Quantifying Test Risk Using Design of Experiments

Authored by:
Michael Harman, STAT COE
Luis Cortes, STAT COE
Raymond Hill, Ph.D., AFIT/ENS

16 April 2015

Revised 30 September 2018

The goal of the STAT 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 Center of Excellence
2950 Hobson Way – Wright-Patterson AFB, OH 45433

Table of Contents

Revision 1, 30 Sep 2018: Added reference to the DOT&E TEMP Guidebook and slightly edited associated content; otherwise formatting and minor edits.

Executive Summary

The intended outcome of test is to provide information to decision makers. When test resources are constrained, the test planner must be ready to identify and quantify the risk of missing intended outcomes due to executing fewer test points than desired or planned. Department of Defense leadership has advocated for a re-invigoration of technical rigor in test design and this can be leveraged to address risk under these circumstances. Design of experiments methodology supports efficient test designs and can be used to quantify risk associated with test design changes. Assessing design metrics such as model size, signal to noise ratio, power, confidence, and their combined effects due to these design changes facilitates a quantifiable assessment of risk to the test program and helps identify alternate test designs that may be valuable and executable even with reduced resources. Quantifying these changes and determining the quality of data that can be collected is a critical task for the test designer.

Keywords: Test risk, design of experiments, decision-quality information, scientific test and analysis techniques

Introduction

During a T&E leadership panel discussion at a working group, a question was posed about how decisions should be made regarding risk in testing. The first panel member to respond commented that in order to make a decision regarding risk one must first understand the risk and its impact. We thought this was a solid start to a good answer. The panel member talked for a few minutes more but never got more quantitative in his response. As the microphone moved down the table none of the other panel members said anything more quantitative. We were disappointed to see this missed opportunity because, in our minds, the design of experiments methodology provides a solid platform from which to quantify test designs and address changes to the designs when and if they occur.

Department of Defense (DOD) leadership has increasingly emphasized the necessity of technical rigor to counter funding and resource constraints. In 2010, the Director, Operational Test and Evaluation (DOT&E) issued guidance on the use of design of experiments (DOE) (Gilmore, 2010) to better quantify and characterize system performance. Since its 2015 release, DoD 5000.02 calls for program managers to “use scientific test and analysis techniques (STAT) to design an effective and efficient test program” (OSD 2015). In 2015, Deputy Assistant Secretary of Defense for Developmental Test and Evaluation (DASD(DT&E)) introduced guidance in the Defense Acquisition Guidebook to better map requirements to acquisition decisions using a new developmental evaluation framework concept. And as a final example of this emphasis, the DOT&E TEMP Guidebook has a section dedicated to STAT guidance, stating “The authors of the TEMP should employ scientific test and analysis techniques to develop a defensible analytical basis for the size and scope of the T&E program.”

Overall, these initiatives serve to forcefully inject rigor into DOD testing. But what happens when these statistical methods have been employed (resulting in an efficient design type and size) only to be faced with additional resource cuts after the test program has begun?

Some typical situations where external forces affect test planning are listed below.

- Limited resources due to real world events
- New test priorities that force resources away from certain events
- Budget cuts late in the acquisition program which can only be accommodated through decreased testing

Regardless of the source of the impact, T&E leadership must understand the answer to the question ***"what is the test design risk if we reduce the number of events?"*** The corollary is "***what should the community do (in terms of planning and documentation) when decision authorities choose to reduce testing?"*** The answers to those questions are broad and specific to each program. However, one critical aspect cannot be overlooked; in order to accept risk, one must first quantify it. We will use a hypothetical, but realistic, case study to outline a process that addresses test design risk from a STAT perspective in a methodical and rigorous way to illuminate options available to leadership. Background on the DOE principles discussed herein can be found in Montgomery (2013) and in Hill et al (2014).

Background

Scientific Test and Analysis Techniques (STAT) are mathematical and statistical tools and processes used to enable the development of efficient and scientifically rigorous testing throughout the acquisition life cycle. The design and analysis of experiments, or design of experiments (DOE) for short, is the cornerstone statistical methodology in the STAT portfolio. DOE is the integration of well-defined, rigorous, and structured strategies for gathering empirical knowledge about systems and processes using statistical methods for planning, designing, executing, and analyzing a test or a sequence of tests. Montgomery (2011, 2017) provides guidelines for DOE (see **Figure 1**).

