government, prior to initial operational test and evaluation (IOT&E). The procedure stipulates that all the test conditions be executed, then passed on to the defense contractor for assessment. The test execution schedule requires hundreds of test sessions over the course of 6-9 months leading up to IOT&E, with the final outcome being a pass/fail, well after the last test event is conducted. Suppose instead a sequentially phased designed experiment approach with specific desired outcomes (screen large effects, initial model, augment to refine factor and interaction influences, refine model, augment to increase estimation precision, enhance model, augment to fully characterize, final model, validate) be used with reporting at the end of each phase and corrective actions taken. Not only will issues be identified much earlier, but tremendous test point savings can ultimately be realized, resulting in an on-time, on-budget, with more capable system assessment.

References

Box, G. E. P. 1992, “George’s Column: Sequential Experimentation and Sequential Assembly of Designs” *Quality Engineering*, 5:2, 321-330.

Box, G. E. P. & Hunter, J. S. 1961a, "The 2^{k-p} Fractional Factorial Designs Part I", *Technometrics*, 3:3, 311-351.

Box, G. E. P., and Hunter, J. S. 1961b, “The 2k-p Fractional Factorial Designs, Part II”, *Technometrics*, 3, 4, 449-458.

Box, G. E. P., Hunter, J. S., & Hunter, W. G. 2005, *Statistics for Experimenters*, *2nd ed*., John Wiley and Sons, Inc., Hoboken, New Jersey.

Daniel, C. 1962, "Sequences of Fractional Replicates in the 2^{p-q} Series," *Journal of the American Statistical Association*, 57:298, 403-429.

Emanuel, J. T., & Palanisamy, M. 2000, “Sequential Experimentation using Two-Level Fractional Factorials,” *Quality Engineering*, 12:3, 335-346.

Fisher, R. A. 1935, *The Design of Experiments*, Oliver & Boyd, Oxford, England.

Fisher, R. A. 1952, “Sequential Experimentation,” *Biometrics*, 8: 183-187.

JMP Statistical Software 2013, SAS Institute, Cary, North Carolina.

John, P. W. M. 1966, "Augmenting 2^{n-1} designs", *Technometrics*, 8:3, 469-480.

Jones, B., & Nachtsheim, C. J. 2011. A Class of Three-level Designs for Definitive Screening in the Presence of Second-order Effects. *Journal of Quality Technology*, *43*(1), 1-15.

Kimball, A. 1957, "Errors of the Third Kind in Statistical Consulting." *Journal of the American Statistical Association.* 52: 278: 133-142.

Li, W. & Lin, D. K. J. 2003, "Optimal Foldover Plans for Two-Level Fractional Factorial Designs", *Technometrics*, vol. 45, no. 2, pp. 142-149.

Mee, R. W. & Peralta, M. 2000, "Semifolding 2^{k-p} Designs", *Technometrics*, vol. 42, no. 2, pp. 122-134.

Montgomery, D. C. & Runger, G. C. 1996, "Foldovers of 2^{k-p} Resolution IV Experimental Designs", *Journal of Quality Technology*, vol. 28, no. 4, pp. 446-450.

Misra, H., Ríos, A. J., Simpson, J. R. & Vázquez, J. A. 2013, "The Quarterfold, a Sequential Augmentation Procedure for Resolution IV Fractions", *Quality Engineering*, 25:2, 118-135.

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

Nguyen, N. K., & Lin, D. K. J. 2011, "A Note on Small Composite Designs for Sequential Experimentation", *Journal of Statistical Theory and Practice*, 5, 1, 109-117.

Rios, A. J., Simpson, J. R. & Vazquez, J. A. 2011, "Sequential Experimentation Approach for Augmenting of Resolution III Fractions," *Communications in Statistics-Theory and Methods*, 40, 2337-2357.

Scuillo, C. 2007, "Augmenting Second Order Designs for Validation and Refinement," *Master's Thesis*, Florida State University, Tallahassee, Florida.

Design Expert Software 2013, Stat-ease Corporation, Minneapolis, Minnesota.

Vining, G. Geoffrey 2011, "Technical Advice: Design of Experiments, Response Surface Methodology, and Sequential Experimentation," *Quality Engineering*, 23:2, 217-220.

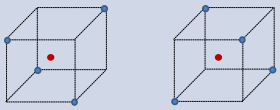
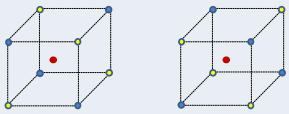
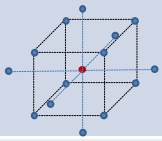
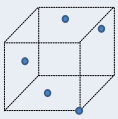
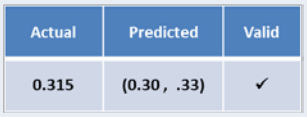
Appendix A –Sequential Design and Analysis Tutorial in 4 Stages

Many test planners think in terms of one-stage screening test designs as sufficient for complete discovery. The primary purpose of this section is point to the various needs of any test project that can't be satisfied by a single shot design. In fact, a wise test strategy involves placing points in a design stage for the express purpose of determining which kinds of additional points are needed in latter stages. We will explore a general test strategy by means of an example, starting with a realistic number of factors with a mix of categorical and numeric types. The general sequence and flow of testing is to:

- Perform a screening design (assuming the intent is to detect main effects and some two factor interactions) to obtain an initial error estimate (including pure error), and detect lack of fit
- Augment for decoupling aliased interactions associated with significant effects and increase the precision of error estimation (not unlike 2^k replicated in blocks)
- Augment for second order, assuming a higher order model is required
- Augment for optimization, with the possibility that the optimum is located outside the region of experimentation, add points near the region of the optimum and extend the ranges of some of the factors
- Augment for validation and possible further model refinement

Table 1 displays the four blocks designed specifically for this problem. In general there will be four stages of testing: screening, decoupling, response surface methods (RSM), and refinement.

Table 1. Sequential Design Stages for Appendix Example

Block	Design	Model
1: Screen – linear model		$Y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \ some \sum_{i<j} \beta_{ij} x_i x_j + \varepsilon$
2: Decouple – interaction model		$Y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i<j} \beta_{ij} x_i x_j + \varepsilon$
3: RSM – 2nd order model		$Y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i<j} \beta_{ij} x_i x_j + \sum_{i=1}^{k} \beta_{ii} x_i^2 + \varepsilon$
4: Refine – add precision and validate		

Actual	Predicted	Valid
0.315	(0.30 , .33)	✓

Problem Specifics

The approach for this example is to assume a single response, six factors of interest with a mix of categorical and numeric factor types, 2-level categorical factors, generic level coding, and a desire to potentially fit a 2nd order model. To illustrate the features of sequential learning via design-analyze-design, we will use a Monte Carlo approach to generate the data from a true assumed underlying system polynomial model with noise. Noise will be added to make the data realistic and also challenge the design and statistical modeling process.

