OFFICE OF THE SECRETARY OF DEFENSE
1700 DEFENSE PENTAGON
WASHINGTON, DC 20301-1700

OPERATIONAL TEST
AND EVALUATION

JUL 23 2013

MEMORANDUM FOR COMMANDING GENERAL, ARMY TEST AND EVALUATION COMMAND
DIRECTOR, MARINE CORPS OPERATIONAL TEST AND EVALUATION ACTIVITY
COMMANDER, OPERATIONAL TEST AND EVALUATION FORCE
COMMANDER, AIR FORCE OPERATIONAL TEST AND EVALUATION CENTER
COMMANDER, JOINT INTEROPERABILITY TEST COMMAND

SUBJECT: Best Practices for Assessing the Statistical Adequacy of Experimental Designs Used in Operational Test and Evaluation

Recent discussions within the test community have revealed that there are some misunderstandings of what DOT&E advocates regarding the appropriate use of statistical power when designing operational tests. I, as well as others in the test community, have observed that power calculations based on a single-hypothesis test on the overall mean are being used inappropriately by both government and industry in attempt to right-size a test. The purpose of this memorandum is to make clear what I view are best practices for the use of power calculations, as well as other statistical measures of merit that should be used to determine the adequacy of a test design.

Single-hypothesis test power calculations are generally inappropriate for right-sizing operational tests because they are not consistent with the goal of operational testing: to characterize a system's performance across the operational envelope. Furthermore, such estimates of power are unable to distinguish between both good and flawed test designs because they focus solely on the number of test points and ignore the placement of those points in the operational envelope. More informative power estimates exist. Power calculations that estimate the ability of the test to detect differences in performance amongst the conditions of the test (factors) will distinguish between good and flawed designs.

These "factor-level" power calculations are inherently related to the goal of the test; they not only describe the risk in concluding a factor is not important when it really is, but they are also directly related to the precision we will have on the quantitative estimates of system performance. The latter is key in my determination of test adequacy; without a measure of the expected precision we expect to obtain in the analysis of test data, we have no way of determining if the test will accurately characterize system performance across the operational envelope. A test that has low power to detect factor effects might fail to detect true system flaws; if it does, we have failed in our duty as testers.

In some cases a single hypothesis test might be warranted: tests where it is impossible to vary factors, or in cases where the threshold requirement must be met and we must be confident in correctly rejecting the null hypothesis (e.g., if the system's performance is not statistically significantly above threshold we must assume it is failing). In those cases, a series of operating characteristic curves for various test sizes is warranted to determine test risk. In all other cases, an estimate of the precision of the test must be provided for assessing the adequacy of characterizing the system's performance across conditions. This will typically be in the form of power estimates to detect factor effects; however, other methods might be acceptable provided they capture the accuracy of the test's ability to characterize performance.

As part of the community's discussions about statistical power, some have also noted potential difficulties in determining the appropriate effect size for a hypothesis test, since in many cases we have limited knowledge of the underlying variance of the system's performance. Because of this lack of knowledge, some assert we run the danger of either over or under-estimating the required sample size. However, we are able to estimate standard deviations in the data using historical or legacy system performance, modeling and simulation, as well as engineering judgment. Furthermore, a program's development typically consists of several phases, enabling testers to modify and learn about system performance (including the standard deviation) over the continuum of integrated and operational testing. I have also observed that an effect size selected for a single hypothesis test is not typically meaningful, and is another reason why I do not advocate this method for sizing tests. However, the solution is to ensure that effect sizes for observing important factors are properly motivated using all available sources of data and based on what is operationally relevant to the warfighter.[1]

Although important, power is not the only measure of the adequacy and merit of a test design. Test designs should be constructed based on the goal of the test, which should include the ability to discern factor effects (e.g., main effects, interactions, in a statistical model), with minimal factor aliasing, and efficiency in covering the design space (that is, the operational envelope). Design excellence consists of having enough test points placed in the right locations in the operational envelope to answer the questions of interest for the test. When proven strategies of point-placement are employed (experimental designs), statistical power is a needed and superb tool to "right-size" the number of test points needed. Neither placement nor number of points can be neglected. I encourage the use of power to assess test design but also advocate other quantitative measures (discussed in the attached document) where appropriate.