Figure 2 illustrates the place of DOE in the acquisition process. Capability requirements are decomposed into technical and operational requirements, which are expressed in the form of measurable and testable technical performance measures (TPM) that make-up the backbone of the T&E strategy—the Developmental and Operational Evaluation Frameworks. ***Some of these TPMs can become the "response variables" in step 2,*** **Figure 1.**

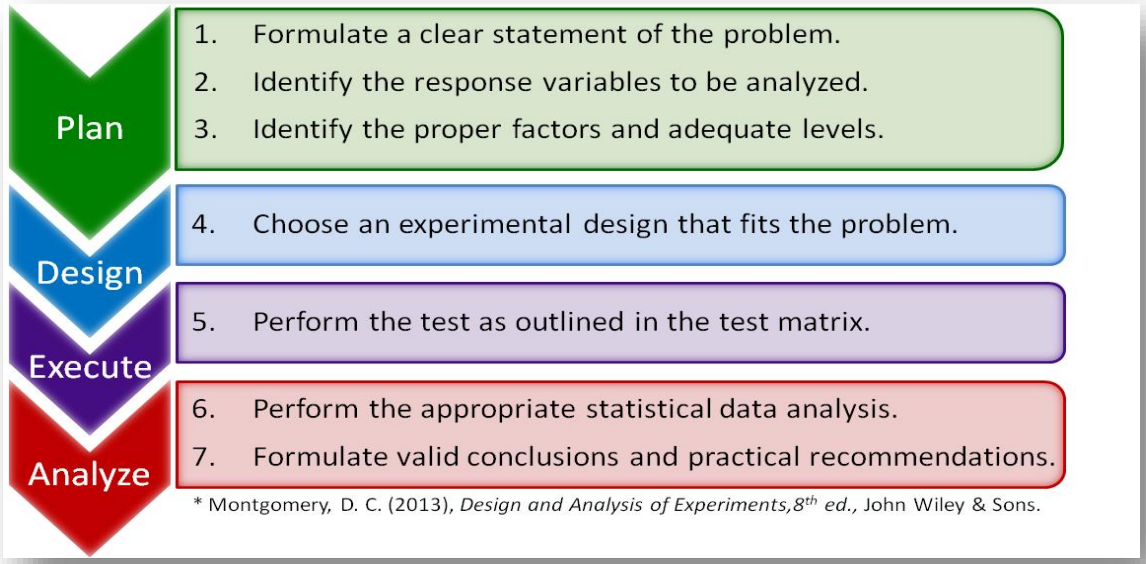


Figure 1: Guidelines for the Design and Analysis of Experiments

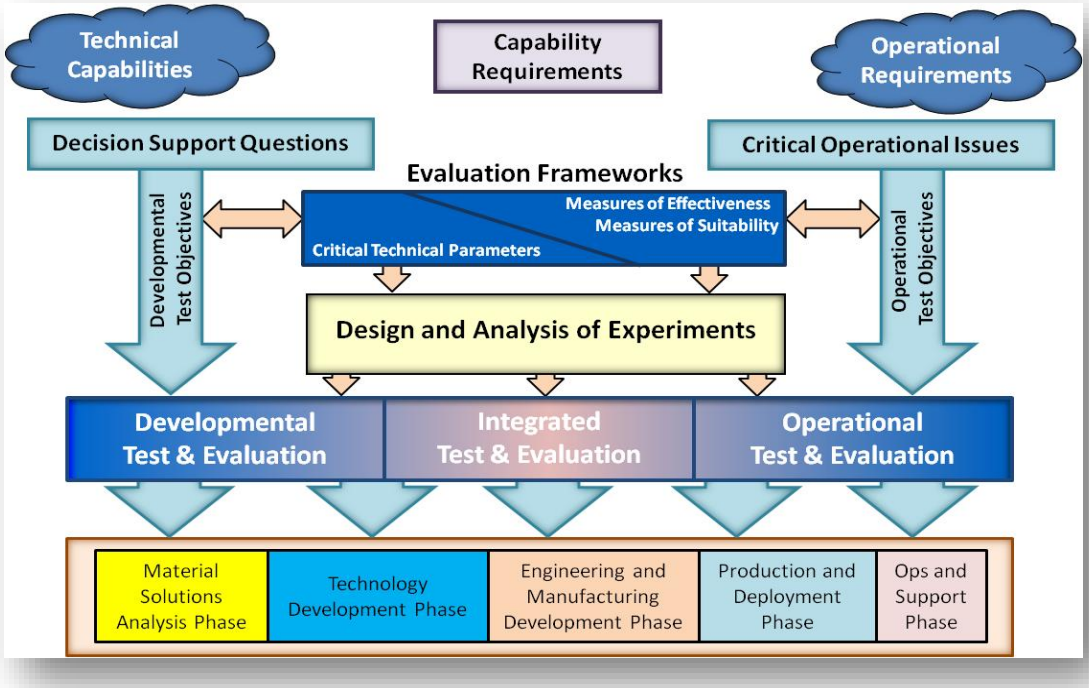


Figure 2: Design of Experiments and the Acquisition System

A well-designed test should produce the maximum amount of information with the minimum amount of resources to inform its objective(s). Often, the ultimate objective of a test is to inform the assessment of the system's performance, interoperability, reliability, and cyber security. DOE can help generate test efficiencies, improve the fidelity of test results, illuminate risks, enable better-informed decisions, and ultimately, enable fielding a more effective, suitable, and survivable system.

DOE adds rigor and discipline to T&E and facilitates a comprehensive understanding of the tradeoffs in the techno-programmatic domains: risks, cost, and the quality and utility of the information to be collected. **Figure 3** illustrates the tradeoff space, which