Suppose the following six factor scenario resulted from process decomposition:

Factor	A	B	C	D	E	F
Type	Numeric	Numeric	Categorical	Numeric	Categorical	Numeric
Levels	2+	2+	2	2+	2	2+

Coded units with initial low=-1 and high=+1 will be used throughout for simplicity

Response: Y is continuous and assumed to be a target score, so greater values are better

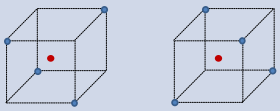
It will be assumed the system model is quadratic with 4 significant factors and truth model:

$$Y = +79.0 - 6.0\,x_B + 4.0\,x_D - 7.5\,x_F + 5.0\,x_A\,x_F - 5.5\,x_B\,x_D + 4.5\,x_D\,x_F + 4\,x_B^2 - 9x_D^2 + \varepsilon, \text{ where } \varepsilon \sim N(0, 3^2)$$

Because this example is a lone illustration of sequential assembly, care is taken to make the truth model characteristic of a 'standard' 2nd order model encountered in practice. Some aspects of the above model that are consistent with many historical systems are:

- Effect sparsity: 4 of the 6 factors are significant, and system driven primarily by main effects and 2-factor interactions, and six total main effects + 2-factor interactions
- Model term heredity: 2 of the 3 significant 2-factor interactions abide by strong heredity (both main effects of the interaction are also significant), while one interaction has weak heredity (only one main effect in the interaction is significant)
- Quadratic behavior: Often some subset of the numeric factors have significant curvature, and in this case 2 of the 4 numeric factors are 2nd order

Block 1: Screening Factors (ME + some 2FI)

Block	Design	Model
1: Screen – linear model		$Y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \; some \sum_{i<j} \beta_{ij} x_i x_j + \varepsilon$

An initial screening design is needed to determine not only the factors with significant main effects, but to hopefully uncover the majority of two-factor interaction relationships and test for curvature to determine whether additional design points should be added to estimate the 2^{nd} order effects. It is decided that an Initial 2^{6-2} fraction with centers is not only efficient, but should effectively meet the objectives. Beyond the scope of this tutorial is a power analysis, a critical step before deciding whether this screening design has sufficient power to detect certain prescribed effect sizes in the presence of noise. The specific design chosen is a classical resolution IV quarter fraction in 16 runs, plus 1 pseudo-center (2^2 factorial in C and E) set of 4 center runs (giving 1 degree of freedom (df) for curvature and 3 df for lack of fit testing), for a total of 20 points

The fractional factorial alias structure has defining relation I = ABCE = ADEF = BCDF. The alias chains for the main effect and two-factor interaction effects are provided in Table 2.

Table 2 Alias Chains and Screening Design for Block 1 Tests Including Simulated Data

[A] = A + BCE + DEF
[B] = B + ACE + CDF
[C] = C + 0.8 * ABE + 0.8 * BDF
[D] = D + AEF + BCF
[E] = E + 0.8 * ABC + 0.8 * ADF
[F] = F + ADE + BCD
[AB] = AB + CE
[AC] = AC + BE
[AD] = AD + EF
[AE] = AE + BC + DF
[AF] = AF + DE
[BD] = BD + CF
[BF] = BF + CD

Run	Factor 1 A:A	Factor 2 B:B	Factor 3 C:C	Factor 4 D:D	Factor 5 E:E	Factor 6 F:F	Response 1 Target Score
1	-1.00	-1.00	Level 1 of C	1.00	Level 1 of E	1.00	88
2	1.00	1.00	Level 2 of C	-1.00	Level 2 of E	-1.00	90
3	0.00	0.00	Level 1 of C	0.00	Level 2 of E	0.00	75.3
4	1.00	-1.00	Level 2 of C	-1.00	Level 1 of E	1.00	78.5
5	0.00	0.00	Level 2 of C	0.00	Level 2 of E	0.00	78.2
6	1.00	-1.00	Level 1 of C	1.00	Level 2 of E	1.00	106
7	1.00	1.00	Level 2 of C	1.00	Level 2 of E	1.00	75.5
8	-1.00	-1.00	Level 2 of C	1.00	Level 2 of E	-1.00	112
9	-1.00	-1.00	Level 1 of C	-1.00	Level 1 of E	-1.00	97.2
10	1.00	-1.00	Level 2 of C	1.00	Level 1 of E	-1.00	103
11	1.00	1.00	Level 1 of C	1.00	Level 1 of E	-1.00	76
12	-1.00	-1.00	Level 2 of C	-1.00	Level 2 of E	1.00	64.7
13	-1.00	1.00	Level 2 of C	-1.00	Level 1 of E	-1.00	97
14	-1.00	1.00	Level 1 of C	1.00	Level 2 of E	-1.00	88
15	0.00	0.00	Level 1 of C	0.00	Level 1 of E	0.00	78
16	0.00	0.00	Level 2 of C	0.00	Level 1 of E	0.00	80.9
17	-1.00	1.00	Level 1 of C	-1.00	Level 2 of E	1.00	71.5
18	1.00	-1.00	Level 1 of C	-1.00	Level 2 of E	-1.00	85.1
19	1.00	1.00	Level 1 of C	-1.00	Level 1 of E	1.00	72.8
20	-1.00	1.00	Level 2 of C	1.00	Level 1 of E	1.00	70.9

Classical 2^{k-p} fractional factorials not only contain attractive design point location symmetries, but also project to full factorials in fewer factor subsets. Because the 2^{6-2} is resolution IV, it will project into a full factorial in all 4-factor subsets of the 6 total factors (there are 15 total), except the 3 from the defining relation (ABCE, ADEF, BCDF). Figure 6 shows one of the 12 4-factor full factorial projections, while Figure 7 shows the ADEF projection, resulting in a 2^{4-1} fraction. This fraction has 2-factor interaction (2FI) aliased with another 2FI. Once data is collected and initially analyzed, those factors with significant main effects build evidence in favor of particular 2FI (due to the heredity principle) within an alias chain, assuming the effect associated with that alias chain is large. So, although 2FI are aliased in this screening design, there is a reasonable chance that the correct ME and 2FI model can be ascertained from this initial experiment.