As always, I am open to further discussion and suggestions for improving the rigor and adequacy of testing in the Department. To that end, the attached document provides an extended discussion on different statistical methods for assessing test adequacy and comparing between designs. The type of method that is appropriate depends on the goal of the test and the experimental design methodology used. There is no one-size-fits-all solution; rather there is a collection of useful tools that apply in various combinations for different test goals and designs. I expect my staff, and encourage the test community at large, to use a variety of tools tailored to

[1] See for example Russ Lenth's "Guidelines for Estimating Sample Size", 2001 for effective strategies to unearth reliable estimates of delta and sigma.

assess the statistical adequacy of a test design; nevertheless, power/confidence will continue to be chief among that toolset.

J. Michael Gilmore
Director

Attachment:
As stated

Best Practices and Statistical Measures of Merit for Assessing Test Designs

Test Goal

A clear test goal is essential in determining which statistical measures of merit are useful in assessing the adequacy of the test design. Table 1 summarizes common general classes of test goals, the associated test phase when that goal is most likely, and potentially useful experimental designs for achieving the specified goal. Table 1 is far from comprehensive or prescriptive; rather, it provides a general framework.

Table 1: General Classes of Test Goals

Test Objective[1]	Likely Applicable Test Phase	Potentially useful Experimental Designs[2]
Characterize performance across an operational envelope Determine whether a system meets requirements across a variety of operational conditions	DT and OT	Response surface designs, optimal designs, factorial designs, fractional factorial designs
Compare two or more systems across a variety of conditions	DT and OT	Factorial or fractional factorial designs, matched pairs optimal designs
Screen for important factors driving performance	CT and DT	Factorial or fractional factorial designs
Test for problem cases that degrade system performance	Primarily DT, OT for Business Systems	Combinatorial designs, Orthogonal Arrays, Space filling designs

---

1 The *NIST Engineering Statistics Handbook* discusses several goals of tests including prediction, characterization, and optimization. http://www.itl.nist.gov/div898/handbook/

2 Douglas Montgomery's text, *Design and Analysis of Experiments,* describes simple comparative experiments, factorial and fractional factorial designs, response surface designs and robust parameter design.

Myers and Montgomery's text, *Response Surface Methodology,* provides a description of factorial and fractional factorial designs, response surface designs, optimal designs, and robust parameter design.

Phadke's text, *Quality Engineering Using Robust Design,* provides an overview of Taguchi designs (Robust Parameter Designs) and Orthogonal Arrays.

Meeker and Escobar's text, *Statistical Methods for Reliability Data,* provides an overview of accelerated life tests.

Santner's text, *The Design and Analysis of Computer Experiments* provides an overview of space filling designs.

NIST provides a general overview of combinatorial testing and many useful resources: http://csrc.nist.gov/groups/SNS/acts/index.html.

Optimize system performance with respect to a set of conditions	CT and early DT	Response surface designs, optimal designs
Predict performance, reliability, or material properties at use conditions	CT and early DT	Response Surface Designs, Optimal Designs, Accelerated life tests
Improve system reliability or performance by determining robust system configurations	CT and early DT	Response surface designs, Taguchi designs (Robust Parameter Designs), Orthogonal Arrays

Notes: DT = Developmental Test; OT = Operational Test; CT = Contractor Test.

