

REPORT DOCUMENTATION PAGE					Form Approved OMB No. 0704-0188	
The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.						
1. REPORT DATE (DD-MM-YYYY) 08/19/2020		2. REPORT TYPE Technical Report			3. DATES COVERED (From - To)	
4. TITLE AND SUBTITLE Modern Test Design and Analysis Playbook: An integration of scientific and statistical methods to catalyze efficiency and effectiveness in evaluation planning, test design, test execution, data analysis, and reporting.				5a. CONTRACT NUMBER W56KGU-18-0004-0001		
				5b. GRANT NUMBER		
				5c. PROGRAM ELEMENT NUMBER		
6. AUTHOR(S) Cortes, Luis A.; Ortiz, Francisco;				5d. PROJECT NUMBER		
				5e. TASK NUMBER		
				5f. WORK UNIT NUMBER		
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) The MITRE Corporation 7515 Colshire Drive McLean, VA 22102				8. PERFORMING ORGANIZATION REPORT NUMBER PRS-20-1492		
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) Naval Surface Warfare Center, Port Hueneme Division				10. SPONSOR/MONITOR'S ACRONYM(S) NSWC PHD		
				11. SPONSOR/MONITOR'S REPORT NUMBER(S)		
12. DISTRIBUTION/AVAILABILITY STATEMENT DISTRIBUTION STATEMENT A. Approved for public release: distribution unlimited.						
13. SUPPLEMENTARY NOTES						
14. ABSTRACT This document provides a cohesive framework and recommendations to integrate design and analysis of experiments methodologies into the fabric of capability-based test and evaluation to catalyze efficiency and effectiveness in evaluation planning, test execution, data analysis, and reporting to produce decision-quality information at the right cost.						
15. SUBJECT TERMS Decision Support (General); decision support; test design; capability-based test and evaluation; evaluation; data analysis;						
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES 101	19a. NAME OF RESPONSIBLE PERSON Susan Carpenito	
a. REPORT	b. ABSTRACT	c. THIS PAGE			19b. TELEPHONE NUMBER (Include area code) 781-271-7646	



Sponsor: NSWC PHD
Division: Naval Programs Division
Department No.: P631
Contract No.: W56KGU-18-0004-0001

The views, opinions and/or findings contained in this report are those of The MITRE Corporation and should not be construed as an official government position, policy, or decision, unless designated by other documentation.

©2020 The MITRE Corporation.
All rights reserved.

The MITRE Corporation
7515 Colshire Dr.
McLean, VA

Modern Test Design and Analysis Playbook

**An integration of scientific and statistical methods
to catalyze efficiency and effectiveness in
evaluation planning, test design, test execution,
data analysis, and reporting.**

**Luis A. Cortes, PhD, PE
Francisco Ortiz, PhD**

19 August 2020

This page intentionally left blank.

Approved By

Emaan Osman
Department Head, P631

Date

David Coleman, Ph. D.
Chief Engineer, P631.

Date

ACKNOWLEDGEMENT

The authors wish to thank the following individuals for their contributions: Dr. Michael Sheehan, Dr. Suzanne Beers, Dr. David Coleman, Dr. Jason Daly, Mr. Howard Lazoff, Ms. Elizabeth Torres, Ms. Janet Manu, Dr. Lillianne Troeger, and Mr. Timothy Mull (The MITRE Corporation); Dr. David Lechner (Northrup Grumman Corp.); Mr. Kenneth White, and Mr. John Cunnick (Naval Surface Warfare Center Port Hueneme Division).

Executive Summary

The Department of Defense makes significant investments in the acquisition and sustainment of weapon systems, defense business systems, and national security systems. Because delivering cost effective, safe, interoperable, survivable, operationally effective, operationally suitable, and lethal capabilities at the speed of relevance is imperative, every acquisition decision in the lifecycle must be well informed. *Capability-based test and evaluation* is one of the most effective and efficient means for gaining knowledge in engineering, design, and production to inform decisions about a capability. A well-designed capability based test and evaluation strategy aims at maximizing the amount of decision-quality information obtained while minimizing the amount of resources used. This document provides a cohesive framework and recommendations to integrate *design and analysis of experiments* methodologies into the fabric of capability-based test and evaluation to catalyze efficiency and effectiveness in evaluation planning, test execution, data analysis, and reporting to produce decision-quality information at the right cost.

While the concept of capability based test and evaluation has been widely accepted, its implementation by the Navy is on its infancy. The MITRE Corporation has been collaborating with Naval Surface Warfare Center Port Hueneme Division and the Navy Cross-SYSCOM Capability-Based Test and Evaluation Leadership Working Group in the development and formulation of artifacts for the successful implementation of capability-based test and evaluation across the Navy. Some of the artifacts include an acquisition crosswalk, critical path roadmap, implementation guide, and revision to the Navy test and evaluation guidebook. This document provides strategies for the full integration of design and analysis of experiments methodologies into the capability-based test and evaluation signature process—the mission-based test design.

The use of design and analysis of experiments can simultaneously help generate more efficient (faster and more economical) and more effective (more informative) test designs than other traditional methods of test design while providing a better understanding of the fulfillment of the system capabilities and a better understanding of the risks associated with the decisions to be made. Design and analysis of experiments methodologies integrate well-defined and structured scientific strategies and statistical techniques for first gathering empirical knowledge about a system and then transforming that knowledge into decision-quality-information. Design and analysis of experiments methodologies drive testers to determine exactly what questions need to be asked to provide information to decision-makers, how many resources are needed to answer those questions, and which analysis techniques are required to interpret the data to answer the questions.

Many programs have reported the benefits of using design and analysis of experiments for test. For instance, one Program Office reported a simultaneous 60% reduction in test cycle time and a ten-fold increase in number of technical parameters measured relative to the original test strategy proposed by systems engineers. Another Program Office reported a reduction of 80% in their test cycle time relative to the original test strategy. Reaping these kinds of benefits is not uncommon.

Design and analysis of experiments increases mission impact and leverages statistical engineering methods and know-how into rigorous, statistically driven strategies for planning, executing, analyzing, and reporting tests that result in:

- Well-defined mission-oriented objectives within the scope of the decision to be informed
- Testable mission-oriented response variables linked to capability evaluations

-
- Complete coverage of the test space mapped to capability attributes and a comprehensive list of factors and conditions that influence the attributes
 - Efficient and effective test designs with good statistical properties
 - Disciplined test execution protocols and well-developed run matrices
 - Inferential statistical techniques and standardized evaluation criteria to analyze the data
 - Defensible and reliable decision-quality products to inform decisions

The positive outcomes of DOE based test design are far reaching and provide substantial benefits to the test strategy. MITRE recommends Naval Surface Warfare Center Port Hueneme Division: (1) adopts and starts implementing capability-based test and evaluation; (2) adopts and starts implementing the test design framework outlined in the document; (3) incorporates the test design framework into the digital engineering transformation initiative; and (4) continues developing essential playbooks for the successful implementation of capability-based test and evaluation, such as a model-based systems engineering playbook.

Contents

1	Introduction	1
2	Document Scope and Organization	4
3	Top T&E Policy and Guidance in DoD and DoN	4
4	Modern Test Design and Analysis Framework	10
4.1	Identify Key Knowledge Points and Decisions	14
4.2	Identify Key Capability Attributes to be Evaluated	16
4.3	Develop Evaluation Objectives and Scope	20
4.4	Identify the Test Space (Factors and Levels)	23
4.5	Construct Test Designs	34
4.6	Coordinate and Execute the Test Strategy	54
4.7	Perform Exploratory Data Analysis	56
4.8	Inform Key Knowledge Points and Decisions	64
5	Summary and Recommendations	66
	References	71
	Appendix A	77
	A1 Top DoD T&E Policy	79
	A2 TOP DoN T&E Policy	84
	Appendix B	87
	B1 Screening Designs	88
	B2 Response Surface Designs	89
	B3 Small Response Surface Designs	89

List of Figures

Figure 1. Design and analysis of experiments in the lifecycle	3
Figure 2. CBT&E in capability-based acquisition.....	6
Figure 3. CBT&E approach	7
Figure 4. Mission-based test design (MBTD) process.....	8
Figure 5. T&E process flow.....	10
Figure 6. Evaluation and test design and analysis process (PEA&R + N)	12
Figure 7. Generic crosswalk evaluation framework (XEF) template	15
Figure 8. Integrated evaluation framework template	15
Figure 9. Generic functional flow block diagram.....	18
Figure 10. Generic decomposition of mission requirements	19
Figure 11. Framework for sequential evaluation and test.....	21
Figure 12. Cause-and-effect diagram for potential factors	25
Figure 13. Cause-and-effect diagram organized by factor type.....	26
Figure 14. Commonly used factor types	27
Figure 15. DOE trees	28
Figure 16. Response surfaces for categorical and continuous factors	28
Figure 17. Visualization of main effects and two-factor interactions.....	30
Figure 18. Components of main effects and interactions for the 2^3 design	32
Figure 19. IPO diagram illustration	33
Figure 20. Sequential ensemble of test designs	36
Figure 21. Projection and confounding of 2_{III}^{3-1} main fraction	37
Figure 22. Sparsity of effects and projection properties	37
Figure 23. Some features of traditional response surface designs	41
Figure 24. Tradeoff space in selecting a test design	42
Figure 25. Power vs. number of runs for fixed α and $\delta\sigma$	44
Figure 26. Power vs. number of runs for different α and $\delta\sigma$	45
Figure 27. Partial list of test design metrics for some screening designs	48
Figure 28. Example of 2^4 completely randomized and split-plot designs	49
Figure 29. Low order vs. a higher order nonlinear response surface.....	49
Figure 30. Linear regression vs. Gaussian process models	51
Figure 31. Coverage of faults due to t-way interactions.....	53
Figure 32. Graphical methods for identifying significant factors and interactions	57
Figure 33. Graphical methods for validating DOE assumptions	60
Figure 34. Graphical methods for diagnostics	61
Figure 35. Contour maps and response surfaces comparisons	62
Figure 36. Evaluation and test process flow	67

List of Tables

Table 1. Top Level DoD T&E Policy	5
Table 2. Top Level DoN T&E Policy	5
Table 3. Content of memos from DOT&E	9
Table 4. DoD T&E process goals	10
Table 5. Comparison of DOE process as outlined by authoritative sources.....	11
Table 6. Strategies for identifying evaluation measures	19
Table 7. Classes of Evaluation Objectives.....	22
Table 8. Strategies for formulating evaluation objectives	23
Table 9. Factor definitions	24
Table 10. Illustration of test capacity to meet factor requirements.....	26
Table 11. Illustration of categorical and continuous factors.....	27
Table 12. Design matrix for a 2^3 factorial design	30
Table 13. Effect of A for different settings of factors B and C	31
Table 14. Strategies for identifying the test space	33
Table 15. Sample size required to maintain 80% power as a function of $\delta\sigma$	43
Table 16. Power for sample size $N = 13$ as a function of $\delta\sigma$	44
Table 17. General test design evaluation criteria.....	47
Table 18. Common Types of Space Filling Designs	50
Table 19. Factor and level list for software testing application.....	52
Table 20. Efficiency to cover t-way interactions (network security).....	53
Table 21. Strategies for selecting an experiment design.....	54
Table 22. Strategies for test execution.....	55
Table 23. ANOVA table for Response 1	58
Table 24. Fit statistics summary	59
Table 25. Strategies for data analysis	63
Table 26. Contributions of the test design framework to the test strategy	64
Table 27. Strategies for informing the evaluation	65

This page intentionally left blank.

Modern Test Design and Analysis Playbook

1 Introduction

The Department of Defense (DoD) acquires and sustains weapons systems, defense business systems, and national security systems. Because delivering an effective, suitable, survivable, sustainable, and affordable capability to the users is imperative, every decision must be well informed. Misinformed decisions could expose the programs to cost, schedule, and performance risks, and can create a progressive knowledge deficit throughout the lifecycle that leaves decision makers with less and less knowledge to make informed decisions about the capabilities and on how and when to proceed to subsequent acquisition phases. Capability based test and evaluation (T&E) is one of the most effective and efficient means of gaining knowledge in engineering, design, and production to inform decisions about a system's capability.

Test, in native context, is the process of submitting a system to deliberate and planned changes in conditions to observe and measure how those changes influence the attributes of a capability or system (such as physical properties, quality, performance, reliability, etc.). Likewise, **Evaluation** is the rigorous and systematic interpretation of the effects of those changes through examination, analysis, assessment, and comparison of the empirical observations to standards, requirements, and specifications. Simplistically T&E can then be defined as the process of submitting a capability or system to methodical changes in conditions to generate empirical observations so we can examine, analyze, and assess the effects of those changes and compare the observations to requirements, specifications, and standards. Draft *DoD Test and Evaluation Instruction* (DoDI 5000.UF), under review for legal sufficiency, describes the purpose of T&E—to enable the DoD to acquire systems that support the warfighter in accomplishing their missions. T&E provides systems engineers and decision-makers with knowledge to help managing risks, measuring technical progress over the lifecycle, and characterizing operational effectiveness, operational suitability, interoperability, survivability, and lethality.

The tradeoff in T&E is to determine the true capability of the system, described by its technical and operational attributes, with a minimum number of samples that yield valid conclusions while minimizing the risks of making bad decisions. From a T&E perspective, the ultimate risks that need to be managed are the risk of considering that a system provides a capability when it does not or that a system fails to provide a capability when it does.

In dealing with this tradeoff, test designers often face several critical questions: What information is required by decision-makers? How will the information be generated? How many samples or tests are required? How will the test space be sampled? What are the conditions in which those samples are obtained? How are those conditions determined? What are the tests measuring? How are the tests executed? How are the test results measured? What happens if some tests cannot be performed? What results are obtained from the samples? What techniques are used to analyze the data? What conclusions can be drawn from the data? How are the results and conclusions communicated? How good was the test? How good is the evaluation?

A well-designed test strategy should illuminate those tradeoffs, answer the questions above, and produce the maximum amount of information with the minimum amount of test resources (time, events, or samples) to inform its objectives. Often, conventional test design and analysis techniques are insufficient to answer those questions effectively and efficiently. Statistical test

design methods can replace conventional techniques and provide the maximum amount of information with the minimum amount of test resources.

The need for T&E to capture the full benefits afforded by state-of-the-art statistical methods has been recognized for decades. In 1995, the Military Operations Research Society (MORS), in collaboration with the International Test and Evaluation Association (ITEA), held the Joint mini symposium *How Much Testing is Enough* and concluded that statistical methodologies, particularly **design and analysis of experiments** (DOE), can increase the efficiency and economy of information collection, processing, and evaluation. In 1998, the study *Statistics, Testing, and Defense Acquisition* by the National Research Council (NRC) echoed that conclusion and recommended the Test and Evaluation Master Plan (TEMP) include: (1) the purpose and scope of the test; (2) explicit identification of the test factors and methods for handling them; (3) definition of the test environment and specification of constraints; (4) comparison of variation within and across test scenarios; and (5) specified, consistent data recording procedures.

Statistics is the mathematical science dealing with the planning, collection, analysis, organization, explanation, and presentation of data. A common goal of statistics is to investigate and establish a relationship between events. Clearly, it seems practical to exploit the tremendous power of combining statistical methods and testing with the specific goal of acquiring deeper knowledge and understanding of a system and of maximizing the usefulness and accuracy of the information. DOE-based test design and analysis allows a program to exploit the power of combining sound test practices and statistical data analysis.

The definition of test is very similar to the definition of experiment. An experiment is commonly defined as a test or series of sequential tests in which deliberate and systematic changes are made to the input variables that could influence the behavior of a system so we can observe and exploit the changes in the output responses given natural variation embedded in the system. Mathematically, the observed changes in the response of the system due to the deliberate changes made to the input variables can be described by equation (1), which is referred to as the effects model (two-way example):

$$y_{ijk} = \mu + \alpha_i + \tau_j + (\alpha\tau)_{ij} + \varepsilon_{ijk} \quad (1)$$

In equation (1), y_{ijk} is the response for the ijk^{th} observation, μ is the overall mean, α_i is the effect of the i^{th} level of factor A, τ_j is the effect of the j^{th} level of factor B, $(\alpha\tau)_{ij}$ is the effect of the interaction between factors A and B, and ε_{ijk} is the random error component. The equation shows how the overall mean, which is a constant, changes with the changes attributed to the different levels of the input variables.

An experiment is carried out to support, refuse, or validate a hypothesis, which is essentially the objective of a test. Like tests, experiments provide insight into cause and effect by demonstrating an outcome when conditions are manipulated. Thus, the same principles that are applicable to experiments can be applied to T&E.

DOE is the systematic integration and exploitation of well-defined, rigorous, disciplined, and structured scientific strategies into a foundational framework for gathering empirical knowledge about a system or process using statistical methods for planning, designing, executing, and analyzing an experiment or test (or series of progressive tests). The core principle is to make deliberate and systematic changes to the input variables of a system to observe and exploit the

changes in the output response. As implied by its name, DOE recognizes two aspects to any test—the design of the test and the analysis of its data. DOE is also defined as “a systematic, rigorous approach to engineering problem-solving that applies principles and techniques at the data collection stage to ensure the generation of valid, defensible, and supportable engineering conclusions. In addition, all of this is carried out under the constraint of a minimal expenditure of engineering runs, time, and money.” (NIST/SEMATECH e-Handbook of Statistical Methods, Section 4.3.1)

The application of DOE methodologies can simultaneously help generate test efficiencies, improve the fidelity of test results, illuminate risks, enable better-informed decisions, and ultimately enable the fielding of an effective, suitable, and survivable capability. The use of DOE in T&E maximizes the utility of the information so decision makers can obtain a deeper knowledge of the system’s capabilities across the entire technical and operational space and of the risks associated with the decisions to be made. DOE provides an explicit relationship between the level of knowledge desired to inform decisions and the cost to achieve that level of knowledge.

DOE is not only applicable to T&E. The use of DOE methodologies can help produce defensible artifacts that directly tie levels of knowledge required to inform technical and operational decisions with the utility of information. DOE has utility in activities and processes such as performance assessment, tactical decision making, analysis of alternatives, system or mission area qualifications, product improvement, regression testing, trade-off analysis, and sensitivity analysis.

Figure 1 illustrates key knowledge points in the lifecycle of a major capability acquisition where it is practical to inject DOE. Because DOE facilitates the understanding of the tradeoff between cost, risk, and quality of information obtained from test, it provides decision makers with a unique management tool to help them determine how much knowledge they can afford.

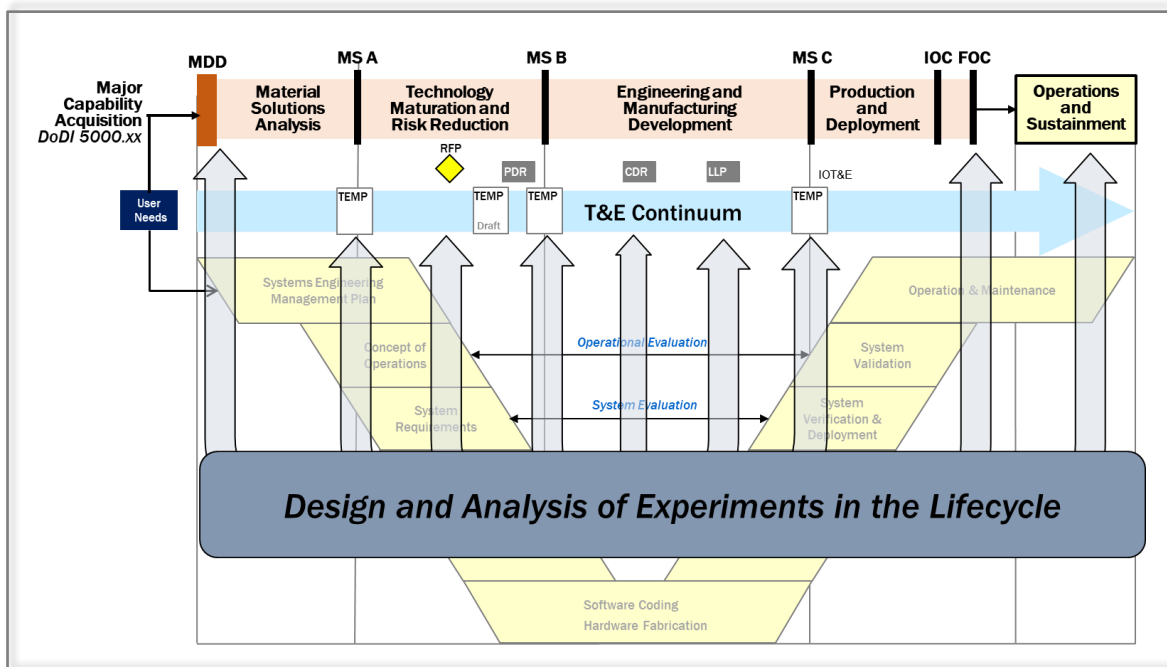


Figure 1. Design and analysis of experiments in the lifecycle

2 Document Scope and Organization

This playbook is not a comprehensive test planning guide. The primary objective of the document is to provide a path to integrate design and analysis of experiments methodologies into the fabric of T&E to create a cohesive strategy for developing rigorous and economical test designs that produce decision-quality information for decision-makers. Since all acquisition Programs are different and there is no “one size fits all” solution, the content provided in this document is expected to be tailored to fit specific program’s needs.