involves balancing the amount of information to be obtained from the test, the funding available, and the risks associated with detecting the influence of test factors or conditions that drive system performance.

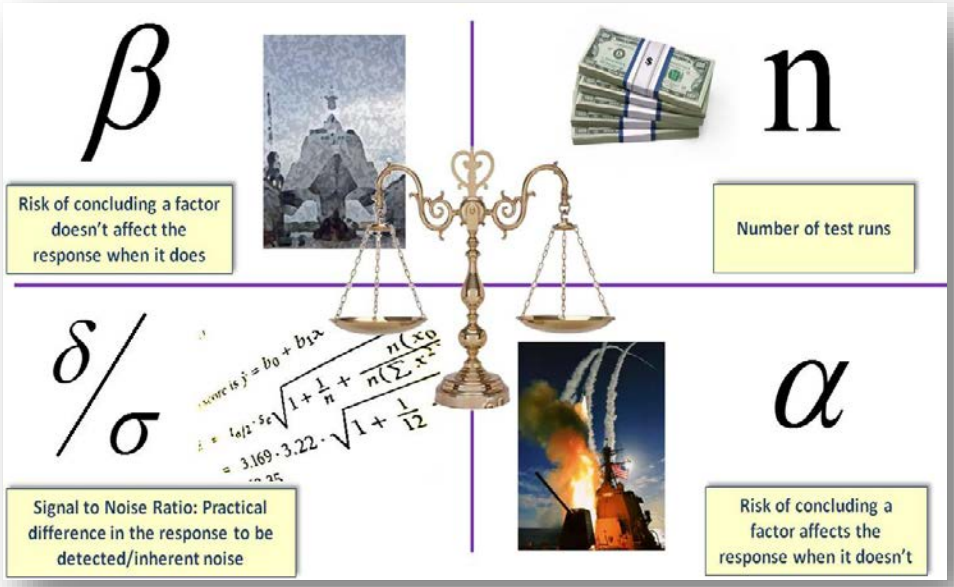


Figure 3: Trade Space in Designing Experiments

The DoD Test Management Guide (DOD 2012) lists some of the benefits gained by the testers when using DOE:

- Understanding the likelihood of successfully or mistakenly identifying performance drivers
- A sound method for determining the number of tests needed to obtain required information
- The ability to make informed trade-offs of test costs versus the quality of information gained
- A rigorous method of determining the conditions that provide the most useful data
- The ability to identify interactions between test factors

Case Study

Description

Consider the hypothetical situation where a test has been designed but not executed; all points can be considered for re-design. We will not address changes to a test already underway. For the purposes of this example we also assume no difference in cost from test point to test point. This is not always true, but it keeps this example from becoming needlessly complicated in the cost realm. The objective of this design is to screen the ME (main effects) and 2FI (two-factor interactions) that have an effect on the response(s) of interest.