Characterize

A common goal in testing, and arguably the most important and commonly used goal for operational testing, is to characterize performance across a variety of operational conditions. It is important to note that if we are able to characterize performance with sufficient precision across a variety of conditions, then we are also able to determine whether the system meets a specified requirement at a similar level of precision across those same conditions. Multiple classes of test designs may be useful when characterization is the primary test goal including factorial and fractional-factorial designs, response surface designs, and optimal test designs. The appropriate test design will depend on the complexity of the operational envelope and expected performance variation across the operational envelope. Some conditions (levels of the factors) might be difficult to obtain, making some test designs more suitable (e.g., optimal over factorial). However, it should be noted that in most cases Taguchi designs, factor covering arrays, and combinatorial test designs are inappropriate for characterization because they provide low power for detecting differences in performance across the operational envelope. Power is an extremely important measure when the goal of the test is characterization. Other important measures are discussed below.

Compare

Direct comparison between two or more systems is a common test goal. A variety of test designs are useful in comparing among multiple systems. The best comparisons can be made using a matched design where the systems (or processes) being comparing are subjected to the same tests across all conditions. This approach controls for unwanted variability in the comparison; however, other types of test designs for comparisons exist. Power for detecting performance differences among systems is an extremely important metric for this test goal. Low power tests will mostly likely result in an inability to draw conclusions about differences in performance across systems after the testing is completed.

Screen

Screening is an important test goal prior to Initial Operational Test and Evaluation (IOT&E). As I highlighted in my initiatives, an important part of integrated testing is to identify the key factors early that affect system performance. Identifying these key factors and screening out unimportant factors is essential to constructing the initial operational test. Factorial designs and fractional factorial designs are extremely useful design tools in screening for the most important factors. When the number of factors and levels under consideration is extremely large, optimal designs and orthogonal arrays can also be useful.

Problem Cases

Testing for problem cases is typically unique to cases where the outcome of the test is deterministic – these experimental designs are common for software testing. Here we are interested in finding what combinations of factors and conditions result in problems (or failures). Combinatorial tests based on orthogonal arrays provide an efficient methodology for covering use cases such that faults caused by certain combinations of factors (two-way, three-way, etc.) can be quickly detected. Because the system's performance and test outcome is not stochastic in nature, statistical power is not a meaningful measure of merit of these test designs. Rather, the strength of the design is defined in terms of covering as many of the various combinations of input conditions (two-way, three-way, etc.) that cause faults.

Testing for problems and the corresponding test designs are not necessarily limited to determinist outcomes. For example, one may develop a test plan consisting of only the most stressing cases to search for problems. This is a risky approach to operational testing. While the test may find problems, it will not be able to characterize performance of the system due to confounding between the test factors. This confounding will limit our ability to determine the causes of problems or draw conclusions about performance in conditions other than those exact cases tested. Because of these limitations, I do not recommend these test designs for the operational testing of systems we know have stochastic response variables.

Optimize

Process optimization is not a common test goal of IOT&E. However, it is extremely useful in system design and manufacturing. Additionally, it can be useful in the development of tactics, techniques, and procedures (TTPs).

Predict

Interpolation and extrapolation comprise the two general classes of prediction. In OT&E we often wish to predict performance in areas within the operational envelope. In these cases an experimental design that provides flexibility in defining the statistical model is useful in producing predictions with reasonable precision. These designs include response surface designs

and optimal designs. The other class of prediction is based on extrapolation. These cases are typically riskier, and the validity of the model must be carefully scrutinized since estimates of system performance will be made outside the range where data were collected. Predicting outside of conditions tested into unproven areas of the operational envelope is an example of this type of prediction. Accelerated life tests for predicting reliability at use conditions are another type of extrapolation. In general, models used in operational testing must be rigorously validated and accredited using live test data collected across the full range of operational conditions; Design of Experiments (DOE) does not provide a "magical" means to extrapolate data and predict performance in untested regions of the operational envelope.