Block 1 Design Characteristics:

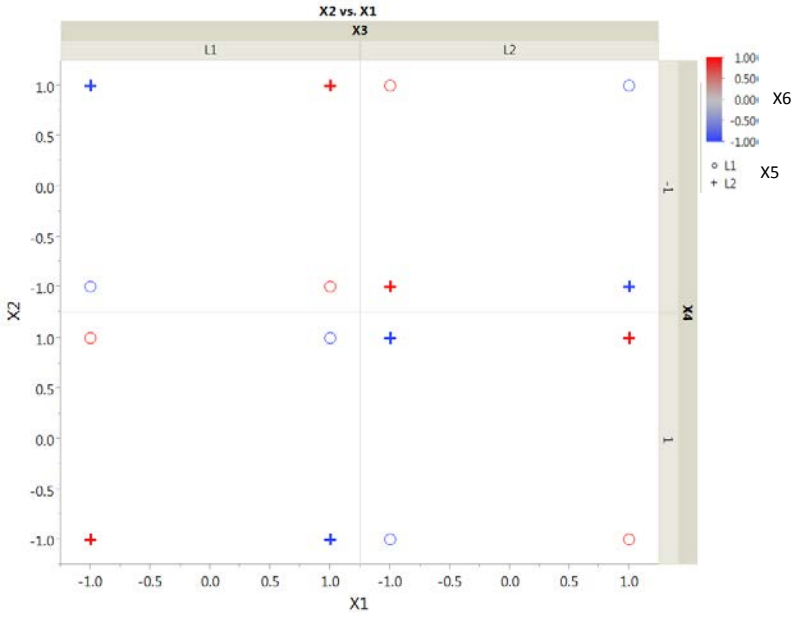


Figure 6. 2^{6-2} Design, 16 points from a 6D perspective – here a ABCD (or X1 X2 X3 X4) projection to 24, with alternating coverage in EF (X5 X6) using color and symbols. Center points not shown.

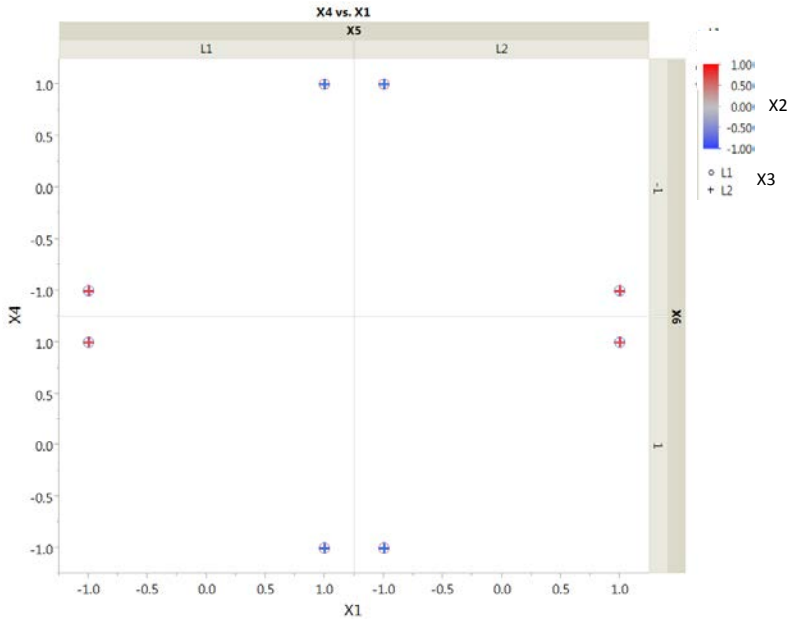


Figure 7. 2^{6-2} Design Points – Projection to 2^{4-1} in ADEF. So AD=EF, for example. Center points not shown.

Block 1 Analysis

The data for the 20-run Block 1 design is simulated and analyzed. Please note that the following discussion describes the findings and decisions based on a single simulation of the truth model plus noise. During construction of this sequential assembly tutorial, in some blocks multiple simulations are

performed to ensure that the single example dataset simulated is representative of other similar realizations from the truth model. The graphs in Figure 8 show the effects chosen for the model, along with the 2FI model graphs. The aliases in red show the chains for the significant interaction effects.

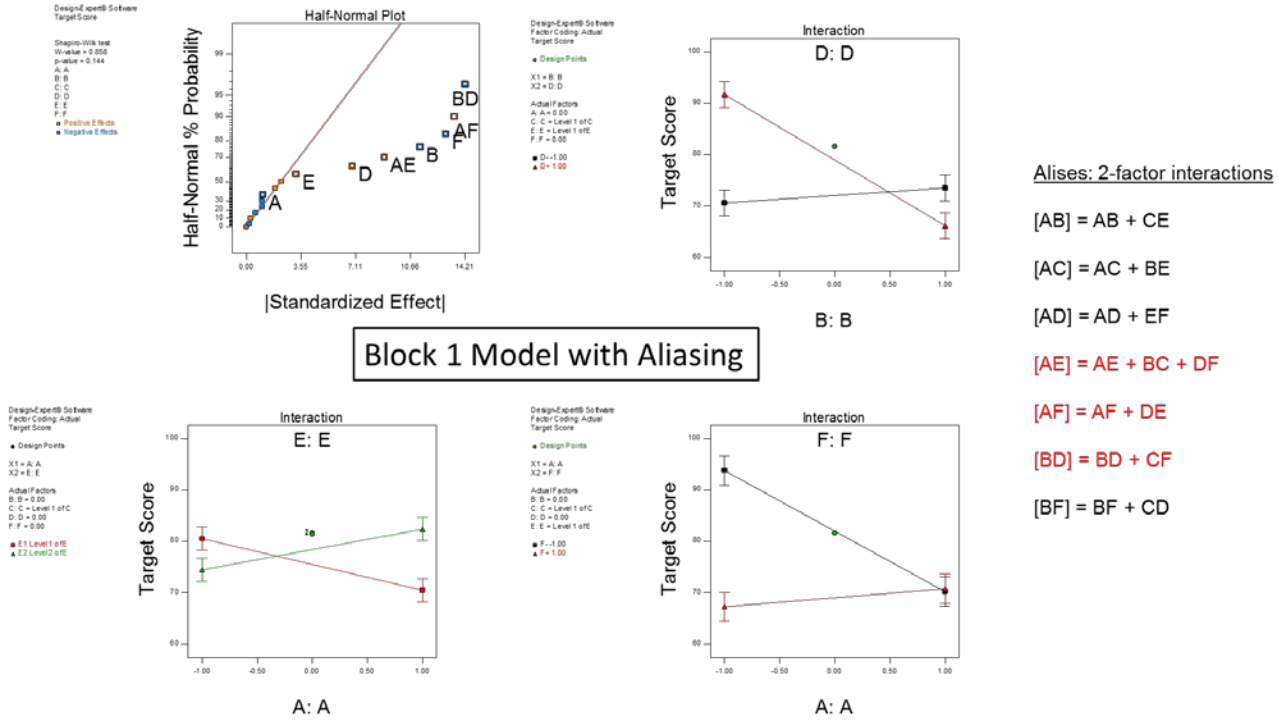


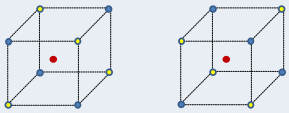
Figure 8. Analysis of Block 1 data from the 2^{6-2} fractional factorial plus centers. The normal probability plot of effects (top left) shows 6 effects (BD, AF, F, B, AE, and D) appear to be significant, while two others (A and E) are added for hierarchy. No attempt is made to determine correct effects from aliases

Typically, as part of the model building with screening designs, reasoning based on effect heredity is made as to which effects in a chain are contributing to the column (chain) having a large effect estimate. In this case, because it is far more likely for main effects to matter than 3FI, it is safe to assume B, D, and F have significant main effects. Then, based on the significant 2FI chains, the more likely interaction in BD=CF is BD. For the other two chains, DF may be driving the longer 3-effect chain, but it is not clear whether AF or DE is truly important in that chain. An additional set of points (Block 2), involving some form of foldover to decouple the interactions, is warranted.