The secondary objective is to provide a primer to help strengthening T&E staff’s knowledge, skills, and abilities to apply statistical thinking via DOE. While the scope of the document is to provide a standard “how to” for practitioners, the goal is also to help program personnel understand the basics of the DOE process, its value, and its benefits. It is presumed that the reader has a basic knowledge of T&E, a basic understanding of the defense acquisition system, and an elementary knowledge of statistics.

Section 3 provides an overview and flow down of top-level T&E policy and guidance for DoD and Department of the Navy (DoN) with emphasis on policy for the use of DOE in T&E. Section 4 starts with the introduction of capability-based T&E (CBT&E) and its core process, model-based test design (MBTD). Following that introduction, the rest of the section provides a step-by-step discussion of the DOE methodology. Section 5 provides a summary and recommendations. Appendix A provides additional details on DoD and DoN policy and guidance. Appendix B provides a catalog of the most popular screening and response surface designs.

3 Top T&E Policy and Guidance in DoD and DoN

The Defense Acquisition System (DoDD 5000.01) describes the overarching policies and principles that govern the management of capability acquisitions while the *Operation of the Adaptive Acquisition Framework* (DoDI 5000.02) provides the policies to support the Defense Acquisition System. The Adaptive Acquisition Framework (AAF) shifts acquisition policy from the traditional ‘one size fits all’ philosophy of the Defense Acquisition System to a tailorable six-pathway acquisition strategy intended to provide capability to the users while developing or employing advanced acquisition strategies that match the characteristics of the capability being acquired. Draft *Test and Evaluation Instruction* (DoDI 5000.UF), under review for legal sufficiency, establishes T&E policy and procedures across the AAF for developmental test and evaluation (DT&E), operational test and evaluation (OT&E), live fire test and evaluation (LFT&E), and integrated test and evaluation (IT&E). A brief description of these policies is provided in Table 1 while additional details are provided in Appendix A.

DoDI 5000.UF requires Programs to identify how design and analysis of experiments (also known as scientific test and analysis techniques) “will be used to design an effective and efficient test program that will produce the required data to characterize system behavior and combat mission capability across an appropriately selected set of factors and conditions”.

Table 2 provides the flow down of key T&E policies for the DoN. *Defense Acquisition System and Joint Capabilities Integration and Development System Implementation* (SECNAVINST 5000.2F) prescribes specific policies and procedures, including T&E, for the operation of the Joint Capabilities Integration and Development System (JCIDS) and Defense Acquisition system within the Department of the Navy (DON).

Table 1. Top Level DoD T&E Policy

Policy	Title	Effective Date	Description
DoDD 5000.01	<i>The Defense Acquisition System (Change 2)</i>	31 August 2018	Provides guiding principles for the management of capability acquisitions
DoDI 5000.02	<i>Operation of the Acquisition System</i>	23 January 2020	Provides the policies to support the Defense Acquisition System
DoDI 5000.UF	<i>Test and Evaluation Instruction</i>	In coordination	Establishes T&E policy and procedures across the Adaptive Acquisition Framework (AAF) for DT&E, OT&E, LFT&E, and IT&E

Table 2. Top Level DoN T&E Policy

Policy	Title	Effective Date	Description
SECNAVINST 5000.2F	<i>Defense Acquisition System and Joint Capabilities Integration and Development System Implementation</i>	26 March 2019	Prescribes specific T&E policies and procedures for the operation of the Joint Capabilities Integration and Development System (JCIDS) and the Defense Acquisition System within the Department of the Navy
SECNAVINST-M 3960.1	<i>Department of the Navy (DoN) Test and Evaluation (T&E) Guidebook</i>	30 July 2020 (Draft)	Supplements chapter 5 of SECNAVINST 5000.2F with policy, guidance, process, and best practices for Navy DT&E, OT&E, and LFT&E
NAVSEAINST 3960.2E	<i>Test and Evaluation Policy is in draft and will NAVSEAINST 3960.2E managed by NAVSEA</i>	Will replace NAVSEAINST 3960.2D of 22 April 1988)	Provides policy and procedures for acquisition related T&E activities of programs and systems
PEOIWSINST 3960.2	<i>Planning, Execution, Analysis, and Reporting of Integrated Combat/Weapon System Test and Evaluation Events</i>	27 April 2015	Provides a standardized process for the planning, execution, analysis, and reporting (PEA&R) of PEO IWS integrated combat and weapon system T&E events
COMOPTEVFORINST 3980.21	<i>Operational Test Director's (OTD) Manual</i>	3 May 2019	Prescribes policy and procedures for the planning, conduct, and reporting of OT&E
NSWC PHD TP-60	<i>Surface Ship Weapon/Combat Systems Test and Evaluation Guide</i>	31 July 2017	A resource for the organization's workforce to support T&E throughout the lifecycle, including Combat System Ship Qualification Trials (CSSQT)
CBT&E Implementation Plan	<i>Implementing Capabilities-Based Test and Evaluation (CBT&E): A Guide for Test Professionals</i>	5 August 2020 (Draft)	Describes the need for CBT&E, provides definition, and provides recommendations on how to implement CBT&E

Department of the Navy (DoN) Test and Evaluation (T&E) Guidebook (SECNAVINST-M 3960.1) supplements chapter 5 of SECNAVINST 5000.2F with policy, guidance, process, and best practices for Navy DT&E, OT&E, and LFT&E, and IT&E. The instruction requires Programs to use CBT&E to “integrate DT&E, OT&E, and LFT&E into a single T&E continuum for the six pathways identified in the AAF”. Similarly, the instruction requires Systems Commands (SYSCOMs) and Warfare Centers “to manage assigned infrastructure, resources and policy to ensure efficient and effective CBT&E that supports the DT&E and LFT&E requirements of systems within the SYSCOM’s domain” and to “integrate scientific test and analysis techniques into DT/OT/CBT&E process”. DoN T&E Guidebook also requires programs to include in the TEMP a decision support framework that correlates “PM decisions and program events to be supported, data required to support the decisions, timeline for delivery, and resources required to generate the necessary data”.

CBT&E is a strategic initiative to ensure the right capability is delivered to the warfighters at the speed of relevance. CBT&E fundamentally changes the way DoD tests and evaluates warfighting capabilities by shifting DT&E focus from specification compliance to assessing mission capability in a systems-of-systems (SoS) context in operationally relevant environments while also validating individual technical requirements. CBT&E focuses on the system’s contribution to mission success. DoN T&E Guidebook defines capability as “the ability to achieve a desired effect under specified standards and conditions through combinations of ways and means to perform a set of tasks”. *Implementing Capabilities-Based Test and Evaluation (CBT&E): A Guide for Test Professionals* (Draft Revision 4) formally defines CBT&E as “the evolution of DT to increase the focus on mission-level requirements through the developmental cycle”.

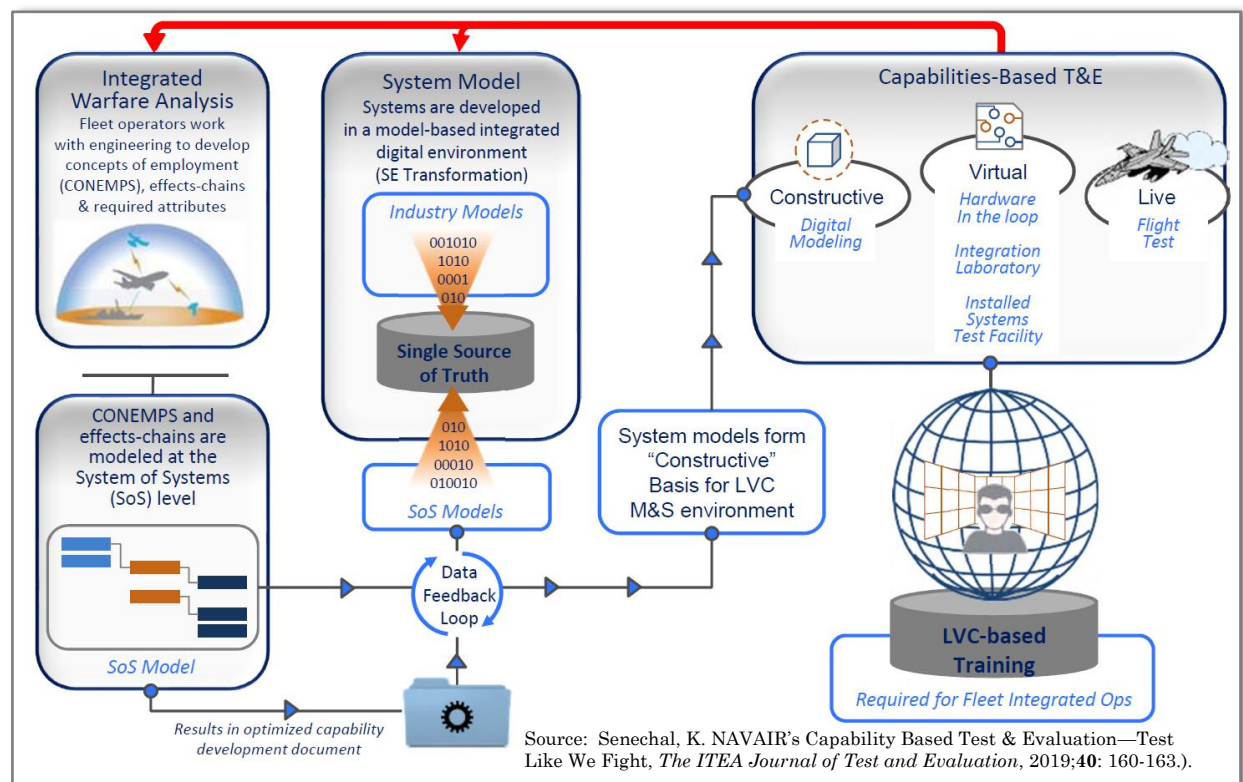


Figure 2. CBT&E in capability-based acquisition

Figure 2 shows the place of CBT&E in capability-based acquisition. The strategy envisions an upfront integrated warfare analysis that uses Concepts of Employment (CONEMP), effects chains, and systems attributes that are modeled at the system-of-systems (SoS) level in a digital environment that links warfare analysis, requirements definition, systems engineering, T&E, etc. to a single source of truth. When used effectively, CBT&E not only improves test efficiency, but also accelerates the identification of deficiencies that impact mission requirements and provides the necessary information to decision makers so they can pivot away from systems that do not meet mission needs.

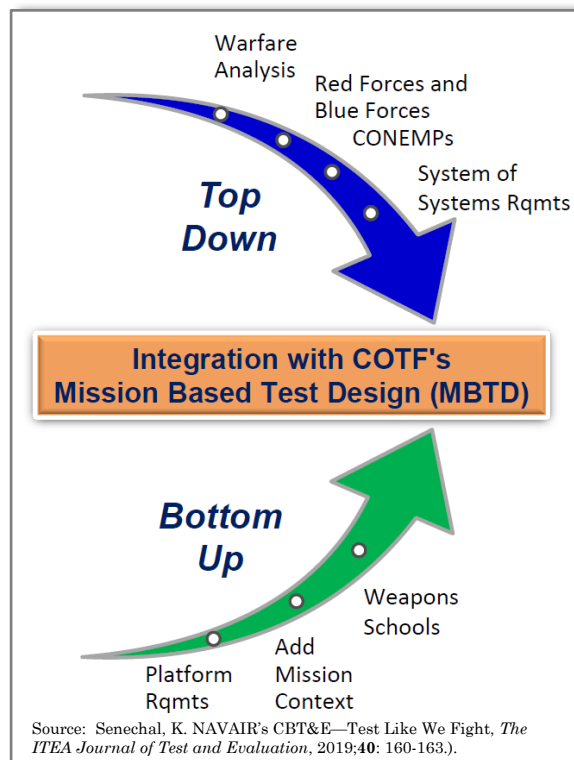


Figure 3. CBT&E approach

CBT&E takes advantage of a two-prong approach as illustrated in Figure 3. The “top-down” approach leverages warfare analysis products (warfare analysis, CONEMPs, and SoS requirements) that were derived at the campaign level. When products are unavailable for implementing the “top-down” approach as is the case for some missions, CBT&E relies on a “bottom-up” approach that requires a set of more traditional elements (platform requirements, mission context, and weapons schools) to establish how the systems will be used.

The heart of the CBT&E process is Commander, Operational Test and Evaluation Force (COMOPTEVFOR) mission-based test design (MBTD), which is part of the Integrated Evaluation Framework (IEF). The MBTD process, illustrated in Figure 4 and documented in COMOPTEVFORINST 3980.2, focuses testing on the mission chains, tasks, and subtasks that must be performed to complete the mission.

Depending on its stage of development, a program that employs CBT&E principles may have artifacts that are useful for the preparation of test designs. Those include:

- Warfare analysis products such as platform mission task (PMT), traceable to Navy Required Operational Characteristics/Projected Operational Environment (ROC/POE), that map requirements, conditions, and system deficiencies to the accomplishment of mission tasks (a PMT is a model that outlines a set of steps that technology and human resources in a platform needed to perform to deliver an expected behavior under a set of conditions)
- Applicable tactical tasks (such as naval tactical tasks, Marine Corps tasks, or Coast Guard tasks) that align with PMT views (A PMT view is a representation of an applicable PMT that a system or SoS performs).
- DODAF views and associated system architecture that captures the system as planned or designed vs. as built

- A decision support framework that hierarchically links key decisions or knowledge points to be informed over the lifecycle to the data needed to inform those decisions
- An IEF, including an MBTD endorsed by resource sponsors

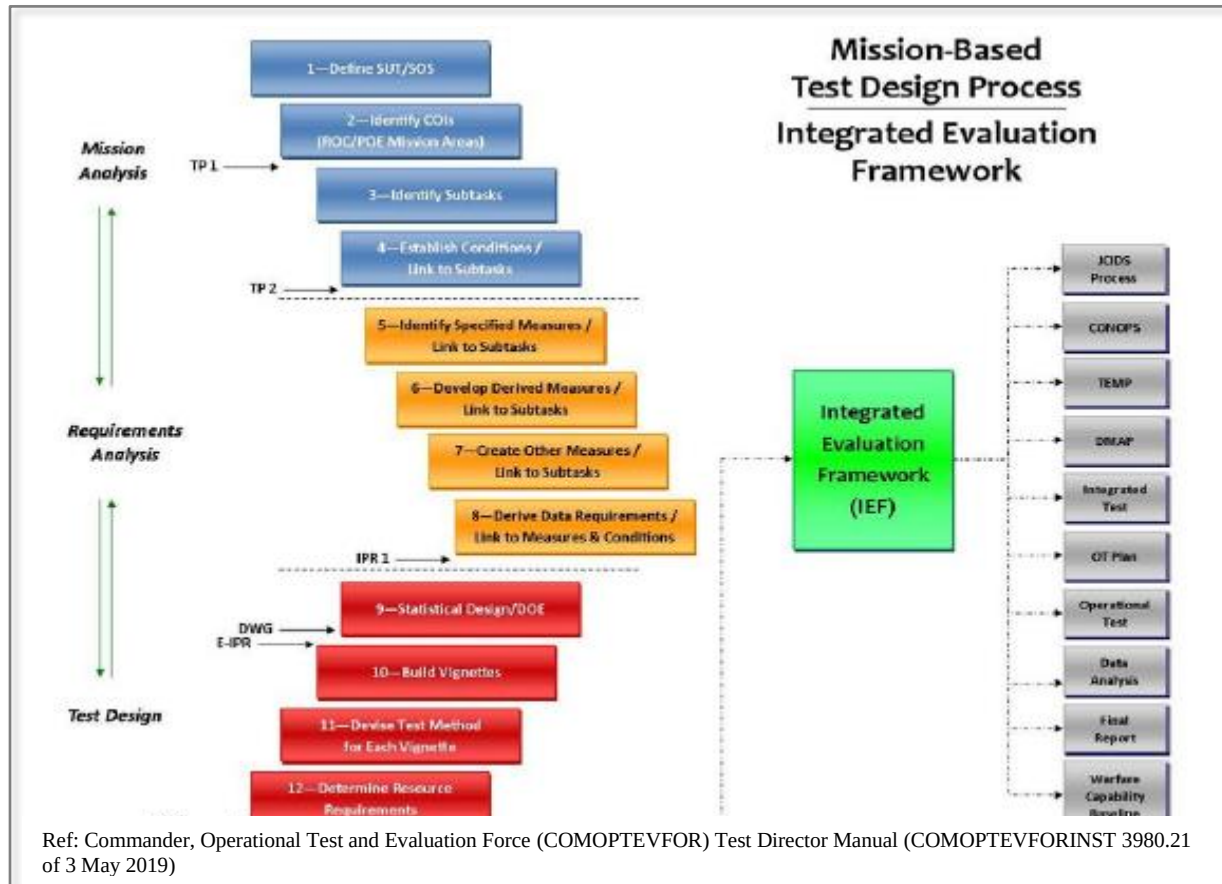


Figure 4. Mission-based test design (MBTD) process

CBT&E key enablers include:

- *Mission engineering* - the process of determining warfighter expectations for a new or modified capability.
- *Model-based systems engineering (MBSE)* – the formalized application of graphical modeling to support and automate the systems engineering process throughout the acquisition lifecycle: requirements development, analysis, design, implementation, verification, and validation. MBSE focuses in the digital representation of a weapon system’s capability model to inform evaluations, and ultimately decisions, throughout the system's life cycle. MBSE provides a better approach for requirements cascading and verification, reduces test cycles, lowers development risks, and produces executable models rather than paper models.
- *Statistical engineering* - the amalgamation of statistical, computational, scientific, and engineering methods and techniques for the collection, organization, interpretation, visualization, analysis, and assessment of data applied to problems that are narrowly or

poorly defined, broad in scope, large, complex, or require significant upfront work to figure out a structure. Statistical engineering leverages the scientific understanding of how things work with a statistical capacity to figure out why when they do not and guides the understanding in knowledge and decision-making to generate better results. DOE is at the heart of statistical engineering and an integral part of the Navy's MBTD.

Table 3 summarizes the content of memos issued by Director, Operational Test and Evaluation (DOT&E) between 2010 and 2017 to provide guidance and set expectations on elements of DOE that must be addressed by the programs. Those elements are captured in DoDI 5000.UF.

Table 3. Content of memos from DOT&E

Title	Effective Date	Content
<i>Guidance on the use of design of experiments (DOE) in Operational Test and Evaluation</i>	19 October 2010	Sets expectations on elements of DOE that must be addressed in the TEMP
<i>Case studies for the use of design of experiments (DOE) in Developmental Test and Evaluation (DT&E)</i>	21 January 2011	Examines the applicability of DOE to DT&E activities, from early engineering analysis, through incremental development, to final verification of requirements
<i>Flawed application of design of experiments (DOE) to Operational Test and Evaluation (OT&E)</i>	26 June 2013	Reinforces the expectations set forth in the October 2010 memo and addresses areas for improving the use of DOE
<i>Best practices for assessing the statistical adequacy of experimental designs used in Operational Test and Evaluation</i>	23 July 2013	Identifies statistical figures of merit that should be used to determine the adequacy of a test design
<i>Inadequacy of recently proposed test design for the P-8A Increment 2 and Multi-static Active Coherent (MAC) Programs</i>	24 September 2013	Identifies deficiencies in the test design concept for evaluating the P-8A Increment 2 and MAC systems
<i>Guidance on the validation of models and simulation used in Operational Test and Live Fire Assessments</i>	14 March 2016	Sets expectations for the TEMP and Test Plan to describe the modeling and simulation (M&S) capability, its intended use, and its verification, validation, and accreditation (VV&A) approach
<i>Clarification on guidance on the validation of models and simulation used in Operational Test and Live Fire Assessments</i>	17 January 2017	Re-emphasizes the expectations for the M&S validation strategy, including the quantitative comparison of live and M&S outcomes using statistical methods and a comprehensive strategy to assess M&S output across the operational domain for which the M&S is accredited

NAVSEAINST 3960.2E and PEOIWSINST 3960.2 prescribe standardized procedures for the planning, execution, analysis, and reporting (PEA&R) of integrated combat and weapon system T&E events for programs and systems managed by NAVSEA, affiliated PEOs, and NAVSEA Warfare Center Divisions.

4 Modern Test Design and Analysis Framework

The *DoD Test and Evaluation Management Guidebook (TEMG)* is a generic desk reference to assist program management and T&E personnel in executing their responsibilities. The TEMG describes the standard DoD T&E process, which is illustrated in Figure 5. Table 4 lists the goals for each phase.