The design for this case is a 7 factor, thirty-six (36) run screening optimal design with the following statistical characteristics:

- Four 2-level continuous factors; three 2-level categorical factors
- All (7) main effects (ME) and all (21) two-factor interactions (2FI) are modeled
- 5% α (confidence = 100% - α; 95% in this case)
- Signal to Noise Ratio, $SNR \equiv \delta/\sigma = 2/1$
- Power > 99% for all terms (Power = 100% - β)
- Average prediction variance = 0.43
- 36 runs with no center points (corners only)
- 10% more runs (4) were added for expected inefficiency/retest; 40 points total

For a dramatic effect, assume a cut of 50% (20 points) is leveraged against this design and program leadership has asked the T&E team to quantify the impact of the cut. Specifically, what is our ability to effectively execute this test, to identify the ME and 2FI that affect the response(s), to have the basis for a meaningful evaluation, and to provide adequate information to decision makers? The two questions that immediately come to mind are impact on the program and impact on the test design. These questions can be assessed using test design metrics and performing a thorough trade-off analysis.

Trade-offs

Modify the Test Objective

Risk: Test delays can be introduced
Impact: Test strategy can be affected

When the number of runs is capped, the resulting design may not be adequate to address the original goals. In this case, the goal was to evaluate all ME and all 2FIs. If it is critical to evaluate all ME and 2FIs and the resources are not available, fundamental changes to the plan must be considered. For instance, the test team may decide to evaluate only the ME for this test while negotiating for additional test resources for later use in sequentially expanding the design to consider the 2FI.

Reallocate Resources from Other Tests

Risk: This test design may be more rigorous than others
Impact: The return on investment of this design may be higher than that of the other tests

Design of experiments is a rigorous technique for gathering knowledge to inform the acquisition system. The selection of a test design, which often results from the collaboration between system engineers, testers, subject matter experts, statisticians, and the decision makers themselves, establishes the number of test points. The process involves a-priori knowledge of the risks the decision makers are initially willing to take plus a careful analysis of the requirement and consideration of the factors that affect system performance. This needs to be communicated to leadership. In this case, a 50% reduction in test points may represent a compromise in the decision-quality information produced—i.e., the ability to test the significance of all ME and 2FI with sufficient statistical power. The community must consider finding savings somewhere else rather than in cutting runs here.

Select an Original Design

Risk: Sometimes it is hard to recover from inadequate designs
Impact: The quality of the information can be affected

A critical step in the DOE process (step 4, **Figure 1**) is to select an experimental design that fits the objective of the tests. There are several types of experimental designs, such as factorial designs, optimal designs, response surface designs, space filling designs, etc. Each type of design has its own advantages, limitations, and statistical properties. In this case, the test team selected an optimal design with a full understanding of the design limitations and full concurrence from the program.

Change the Test Design Type

Risk: Other approaches may result in a less effective design
Impact: The new test design may not address the test objectives

There are a variety of test designs that may fit test objectives. However, the character of some of the factors that drive system performance, the structure of the test design, or the conditions in which the test is being conducted may afford an opportunity to conduct the test in a different way and to select a different test design. For example, a full factorial could be reduced to a fractional factorial (which reduces the potential number of runs by at least 50%). Likewise, a completely randomized design can be replaced by a split-plot design if the desired precision to evaluate the effects of at least one factor is less than the precision required to evaluate the other factors. Each of these alternatives can help reduce the overall test time and potentially reduce the cost of a test. However, each design carries their own challenges, which are very well understood by an experienced statistical test designer.

Reduce the Number of Factors

Risk: Existing factor effects may be missed (a type II error)
Impact: All true underlying effects that impact system performance may not be detected/modeled.

Planning a test using design of experiments is a rigorous process that, when performed correctly, can result in a design that provides a significant amount of information. Planning involves multiple sessions with systems engineers and subject matter experts that culminate with the identification of factors and the corresponding levels that should influence the performance of a system. If the number of runs is reduced, the original factors can be ranked by expected significance. Then, the factor expected to be the least significant can be measured but not explicitly controlled in the test. This removal of a design factor reduces the dimension of the design and can therefore accommodate this significant test size reduction.