Regarding the test for curvature from the Block 1 design, the F-test shows curvature with a p-value of 0.0985, indicating moderate significance. Because an additional experiment will be performed, additional centers will be added and the curvature will be checked again before deciding to augment for a 2nd order model.

Block 2: Decouple, Increase Power and Pure Error (PE) Estimate

The second set of runs is constructed with two purposes and one expected additional benefit. The primary purpose is to foldover the screening design to decouple the 2FI effects. The other purpose is to add an additional set of 4 center runs to enable another curvature test. Both augmentations have the added benefit of increasing the power of the test for model effects and for curvature. The approach to analysis will be to combine the runs from Blocks 1 and 2 for modeling and to use the block effect to remove any background variation due to the different blocks.

Block	Design	Model
2: Augment – interaction model		$Y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i<j} \beta_{ij} x_i x_j + \varepsilon$

A semifold of the original design is selected due to the relative efficiency of such a strategy (only 8 runs) and because reasonable inferences regarding which 2FI are ultimately driving performance had been made during screening. The decision is made for the semifold to foldover on F and hold C at the low level (-1). The rationale is that F has the largest main effect and appears in most alias chains with significant effects, while C doesn't appear to be involved. Another psuedo-center (4 points) is added, to provide the curvature check and 3 additional error degrees of freedom. Figure 9 shows a 4-factor projection of the augmented Block 2 design in ADEF to show that 2^{4-1} is now a 2^4 full factorial.

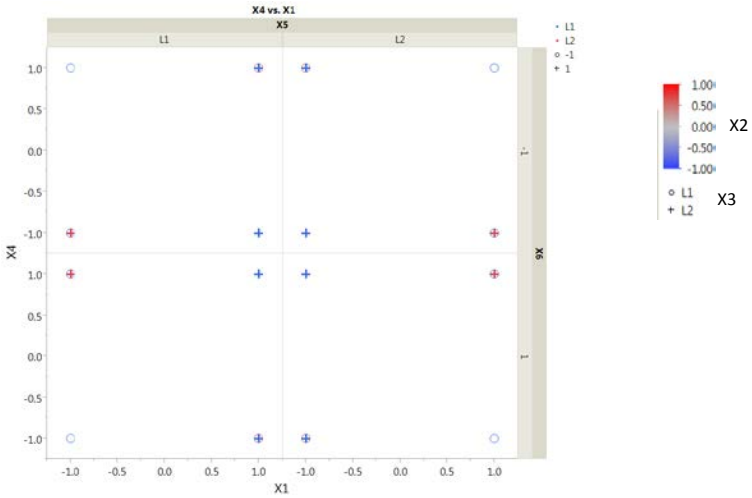


Figure 9. 2^{6-2} Design with semifold augmentation – Projection to 2^4 in ADEF (X1 X4 X5 X6) with coupled interactions in the original design are decoupled after the semifold. Center points not shown.

Block 2 Analysis

The original 20 runs from Block 1 remain as the original data, while the 12 Block 2 runs are simulated, resulting in an executed 32-run design to be analyzed. All the 2FI of interest have been successfully

decoupled, allowing the normal probability plot of effects to show the likely model, at least in terms of main effects and interactions. Figure 10 shows that plot along with the 2FI plots.

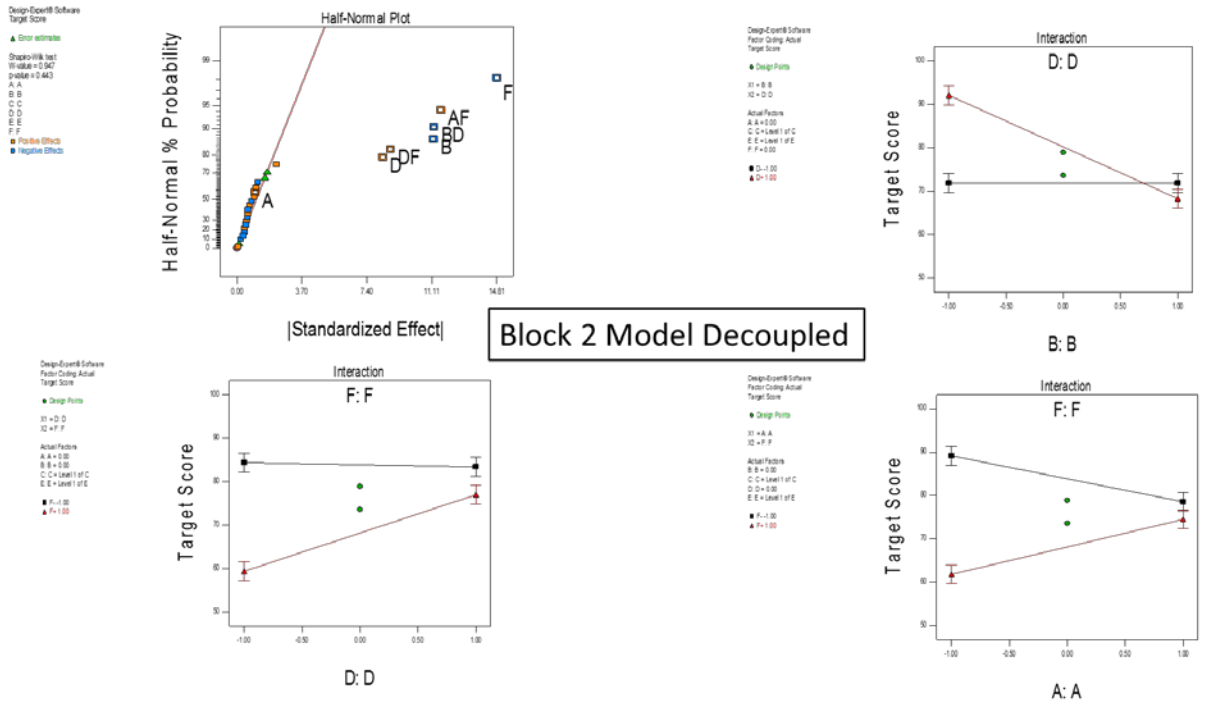


Figure 10. Analysis of Block 2 data from the original screening design, semifold plus centers. The normal probability plot of effects (top left) now shows F, AF, BD, B, DF, and D appear to be significant, while A is added for hierarchy.