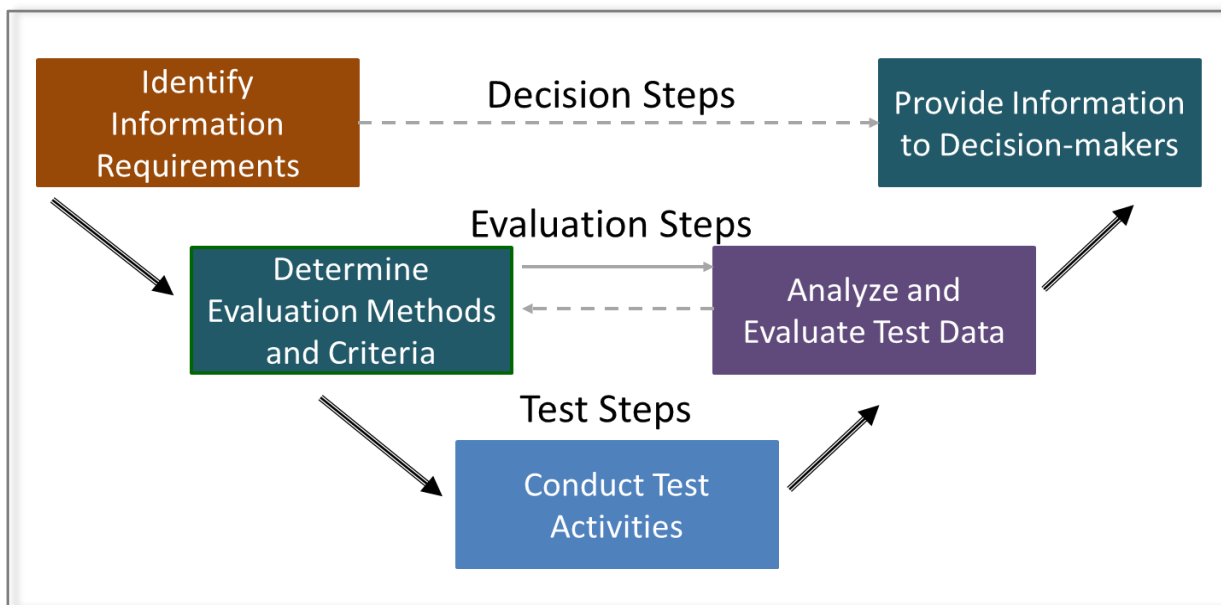


Figure 5. T&E process flow

Table 4. DoD T&E process goals

Phase	Goal
Identify information requirements	Identify the information required by decision-makers
Determine evaluation methods and criteria	Determine evaluation methods to obtain the information
Conduct test activities	Coordinate test activities and plan and conduct the test
Analyze and evaluate test data	Study the test results to develop and support conclusions and recommendations
Provide information to decision-makers	Use conclusions and recommendations to make informed decisions

The DOE methodologies, techniques, and tool have been well established within the experiment design community of practice. Table 5 contrasts the DOE process as outlined by three authoritative sources: the TEMG, the textbook *Design and Analysis of Experiments* by Dr. Douglas Montgomery, and the *NIST/SEMATECH e-Handbook of Statistical Methods*. The content of the three processes is, for all practical purposes, identical. The only differences between them are the name of some steps and the way in which the steps are organized.

Table 5. Comparison of DOE process as outlined by authoritative sources

Test and Evaluation Management Guide	Montgomery	NIST/SEMATECH e-Handbook of Statistical Methods
Determine the purpose of the test to include output variables (results) that will be measured	Recognition of and statement of the problem	Set objectives
	Selection of the response variable	Select process variables
Determine the input variables (factors) that are expected to have an impact on results	Choice of factors, levels, and range	
Determine the levels for each factor		
Determine constraints that must be considered when creating the test matrix		
Create the test matrix and review it with interested parties and subject matter experts	Choice of experiment design	Select an experimental design
Execute test	Perform the experiment	Execute the design
Analyze test data	Statistical analysis of the data	Check the data are consistent with experimental assumptions
		Analyze and interpret results
Make decisions based on analysis.	Conclusions and recommendations	Use and present results

The TEMG identifies the following benefits from the application of DOE:

- An understanding of the likelihood of successfully identifying, as well as the likelihood of mistakenly identifying, significant performance drivers
- A sound method for determining and justifying the number of tests planned and assets needed to obtain required information
- The ability to make informed trade-offs of test costs (assets) versus information gained
- A rigorous method of determining specific scenarios that provide the most useful data for characterizing system performance including deficiencies and for identifying drivers of system performance
- The ability to identify interactions between test factors, which is seldom possible with other methods of developing test matrices

Because both the DOE and the T&E process incorporate the scientific method, it is very easy to map the DOE process into the fabric of the T&E process and within the paradigm of the PEA&R process. Figure 6 was constructed with that in mind. In Figure 6, the chevrons correspond to the pillars of the PEA&R process with the addition of a chevron for Needs, the middle blocks

correspond to the objectives or goals of the T&E process, and the blocks to the right corresponds to the seven steps of DOE augmented with one necessary step to complete the T&E process.

It is of outmost importance to perform those steps as well as possible to ensure the success of the test. It is also important to remember that test incorporates the scientific method, which suggests that some of the steps are iterative rather than a one-step-forward-at-a-time process.

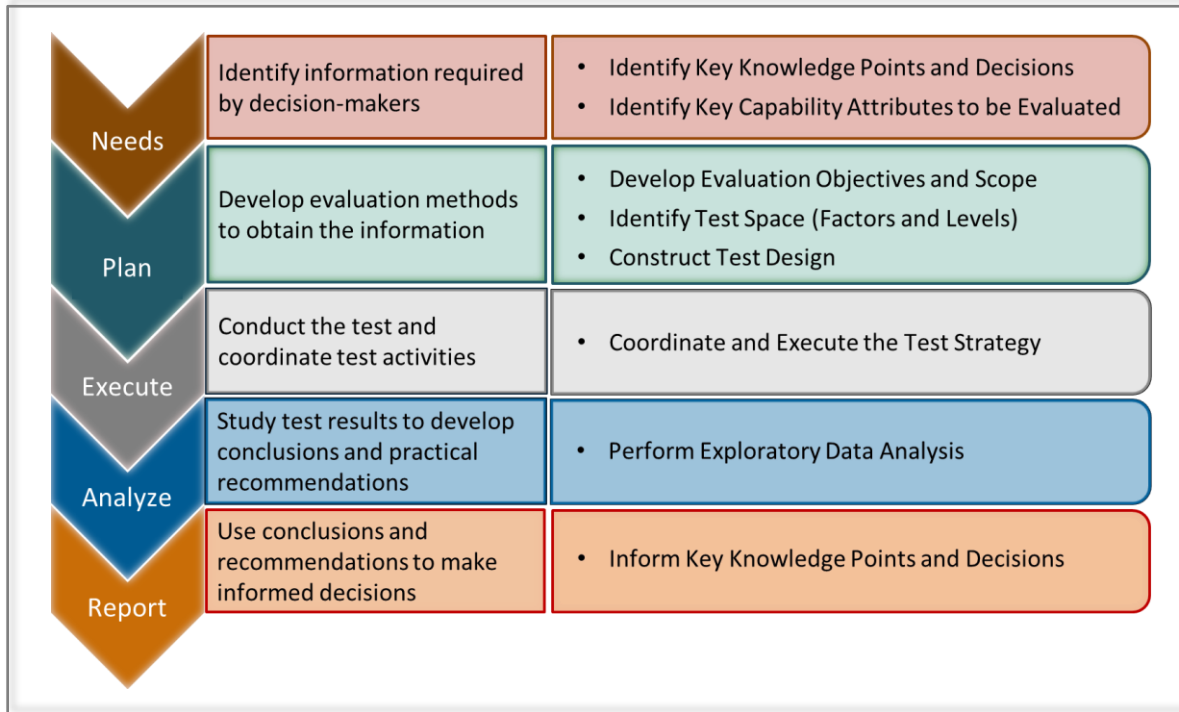


Figure 6. Evaluation and test design and analysis process (PEA&R + N)

Figure 6 suggests that an efficient and effective evaluation and test strategy requires involvement of numerous experts from a cross-section of various communities that include: program leadership, program management, policy, mission engineering, intelligence, targets, systems engineering, test execution, systems, test facilities and range operations, modeling and simulation, instrumentation, the environment, system operation, and data analysis. Similarly, an effective an efficient strategy may require involvement from oversight organizations like DOT&E and Director, Developmental Test, Evaluation, and Assessment (D(DTE&A), resource sponsors and CD&I, Operational Test Agencies, Program Executive Office, Systems Commands, Technical Authorities, Warfare Centers, and Systems Engineering and Technical Assistance (SETA) contractors.

For the evaluation and test strategy to be successful, they all need a common and clear understanding of what is to be tested, why it is going to be tested, how will the test be carried out, under which conditions the test will be carried out, who will carry out the test, what data is to be collected, how the data is to be analyzed, and what the criteria is for evaluating the data. While these types of questions are typical of every test, the supporting answers may not be obvious as one may think, and may involve having intimate understanding or knowledge of:

- Decisions to be informed
- Scope and objectives of the evaluation

-
- Capabilities that require testing
 - Type of responses expected
 - Character of the factors and the ranges over which they could be varied
 - Different types of experiment designs
 - Debarred regions of test space or disallowed factor combinations
 - Interactions between factors
 - Live-virtual-constructive (LVC) planning and execution practices
 - Availability of LVC assets and resources
 - Control mechanisms for potential sources of natural and induced variation
 - Following proper experimental protocol
 - Number and level of proficiency of operators required to carry out the test
 - Type of information to be extracted from the test
 - Statistical analysis techniques

Since it is almost impossible to find a single or even a few experts with all that knowledge, a way to deal with those questions is for the program to charter a DOE Working Group (DOE WG) of functional and subject matter experts knowledgeable in each of those areas. The mission of the DOE WG is to assist program's in the successful incorporation and implementation of DOE methodologies into the test strategy and TEMP. Generally, the chief developmental tester (CDT) serves as DOE WG lead with the assistance of the lead developmental test office (LDTO) representative.

Chartering a DOE WG facilitates communications, fosters a common understanding of the test strategy among the test community, affords the opportunity for testers and analysts to get engaged in and contribute to the development and coordination of the test strategy early in the process. The DOE WG also assists the Program in:

- Ensuring test objectives are clear, testable, measurable, unbiased, and relevant
- Linking response variables and decision support questions, evaluation objectives, and mission tasks
- Identifying the type of tests required
- Preparing, defining, and implementing the test design strategy
- Refining the response variables for evaluation via DOE
- Informing programs on risk related to fulfilling entry and exit criteria for key knowledge points, including decision milestones

The following sections provide discussions on each of the steps illustrated in Figure 6. Throughout the discussions, references are made to, and links provided to, sections of the *NIST/SEMATECH e-Handbook of Statistical Methods* ([NIST Handbook Section X](#)).

4.1 Identify Key Knowledge Points and Decisions

The main purpose of this phase is to identify key decisions or knowledge points across the acquisition lifecycle to be informed by evaluating system capabilities.

The strategy is to drill down the following process:

- Identify knowledge points and decisions across the lifecycle
- Identify technical and operational capabilities to be evaluated to gain knowledge required to inform decisions
- Identify mission area measures and technical parameters (response variables) to quantify the systems capabilities
- Identify LVC events required to generate the data for estimating the response variables
- Document all the information above in a decision support framework, such as the crosswalk evaluation framework (XEF), and use it to guide evaluation planning and test design

The knowledge points and decisions should represent major turning points in the acquisition and sustainment strategy made by decision-makers such as Milestone Decision Authority (MDA), Program Manager (PM), Principal Assistant Program Manager (PAPM), Assistant Program Manager (APM), and Certification Authorities (CA) that need T&E information to make decisions in an informed manner. The PM identifies those decisions, which may include programmatic, acquisition, technical, or operational decisions such as decisions involving significant resource expenditure, manufacturing readiness decisions, integration readiness decisions, technical readiness decisions, key integration points, entry into next test phases, milestone decisions, etc. Those key knowledge points or decisions as well as the data required to support them, timeline for delivery, and resources required to generate the necessary data are documented in the XEF.

The XEF is a strategic-level planning tool that:

- Identifies key programmatic, acquisition, engineering, technical, and operational decisions to be informed by evaluations of system's capabilities grouped in terms of functional evaluation areas such as performance, interoperability, cybersecurity, reliability, etc.
- Describes the essence of the decisions to be made expressed in forms of statement of objectives or questions
- Describes how those decisions will be informed with knowledge gained from the independent evaluation of the system's capabilities
- Maps the information sources to the decisions supported

Figure 7 illustrates a generic XEF template. The XEF links key decisions (in color green), the essence of the information or knowledge needed to inform each key decision in terms of decision support questions (DSQ) or statements of objectives (SOO) (in color orange), system capabilities expressed as mission tasks performed by the system and grouped by functionality (in color yellow), top-level operational and technical parameters that serve as means to quantify system capabilities (in color blue), and the LVC events (in color purple) to generate the data for quantifying the operational measures or technical parameters that will be used to evaluate system capabilities and inform knowledge points or decisions. The PM is responsible for developing the XEF in

consultation with the chief developmental tester (CDT), the chief engineer, and the OTA representative.

Platform Mission Task/Subtask (PMT)	Measures		Knowledge Points or Decisions							
	Operational Measures (Ref. to IEF)	Technical Parameters (Ref. to TRD/SRD)	Decision 1		Decision 2			Decision ...	Decision n	
			Objective 1	Objective 2	Objective 3	Objective 4	Objective 5	Objective 6	Objective ...	Objective n
			Data Sources (Live, Virtual, Constructive)							
Performance										
PMT 1.1	M1 (E 2.1.1)	TP 1 (3.x.x.5)	Const. (DT 1)		Virtual (DT 1)				Live (DT 3)	Live (OT 1)
	M2 (E 2.1.2)	TP 2 (3.x.x.6)	Const. (DT 1)		Virtual (DT 2)				Virtual (IT 3)	Virtual (OT 4)
PMT 1.2	M3 (E 2.1.3)	n/a				Live (DT 2)			Live (IT 3)	
	n/a	TP 4 (3.x.x.8)				Virtual (DT 2)			Virtual (IT 3)	Live (IT 4)
Interoperability										
PMT 1.3	M5 (E 3.4.5.1)	TP 9 (3.x.x.1)				Live (DT 2)		Live (DT 2)		
	M6 (E 3.4.6.1)	TP 10 (3.x.x.2)		Const. (IT 1)		Virtual (DT 2)		Virtual (DT 2)		Live (OT 4)
PMT 2.1	M7 (E 3.4.6.2)	TP 11 (3.x.x.3)		Const. (IT 1)				Virtual (IT 2)		Virtual (IT 4)
	n/a	TP 12 (3.x.x.4)						IT 2		Live (OT 2)
Cybersecurity										
PMT 3.1	M13 (E 5.4.5.1)	TP 5 (3.x.x.9)			SW Dev Assess.		SW Dev Assess.	SW Dev Assess.		
	M14 (E 5.4.6.1)	TP 6 (3.x.x.10)	Cont Assess.		Cont Assess.	Cont Assess.	Cont Assess.			
PMT 3.2	M15 (E 5.4.6.2)	TP 7 (3.x.x.11)				Vuln Assess.			Vuln Assess.	
	M16 (E 5.4.6.3)	TP 8 (3.x.x.12)				Adver Assess.			Adver Assess.	
Reliability										
PMT 4.1	M9 (S 1)	TP 13 (4.x.x.1)		M-demo 1						OT 3
	M10 (S 2)	TP 14 (4.x.x.2)		M-demo 1				IT 2		OT 3
	M11 (S 3)	n/a				M-demo 2		IT 2		
PMT 4.2	n/a	TP 16 (4.x.x.3)				M-demo 2		IT 2		

Figure 7. Generic crosswalk evaluation framework (XEF) template

Critical Tasks and Critical Measures		
Task Title	Critical Measures	
Task x.y.a.a - Receive Information	M1 (KPP) (RV)	Time to receive information
Task x.y.a.b - Process Information	M2 (KPP)	Time to process information
	M3 (KPP)	Percentage of the information processed
	M4 (KPP)	Accuracy of the information processed
Task x.y.a.c - Correlate Information	M5 (KPP)	Accuracy of the information processed
	M6 (KPP) RV	Time to correlate to similar information
Task x.y.b.c - Produce Reports	M7 (KPP) RV	Time to produce reports
Task x.z.c - Display Reports	M8	Time to display information

Figure 8. Integrated evaluation framework template

Figure 8 is a generic template from the integrated evaluation framework (IEF), which is used by the operational test community. The IEF is a tactical-level tool that describes the overarching approach for evaluating the system's operational effectiveness, operational suitability, and survivability. The IEF relies on the MBTD process to link mission-based capabilities in the form of platform mission tasks and subtasks, critical operational issues (COI), and key mission-oriented

response variables or measures (often times key technical parameters) needed to inform operational test evaluations. The MBTD details the missions, tasks, and subtasks to be supported by the system, the conditions under which these elements must be performed, the data required to support performance evaluation, test methods, and test events to be accomplished, and test resource requirements.

4.2 Identify Key Capability Attributes to be Evaluated

The XEF documents the top-level operational measures and technical parameters (also referred to as response variables) that serve as means to quantify system capabilities. This phase of the process is what is considered the first proper step of DOE.

The strategy for this phase is to:

- Identify the response variables associated with the capabilities to be evaluated
- Identify the data type for each response variable
- Draw functional flow block diagrams and mission decomposition diagrams
- Assess the metrology system that will be used to measure each response
- Baseline sources and type of variability that could influence each response
- Identify the expected range of values for each response

The XEF typically relies on capturing technical attributes listed in the DoD architectural framework (DODAF), systems engineering plan (SEP), or the systems requirements document (SRD), which should capture top-level requirements from the JCIDS process, the initial capabilities document (ICD), and the capability development document (CDD). It also includes some of the IEFs measures obtained by decomposing the Navy's required operational capabilities and projected operational environment (ROC/POE) mission threads into tasks and subtasks that are associated with mission-oriented response variables. The attributes and responses from the XEF and IEF make-up the backbone of the T&E strategy and are used to inform decisions or assess progress towards achieving a desired capability. Technical attributes and measures often include:

- *Key performance parameters (KPPs)* - capabilities that, if not demonstrated, are reason for program reassessment or possible termination.
- *Key system attributes (KSAs)* - system or subsystem capabilities with sufficient priority for cost, schedule, or performance insight but not as critical as KPPs.
- *Critical technical parameters (CTPs)* - a subset of the program's technical parameters that describe measurable critical system characteristics that enable the attainment of required operational capabilities when achieved. While programs may have thousands of technical parameters that must be assessed to support specification compliance and system evaluations, not all of them are CTPs. CTPs are linked to operational requirements.
- *Measures of effectiveness (MOEs)* - measures the degree of a system's capability to achieve mission success considering the total operational environment.

-
- *Measures of suitability (MOSs)* - measures the extent to which a system is sustainable and integrates into the operational environment. MOSs include measures for reliability, maintainability, survivability, etc.
 - *Measures of performance (MOPs)* - measures that characterize physical or functional attributes related to the execution of the mission or function. They quantify technical or performance requirements directly derived from MOEs and MOSs.
 - Any other technical performance measures or operational measures required by the CDT or OTA lead to complete analyses of capability areas like interoperability, cybersecurity, reliability growth, maintainability, etc.

The attributes and measures are typically associated with two types of issues that help determine how test results will be judged and enrich decision-making process—critical technical issues (CTI) and critical operational issues (COI). They are normally stated in the form of questions and often require enough associated attributes or measures to adequately determine an answer. As it will be seen in subsequent sections, the more accurately the questions are expressed, the more effective and efficient the overall evaluation process.

CTIs are issues that must be examined in DT&E to evaluate technical parameters, systems characteristics, manufacturing issues, engineering specifications, etc. CTIs help inform how well a system is designed and manufactured when performing mission essential tasks. COIs are issues that must be examined in OT&E to evaluate the capability of a system to perform its mission. COIs help inform how well a system performs its mission under a given set of conditions and how ready, supportable, survivable, and safe the system is to sustain effective performance in combat and peacetime operations. COIs address the overall system's operational capability when examined in realistic operational mission environments and operated by warfighters.

The terms “technical parameters”, “technical attributes”, “response variables”, “output variables”, “evaluation measures”, and “mission-oriented response variables” are used interchangeably to describe the responses that are observed or measured in the test and which are used to inform key knowledge points or decisions; however, they are not interchangeable with factors or conditions. Regardless of the name used, it is important the response variables provide the information needed about the capability to be evaluated. All response variables need to be adequately documented. From this point forward, the preferred terms are response or response variable.

The data type of the response can heavily affect the size of a test design and the quality of its corresponding analysis. Sometimes the response variables are binary in nature (i.e. pass/fail). Other times the response variables are continuous over a range of values. For instance, the objective of an Offensive Anti Surface Warfare (OASUW) test could be to evaluate the Probability of hit (P_{hit}) or the miss distance of a missile. P_{hit} is a binary measure whereas miss distance is a continuous (yet bounded) measure. Test designers often recommend, whenever possible, to use continuous response variables rather than binary response variables because it would produce more informative and more economical designs. The evaluation of a simple test design with off-the-shelf statistical software can show that using a binary response variable over a continuous response variable can result in much more than a 100% increase in sample size (given the same objectives and design parameters) to compensate for the lack of information in a binary response variable. Similarly, the model produced with a continuous response variable will likely be more informative than a model produced with a binary response variable. Additionally, a continuous response variable makes it easier to detect significant shifts in performance due to varying the inputs factors.

Objectives involving multiple types of response variables are commonplace. For example, there are situations in which the interest is to evaluate a function or a capability with both a binary measure and a continuous measure (such as P_{hit} and miss distance as in the OASUW example above). In those cases, the test design should be sized to the binary response.