Improve

Improve (unlike optimize) refers to tests that are specifically designed to make processes or systems robust to uncontrollable conditions. These types of experiments are used in designing systems to ensure robust performance across all operating conditions. Additionally, these designs are useful in design for reliability efforts. In these types of experiments the tester controls both controllable factors (that is factors that can be controlled in the manufacturing process) and uncontrollable factors (often referred to as noise factors, e.g. humidity, operating conditions, etc.). The goal of the test is to determine the settings of the controllable factors that result in robust performance across all levels of the noise factors. This test goal is definitely an important one, but one that typically arises during the manufacturing process, and not in OT where we want to characterize system performance across all conditions. Taguchi designs (Robust Parameter Designs) were originally developed to address the "Improve" test objective. In the Taguchi thinking, interactions and statistical power are not important because the goal of the test is only to find the most robust setting of the controllable factors. However, the research community has identified many improvements over traditional Taguchi designs based on orthogonal arrays over the years.[3] My office does not currently see the benefit of applying these test designs to operational testing; however, test designs constructed using these techniques are not necessarily disallowed. My expectation is that if a Taguchi-based orthogonal array design (or similar) is used that the statistical measures of merit needed to assess any other test design must be provided.

[3] For more information on the academic debate over Taguchi's robust parameter design see, *Taguchi's Parameter Desgin: A Panel Discussion,* and Pignatiello and Ramberg's article, *Top Ten Triumphs and Tragedies of Genichi Taguchi.*

Statistical Measures of Merit

In 2009, I highlighted in my initiatives the importance of assessing statistical confidence, power, and some measure of how well the test spans the operational envelope. Since that time there has been a large emphasis on statistical confidence and power, which are essential for assessing the statistical adequacy of any test plan. They inform us of the risks of making an incorrect decision for a proposed test design. However, it is particularly important to note that no single statistical measure of merit, or group of measures, can completely characterize the quality of a test design. In my assessment of test designs, I start with careful scrutiny of the choice of response variables, factors, levels, and the choice of statistical model which will best ensure the testing will result in an adequate characterization of system performance. These critical pieces to a test design are inherently difficult to quantify, making engineering and operational expertise essential constructing a test. Notwithstanding the recent emphasis on statistical power, the above elements remain paramount for assessing test rigor and adequacy.

Assuming the proper choice of the important factors and responses, statistical measures of merit provide a quantitative means to evaluate the quality of an experiment, and/or for comparing different experimental designs. They can be used to characterize the quality of the prospective test length (sample size) and design choice, by considering the implications on knowledge gain (precision), cost, and risk.

In addition to power and confidence, there are other measures and tools available that provide valuable insights when assessing the statistical rigor of a test design. Table 2 provides a summary of the most commonly used statistical measures of merit that should be considered when planning an experiment. The type of method that is appropriate is dependent on the goal of the test and the experimental design methodology used. Again, there is no one size fits all solution.

Table 2: Statistical Measures of Merit

Statistical Measure of Merit	**Experimental Design Utility**	**Usage**
Statistical Model Supported (Model Resolution/Strength)	Describes the flexibility of the empirical modeling that is possible with the test design	Match to the design goal, and expected physical response of the system. (Second order is normally adequate for characterization.)
Confidence	The true negative rate (versus the corresponding risk, which is the false positive rate). Quantifies the likelihood in concluding a factor has no effect on the response variable when it really has no affect.	Maximize
Power	The true positive rate (versus the corresponding risk, which is the false	Maximize

	negative rate). Quantifies the likelihood in concluding a factor has an effect on the response variable when it really does.	
Correlation Coefficients	Describes degree of linear relationship between individual factors.	Minimize correlation between factors
Variance Inflation Factor	A one number summary describing the degree of collinearity with other factors in the model (provides less detail then the individual correlation coefficients).	1.0 is ideal, aim for less than 5.0
Scaled Prediction Variance	Gives the variance (i.e., precision) of the model prediction at a specified location in the design space (operational envelope).	Balance over regions of interest
Fraction of Design Space	Summarizes the scaled prediction variance across the entire design space (operational envelope).	Keep close to constant (horizontal line) for a large fraction of the design space
Optimality Criteria	Provides rank ordering of designs based on individual optimality criteria	Useful for comparing between optimal designs