Not only do all the model terms show significance (all p-values < 0.01), but the interaction plots show lines with noticeable line slope differences, making interpretation meaningful. Although the journey appears to be complete for the lower order terms, the test for curvature (Table 3) now shows that second order terms are conspicuously absent.

Table 3. Block 2 Analysis Test for Curvature

But, curvature is significant!

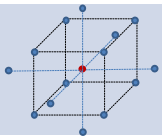
ANOVA Summary		
	Adjusted	Model
	F-value	p-value
Model	80.65	< 0.0001
Curvature	10.96	0.0032

The following stage of testing (Block 3) focuses on providing the test points to allow proper fitting of the higher order model called out by the test for curvature.

Block 3: Augment for 2nd Order

The objectives for this stage of testing are to add points efficiently that will permit adequate estimation of the full second order model and to continue enhancing error estimation and improve decoupling of

interactions. One choice here would be to add axial points, as in a central composite design augmentation, but by using an I-optimal algorithm for all points. Those points will serve to estimate all the terms in the 6-factor second order model (recall 2 factors are categorical, so only 4 pure quadratics need to be fit). Understanding a minimum of 8 points are needed for the 4 quadratics, 12 points are added to Stage 3 to enhance model estimation.

Block	Design	Model
3: RSM – 2nd order model		$Y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i<j} \beta_{ij} x_i x_j + \sum_{i=1}^{k} \beta_{ii} x_i^2 + \varepsilon$

Design metrics are used to evaluate and compare various choices (point types, optimality algorithm, software package) for generating the additional 12 points. The most important metrics are: VIF max and average, median (50 percentile) standard error of the mean, 95 percentile standard error of the mean, G-efficiency, D-optimality criterion, and I-criterion. The final design, including all 44 points of the 3 blocks, is shown in Figure 11.

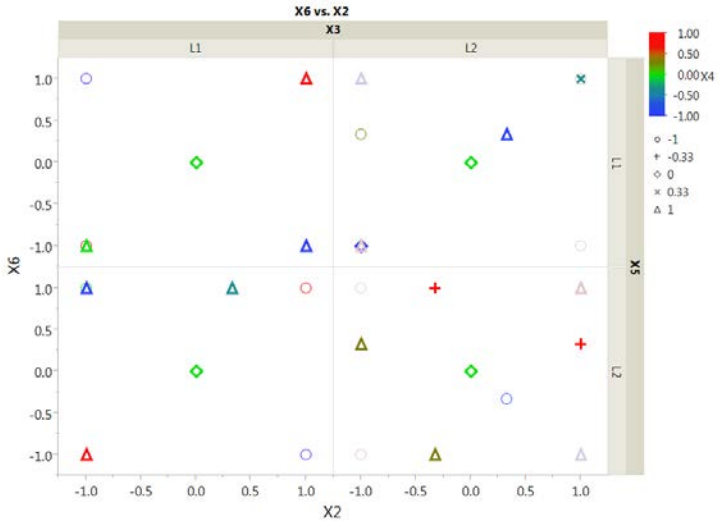


Figure 11. Block 3 Design, with original 2^{6-2}, augmented with semifold, then augmented via I-optimal points for full second order model.

The points from Block 3 include a mix of interior and edge points. The set of 12 points used here represent the best set of points in terms of the design metrics described above. The next step is to simulate 12 new observations and analyze the full 44-point design.

Block 3 Analysis – Second order design analysis and optimization

Clearly, from the previous analyses, the focus is now on second order modeling. In the 6-factor design with 2 categorical factors there are 25 model terms, excluding the intercept. In terms of model building, there are two natural choices. The first is to fit the full 2nd order model with all main effects, 2FI, and

pure quadratics. The other is to trim the model to include only the significant model terms and maintain hierarchy. Both models are fit and Table 4 provides a summary comparison.

Table 4. Model Selection Choices for the Quadratic Model

Criterion	Full Quadratic	Reduced 2nd Order
Model df	25	9
R^2	0.972	0.953
Adj R^2	0.927	**0.940**
Predicted R^2	0.701	**0.908**
MSE	10.766	**8.350**
FDS		
50%	0.679	**0.388**
95%	0.910	**0.511**
VIF max	2.95	**1.88**
Residuals	no transform, no outliers	no transform, no outliers

Summary statistics clearly show that the reduced second order model is the better choice overall. The first consideration is model df, or parsimony, which favors simpler models, so all other metrics equal, the simpler model is preferred. The comparison statistics show the reduced model outperforms the full model in nearly every category, showing superior adjusted and predicted R^2, as well as smaller MSE and max VIF, plus better prediction uncertainty.

The 9-term model is then used to assess the fit graphically and determine the general shape of the surface (Figure 12). For this example we will assume that one of the objectives is to learn about factor conditions giving better (higher) responses for target score. As such, the region circled in red, at the perimeter of the experimental region, is most promising.

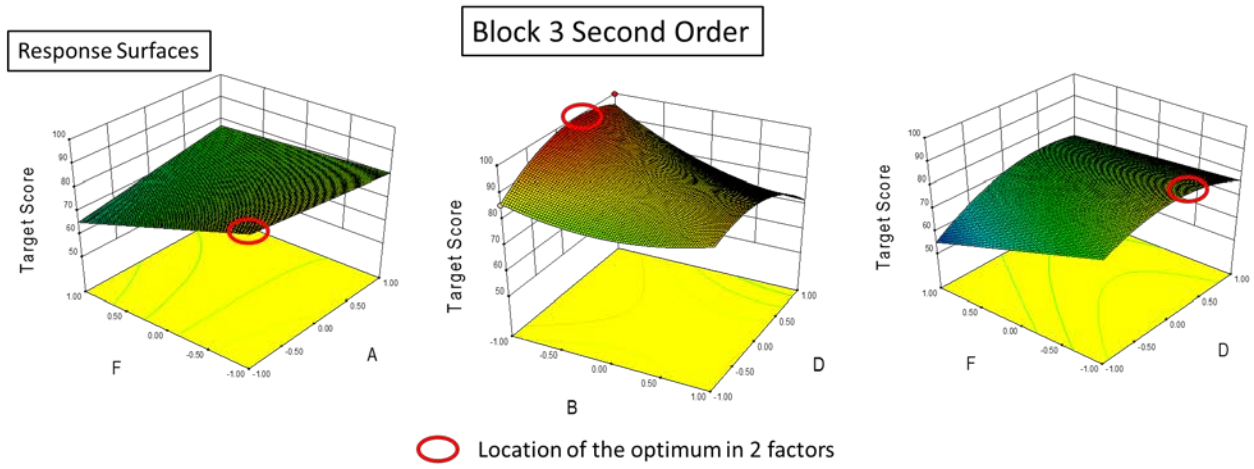
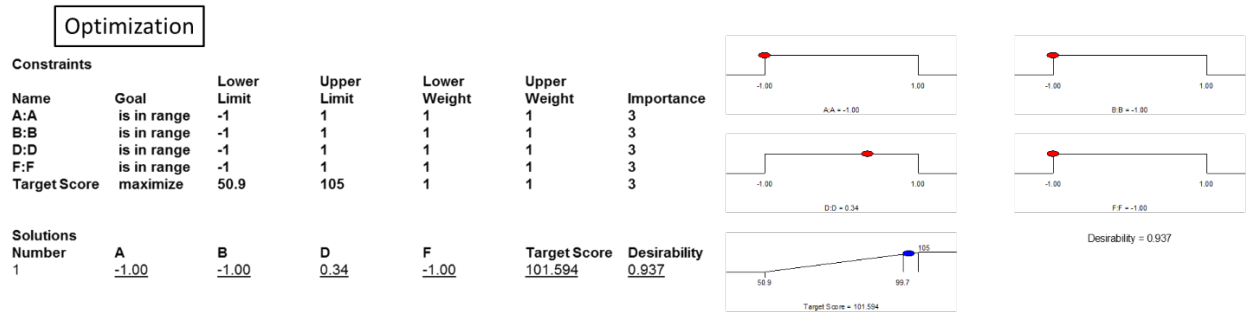