A useful tool for identifying responses are functional flow block diagrams. A functional flow block diagram, like the generic diagram illustrated in Figure 9, shows the sequential relationships of the functions that must be performed by a system. In other words, it shows how the principal encapsulated functional modules or components of a system (represented by blocks) are decoupled and interconnected. The functions, viewed as individual systems, enable the visualization of an SoS. In both cases, the functional flow block diagrams help understand how the components fit into the whole. Each function can be performed in parallel, or in sequence and can be viewed in terms of inputs, outputs, and interfaces. The diagram shows the sequence of *what* must happen but does not provide any information about *how* the functions are performed. The responses are illustrated with the letter R and the factors and conditions that influence the responses, whether internal or external to the functions, are illustrated with the letters A - L. It is precisely the objective of a test to find the relationship between responses and factors. Functional block diagrams can be built from documents such as DoD architectural framework (DODAF) systems view (SV) four and five. SV-4 describes the functions performed by systems and the flow among those functions. SV-5 maps the systems functions to operational activities.

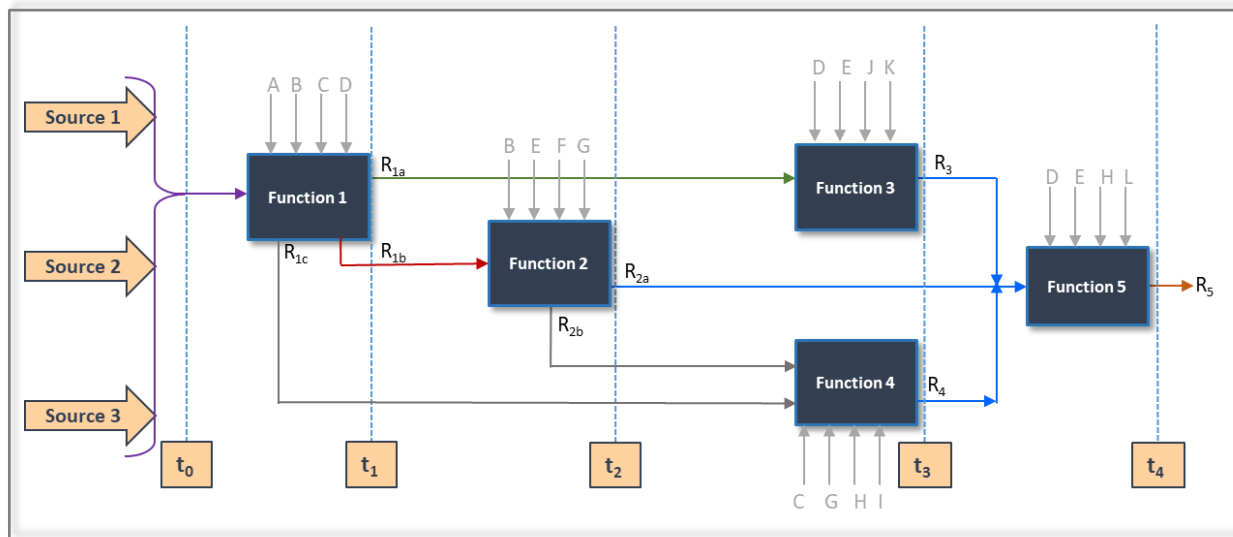


Figure 9. Generic functional flow block diagram

Another useful tool for identifying responses is the decomposition of mission requirements into tasks and subtasks as illustrated in Figure 10. At the platform level, the mission is decomposed into tasks and subtasks and the functions (or their associated responses) are allocated to the subtasks. Often, the responses used to quantify the capability of a system via a functional block diagram are identical to the responses used to quantify capability via PMT views.

It is also critically important to define the metrology system prior to carrying on a test—what, how, where, and when to measure. The methods of measuring, whether directly or indirectly, often produce unwanted variability in the responses. Unwanted variability may have contributions from

the system, materials, parts, or operators. In any case, identifying, separating, and removing sources of unwanted variability leads to improved accuracy in measuring the responses. It is also important to understand the consequences of how measurements are made and the risk the variability in the response creates. The attributes of the measurement system that need to be considered are accuracy, precision, stability, and capacity. Accuracy refers to the closeness of the measured values to a standard. Precision refers to the closeness of the measured values to each other. Stability refers to the degree to which the measured values are robust to wild fluctuations. Capacity refers to the extent to which the measures can be obtained.

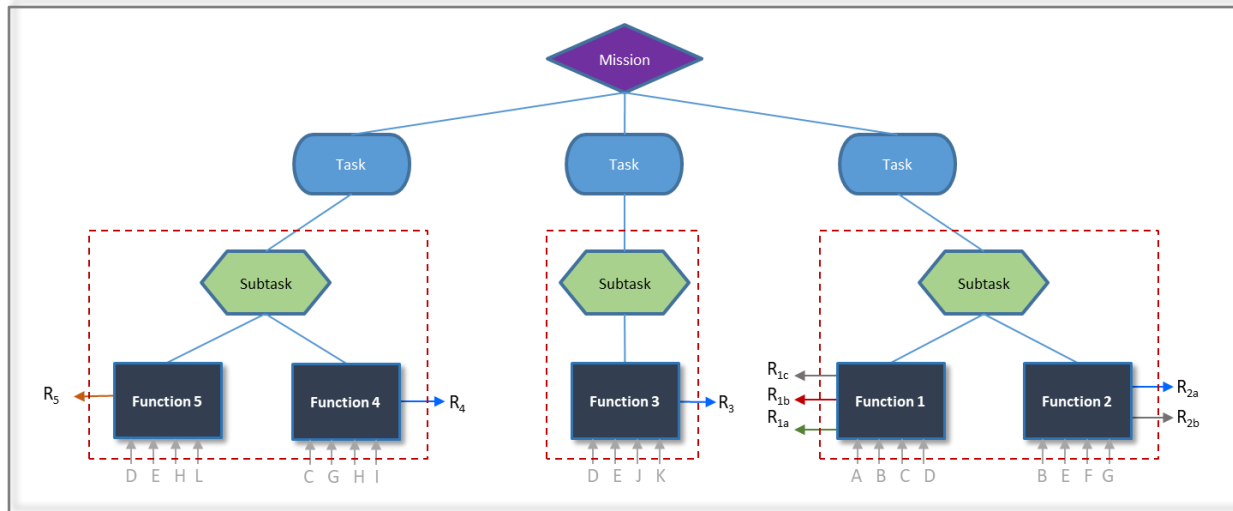


Figure 10. Generic decomposition of mission requirements

Table 6. Strategies for identifying evaluation measures

Strategy	Outcome
Identify the response variables associated with the capability to be evaluated	Leverages the traceability of requirements to produce a comprehensive set of testable and measurable responses and a factor space from which to generate the run matrix
Identify the expected range of values for each response	Helps in understanding the expected distribution of the response variables for different vignettes (such as for easy, nominal, and stressful conditions)
Identify the data type (e.g. continuous, discrete, ordinal nominal) for each response variable	Informs the selection of a test design, its size, and the analysis technique that should be used
Identify the sources and type of variation (e.g. stochastic, deterministic) that could influence each response variable	Facilitates the selection of both a good test design that fits the problem and of a proper execution strategy
Identify and assess the metrology system that will be used—accuracy, precision, stability, and capacity—for measuring each response variable	Verifies measurements have the precision and accuracy required by the objectives and permits the identification and potential removal of variability that otherwise would be built-in in the measurements

4.3 Develop Evaluation Objectives and Scope

Formulating evaluation objectives is the first step in planning a test ([NIST Handbook Section 5.3.1](#)). This step, as fundamental as it sounds, sometimes is undermined because either the overall evaluation path is not completely defined, the type of test required is unclear, there are different perspectives among stakeholders about the evaluation objectives, or it is just too hard to write down the objectives. A well stated objective contributes substantially to a better understanding of the system's capability and to a better evaluation.

The strategy for this phase is to:

- Formulate clear, testable, measurable, and unbiased evaluation objectives of practical consequences
- Develop a sequential evaluation and test framework
- Identify the purpose of each objective: screening, characterization, etc.

A well stated objective should be clear, testable, measurable, unbiased, and relevant:

- Clear – an objective that is concise, detailed, and well defined in terms of achieving desirable outcomes
- Testable – an objective for which a feasible test can be designed
- Measurable – an objective that allows measurement of the outcomes with a number, which is essential for evaluating the maturity of a capability
- Unbiased – an objective that is based on facts and that is not influenced by interpretation, which suggests that the objective should be developed by a diverse group of stakeholders with different viewpoints
- Relevant – an objective that leads to a desired outcome or that aids in a decision

Generally, evaluation objectives are documented in the evaluation frameworks. Sometimes the objectives are derived from a decomposition of mission capabilities into tasks and subtasks. Other times they are derived from the need to verify or validate systems engineering requirements, operational requirements, correction of deficiencies, new or added capabilities, continuing programmatic assessment, M&S capabilities, and tactics, techniques, and procedures (TTP).

Evaluation objectives can be viewed in several ways. One way is to formulate them as state or process objectives. In a state problem, the interest is in the given state of a capability at a time. In a process problem, the interest is in the change of a capability over time. For instance, the interest could be in measuring the influence that several factors have on a response variable (like detection range of a radar system) during a specific test (state problem) or in measuring changes in that response variable from test to test to evaluate the maturity of that capability over time (process problem). Viewing the development of objectives through this lens enables a *sequential evaluation* framework shown in Figure 11 to efficiently integrate successive phases of test (a concept behind IT&E) to immediately learn crucial information about the capability before proceeding to the next phase of test.

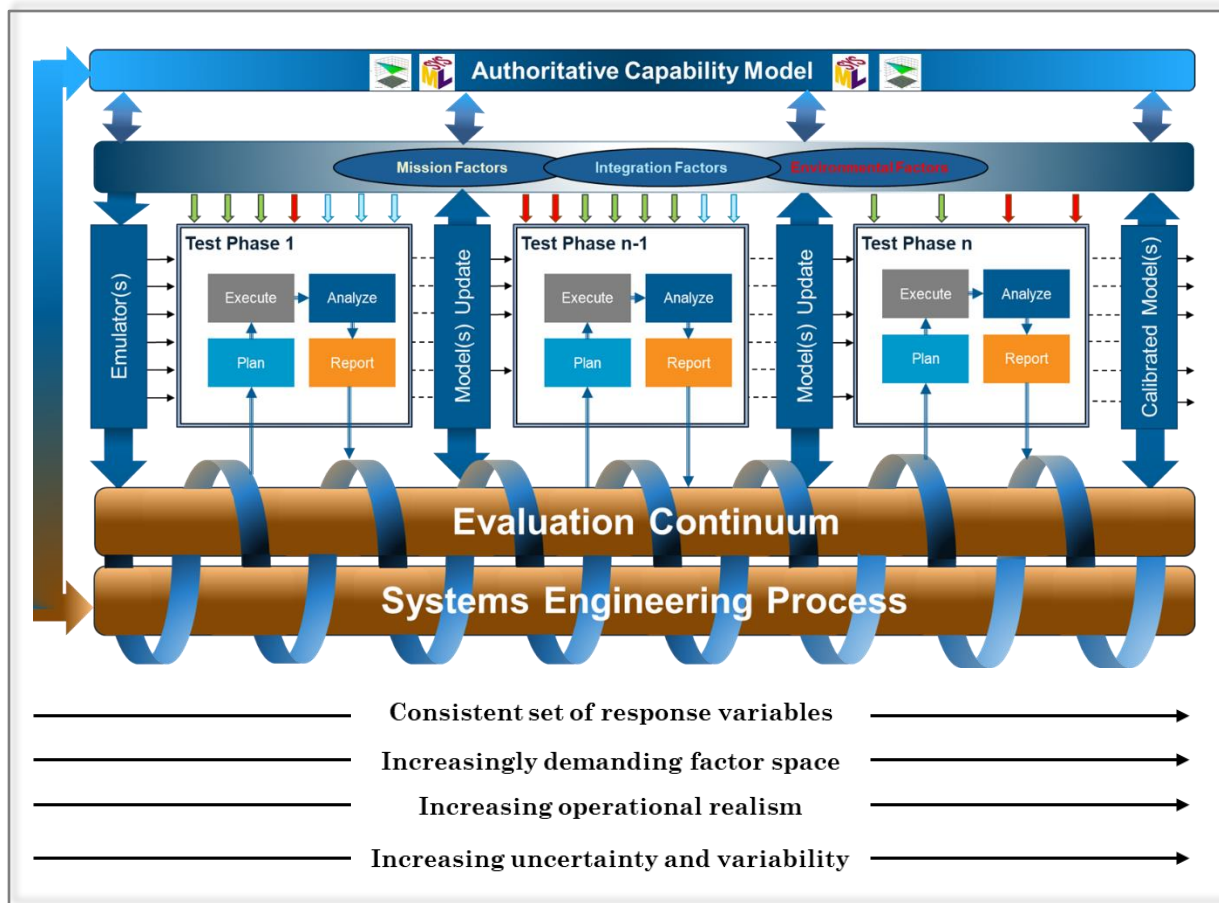


Figure 11. Framework for sequential evaluation and test

Figure 11 shows that as the test phases mature over time, we seek efficiencies and effectiveness by dropping factors from consideration that proved to be not significant, by adding other factors of interest, by adding new response variables, or by executing a *sequential ensemble* of small tests (within each phase) rather than a large test (sequential testing and sequential ensemble of test designs are further discussed in Section 4.5). Regardless of the type of evaluation objective, a best practice is to follow the sequential ensemble strategy.

Figure 11 also attempts to illustrate how a capability model (a.k.a. authoritative source of truth in model-based systems engineering (MBSE) language) matures over time through a series of updates that occur at the end of each phase of a sequence of tests (e.g. DT&E, IT&E, TECHEVAL, IOT&E, CSSQT, or tests for M&S VV&A). The initial capability model could be a mathematical model, a crude physics model, or a DODAF SV coded in systems modeling language (SysML). SysML is an MBSE general purpose language that can be used for the analysis, design, and verification of complex systems using graphical notations as well as to gain efficiencies and effectiveness in other activities of the T&E ecosystem. A consistent set of response variables, critical to the mission, are evaluated through the different phases of test while the factor space (grouped under the headings of mission factors, integration factors, and environmental factors) is adapted to both the stage of development of the capability and test resources (it may not be necessary or practical to follow this approach all the time). The complexity of the factor space

typically increases over time, as shown in the bottom arrows. The factors and conditions are screened to determine their significance (a typical DOE outcome) and to update the model. These model updates are consistent with the Model-Test-Model (MTM) paradigm in which live test results and prior model results can be used to inform the next step in model evolution until the model acceptability criteria can be achieved. Evaluation Continuum represents the integration of XEF and IEF planning steps as well as independent DT and OT evaluations and reporting activities.

Evaluation objectives can also be classified as screening, characterization, comparison, validation, and demonstration. This way of thinking accommodates the selection of test designs discussed in Section 4.5 and the sequential ensemble mentioned earlier. Table 7 provides a synopsis of the objectives and a cross reference to some of the test designs discussed in Section 4.5.

Table 7. Classes of Evaluation Objectives

Evaluation Purpose	Goal	Utility	Example Test Designs (Refer to Section 4.5)
Screening	Identify factors that influence the performance of a system. Practical when there is no historical information about which factors influence performance, such as when testing a system for the first time.	Performance Assessment	Plackett-Burman, Fractional Factorial, Factorial, Definitive Screening
Characterization	Develop a precise response surface or mathematical model of the response(s). Practical when there is a need to have a complete description of the nature of the system. Often involves predicting performance.	Tactical Decision Making Predictive Analytics, TTPs	Response Surface, Optimal
Comparison	Compare legacy and new systems, or different systems exposed to the same conditions systematically varied, or the same system before and after configuration changes under the same conditions systematically varied, or the application of different levels of TTPs.	Analysis of Alternatives Product Improvement	Split-plots
Validation	Ascertain the representation or model of a system is consistent with reality or another scientific model. Validation, to some extent, is a special case of a comparison objective.	M&S VV&A	Response Surface Match + other statistical techniques
Confirmation	Confirm a system is operating consistently over time. This type of objective is practical for regression testing.	Regression Testing	Plackett-Burman, Fractional Factorial, Factorial, Optimal
Optimization	Select factor settings that maximize or minimize the response(s) or that minimize the sensitivity to changes in the response(s) due to operating conditions found in the field.	Trade-off Studies	Response Surface
Demonstration	Show, prove, or provide evidence of a capability. Unlike the previous types of objectives, this type involves a limited number of runs (mostly one) selected from a comprehensive design; however, they are limited in the information they provide.	Proof of Concept	Limited Scope

Table 8 provides strategies that can help in this step of the process. Generally, the formulation of evaluation objectives guides the test design strategy and helps with the selection of responses, factors, and test designs. Often, all these need to be considered almost simultaneously when planning a test. As an example of an objective formulated with the perspective described above in mind, consider a test where the interest is to determine the influence that several factors have on the detection task of an Anti-Air Warfare (AAW) mission. The objective can be stated as “Characterize the capability of the air surveillance radar system to detect threat representative targets (of various radar cross sections, traveling at different altitudes and different speeds) in both clear and electronic attack environments and under operationally representative environmental conditions”.

Table 8. Strategies for formulating evaluation objectives

Strategy	Outcome
Develop a capability-based T&E strategy to inform systems engineering and key acquisition decisions	Produces a test strategy that spans across the entire acquisition lifecycle
Use the XEF and IEF to guide the evaluation strategy	Correlates decisions to be informed, evaluation objectives, and the data needed to inform the evaluation objectives
Develop a sequential test design strategy that maximizes opportunities for IT&E	Enables test efficiencies—one integrated test plan, one integrated test, and independent evaluations
Introduce DOE methodologies as early as possible in the acquisition process	Enables adoption of statistically based test design methods that improve the overall effectiveness and efficiency of T&E
Formulate clear, testable, measurable, unbiased, and relevant evaluation objectives—answer the “what, when, who, where, why, and how” questions pertaining to the objectives	Facilitates the identification of technical and operational parameters that will be used to evaluate the system
Identify the purpose for each objective (screening, characterization, comparison, validation, confirmation, optimization, and demonstration)	Enables the selection of an appropriate test design

4.4 Identify the Test Space (Factors and Levels)

Once the XEF and IEF are developed, developmental and operational testers use DOE methodologies to produce effective and efficient test designs. The XEF lays the decision support context for a capability-based test design while the IEF lays the mission context for a mission-based test design. Those test designs will produce the data to characterize system behavior and combat mission capability across the selected factors and conditions. Since the XEF and IEF already identify the system attributes that will be evaluated, the next step in the process is to define the test space—the correct and complete identification of potential factors that can influence system performance. Factors are also referred to as conditions.

The strategy for this phase is to:

- Identify all potential design factors and all nuisance factors that could influence the response variables, their type, and how to control them
- Use factors that make sense and that are consistent with the evaluation objectives
- Identify potential interactions between factors
- Select factor levels that are practical and that can produce the maximum amount of information with the minimum number of runs
- Identify combinations of factors that are not feasible (constraints and disallowed combinations)
- Establish if blocking or other randomization restrictions are involved

Factors, or conditions, are variables that can take two or more values called levels. Factors are variables that influence the performance of a system or the accomplishment of its mission. Factors can be described in various ways, including by the content of their information, by the variable type used to describe the factor and its levels, or simply by how easy it is to change levels during an experiment. Some of these descriptions are mutually exclusive, which means that factors can be described simultaneously in different ways. The types of factors used in a test can heavily affect the size of a design and the quality of its analysis. Therefore, it is extremely important to identify all factors that could influence the responses, even as background, and to get consensus on which are of interest and to be considered in the test. Factors are classified as potential design factors or nuisance factors, depending on how much interest is on them, as illustrated in the Table 9.

Table 9. Factor definitions

Factor Class	Factor Type	Definition
Potential Design Factors (of interest)	Controllable Factors	Factors for which their levels are systematically varied throughout the test and for which we want to estimate their effects on the responses (i.e. target altitude). Analysis of variance can be used to estimate their effects.
	Uncontrollable Factors	Factors for which their levels are allowed-to-vary naturally and for which we cannot estimate their effects, but for which we can estimate correlation (typically via analysis of covariance). Also called recordable factors (i.e. sea state).
Nuisance Factors (not of interest)	Controllable Factors	Factors for which the levels are set to a specific value and for which there is no interest in estimating their effects on the responses. Their effects are often small. They are referred as held constant factors and as blocking factors (i.e. CIC operator experience level).
	Uncontrollable Factors	Factors for which the levels are not systematically varied and for which there is no interest in establishing their effects on the responses (i.e. ammunition lot).

Typically, potential design factors are factors of interest that may have a significant influence in the outcome of the test while nuisance factors tend to obscure the outcome of the test even though there could be no interest in them. Although nuisance factors can have a large effect on the response, their effect can be blocked (i.e. controlled) through statistical techniques. Often, the objective of a test is to determine the settings of the design factors that: (1) influence the responses;

(2) minimize the variability in the responses; (3) minimize the effects of the nuisance factors; or (4) optimize the response. The relationship between the factors and the responses may be partially known or unknown a priori.

The process for identifying the potential design factors that need to be considered in the test design starts with the formulation of a cause-and-effect diagram, or “fishbone diagram”, as shown in Figure 12. The fishbone diagram organizes those potential factors into the traditional causes of materials, manpower, measurements, methods, mother nature, and machines that can affect the responses.