Statistical Model Supported[4]

The statistical model supported by the test design is a primary consideration in determining test adequacy that is often overlooked. However, the statistical model is very important as it provides a snapshot of the knowledge gained about the behavior of the response across the operational envelope. The following types of statistical models are useful in thinking about test adequacy:

- First order models, allow for the estimation of main effects only (shifts in the mean for categorical factors or linear relationships for continuous factors).
- Second order models, allow for the estimation of main effects, two-way interaction effects, and quadratic terms for continuous factors.
- Third order models, allow for the estimate of all second order model terms plus three-way interaction effects, and cubic terms for continuous factors.

Model complexity can extend to any order. Additionally partial order models are possible; one example of a reduced second order model is a model that contains main effects and two-way interactions, but not quadratic terms. Larger-order models result in more flexible modeling; a flexible model allows for a closer fit to the observed data. However, in operational testing, when the goal is to characterize performance, second order models tend to be adequate for describing major changes in performance across the operational envelope due to the principle of *sparsity of effects* which notes that most systems are dominated by a few main effects and low-order interaction effects. On the other hand, if the test goal is to screen for important factors

[4] Myers and Montgomery's text, *Response Surface Methodology,* provides a more detailed discussion of model order, the principle of sparsity of effects, and design resolution.

a lower-order model may be appropriate. For prediction of a complex response surface, higher order models may be necessary.

For two-level full and fractional factorial experiments, the order of the statistical model is often discussed in terms of their design "resolution." A design with greater resolution can accommodate higher order model terms than a design with lower resolution. The lower the resolution, the more terms in the model are confounded with other terms, making the cause of observed performance differences amongst the different test conditions difficult to resolve. Resolution III, IV and V designs are particularly important because they address second order models, which are commonly used in operational testing. Definitions of these designs are shown below:

1. **Resolution III designs**. Main effects may be indistinguishable from some two-factor interactions
2. **Resolution IV designs**. All main effects can be estimated independently but some two-factor interactions may be indistinguishable from other two-factor interactions.
3. **Resolution V designs**. All main effects and two-factor interactions can be estimated independently from each other.

Confidence and Power[5]

Statistical confidence and power are two extremely important measures that inform us of the risks of making an incorrect decision based on test results. Confidence and power are only meaningful quantities in the context of specific hypothesis tests. In DOE we are interested in multiple hypothesis tests, one for each model term considered. One minus the confidence tells us about the level of risk of false positives (that is concluding a factor significantly affects performance when it truly does not) that we are willing to accept in both test planning and analysis. This risk is often referred to as a Type I error. Power tells us about the probability of detecting significant test outcomes. Power is a function of the statistical confidence level, the effect size of interest, the variability in the outcomes, and the number of tests. It not only describes the risk (1–Power) in concluding a factor does not have an effect on the response variable when it really does (Type II error), but also is directly related to the precision we will have in reporting results. The latter is key in my determination of test adequacy; without a measure of the expected precision we expect to obtain in the analysis of test data, we have no way of determining if the test will accurately characterize system performance across the operational envelope. A test that has low power might fail to detect true system flaws; if it does, we have failed in our duty as testers. Estimating statistical power is not the uncertain and unreliable task suggested by some; instead it is a mature and proven discipline backed by more

[5] References on statistical confidence and power are numerous; see for example Vining and Kowalski, *Statistical Methods for Engineers,* for a practical explanation of statistical confidence and power.

than 50 years of practice in all fields of science, medicine, engineering, and the social sciences. Additionally, in a limited resource test environment, power analyses allow decision makers to see the tradeoff between risk and test resources.

While avoiding both types of errors is ideal, striking the proper balance between the two risks can be crucial. Adjusting the confidence levels in a test changes the distribution of risk between Type I and Type II errors; increasing the confidence level will decrease the false positive rate (Type I error), but also increase the false negative rate (Type II error).