Figure 12. Response surface plots showing location of best performance within the region of experimentation, across different factor pairs (top), and then in BD again, but rotated.

An optimization (maximize target score) of the model function (Figure 13) confirms this finding, showing that extremes in A, B and F of -1 in coded units, combined with a 0.4 coded unit setting for D gives maximum predicted target scores. One possible strategy for next tests is to move slightly outside the original 6D hypercube and investigate possible continued improvement in target score. Clearly, one should ensure these new factor level settings (out to -1.5 in coded units) are feasible.



Optimization

Constraints

Name	Goal	Lower Limit	Upper Limit	Lower Weight	Upper Weight	Importance
A:A	is in range	-1	1	1	1	3
B:B	is in range	-1	1	1	1	3
D:D	is in range	-1	1	1	1	3
F:F	is in range	-1	1	1	1	3
Target Score	maximize	50.9	105	1	1	3

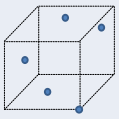
Solutions

Number	A	B	D	F	Target Score	Desirability
1	-1.00	-1.00	0.34	-1.00	101.594	0.937

Figure 13. Optimization performed on Block 3 surface, showing settings of the 4 significant factors and the estimated response.

Block 4: Optimize and Validate

In Block 3, the goal is to fit a second order polynomial to the 6-factor test space. This second order model can be used for a number of purposes: to specify which factors matter, characterize the performance, explain the factor interactions and the nonlinear (in B and D) relationships that exist. The second order model can also point to regions of improved response, and in this case, the better conditions are: A, B, and F = -1, D ≅ 0.4, with C and E not being significant. The goals of this final stage, Block 4 of testing, are to: add points to refine the response performance around the region of the optimal, expand some beyond the region of experimentation, add a few points for validation, and strengthen the quadratic model fit.

Block	Design	Model
4: Refine – add precision and validate		Actual: 0.315 · Predicted: (0.30 , .33) · Valid: ✓

Design Augmentation Steps

1. To expand the region of experimentation, build an 8-run 2^{6-3} design in traditional coded (-1, +1) units, then in significant factors A, B, D, F, then modify the low/high for A, B and F to (-1.5, -1) in coded units for A, B, F and set low/high for D to (0.1, 0.7).
2. With the region expansion design (2^{6-3}) as the initial 8 points of Block 4, add 4 more points using the I-optimal criterion to satisfy two objectives: (1) enhance the model fit, (2) serve as validation points. Prior to execution, be sure to randomize all 12 Block 4 points.
3. Before deciding on this two-fold (8-run ridge expansion fraction plus 4 I-optimal validation points) augmentation, compute and assess the combined design (now 56 points) metrics, compared to the Block 3 metrics, for improved statistics. It is anticipated better prediction capability will be secured in the region of optimum.

Figures 14, 15, and 16 show the Block 4 points relative to the existing design from different perspectives. The 8 points expanding the region of interest extend the cube in particular directions, while the validation points help balance the 8 new points relative to the other factors and the existing 2nd order model.

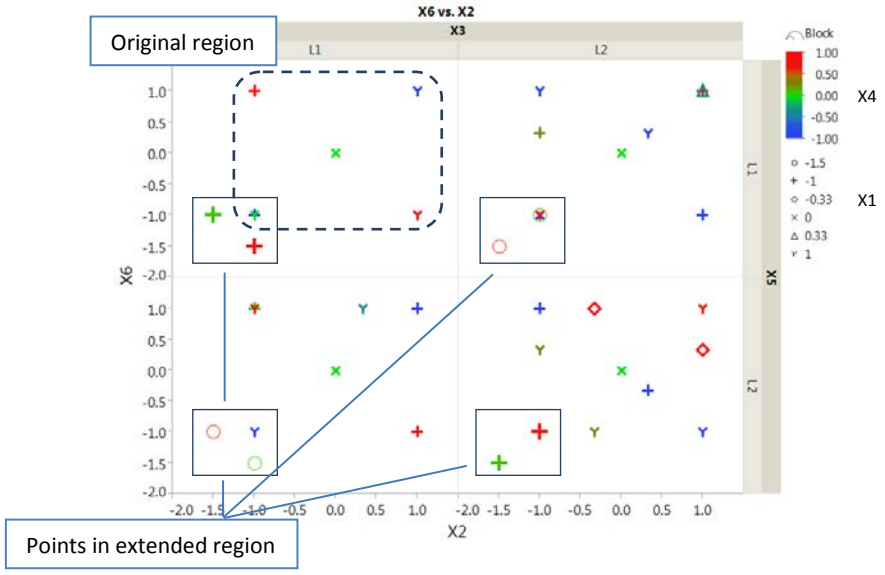


Figure 14. Block 4 Design with 2nd order design augmented to explore region of improved response in X1, X2, X6.