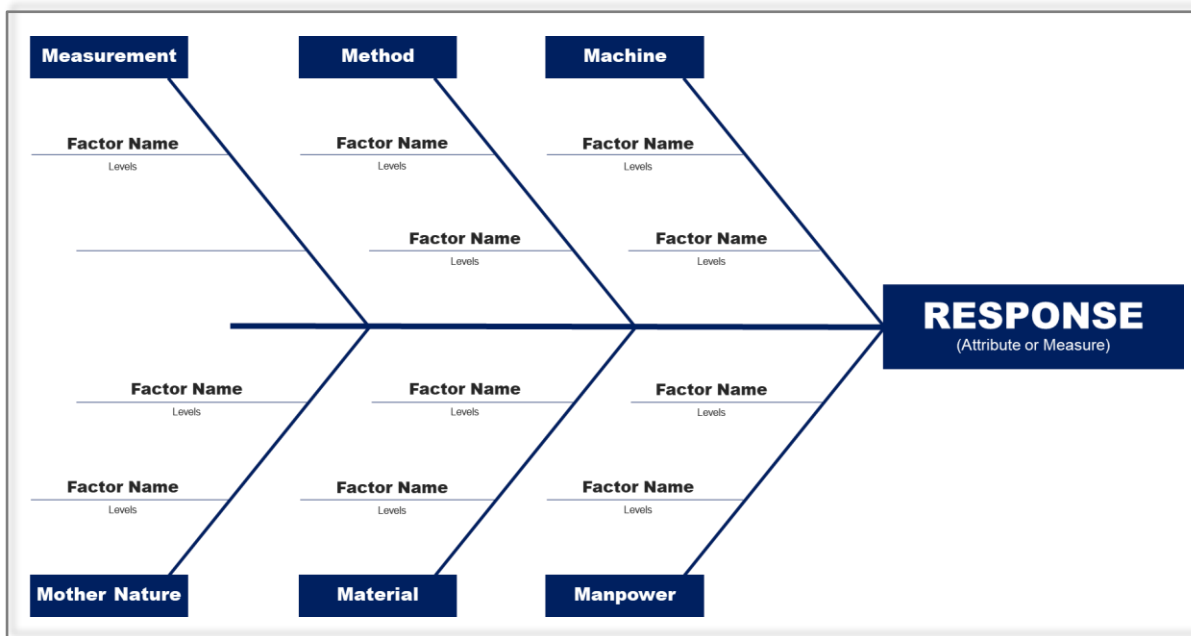


Figure 12. Cause-and-effect diagram for potential factors

Once the potential factors have been identified, the next step is to identify the span or region of interest for each factor (these are known as levels), the test capacity or limits at which the factors can be tested, potential factor interactions, and unfeasible combinations of factor levels. The tasks discussed above are not always easy and may require collaboration of functional and subject matter experts—hopefully as part of the DOE WG.

A good practice is to build a table like Table 10 that correlates the span of the factor space with the available test capacity. The table illustrates the capacity of a test program to test to the maximum of the levels of four fictitious factors. This table can serve as input to the TEMP in terms of the resources needed for test.

Once all the factors and levels have been identified, they are reorganized into the fishbone diagram illustrated in Figure 13. The information in this diagram is the main source of information for creating a test design.

Another way to classify factors is by the content of the data. Factors, as well as responses, can be described as qualitative or quantitative. Figure 14 depicts some commonly used data types, their relationships to each other, detailed description of each type, an example of that type of factor and levels, and the relative amount of information provided by each type of factor.

Table 10. Illustration of test capacity to meet factor requirements

Factor	Operational Space	Test Capacity	Factor Levels	Test Space
Factor A	32 - 125	10 - 140	32 - 125	Suitable
Factor B	20 - 45	0 - 25	20 - 25*	Limitation
Factor C	0 - 90	20 - 95	20* - 90	Limitation
Factor D	5 - 25	5 - 25	5 - 25	Suitable
Factor E	60 - 300	80 - 200	80* - 200*	Limitation

*identifies limitations to the scope of testing

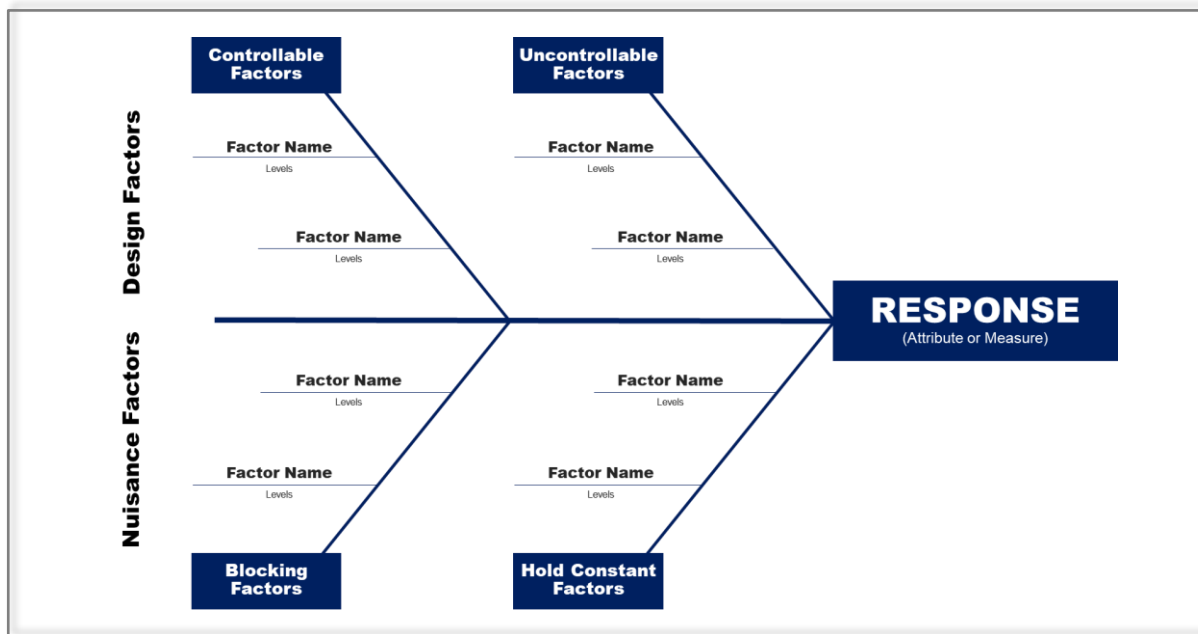


Figure 13. Cause-and-effect diagram organized by factor type

Qualitative factors, commonly referred to as categorical factors, are factors that can be sorted into groups or categories and cannot take all the values within the span of its levels. Categorical factors are considered nominal when they are expressed by natural language or ordinal when they can be ranked in the form of numbers. Categorical factors do not allow for any inference between levels.

Quantitative factors, commonly referred to as numerical factors, are factors that can be measured. Numerical factors are considered discrete when they take values based on counts from a whole set of numbers or continuous when they take any possible value within the span of its levels. Continuous factors are the most informative factors and they allow for inference between levels.

Test designs with continuous factors are more informative and more efficient. To illustrate the richness of information offered by continuous factors over categorical factors, consider the OASUW test in which the objective of determining how three factors—range to target, target speed, and land in background—influence the capability of a system to hit threats. One option would be to consider the response and factors as categorical and the other to consider the response and factors as continuous levels as shown in Table 11.

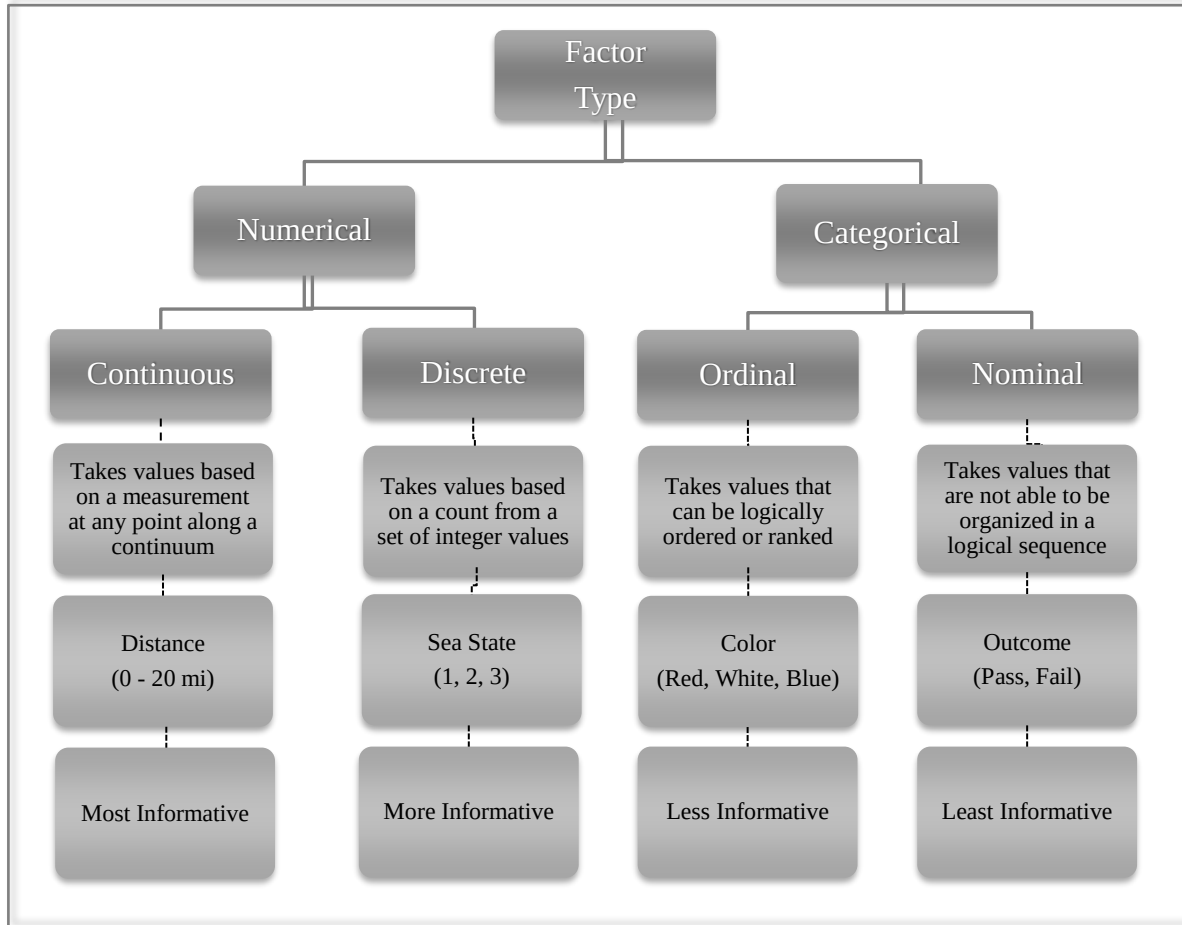


Figure 14. Commonly used factor types

Table 11. Illustration of categorical and continuous factors

Factor*	Factor Descriptor	Categorical Levels (High – Low)	Continuous Levels (High – Low)
A	Range to Target	Medium - Long	25 nm – 45 nm
B	Target Speed	Slow - Medium	10 kts – 20 kts
C	Land in Background	No - Yes	1 nm – 2 nm

* Factors A, B, C are referred to as x_1 , x_2 , and x_3 in Equation 2

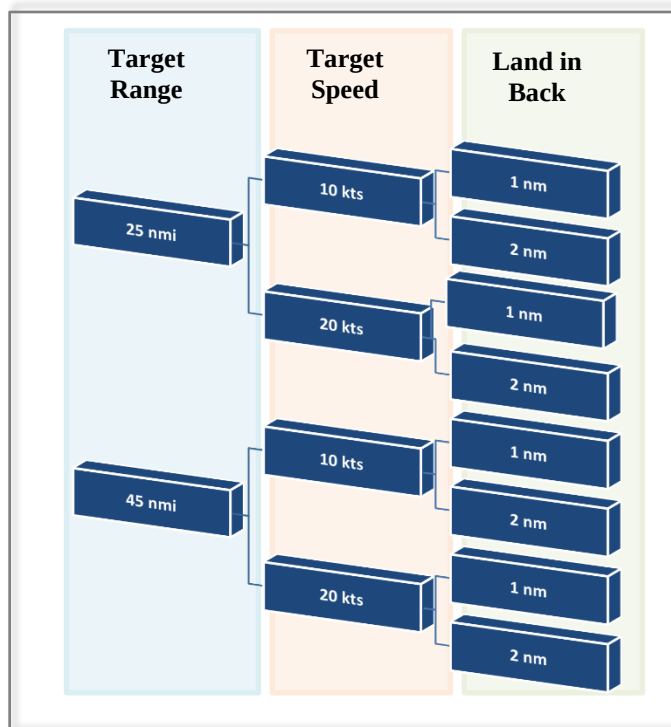
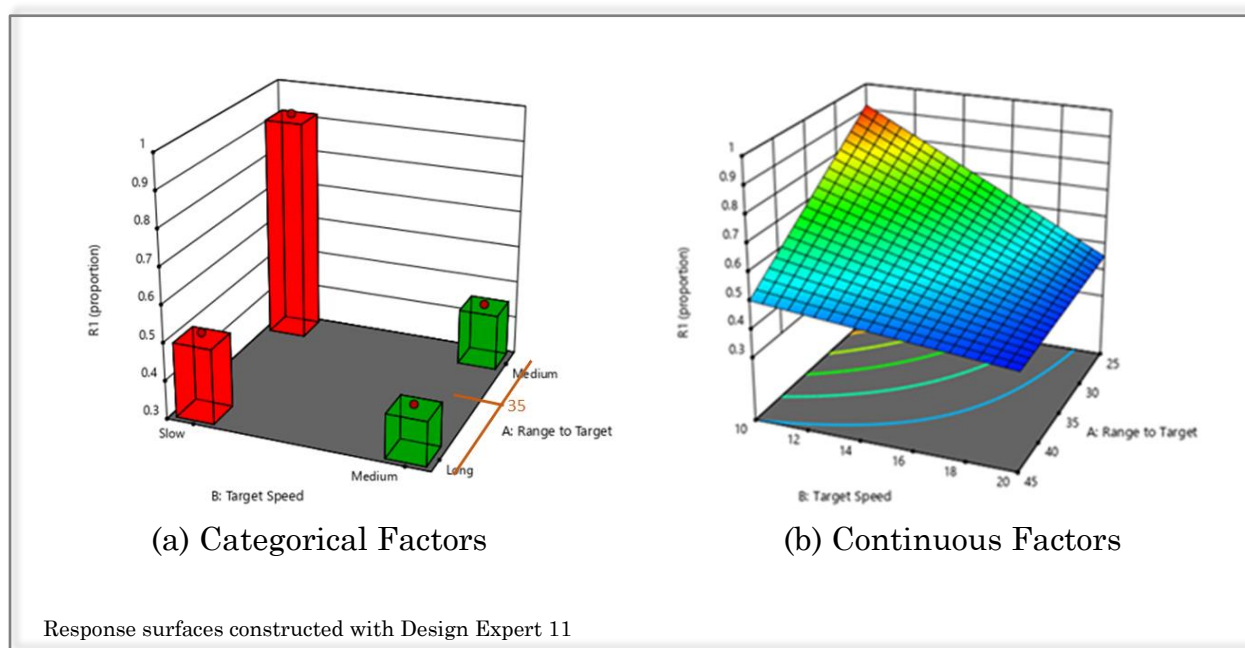


Figure 15. DOE trees

Figure 15 provides a dendritic projection of the factors referred to as the “DOE trees”. In addition to help with the classification of factors, DOE trees show the relationships between feasible combinations of factor levels as a branching or split of a branch. Thus, disallowed combinations are easily recognizable.

For each type of factor in Table 11, the test is (fictitiously) executed, the data is analyzed, and the results are presented graphically in Figure 16. Figure 16 is generally referred to as a response surface—a graph that explores the relationships between factors (in this case A and B) and the response (R1). The design with a continuous response and continuous factors is more informative—it allows for the characterization of R1 throughout the factor space and allows for estimating the response variable in untested regions (i.e. inference between

factor levels) between the high and low levels (such as the center point of the response surface). The reduction in information can lead to an increased difficulty to detect significant effects in the presence of noise, could mask the true underlying system model, and reduce precision often needed for M&S VV&A.



(a) Categorical Factors

(b) Continuous Factors

Response surfaces constructed with Design Expert 11

Figure 16. Response surfaces for categorical and continuous factors

The design with a continuous response and continuous factors is more efficient—it requires a smaller sample size. This is due in part to how the data is modeled—categorical factors require one model term per factor level whereas continuous factors require only one model term for all factor levels. As the number of levels of a categorical factor increases, the more model terms need to be estimated. Two other deficiencies of tests designed with categorical factors (refer to Section 4.5) are that they cannot be augmented to either test for curvature or to estimate higher order effects. Thus, it is more efficient to use continuous factors whenever practicable.

Because the most common objective of a test is to determine the settings of the factors that influence the responses, it is prudent at this point to provide context to the terms effect and interaction. An effect is defined as the change in response produced by a change in factor levels. Effects are referred to as main effects when they are associated with the primary design factors. An effect is mathematically calculated as the difference between the average response at the high-level settings and the average response at the low-level settings of a factor. This simple calculation provides the information of what happens to the response when the factor setting is changed from the one setting to another. The top of Figure 17(a) shows a positive effect of 9 (the difference between settings of factor A from a low average of 1 to high average of 10) while the bottom of Figure 17(a) shows a negative effect of 4 (the difference between settings of factor B from a low average of 8 to high average of 4).

Interactions inform how combination of levels from different factors influence the responses. An interaction occurs when the difference in response between levels of one factor is different at different levels of other factors. Interactions between factors are very common and could contribute significantly to the overall effects on the response. They can place a significant burden on both test design and analysis of data. For planning purposes, the cost of having data available to estimate an interaction is, conceptually, the cost of one run. So, interactions can drive test resources. If they occur and are not considered, the data analysis can produce misleading conclusions. The estimation of interaction effects is identical to the estimation of main effects except that the visualization is slightly different. Figure 17(b) illustrates the case when the effects of the interactions are not present (top) and the case when the interaction is present (bottom).

There are at least two best practices to follow when deciding how many runs should be allocated to interactions. If the underlying system model is understood or subject matter experts have a good understanding of the physical interactions that influence the system, one can estimate only the effects of those influential interactions. If the underlying model is not understood, one can estimate the effects of a small number of interactions following the “sparsity of effects principle”, which states that a system is usually dominated by main effects and a small number of low-order interactions (interactions between a small number of factors, typically not more than three). This is a compromise that is made to reduce the sample size, particularly when the number of samples available is low and the number of potential effects is high.

Now, consider estimating the main effect of factor A for the 2^3 factorial design in Table 11. Figure 18(a) illustrates the typical factorial cube for the design. Table 12 is the design matrix, which is constructed using “+” and “-” to represent the high and low levels for each factor and italic lower case to identify the treatments combinations. The treatment combinations are sorted by the levels (“+” and “-”) of interaction *ABC*. The labels (1), *a*, *b*, *ab*, *c*, *ac*, *bc*, and *abc* indicate the factor levels in a treatment—the high level is represented by the corresponding lower case letter, the low level is represented by the absence of the corresponding letter, and the symbol (1) represents all factors at their lower levels. Each row of the design matrix is then a specific treatment.

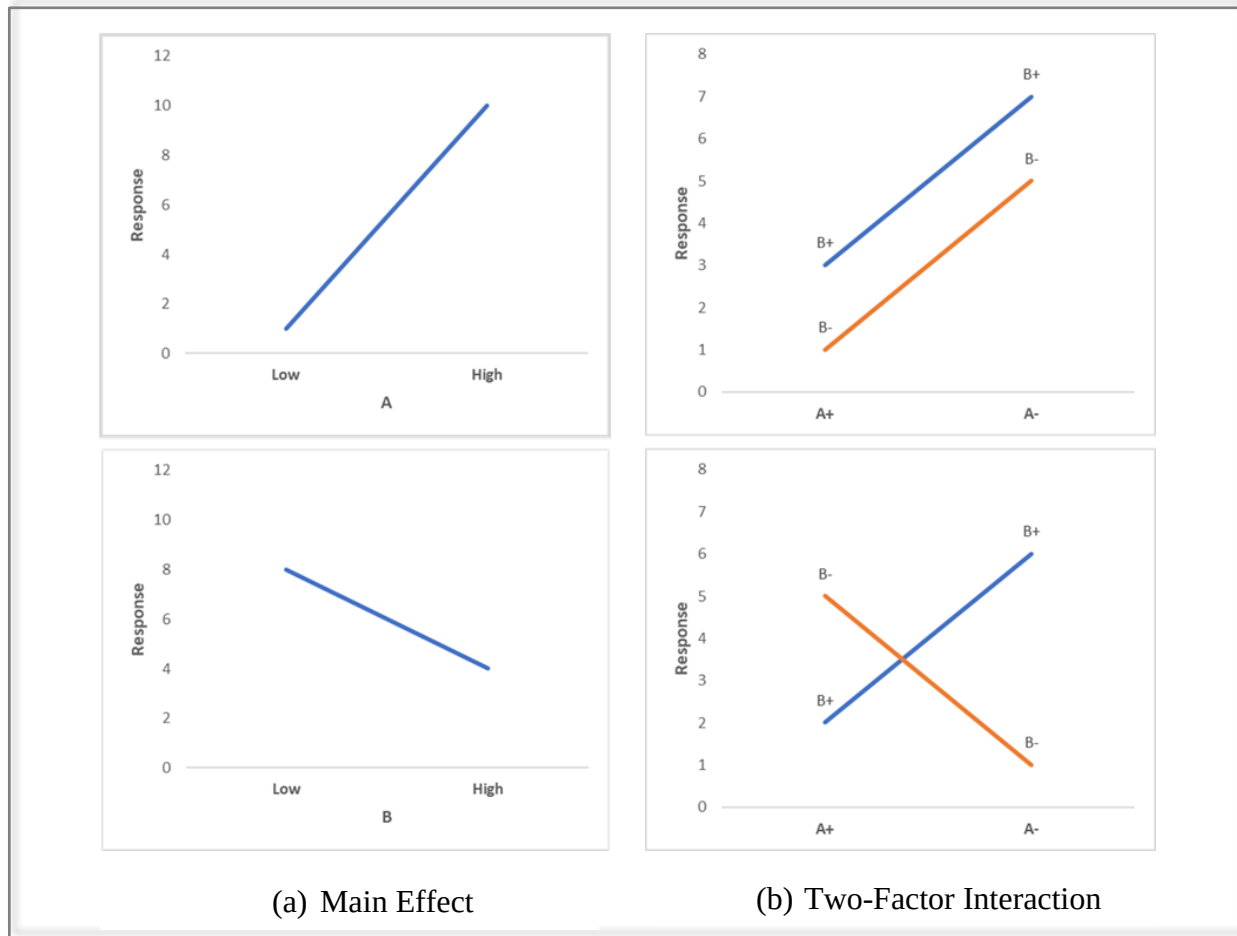


Figure 17. Visualization of main effects and two-factor interactions

Table 12. Design matrix for a 2^3 factorial design

Run	Label	A	B	C	AB	AC	BC	ABC
1	(1)	-	-	-	+	+	+	-
4	ab	+	+	-	+	-	-	-
6	ac	+	-	+	-	+	-	-
7	bc	-	+	+	-	-	+	-
2	a	+	-	-	-	-	+	+
3	b	-	+	-	-	+	-	+
5	c	-	-	+	+	-	-	+
8	abc	+	+	+	+	+	+	+

The highest order interaction, in this case ABC, is referred to as the design generator. Note that the first four rows have a level of “-” while the last four have a sign of “+”. Each of the four combinations with either a “+” or “-” level is a $\frac{1}{2}$ fractional factorial, a key design for gaining

efficiency with sequential ensemble. Each “half” of the fractional factorial could be run independently, and if required, it is augmented with the second half. Combining two separate half fractions of a factorial is called folding.