Collinearity

When designing an experiment, collinearity describes the degree of linear relationship between two or more factors. A well designed experiment minimizes the amount of collinearity between factors. Two or more factors are consider collinear if they move together linearly (as one increases, so does the other). For example, if a test plan only looks at large targets at long ranges and small targets at short ranges, then the target size and the range are perfectly collinear; this effect is often also referred to as complete confounding.

Analysis of data containing highly collinear factors can be misleading, confusing, and imprecise. Variances of coefficient estimates become greatly inflated (making the precision of the test worse) when factors are highly collinear, which leads to model terms (effect of a factor on performance) being deemed non-significant when in fact they were significant (Type II errors). Additionally, using a model containing highly collinear factors to extrapolate or interpolate between design points can yield estimates with large uncertainty.

Two common statistical measures of merit that can be used to help detect collinearity in a test design are correlation coefficients and variance inflation factors (VIFs).[6] Both of these measures are used for planning a DOE and are calculated and monitored prior to executing an experiment. They are functions of the number of runs, the factors and levels in an experiment, and how those factors vary from run to run. They are not a function of the data collected from the test.

Pearson's correlation coefficient is the most commonly used correlation coefficient between two variables and is defined as the covariance of the two variables divided by the product of their standard deviations.[7] Magnitudes near one indicate a strong linear association between the two variables while values near zero indicate little or no linear association. Since most operational tests have more than two factors, it is useful to construct a matrix of correlation

---

[6] Myers and Montgomery's text, *Response Surface Methodology,* provides a more detailed discussion of variance inflation factors. Additionally, the Wikipedia page on variance inflation factors is well written: http://en.wikipedia.org/wiki/Variance_inflation_factor.

[7] Pearson's correlation coefficient is widely discussed in a variety of statistical references. The Wikipedia description is extensive: https://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient.

coefficients to assess the correlation between factor pairs. Statistical software packages make these calculations accessible to all practitioners.

VIFs provide a one-number summary description of collinearity for each model term. Given an experiment with multiple factors, the variance inflation factor associated with the ith factor reflects the increase in the variance of the estimated coefficient for that factor over the variance that would have been obtained if the factors were orthogonal. VIF_i can range from one to infinity. Values equal to one imply orthogonality, while values greater than one indicate a degree of collinearity between the ith factor and one or more factors. The square root of the variance inflation factor indicates how much larger the standard error is (and therefore, how much larger the confidence intervals will be), compared to a factor that is uncorrelated with the other factors. As a rule of thumb, values greater than 5 suggest that collinearity may be unduly influencing coefficient estimates.

Scaled Prediction Variance (SPV)

One reason we conduct tests is to predict future performance of a system within the operational envelope. *Prediction variance* describes the precision involved with making a prediction using an empirical model.[8] Prediction variance is a function of the sample size and correlation in the experimental design, the location in the design space where the prediction is made, and the overall variance in the response. Since the overall variance of the response variable is often not well-known prior to collecting the data, SPV can be used to evaluate test designs; SPV is a measure of the relative prediction variance to a nominal overall variance.

The benefit of SPV is that it can be used to evaluate a designed experiment prior to running the test and collecting data. Multiple designed experiments can be postulated for a single test event and compared using SPV and the best design can be selected. When assessing a design in this way, it is important to consider the full range of values each factor can take. For categorical factors, this is just a matter of considering prediction at each level of the relevant factors. For continuous variables, graphical methods such as contour plots are available.

SPV is relatively straightforward to use in two dimensions (two continuous factors) and for categorical factors. However, in more complex cases simple plots called *fraction of design space* (FDS) plots are useful for investigating prediction variance.[9] The FDS plots the cumulative distribution of the SPV for a given design space. An FDS plot shows the proportion of the design space with SPV less than or equal to a given value. Figure 1 shows the FDS comparing two designs. This chart shows that nearly 80 percent of the Design A space has an SPV below 4.0, while roughly 55 percent of the Design B region has an SPV below 4.0. From

---

[8] Myers and Montgomery's text, *Response Surface Methodology,* provides a more detailed discussion of scaled prediction variance.