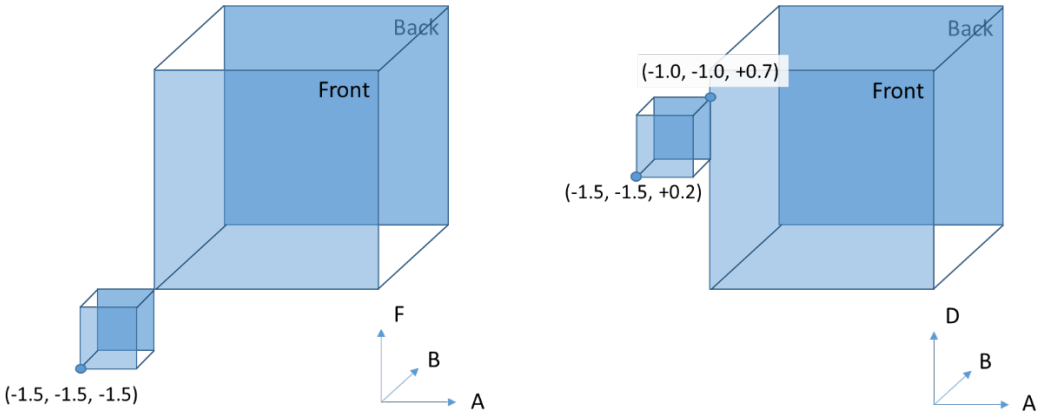


Figure 15 Block 4 Design in 3D showing 2 3-factor subset projections of the extended region of exploration based on optimal response from Block 3, with example points indicated for illustration purposes only.

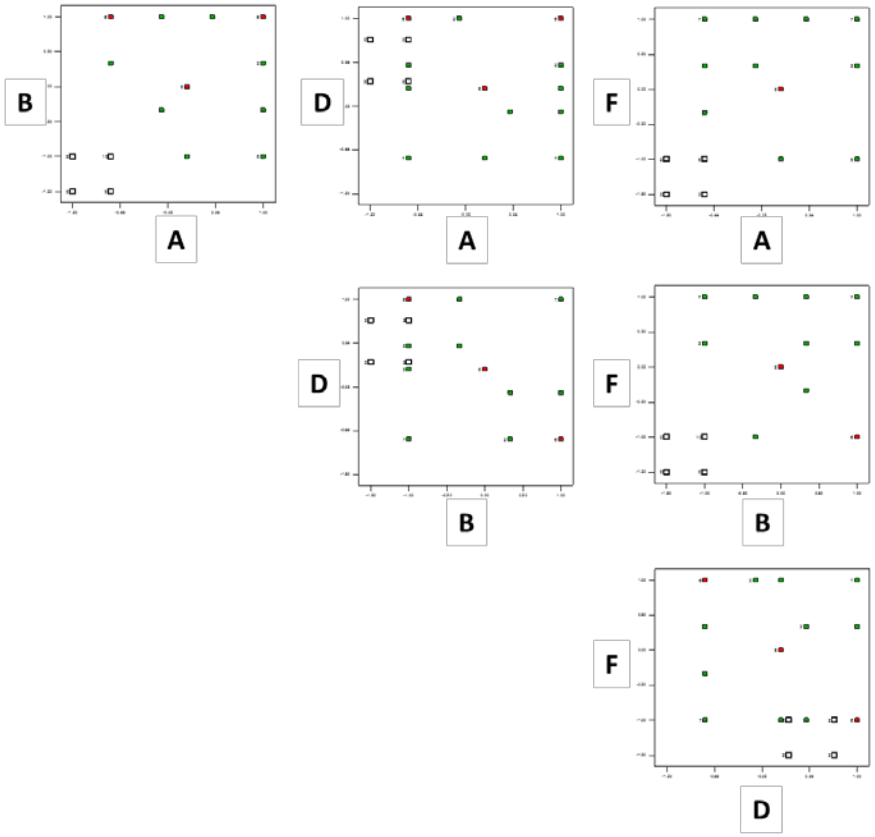


Figure 16. A scatterplot matrix of the 4-variable design in multiples of 2 variables at a time. The extended region is visible in any 2 dimensions.

Block 4 Analysis

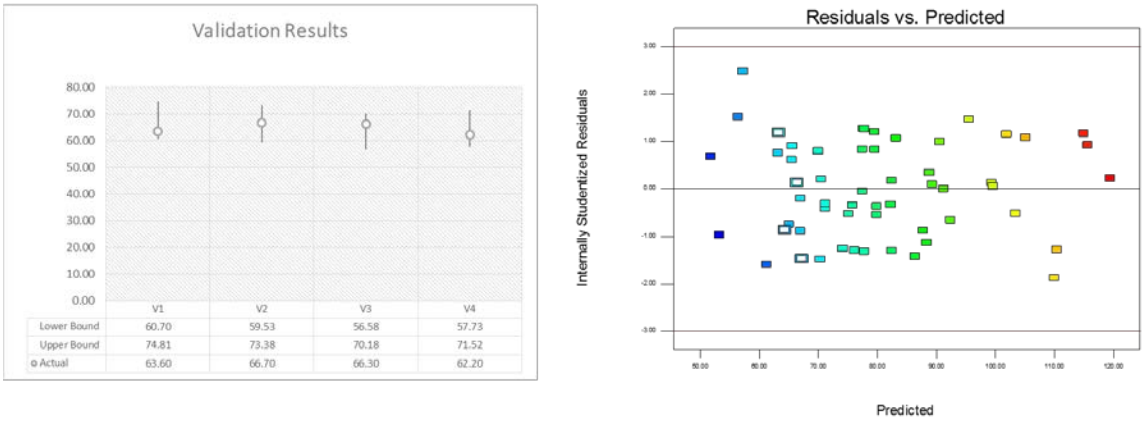
For this last analysis, we execute the final 12 runs and record the simulated values in the target score response. In fitting the empirical model, a couple of adjustments are first made. The software should

know that for each factor with expanded range (A, B, and F) the new low setting of -1. Second, before fitting the model to the new data, be sure to ignore the 4 validation points from consideration (see Step 1 below). They will be used to compare their actual value to the value predicted by the model to assess the ability of the model to effectively predict new observations. The analysis and validation exercise raises some important questions that should be answered. These questions frame the discussion.

Is the model valid? Yes.

The four validation points are used to determine whether the model is valid. Use the following procedure:

1. Set the validation points aside. In Design Expert (Stat-ease, 2013) you choose Ignore, while in JMP choose Exclude.
2. Fit the statistical model to the other 52 points using the same reduced 2nd order model from Block 3.
3. Use a point prediction tool to predict the validation set and compare the prediction intervals to the actual observation (Figure 17a).
4. One final check can involve including the validation points in the fit (select those points to be included). Check the model fit (coefficient estimates, MSE) to see if the models with/without the validation set are different (discussed later in Table 5).
5. Perform a residual analysis of the fit to the superset and see if any of the validation points are outliers. You could even check to see if the validation set represents an outlying set of points (Figure 17b).



a) b)

Figure 17. Plots showing a) the outcome of the validation (left), with the actual (o) falling within the prediction interval captured by the vertical lines; and b) showing the magnitude of the validation point residuals highlighted once the validation points are added back to the model

Are the prediction intervals tighter in the region near the optimum? Yes.

The addition of the range expansion points in fact decrease the standard error of the design for the coded values (-1.5, -1) in A, B, and F, and (0.1, 0.7) in D. Figure 18 shows an approximate 30-40% reduction in the standard error values in the region of the optimum.