Table 13 illustrates how to estimate the main effect of factor A for different settings of the factors B and C when taking n measurements for each treatment. Experimental designs with the same number of observations for each treatment are called *balanced designs* ([NIST Handbook Section 5.3.3.1](#)).

Table 13. Effect of A for different settings of factors B and C

Factor B setting	Factor C setting	Effect of A
High (+)	High (+)	$\frac{abc - bc}{n}$
Low (-)	High (+)	$\frac{ac - c}{n}$
High (+)	Low (-)	$\frac{ab - b}{n}$
Low (-)	Low (-)	$\frac{a - (1)}{n}$

Figure 18(b) illustrates the contributors to main effect A. Main effect A is the average of the treatment combinations where A is kept at a high level (gold plane) minus the average of the treatment combinations where A is kept at a low level (blue plane), which can be expressed as:

$$A = \left[\frac{abc + ac + ab + a}{4n} - \frac{bc + c + b + (1)}{4n} \right]$$

The logic for determining the other effects, the two-factor interactions (2FI), and three-factor interactions (3FI) is similar. Figure 18(c) and Figure 18(d) are visualizations of a 2FI and a 3FI. The equations for the effects and interactions do not provide all the information for a complete analysis. While they allow for estimating the magnitude of each effect and interaction, they do not provide information about experimental error.

Factors can also be described by how easy it is to change levels or settings during the test. Many tests involve situations in which the complete randomization of runs may not be feasible because of: (1) factor level settings that are difficult, impractical, or inconvenient to change; (2) limitations in the resources available to complete the test in homogeneous settings; or (3) both. For example, some tests may involve combinations of resources that are rare-to-find as well as easy-to-find. Some tests may require batch or stream processing for some factors but not for other factors. Likewise, some tests involve factors that need to be estimated more precisely than other factors. Similarly, there are tests specifically designed for comparing outcomes from two different test events. Practitioners refer to all these situations collectively as experiments with hard-to-change factors and easy-to-change factors. The use of both factors that are hard-to-change and factors that are easy to change leads to restriction in randomization and to the need of using a type of design called split-plot design.

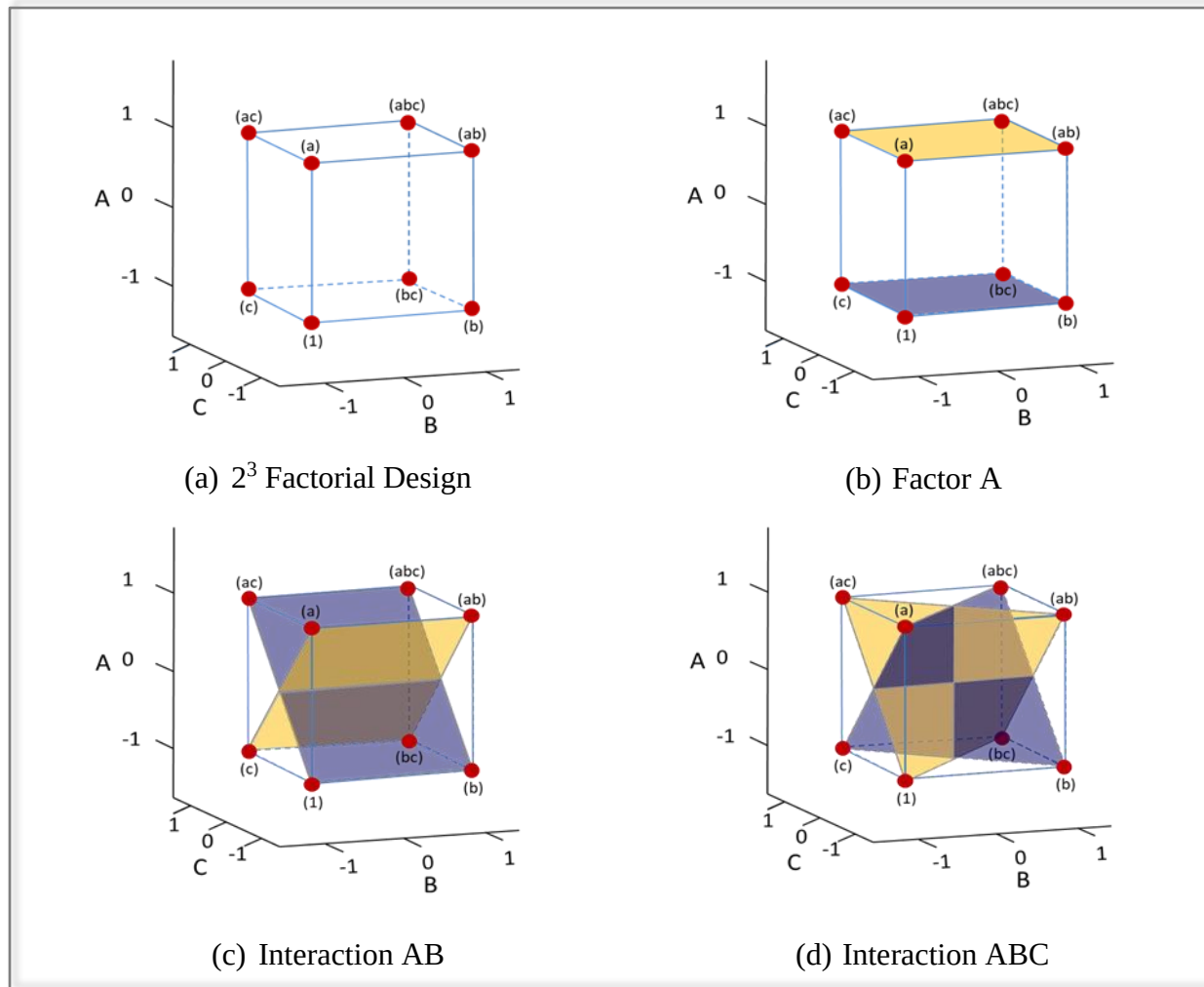


Figure 18. Components of main effects and interactions for the 2^3 design

The terminology used to describe split-plot design comes from agronomic experiments. In those experiments, some factors like irrigation method are restricted to large areas of land called *whole plots*. Whole plots are split into smaller areas of land called *sub-plots*, which allow for the individual application of treatments such as seed variation. Factors associated with whole plots, which are usually considered hard-to-change, are called *whole plot factors*. Factors associated with sub-plots, which are usually considered easy-to-change, are called *sub-plot factors*. One can think of a test that uses a split-plot design as a nesting of two tests—one based on whole plot factors and the other based on sub-plot factors. In theory, the split-plot design is very simple, although it is more complex than a comparable completely randomized design. The most essential characteristic of a split-plot design is that it has two error terms—one associated with whole plots and another associated with sub-plots.

It is important to understand the feasibility of testing specific treatments, or combinations of factors levels, since some combinations may result in undesirable, unsafe, or costly situations. There are some designs (like optimal designs) that are useful for regions involving disallowed combinations.

At this stage of the process, it is useful to prepare an input-process-output (IPO) diagram to capture the relationships between all variables. Figure 19 illustrates a typical IPO diagram.

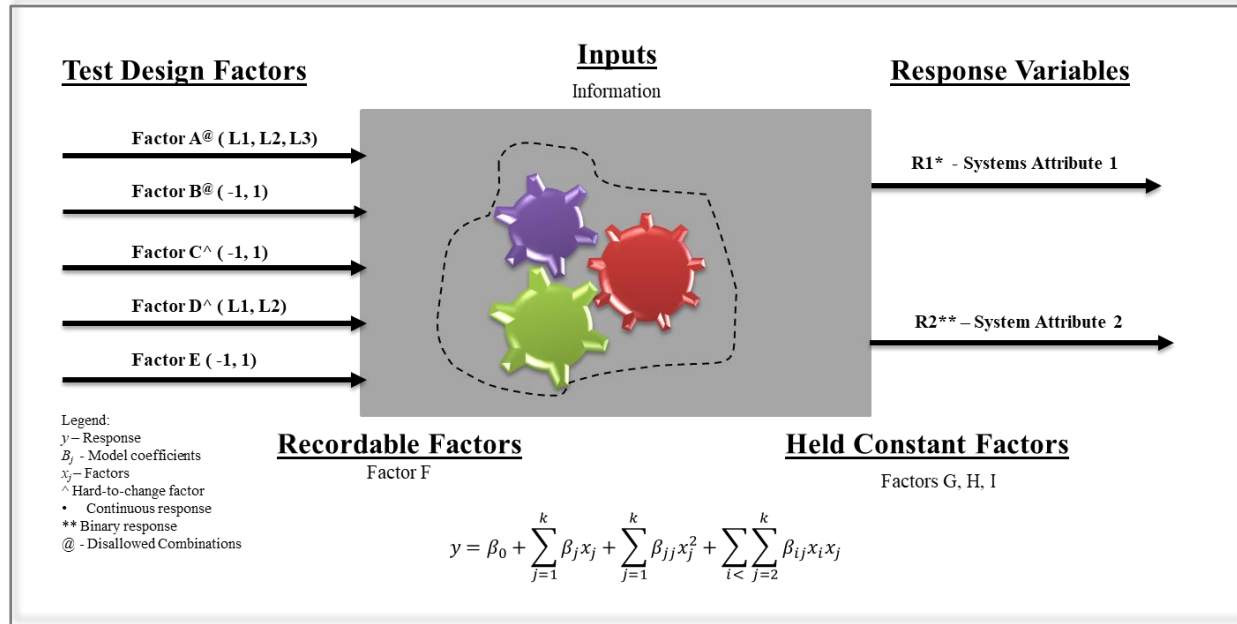


Figure 19. IPO diagram illustration

Table 14. Strategies for identifying the test space

Strategy	Outcome
Use factors that make sense and that are consistent with the evaluation objectives	Maximizes the magnitude and utility of the information for a given test size
Choose appropriate number of levels and settings for each factor	Helps capture the true underlying behavior of the system
Identify the data type (e.g. continuous, discrete, ordinal, nominal) for each factor	Influences the resources needed to conduct a test and the quality of its respective analysis
Understand how all potential factors can affect the response variables of interest— vary factors, fixed factors, blocking factors, nuisance factors, and covariates	Defines the region of test (i.e. operational conditions) and the region of operability
Map the test space – factors and levels—to operationally representative test environments and conditions, in range and size	Ensures data sets produced for each response variable represent the environments the system will encounter during deployment
Maximize the use of continuous factors and minimize the use of categorical whenever possible	Maximizes the amount of information generated while minimizing the resources required
Identify the data type for each factor (e.g. easy-to-change, hard-to-change, hold constant, noise)	Leads to restrictions that influence the type of test design (e.g. split-plot designs) to be selected
Understand the consequences of testing specific combinations of factor levels—costly treatments, debarred regions of space or disallowed combinations, and hard-to-change factors	Ensures the feasibility of the test design execution and influences the type of design that can be selected
Identify the model terms that need to be estimated	Optimizes test design efficiency by removing model terms from consideration

4.5 Construct Test Designs

The fundamental principles of DOE—factorization, replication, randomization, and local control of error (also referred to as blocking)—were formulated 100 years ago. Factorization consists of making intentional, simultaneous changes to the factors that could affect the responses to find and exploit the individual or mutual effect those factors have on the responses themselves. Replication is the application of a treatment (combination of factors levels) to more than one run (i.e. executing a run multiple times) for the purpose of obtaining a valid estimate of the experimental error and a reduction in random error. Replication increases the ability of the design to identify significant factors. In the context used, randomization refers to the random assignment of runs to treatments. Randomization ([NIST Handbook Section 5.3.3.1](#)) averages out the effects of undesirable factors present in the test, generally enables the assumption that the experimental errors are independent and identically distributed random variables, and leads to an unbiased estimate of variance. Local control of error is a technique that allows for minimizing the influence of nuisance factors on the response variables by partitioning the experimental units into homogeneous subsets called blocks. Local control improves the precision of the comparison of factors of interest and reduces or eliminates the component of the variability transmitted from nuisance factors.

DOE is predicated under three fundamental assumptions. First, the model errors are normally and independently distributed random variables with a mean of zero and a variance of σ^2 —often referred to as the normality of the measurements. Second, the variance σ^2 is assumed to be constant at all factor levels—statistically referred to as homoscedasticity or constancy of variance. Third, the implication from the first two assumptions is that the observations are mutually independent and fit the model—often referred to as the independence of the measurements. These assumptions are discussed in the data analysis section.

DOE provides a structured process for effectively and efficiently distributing the test points across the test space, which is one of the issues testers wrestles with the most. A good distribution of the test point across the test space is one of the most important properties of a good test design. There are many different types of designs to choose from ([NIST Handbook Section 5.3.3](#)), each with its own distinguishing features. Some of the designs were developed for specific applications. Considerations that influence the selection of a test design include:

- Nature of the problem to be studied
- Objective of the test
- Number of potential factors and interactions that affect the responses
- Number of runs, including replicates, required vs. available
- Efficiency of the design
- Ease to carry out
- Restricted or debarred regions of interest
- Blocking of influential nuisance factors
- Restriction in randomization

The strategy involved in the construction of a test design is highlighted below:

- Design a sequential ensemble of test designs, if practicable, for each phase of test that is adequate for the event and for the response variable of interest
- Decide on the desired confidence ($1 - \alpha$)
- Decide on the change (δ) that is important to detect for each response
- Estimate the standard deviation (σ) for each response variable
- Choose a model that describes the relationship between factors and response variables
- Select and set-up a set of candidate test designs while considering situations such as hard-to-change factors, constraints, disallowed combinations, and number of runs per test period (blocking)
- Evaluate the candidate designs and compare them based on figures of merit such power ($1 - \beta$), sample size, aliasing among effects, prediction variance, and design efficiency
- Select the design that fits the situation and determine if the sample size is adequate to meet program objectives (may have to negotiate for additional samples, test time, or additional budget)
- Generate a run matrix specifying the number of replicates and blocks

It is common practice to identify many potential design factors during the early stages of planning, particularly for systems that have never been tested. Because the number of runs or treatments increases geometrically with the number of factors, despite of its possible difficulty it is more practical to “ensemble” and carry out a sequence of smaller tests (rather than conducting a large test) and combine their runs to facilitate the estimation of factor effects and interactions of interest. In those cases, each smaller test is designed for a specific purpose and serves as a “steppingstone” for the next phase of the test. This capability to conduct a *sequential ensemble* of tests is an extremely important and useful feature, especially when resource competition is an issue or when dealing with complex systems.

Sequential ensemble of test designs, also referred to as *strategy of experimentation*, is a standard practice in many process industries that creates efficiencies by using a sequential build-up of smaller tests, each with specific objectives, rather than conducting a large single test. Sequential ensemble involves combining knowledge gained during testing, particularly on how the factors influence the responses, with knowledge of experiment design augmentation techniques. Although there are several ways to describe the phases of sequential ensemble, the goal is the same—to create an effective and efficient test design. For this document, sequential ensemble involves four phases—screening, characterization, optimization, and confirmation. Those phases, described below, are consistent with the objectives listed in Section 4.3. Figure 20, prepared with Design-Expert®, illustrates the approach to sequential ensemble and the product from every phase. Note the progression building a response surface and the output model from every phase.

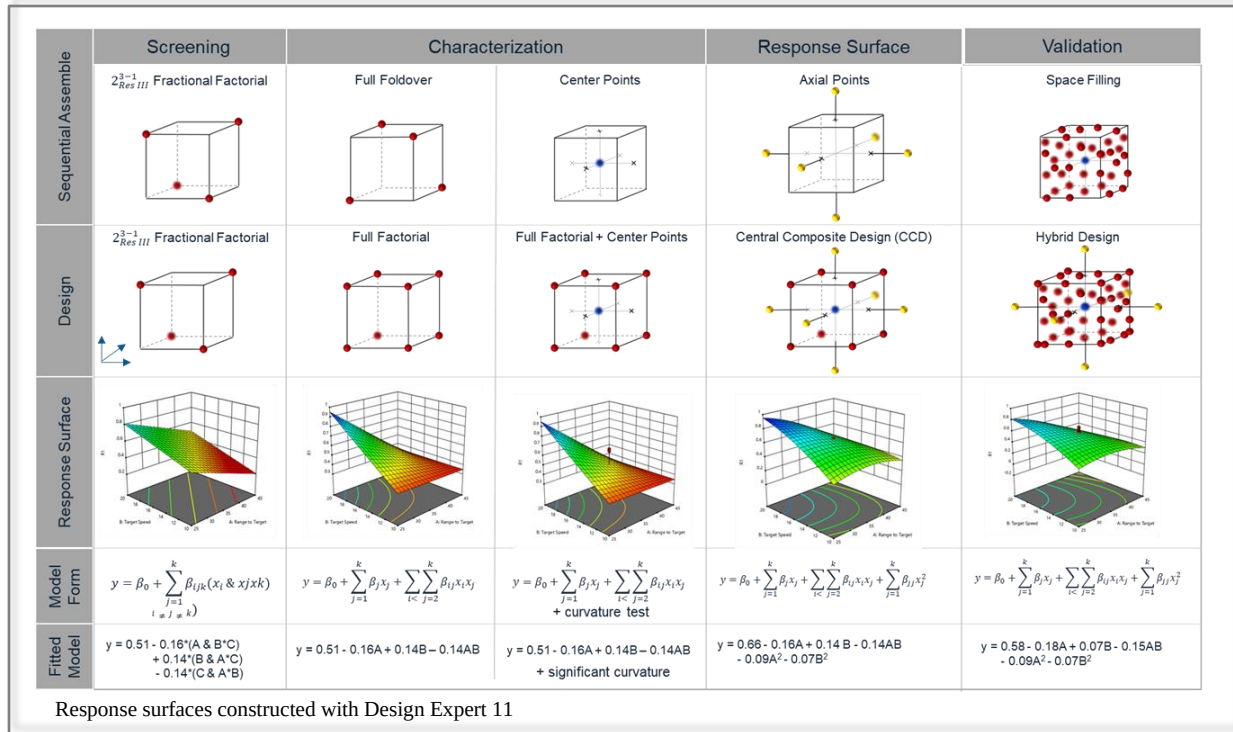


Figure 20. Sequential ensemble of test designs

The first phase of the sequential ensemble, screening, typically consists of running fractional factorial designs (NIST Handbook Section 5.3.3.4) to identify the factors, perhaps from a large number of potential design factors, that have large effects on the responses of interest. Fractional factorial designs are constructed by systematically choosing a fraction of runs from a full factorial design (NIST Handbook Section 5.3.3.3) to efficiently start the process of determining the relationship between the factors and responses. Fractional factorial designs are described mathematically as “ 2^{k-p}_{Res} ” where k refers to the number of factors, $1/2^p$ refers to the fraction, 2 represents the number of levels, and Res refers to the ability of the design to reveal main effects and two-factor interactions (2FI). For example, a 2^{5-1} is a $1/2$ fraction of the 2^5 full factorial that requires only sixteen runs. The use of fractional factorial designs is also based on three key concepts: resolution, sparsity of effects principle, and the projection property.