Reduce the Number of Model Terms

Risk: Fewer terms can be identified in the model
Impact: The model does not provide an adequate representation of the system

At least one design point is needed to estimate an effect. Since our notional case has 9ME + 21 2FI = 30 effects, at least 30 points are required to estimate these effects. Originally, the notional plan was to model all ME and 2FI with 36 points (32 unique points and 4 replicated points). Twenty points are clearly not sufficient to estimate the planned ME and 2FI. To create a model with only 20 points, nine of the 21 2FI must be removed from the model. Eliminating 2FI effects should not be done arbitrarily. A thorough design team discussion should determine which 2FI are believed to be less important. When this is accomplished a design assessment reveals power has fallen to 24% across the board. Now the other metrics must be considered. Once again, a reduction in test points in this case results in a reduction in information on the 2FIs that can affect the system.

Improve the Signal-to-Noise Ratio (SNR)

Risk: The smaller design may only be sensitive to large changes in the response
Impact: It may be more difficult to effectively model the response as a function of the factors

The original SNR was set at 2/1. This means the magnitude of the minimal operational impact in the response (δ) was twice the inherent noise (σ) in the response. Was SNR estimated correctly the first time? A higher SNR will result in more power; however, SNR cannot be set arbitrarily in order to raise the calculated power. If a more precise analysis can justifiably determine that the expected noise is lower (σ is smaller), SNR rises for the same δ value. Similarly, was the assessment of δ sufficient to indicate that the current value is truly the minimum operationally realistic impact the user care about? If δ increases then SNR rises for the same value of σ. If both are better scoped, then SNR can be updated and the design reassessed. In this case, an SNR at 8/1 will raise power back towards 80%; however, this might be unobtainable and/or unrealistic for the system. In summary, reducing system noise, σ, improves the design's signal-to-noise ratio, so if the test must be smaller these parameters could be revisited.

Change Alpha (α)

Risk: Significant factors may not be identified correctly
Impact: The response may be modeled insufficiently or the results may be ambiguous or inconclusive

Increasing α will not change the number of runs required, but it will inflate statistical power by increasing the amount of Type I risk. The original design used α = 5% (95% confidence) to size the design. Confidence is set beforehand and α indicates the proportion of times the design will incorrectly identify a factor as significant to the response. Increasing α (decreasing confidence) will increase the false alarms on significant factors (a Type I error). If we reset SNR at 2/1 and adjust α to 0.2 (80% confidence) power climbs to about 80% for most effects. However, the analysis plan (why we are doing this test) must consider this change.

The resulting data will probably be used to estimate a performance parameter like accuracy or time to execute the mission. Will this decrease in confidence impact our ability to effectively estimate the parameter? Confidence intervals (CI) are calculated for the response at every design point and indicate the range within which the true mean will be seen in the results. Decreasing the confidence level will decrease this range but the risk that the model terms are incorrectly determined also rises. This can potentially render the results ambiguous and/or meaningless when compared to the requirement. DOE software applications vary but most provide some method by which confidence intervals are calculated. Researching CI in the original design and then in the new design should provide a quantitative assessment of how the design size and confidence level impact the analysis.

Balancing Metrics in the New Design

Model size, SNR, and confidence level all contribute to the effectiveness of the design, but the previous sections consider them independently. Modification of all of these metrics should be examined simultaneously to reach an executable design. **Figure 4** shows a summary of potential test program changes along with a minimal risk design solution (design #5). The row colors indicate relative values (green being more desirable) and allow one to quickly search for the "most green" option.