[9] Both Design Expert and JMP provide detailed descriptions of the FDS Plot: http://www.statease.com/news/news0809.pdf and http://www.jmp.com/software/pdf/103044_doe.pdf.

this chart it is clear that Design A is a better design for prediction across the operational envelope.

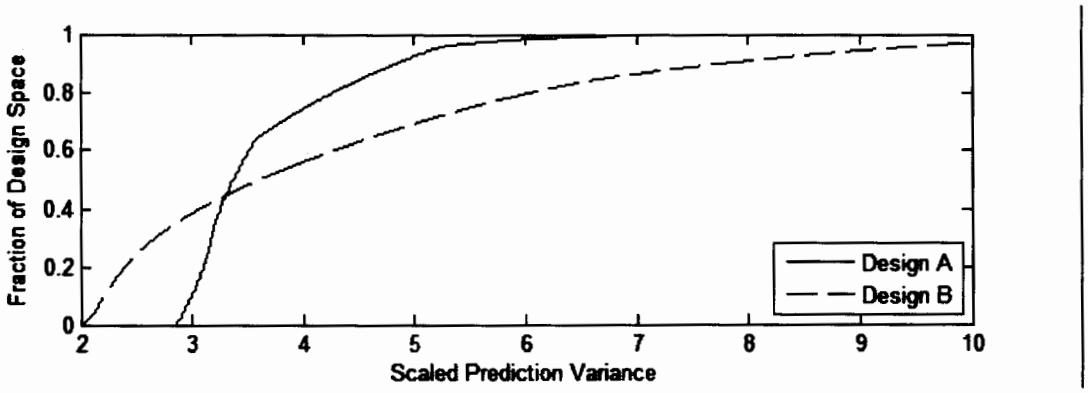

Figure 1: FDS Graph for Candidate Experimental Designs

Optimal Design Criteria

Optimal designs are constructed based on specific mathematical criteria, picking a collection of test points in the operational envelope based on a calculated "score" called an optimality criterion. Therefore, multiple test designs can be compared using their optimality criteria.[10] Several methods exist for optimizing the test point coverage in optimal designs; these include, but are not limited to, D-, I-, and G-optimal criteria. Each method spreads the test points throughout the design space depending on a mathematical formula, and each has different benefits that are tied to the test goal. For example, D-optimal designs spread the points out in such a way as to minimize the overall variance of the parameter estimates while also not letting the covariance between the parameter estimates get too large. I-optimal designs are more focused on prediction, and achieve this by "spreading out" the test point within the design space evenly. I-optimality is directly related to SPV, as it is proportional to the integral of the SPV curve shown in Figure 1. G-optimal designs minimize the *maximum* prediction error over the design space rather than the average prediction error. While many other design criteria have been proposed in the literature, D- and I-optimal designs are perhaps the most popular. This is due in part to their ubiquity in statistical software. While any of these optimality criteria are useful for comparing designs, they should always be used in combination with statistical power, and an assessment of the factor correlation to provide a robust assessment of the test designs adequacy.

[10] Myers and Montgomery's text, *Response Surface Methodology,* provides a detailed discussion optimal designs and optimality criteria.

Summary

In addition to confidence and power there are a variety of tools that are available for assessing the statistical adequacy of a test design. The list provided here is far from comprehensive; however, it does capture the most commonly available tools in statistical design software. The type of tool that is appropriate is dependent on the goal of the test and the experimental design methodology used. There is no one size fits all solution, but rather a collection of useful tools that apply in various combinations to different test goals and designs. In cases where statistical confidence and power do not provide the full picture of the adequacy of the test design these measures and metrics provide amplifying information in assessing test adequacy. Notwithstanding these measures of merit, the choice of response variables, factors, levels, and the choice of statistical model which will best ensure the testing will result in an adequate characterization of system performance, remain paramount.