Resolution (NIST Handbook Section 5.7) is one of the key concepts in the use of fractional factorials. It refers to the degree in which treatment effects (main effect or interactions) are contaminated or confounded with the influence of other treatment effects. For example, resolution III indicates that the main effects are confounded with 2FIs (i.e. it is impossible to differentiate between the main effects and 2FIs) and resolution IV indicates that the main effects are confounded with 3FIs and 2FIs are confounded with other 2FIs (i.e. it is impossible to differentiate between the main effects and 3FIs and between 2FIs). Understanding how to use this feature is extremely useful and can lead to economical designs, particularly when there is an understanding of 2FIs and higher order effects. For instance, the fractional factorial design on the left-hand side of Figure 21 has the confounding structure shown on the right-hand side of the figure, which indicates that instead of estimating main effects A, B, and C the design is estimating the combinations of effects

$A + BC$, $B + AC$, and $C + AB$. However, if any of those interactions did not exist or was negligible, the design would have produced the corresponding main effect.

The projection property states that every fractional factorial design of resolution R contains (i.e. can be projected into) full factorial designs in any subset of $R-1$ factors, as illustrated in Figure 21 for a 2_{III}^{3-1} fractional factorial design. Note that the points of the 2_{III}^{3-1} fractional factorial design correspond to runs 2, 3, 5, and 8 of Table 9, which is considered the principal fraction of the design (the other four points are called the alternate fraction). The projection property is useful when some factors are determined to be not significant, as illustrated in Figure 22. In Figure 22, five factors (A through E) are screened with a 2_{V}^{5-1} fractional factorial design (one-half fraction of a full factorial) and one factor (D) was determined to be not significant. Thus, the 2_{V}^{5-1} fractional factorial can be projected into a 2^4 full factorial design.

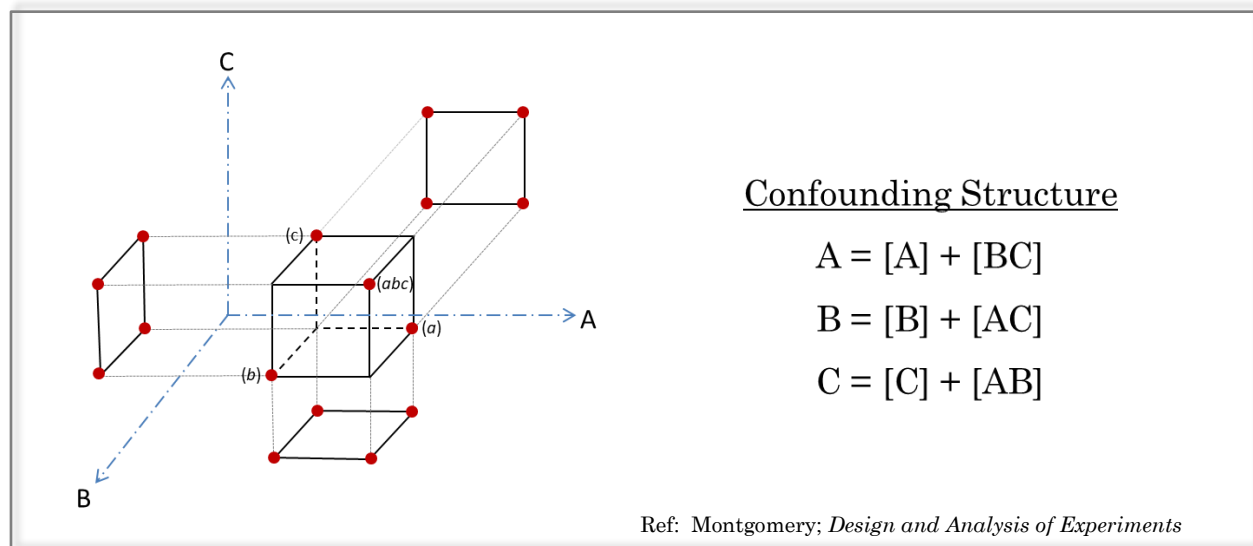


Figure 21. Projection and confounding of 2_{III}^{3-1} main fraction

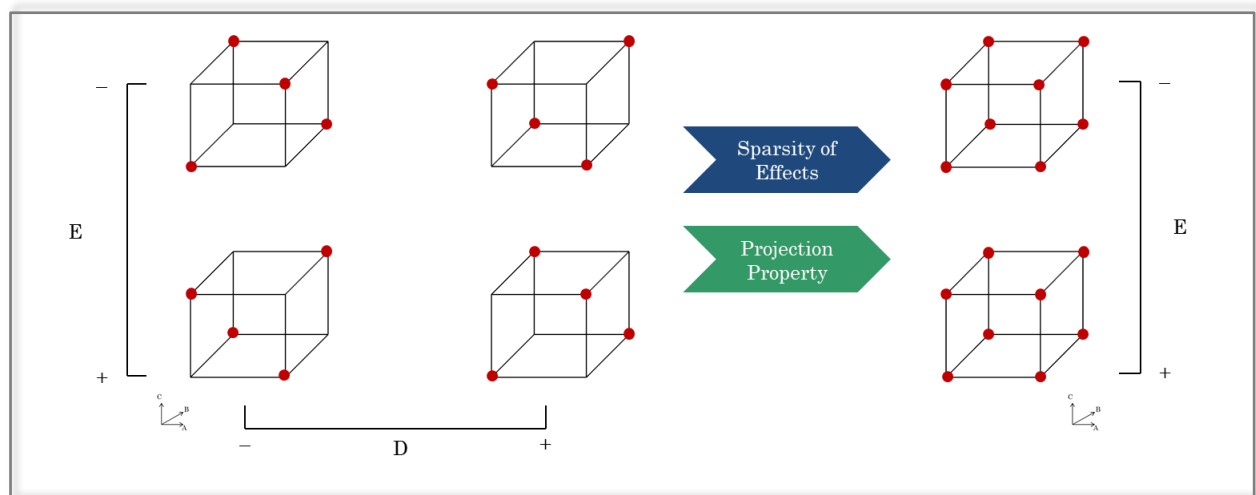


Figure 22. Sparsity of effects and projection properties

The sparsity of effects principle states that system or process is likely to be driven primarily by some of the main effects (effects of combined factors) and low-order interactions. The significant factors are varied in subsequent tests while the insignificant factors are held constant.

The second phase of the sequential ensemble, characterization, typically consists of estimating the significant factors and interactions as well as establishing the presence of curvature on the response surface. Several methods are used to estimate the main effects and interactions:

- *Full foldover* ([NIST Handbook Section 5.3.3.8](#)) – also referred to as mirror image foldover designs, it breaks the confounding between all main effects and 2FIs by combining the runs from the main and alternate fraction, so it doubles the number of runs in order to increase the resolution of the design. It produces a full factorial design. Typically, this method is used when the original fractional factorial design was highly fractionated.
- *Partial foldover* ([NIST Handbook Section 5.3.3.8](#)) – also referred to as a semifold designs, it breaks the confounding between some main effects and interactions by combining the runs from the main fraction with half of the runs from the alternate fraction, so it increases the number of runs by 50% to ensure that some main effects and its 2FIs are clear of other main effects and 2FIs. This method is typically used for resolution IV fractional factorial designs where there is interest in clearing 2FIs involving a specific factor. The product is still a factorial design.
- *Single factor foldover* – it breaks the confounding between a main effect and its interactions and other main effect factors and their corresponding interactions by combining the runs from the main fraction with only runs at different setting of the factor of interest and its interactions. This method is typically used for resolution IV fractional factorial designs where there is interest in clearing a main effect and its 2FIs. The product is still a factorial design.
- *Factorial optimal* – it breaks the confounding between selected main effects and 2FIs by applying an algorithm that chooses additional runs based on the specific factorial model selected. The resulting design will remain a factorial design.

To test the presence of quadratic and higher order effects, the designs are augmented with center points ([NIST Handbook Section 5.3.3.7](#)) and then evaluated for lack-of-fit, which explains how well the model fits the observations (or the variation due to higher order terms not included in the model). Center points also serve to estimate pure error, estimate lack of fit, obtain a more uniform estimate of prediction variance (for response surface designs), and verify the stability of the system.

There are situations, however, in which sequential ensemble is impractical or where a comprehensive experiment is more appropriate for screening factors. In those cases, there are a variety of test designs that are useful. The most popular are the full factorial design ([NIST Handbook Section 5.3.3.3](#)) and the optimal design ([NIST Handbook Section 5.5.2](#)). Other classes of designs include Plackett-Burman ([NIST Handbook Section 5.3.3.5](#)), Oehlert and Whitcomb minimum run resolution IV (MR Res IV), irregular fraction, and general factorial designs. Those designs can be generated with the aid of commercial off-the-shelf statistical software packages.

One of the objectives in DOE is to fit a model to the observations to reveal the underlying mechanism involving system characteristics and operating conditions. These types of models are called empirical models. DOE incorporates linear regression techniques to fit low-order Taylor

series approximations to model the relationship between a response variable and one or more factors. The following Taylor series approximation is commonly used to fit the observations and create a linear regression model for main effects and interactions for a completely randomized design:

$$E(y) = \beta_0 + \sum_{i=1}^k \beta_i x_i + \sum_{j<i}^k \sum_{i=2}^k \beta_{ji} x_j x_i \quad (2)$$

Here, y represents the response, the x 's represents the independent variables or factors, and the β 's represents the regression coefficients. The model error, ε_{ij} , is assumed to be an independent and identically distributed $N(0, \sigma^2)$ random variable and the variance σ^2 is assumed constant throughout the experimental region. The random error incorporates measurement error, variability from uncontrolled factors, variability in the experimental units, general background noise, etc. For k factors, the number of model terms, including the intercept, is:

$$t = 1 + 2k + \frac{k(k-1)}{2}$$

Generally, $t + 1$ is the minimum number of runs required to fit the model.

It is also common practice to use a shorthand notation and substitute the x 's in equation (2) by alphabetic characters and coded factor levels to describe the model. For example, the main effects and two-factor interaction model for $k = 3$ is:

$$y = \beta_0 + \beta_1 A + \beta_2 B + \beta_3 C + \beta_{AB} AB + \beta_{AC} AC + \beta_{BC} BC \quad (2a)$$

In this case, the β 's represents the regression coefficients while A, B, and C represent the main effects associated with those factors (also illustrated in figures 13 and 14), and AB, AC, and BC represent the two-factor interactions. When the equation is expressed in terms of coded factors, the high level of the factors is coded as +1 and the low levels of the factors are coded with -1. In coded factors, the equation is useful to compare the relative impact of the factors by just looking at the model coefficients since the intercept is at the center of the design space. When the equation is expressed in terms of actual factors, as also for equation (2), the equation is scaled to accommodate the levels in the original units as specified for each factor. The intercept, β_0 , is the overall average response of all the runs and the coefficients are adjustments around that average based on the factor settings.

The Taylor series approximation in equation (2) takes the following form to fit observations from a split-plot design:

$$E(y) = \beta_0 + \sum_{i=1}^p \beta_i z_i + \sum_{j<i}^p \sum_{i=2}^p \beta_{ji} z_j z_i + \sum_{i=1}^q \gamma_i x_i + \sum_{j<i}^q \sum_{i=2}^q \gamma_{ji} x_j x_i + \sum_{j=1}^p \sum_{i=1}^q \gamma_{ji} z_j x_i \quad (3)$$

Now x_q represents the vector for the q sub-plot factors, z_p represents the vector for the p whole plot factors, and $k = p + q$. The β 's represent the regression coefficients for whole plot linear terms, and γ 's represent the coefficients for sub-plot linear terms and whole plot by sub-plot interaction terms. The selection, execution, and analysis of a split-plot design may not be a straightforward process as for a completely randomized design.

The third phase is the response surface ([NIST Handbook Section 5.3.3.6](#)) phase. If the curvature is significant, the first-order model (for main effects and interactions) shown in equation (2) may become inadequate to represent the response surface and a higher-order model is needed. Then, the screening designs are augmented to produce a new design that ultimately produces a response surface model that accurately represents the entire design space and can be used to determine the combination of factors levels that optimize the responses. Blocking may be required to eliminate the effects caused by nuisance factors. Typical designs for this phase are response surface designs such as the central composite design (CCD) ([NIST Handbook Section 5.3.3.6.1](#)), the Box-Behnken design (BBD) ([NIST Handbook Section 5.3.3.6.2](#)), the 3^k design ([NIST Handbook Section 5.3.3.9](#)), and the I-Optimal design ([NIST Handbook Section 5.5.2](#)), which are illustrated in Figure 23 for perspective.

For each type of design listed on top, Figure 23 shows the DOE typical cube diagram that represents the test space for three factors, scatter plot projections that represent the coverage of the design, the prediction variance profile that represents the average prediction precision as a function of the test space, and correlation plots that represent the quality of the model term coefficients. Note that the scatter plots are identical for all designs, the prediction variance profile is better for the 3^k design (although it requires the larger number of test points), the most economical designs are the BBD (although it avoids the corner points) and the I-Optimal (although there is more correlation between the model terms as shown by the larger number of squares in the correlation plot with $|r| > 0$). Thus, in conclusion Figure 23 reveals that: (1) none of the designs meet all of the criterion from Table 11—for a design to be good, it must balance the features described in the table; (2) it is important to select a design that fits the problem; and (3) traditional response surface designs can be modified to fit specific needs while preserving desirable properties.

Response surface designs allow for fitting the data to a second- or higher-order models. The second-order models for completely randomized designs is:

$$E(y) = \beta_0 + \sum_{i=1}^k \beta_i x_i + \sum_{j<i}^k \sum_{i=2}^k \beta_{ji} x_j x_i + \sum_{i=1}^k \beta_{ii} x_i^2 \quad (3)$$

Here, the y , x 's, and β 's represent the same as for the main effects and interactions model. The second-order models for split-plot designs:

$$E(y) = \beta_0 + \sum_{i=1}^p \beta_i z_i + \sum_{j<i}^p \sum_{i=2}^p \beta_{ji} z_j z_i + \sum_{i=1}^p \rho_{ii} z_i^2 + \sum_{i=1}^q \gamma_i x_i + \sum_{j<i}^q \sum_{i=2}^q \gamma_{ji} x_j x_i + \sum_{j=1}^p \sum_{i=1}^q \gamma_{ji} z_j x_i + \sum_{i=1}^q \theta_{ii} x_i^2 \quad (4)$$

Here, x_q , z_p , k , β 's, γ 's represent the same as for the split-plot model for main effects and interactions, ρ 's represent the coefficients for whole-plot pure quadratic terms, and θ 's represent the coefficients for sub-plot quadratic terms.

There are situations in which scarce resources—funds, time, material, manpower, equipment—make it impractical to allow the use of traditional designs for fitting second-order models, especially when the number of factors k is high. For those situations, small response surface designs could be attractive since they are more economical. Appendix B provides lists of traditional and small response surface designs.

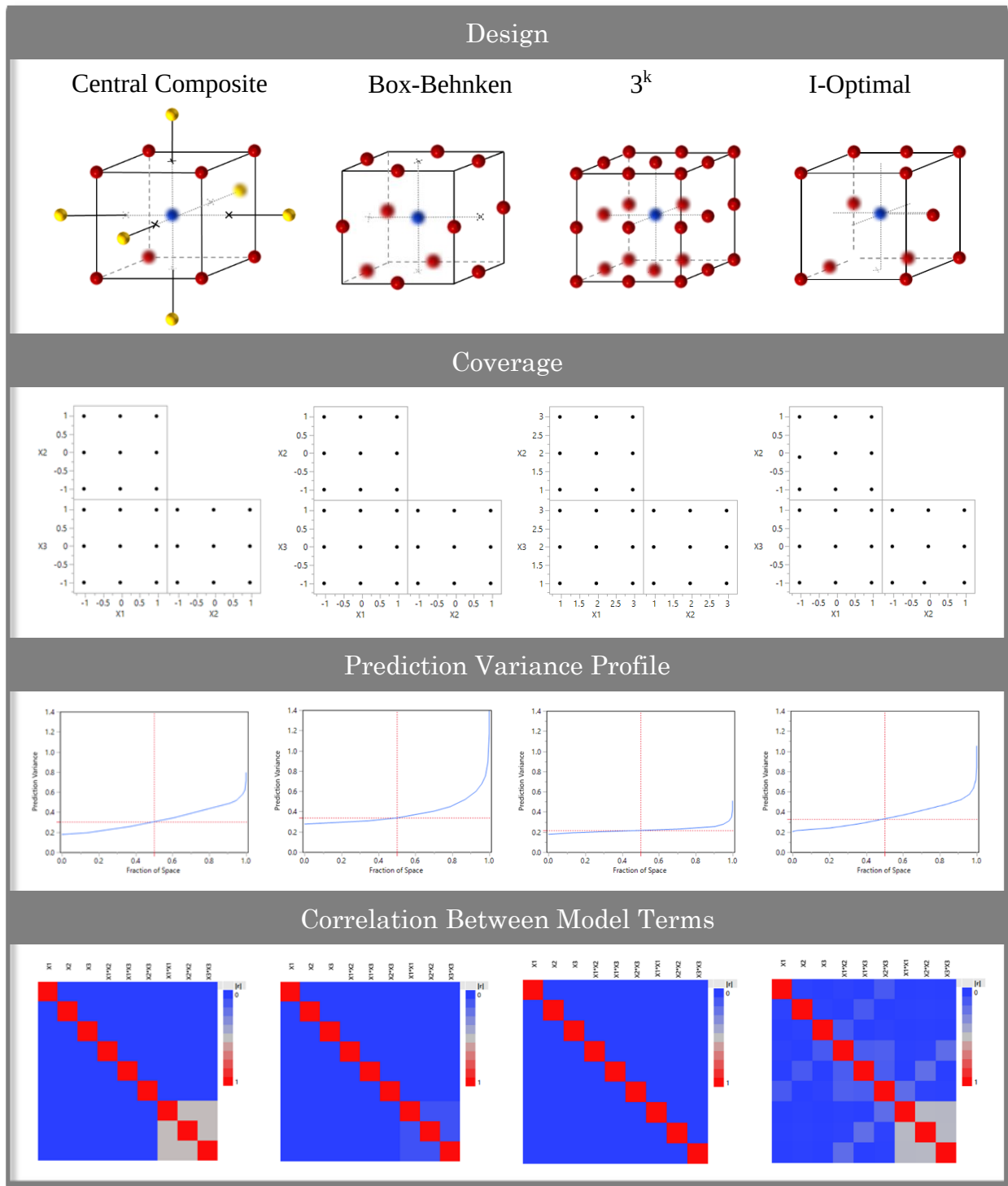


Figure 23. Some features of traditional response surface designs

There are four figures of merit that interact when selecting a test design—confidence, power, statistical signal-to-noise ratio, and sample size. Figure 24 provides a summary. Confidence and power are associated with two measures of statistical risk defined as Type I and Type II errors

(denoted by the Greek letters α and β). Statistical signal-to-noise ratio is associated with the strength of the response signal that needs to be detected in the presence of system noise (denoted by the Greek letters δ/σ). Sample size is the number of runs required by the test given α , β , and δ/σ . Balancing those figures is a critical part of the test design process.

Confidence ($1 - \alpha$) is the probability of not committing a Type I error, or the probability of correctly determining that a factor does not influence the response. Type I error (α) is the risk of declaring that a factor influences the response when it does not (also called probability of false alarm or false positive rate). Confidence is always set before beginning the test and defines the critical value where one will conclude that there is a change in the response attributed to factor effects. For instance, if confidence is set to 80%, there is a 20% chance (denoted as the area α in the center box of Figure 24) that one will conclude there is a change of size δ in the response when in fact there is none. Operational test agencies tend to size the design for a confidence of 80%.