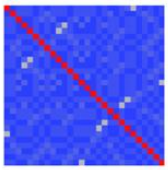
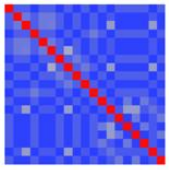
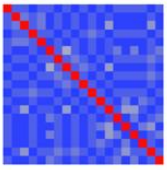
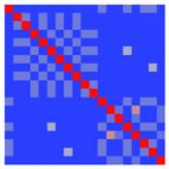
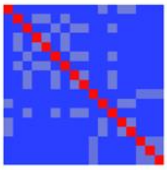
Design #	1	2	3	4	5
Recommended Design					X
Name/Design Type	Original	Reduced Points to 20	Increased SNR	Adjusted alpha	Minimal Risk Design
Factors	7	7	7	7	7
Levels	2	2	2	2	2
Model Supported	ME, 2FI	ME, 2FI (12/21)	ME, 2FI (12/21)	ME, 2FI (12/21)	ME, 2FI (10/21)
Signal to Noise Ratio	2.0	2.0	8.0	2.0	2.0
Alpha	0.05	0.05	0.05	0.2	0.1
# Center Points	0	0	0	0	0
# Repetitions	0	0	0	0	0
Total Runs	36	20	20	20	20
Lowest Power for ME	0.99	0.2	0.72	0.74	0.83
Lowest Power for 2FI	0.99	0.17	0.72	0.74	0.83
Avg variance of prediction	0.43	0.67	0.66	0.66	0.48
FDS Pred Err @95%	0.65	1.20	1.10	1.10	0.70
Aliasing					
Notes	plus 4 points for re-test	Reduced model	Increased SNR to 8 for power	Adjusted alpha to 0.2	Adjusted 2FI and alpha

Figure 4: Design Metric Comparison

It is easy to see that design #5 meets the reduced number of runs, has power above 80%, and prediction variance close to the original design. The aliasing appears to have increased (blue and red squares) but the fewer number of terms has increased the size of each block. The aliasing is on par with the original design and certainly better than designs two, three, and four.

Reporting the Risk

Reporting the risk requires addressing the following items (from a STAT/DOE perspective):

- What information is unobtainable given the resource cut?
- What is the impact on the test program?

- How is the decision-maker's risk increased if less data is collected?

As an example, consider the following responses:

- What information is unobtainable given the cut?
    - 11 of 21 interactions will not be detectable or modeled. Ten of the 21 will be detectable if they exist. We must use our system knowledge to define which ones we think are most likely to exist. Our testing will confirm our assumptions about what we think we know.
- What is the impact?
    - If any more than 10 interactions are EXPECTED TO BE significant, then their contributions will be manifest as noise in the response or will be combined and indistinguishable with other effects.
    - If fewer than 10 interactions are KNOWN to be significant then the test should detect and model them sufficiently.
- How is the decision-maker's risk increased if less data is collected?
    - Confidence was lowered from 95% to 90%. This means that the risk of claiming a factor is significant when it TRULY IS NOT changed by 5%.
    - Signal-to-noise ratio was left at 2.0, meaning that signals between factor levels must be twice as large as the inherent noise to sense a shift in the response. This may result in some factors being deemed "not significant" when in fact they are. Anything that can be done to better estimate the true system noise helps to further refine this test design and quantify its risk.

Conclusion

Testing provides information to decision makers. When testing is curtailed or decreased the ability to provide the information will decline as well. Quantifying this change and determining the quality of data that can be collected is a critical function of the test designer. Applying DOE to test design facilitates a rigorous approach to test point selection and addressing risk when those resources are impacted.

References

Department of Defense (DOD). "Test and Evaluation Management Guide." 6th ed., Dec. 2012.

Gilmore, J. Michael. "Guidance on the use of Design of Experiments (DOE) in Operational Test and Evaluation." Director Operational Test and Evaluation Memo, 19 Oct. 2010.

Director, Operational Test and Evaluation (DOT&E). "Test and Evaluation Master Plan (TEMP) Guidebook." Version 3.1, 19 Jan. 2017.
http://www.dote.osd.mil/docs/TempGuide3/TEMP_Guidebook_3.1a.pdf

Hill, R. R., Ahner, D.K., Gutman, A.J. "What Department of Defense Leaders Should Know About Statistical Rigor and Design of Experiments for Test and Evaluation." *International Test and Evaluation Journal*, vol. 35, no. 3, 2014, pp. 264-271.

Montgomery, Douglas C. "The Principles of Testing." *International Test and Evaluation Journal*, vol. 32, no. 3, 2011, pp. 231-234.

Montgomery, Douglas C. *Design and Analysis of Experiments.* 9th ed., John Wiley & Sons, Inc., 2017.

Office of the Secretary of Defense (OSD). "Operation of the Defense Acquisition System." DOD Instruction 5000.02, 7 Jan. 2015, incorporating Change 3, 10 Aug. 2017.