Standard error of the design

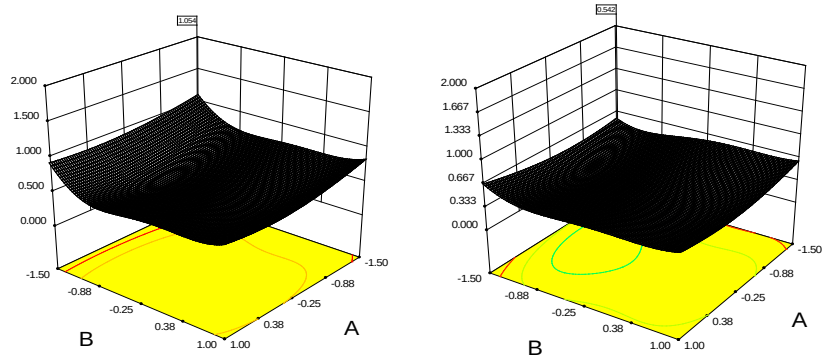


Figure 18. Prediction variance (standard error of design) reduction by adding the validation set which also extends the region of experimentation. Side by side plots show the region from -1.0 to -1.5 in coded units prior to and post augmentation.

Is there a new optimum location and response? Yes.

The new optimum target score now resides outside the original hypercube constrained by -1 to +1 in coded units. The new location now is located -1.5 coded units in 3 of the four significant variables (A, B, F), and at about 0.4 in coded units for factor D.

The table below shows the new optimum location and prediction, as well as the uncertainties of prediction. Recall only 8 points were added for this block. Those 8 were further subdivided into points for region expansion and points to improve the overall model fit. Greater reductions in confidence and prediction intervals can be achieved with a few additional points near the optimum (Table 5).

Table 5. Prediction of Target Score at location = (-1.5, -1.5, C, 0.4, E, -1.5)

Factor	Name	Level	Low Level	High Level	Std. Dev.	Coding
A	A	-1.50	-1.00	1.00	0.000	Actual
B	B	-1.50	-1.00	1.00	0.000	Actual
C	C	Level 1 of C	Level 1 of C	Level 2 of C	N/A	Actual
D	D	0.40	-1.00	1.00	0.000	Actual
E	E	Level 1 of E	Level 1 of E	Level 2 of E	N/A	Actual
F	F	-1.50	-1.00	1.00	0.000	Actual

*** WARNING - One or more factor value(s) is outside of the design space.

								99% of Population	
Response	Prediction	Std Dev	SE Mean	95% CI low	95% CI high	SE Pred	95% PI low	95% PI high	95% TI low	95% TI high
Target Score	120.456	2.98457	3.3208	113.692	127.22	4.46491	111.361	129.551	104.934	135.978

The above interval shows the prediction based on the data including Block 3, while the prediction and intervals below include the Block 4 data also.

Response	Prediction	Std Dev	SE Mean	95% CI low	95% CI high	SE Pred	95% PI low	95% PI high	99% of Population 95% TI low	95% TI high
Target Score	120.413	3.163	1.79828	116.786	124.039	3.63845	113.075	127.75	107.463	133.362

Does the model need to be refined? No.

There is very little change in the model coefficients, before and after validation, which is always important to verify as a part of validation (Table 6).

Table 6. Block 4: Before/After Model Validation Coefficient Estimates

Factor	before	after	% change
Intercept	78.70	78.38	0.32
A	0.66	0.52	N/A*
B	-5.70	-5.93	0.23
D	3.89	3.90	0.01
E	-0.15	-0.17	N/A*
F	-7.05	-7.04	0.00
AF	4.84	5.07	0.24
BD	-6.61	-6.68	0.07
DE	1.16	1.12	0.04
DF	4.43	4.61	0.18
B^2	4.32	4.38	0.06
D^2	-7.06	-7.07	0.01

*included for hierarchy, not significant

Is the value of the optimal improved – probably by moving outside original region? Yes.

Figure 19 shows the change in response values obtained by extending the search region outside the original hypercube. The prediction surface on the right is obtained with the region expansion points.

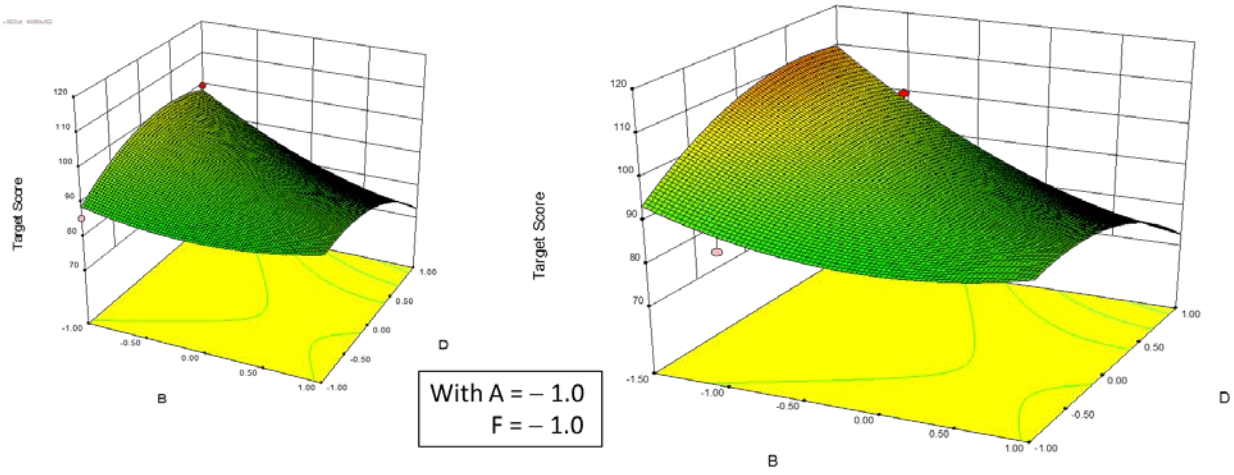


Figure 19. Response surface plot showing the Block 3 optimization vs. the extended region in Block 4.

Summary of Sequential Tutorial

Clearly, the preceding example shows some benefit in sequential testing compared to a single design approach. The single stage design would not have benefitted from the series of analyses that reveal the important factors (4 of the 6), the significant interactions, and the second order nature in two of the factors. A single design would might not have been as efficient in discovering the correct underlying model. Granted, an alternative would have been to build and execute a single second order design plus points for replication and validation. Given the knowledge that the system had high potential for second order, a single stage would have been largely successful. This one-stage design build would also be useful for estimating the total number of test needed, even if sequential testing is used. So why use sequential assembly? If the true model was of lower order, or had fewer significant factors, the one-stage approach would have been wasteful. The sequential approach seeks to test to the right level of fidelity, validate and stop. The blocking illustrated here allows the stages to be conducted over different testing periods, without the variability created by the testing periods to affect any of the analyses or findings. If at all feasible, consider a sequential testing approach as a default.