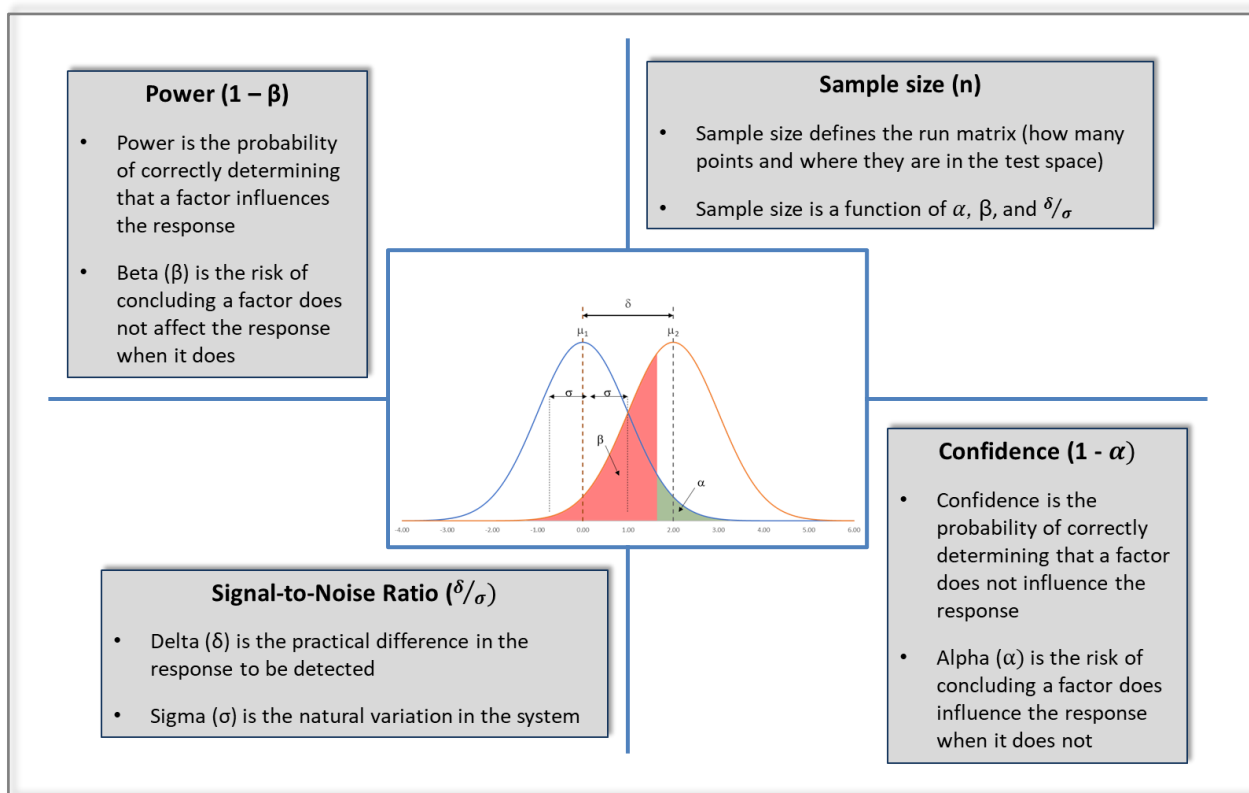


Figure 24. Tradeoff space in selecting a test design

Power ($1 - \beta$) is the probability of not committing a Type II error, or the probability of correctly determining that a factor influences the response. Type II error (β) is the risk of declaring that a factor does not influence the response when it does (also called probability of missed detection or false negative rate). Obviously, the most concerning risk is β ; therefore, extra effort is placed in minimizing β . Power is a measure of how likely a statistical test will detect a change of size δ in the response. For instance, a power of 80% means that there is a 20% chance (denoted as area β in the center box of Figure 24) that one can say there is no change of size δ in the response when in fact there is a change. Power can be increased by increasing the number of replicates. Operational test agencies tend to size the design for a power of at least 80%.

Statistical signal-to-noise ratio (δ/σ) is the ratio of the change in each response variable (δ) that is important to be detected divided by the estimate of the experimental error (σ) for each response variable. An estimate of δ/σ to size the design can be obtained from historical data, an analysis of a similar test event, dry runs, repeatability studies, or from experience.

Sample size (n) depends on confidence ($1 - \alpha$), signal-to-noise (δ/σ), and power ($1 - \beta$). For a desired level of confidence and power, smaller δ/σ require larger sample sizes and vice versa.

As a numerical illustration, consider Table 15 and Table 16 to show the tradeoff and the risks associated with sizing a design for the example in Table 11 using continuous factors and a D-Optimal design for screening main effects and 2FIs. Alpha is set to 0.05 and the objective is to produce a design with a power of 80% or higher.

Table 15 shows changes in sample size required to maintain a power higher than 80% as a function of δ/σ . As δ/σ decreases, the sample size increases dramatically. This could be attributed to the need to detect a smaller shift in the response, larger variability in the response, or both. Table 16 shows the decrease in power for the same sample size as a function of δ/σ . As δ/σ decreases, power decreases dramatically. This means that the design cannot detect the effects caused by the factors on the response, which is a Type II statistical error—the error we want to guard against the most. It is obvious that not estimating δ/σ correctly can present a significant risk to any program since it can result in an incorrect identification of resources that could be difficult to correct later. Table 15 and Table 16 were created with the JMP software package.

Table 15. Sample size required to maintain 80% power as a function of δ/σ

$\delta/\sigma = 2$	$\delta/\sigma = 1$	$\delta/\sigma = 0.5$																																																																								
N = 13	N = 34	N = 128																																																																								
Power Analysis Significance Level 0.05 Anticipated RMSE 1 <table> <tr> <th>Term</th><th>Anticipated Coefficient</th><th>Power</th></tr> <tr><td>Intercept</td><td>1</td><td>0.813</td></tr> <tr><td>Range</td><td>1</td><td>0.838</td></tr> <tr><td>Speed</td><td>1</td><td>0.813</td></tr> <tr><td>Background</td><td>1</td><td>0.813</td></tr> <tr><td>Range*Speed</td><td>1</td><td>0.813</td></tr> <tr><td>Range*Background</td><td>1</td><td>0.813</td></tr> <tr><td>Speed*Background</td><td>1</td><td>0.813</td></tr> </table> Apply Changes to Anticipated Coefficients	Term	Anticipated Coefficient	Power	Intercept	1	0.813	Range	1	0.838	Speed	1	0.813	Background	1	0.813	Range*Speed	1	0.813	Range*Background	1	0.813	Speed*Background	1	0.813	Power Analysis Significance Level 0.05 Anticipated RMSE 2 <table> <tr> <th>Term</th><th>Anticipated Coefficient</th><th>Power</th></tr> <tr><td>Intercept</td><td>1</td><td>0.8</td></tr> <tr><td>Range</td><td>1</td><td>0.799</td></tr> <tr><td>Speed</td><td>1</td><td>0.799</td></tr> <tr><td>Background</td><td>1</td><td>0.8</td></tr> <tr><td>Range*Speed</td><td>1</td><td>0.8</td></tr> <tr><td>Range*Background</td><td>1</td><td>0.799</td></tr> <tr><td>Speed*Background</td><td>1</td><td>0.799</td></tr> </table> Apply Changes to Anticipated Coefficients	Term	Anticipated Coefficient	Power	Intercept	1	0.8	Range	1	0.799	Speed	1	0.799	Background	1	0.8	Range*Speed	1	0.8	Range*Background	1	0.799	Speed*Background	1	0.799	Power Analysis Significance Level 0.05 Anticipated RMSE 4 <table> <tr> <th>Term</th><th>Anticipated Coefficient</th><th>Power</th></tr> <tr><td>Intercept</td><td>1</td><td>0.801</td></tr> <tr><td>Range</td><td>1</td><td>0.801</td></tr> <tr><td>Speed</td><td>1</td><td>0.801</td></tr> <tr><td>Background</td><td>1</td><td>0.801</td></tr> <tr><td>Range*Speed</td><td>1</td><td>0.801</td></tr> <tr><td>Range*Background</td><td>1</td><td>0.801</td></tr> <tr><td>Speed*Background</td><td>1</td><td>0.801</td></tr> </table> Apply Changes to Anticipated Coefficients	Term	Anticipated Coefficient	Power	Intercept	1	0.801	Range	1	0.801	Speed	1	0.801	Background	1	0.801	Range*Speed	1	0.801	Range*Background	1	0.801	Speed*Background	1	0.801
Term	Anticipated Coefficient	Power																																																																								
Intercept	1	0.813																																																																								
Range	1	0.838																																																																								
Speed	1	0.813																																																																								
Background	1	0.813																																																																								
Range*Speed	1	0.813																																																																								
Range*Background	1	0.813																																																																								
Speed*Background	1	0.813																																																																								
Term	Anticipated Coefficient	Power																																																																								
Intercept	1	0.8																																																																								
Range	1	0.799																																																																								
Speed	1	0.799																																																																								
Background	1	0.8																																																																								
Range*Speed	1	0.8																																																																								
Range*Background	1	0.799																																																																								
Speed*Background	1	0.799																																																																								
Term	Anticipated Coefficient	Power																																																																								
Intercept	1	0.801																																																																								
Range	1	0.801																																																																								
Speed	1	0.801																																																																								
Background	1	0.801																																																																								
Range*Speed	1	0.801																																																																								
Range*Background	1	0.801																																																																								
Speed*Background	1	0.801																																																																								

Conversely, Figure 25 and Figure 26 provide graphical illustrations of the tradeoff between the figures of merit for the example in Table 8. Since the example is a 2^3 factorial design with continuous factors at the same coded levels, the power for all model terms is identical and overlay in one curve. The red line in all graphs depict an 80% power threshold.

Figure 25 shows power as a function of number of runs for $\alpha=0.05$ and $\delta/\sigma=1$. The figure shows that with 32 runs one can achieve ~80% power for each model term.

Table 16. Power for sample size $N = 13$ as a function of δ/σ

$\delta/\sigma = 2$	$\delta/\sigma = 1$	$\delta/\sigma = 0.5$																																																																								
$N = 13$	$N = 13$	$N = 13$																																																																								
Power Analysis Significance Level 0.05 Anticipated RMSE 1 <table> <tr> <th>Term</th><th>Anticipated Coefficient</th><th>Power</th></tr> <tr><td>Intercept</td><td> 1 </td><td>0.813</td></tr> <tr><td>Range</td><td> 1 </td><td>0.838</td></tr> <tr><td>Speed</td><td> 1 </td><td>0.813</td></tr> <tr><td>Background</td><td> 1 </td><td>0.813</td></tr> <tr><td>Range*Speed</td><td> 1 </td><td>0.813</td></tr> <tr><td>Range*Background</td><td> 1 </td><td>0.813</td></tr> <tr><td>Speed*Background</td><td> 1 </td><td>0.813</td></tr> </table> Apply Changes to Anticipated Coefficients	Term	Anticipated Coefficient	Power	Intercept	1	0.813	Range	1	0.838	Speed	1	0.813	Background	1	0.813	Range*Speed	1	0.813	Range*Background	1	0.813	Speed*Background	1	0.813	Power Analysis Significance Level 0.05 Anticipated RMSE 2 <table> <tr> <th>Term</th><th>Anticipated Coefficient</th><th>Power</th></tr> <tr><td>Intercept</td><td> 1 </td><td>0.304</td></tr> <tr><td>Range</td><td> 1 </td><td>0.321</td></tr> <tr><td>Speed</td><td> 1 </td><td>0.304</td></tr> <tr><td>Background</td><td> 1 </td><td>0.304</td></tr> <tr><td>Range*Speed</td><td> 1 </td><td>0.304</td></tr> <tr><td>Range*Background</td><td> 1 </td><td>0.304</td></tr> <tr><td>Speed*Background</td><td> 1 </td><td>0.304</td></tr> </table> Apply Changes to Anticipated Coefficients	Term	Anticipated Coefficient	Power	Intercept	1	0.304	Range	1	0.321	Speed	1	0.304	Background	1	0.304	Range*Speed	1	0.304	Range*Background	1	0.304	Speed*Background	1	0.304	Power Analysis Significance Level 0.05 Anticipated RMSE 4 <table> <tr> <th>Term</th><th>Anticipated Coefficient</th><th>Power</th></tr> <tr><td>Intercept</td><td> 1 </td><td>0.112</td></tr> <tr><td>Range</td><td> 1 </td><td>0.117</td></tr> <tr><td>Speed</td><td> 1 </td><td>0.112</td></tr> <tr><td>Background</td><td> 1 </td><td>0.112</td></tr> <tr><td>Range*Speed</td><td> 1 </td><td>0.112</td></tr> <tr><td>Range*Background</td><td> 1 </td><td>0.112</td></tr> <tr><td>Speed*Background</td><td> 1 </td><td>0.112</td></tr> </table> Apply Changes to Anticipated Coefficients	Term	Anticipated Coefficient	Power	Intercept	1	0.112	Range	1	0.117	Speed	1	0.112	Background	1	0.112	Range*Speed	1	0.112	Range*Background	1	0.112	Speed*Background	1	0.112
Term	Anticipated Coefficient	Power																																																																								
Intercept	1	0.813																																																																								
Range	1	0.838																																																																								
Speed	1	0.813																																																																								
Background	1	0.813																																																																								
Range*Speed	1	0.813																																																																								
Range*Background	1	0.813																																																																								
Speed*Background	1	0.813																																																																								
Term	Anticipated Coefficient	Power																																																																								
Intercept	1	0.304																																																																								
Range	1	0.321																																																																								
Speed	1	0.304																																																																								
Background	1	0.304																																																																								
Range*Speed	1	0.304																																																																								
Range*Background	1	0.304																																																																								
Speed*Background	1	0.304																																																																								
Term	Anticipated Coefficient	Power																																																																								
Intercept	1	0.112																																																																								
Range	1	0.117																																																																								
Speed	1	0.112																																																																								
Background	1	0.112																																																																								
Range*Speed	1	0.112																																																																								
Range*Background	1	0.112																																																																								
Speed*Background	1	0.112																																																																								

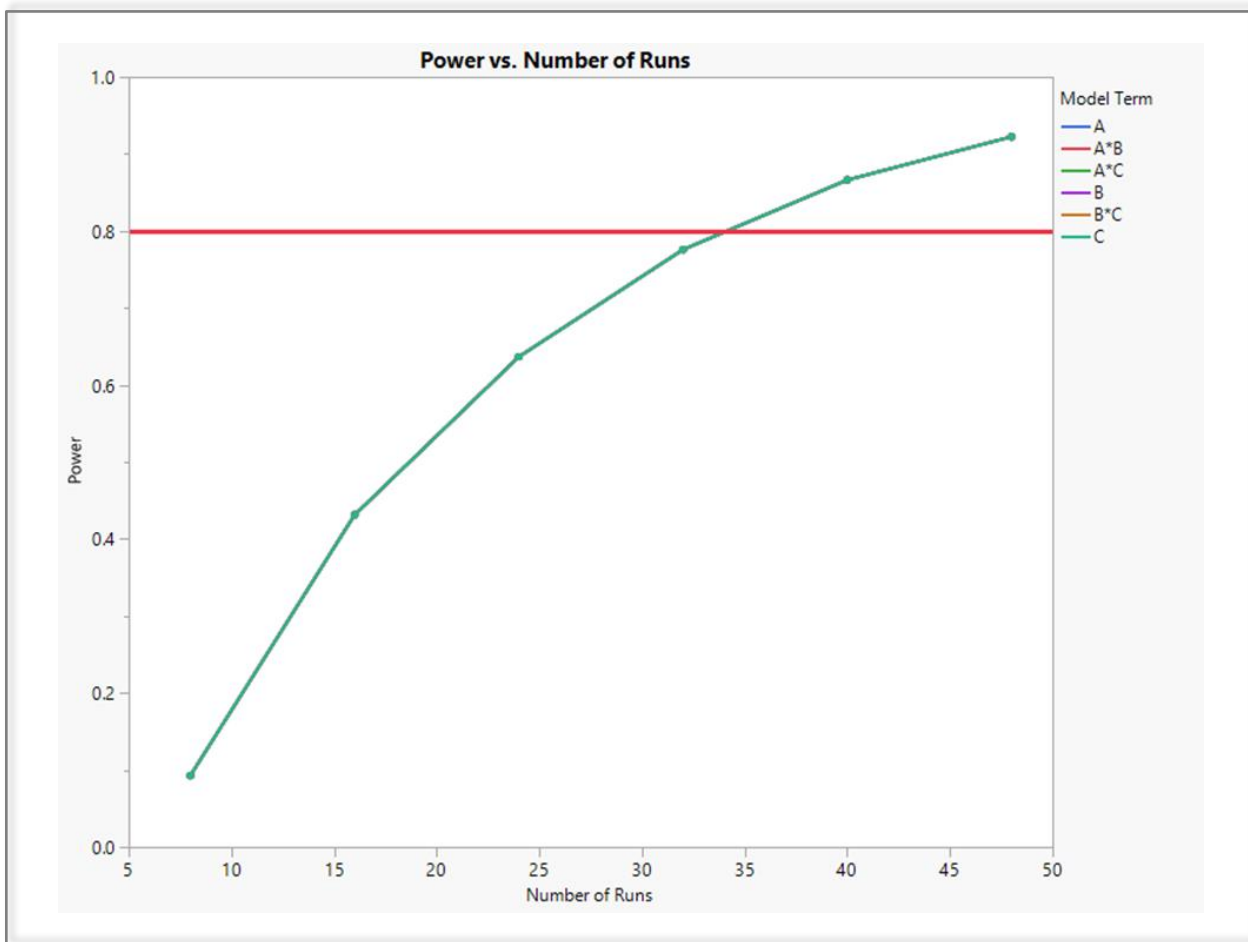


Figure 25. Power vs. number of runs for fixed α and δ/σ

Figure 26 shows power as a function of number of runs for different values of alpha and signal-to-noise. The curves can vary quite significantly with changes in both α and δ/σ . For $\delta/\sigma=0.5$, power is well below the 80% threshold and more than 48 runs are needed to satisfy the power threshold. For $\delta/\sigma=2.0$, a 16 run test design will satisfy the power threshold regardless of the alpha level. Alpha should be set before the beginning of the test design process.

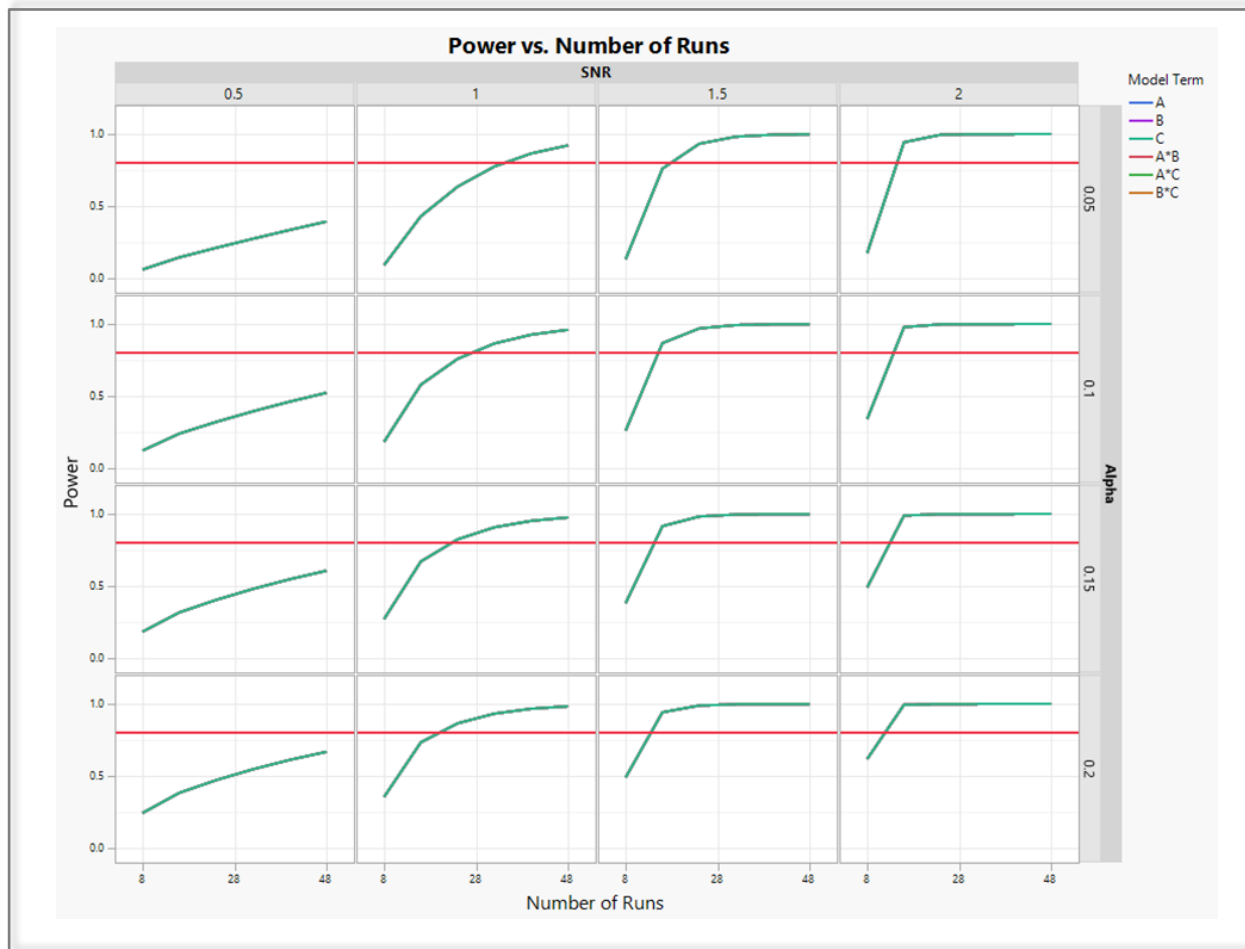


Figure 26. Power vs. number of runs for different α and δ/σ

There are situations in which standard designs are not appropriate or practical, which include:

- A blocking strategy that does not fit a standard blocked design
- Combinations of factor settings that are not feasible or safe
- Design regions of irregular shapes
- Recovery from improper planning
- Nonlinear models are appropriate
- Quadratic terms in the presence of qualitative factors
- Multiple randomization restrictions

For those situations, optimal designs ([NIST Handbook Section 5.5.2](#)) are a useful option. In some cases, they are the only options.

To illustrate the tradeoffs between different types of designs, consider a situation where the objective of a screening design is to determine which of five continuous factors (at two levels each) and ten of their two-factor interactions (2FI) influence a response of interest. In this example, the run to run cost is the same (not always true in a real situation), confidence was fixed at 95%, and S/N was fixed at 2.0. Figure 27, prepared with JMP, provides a comparison of key evaluation properties of three screening designs—a full factorial design, a $\frac{1}{2}$ half fractional factorial design, and a D-Optimal design—executed as completely randomized design (CRD) and with a continuous response variable. Note that even after selecting the desired level of confidence, power, and signal-to-noise ratio for the test designs, there are several different types of designs that can be built to those specs.

Power estimates aside, Figure 28 illustrates typical run matrices for a 2^4 in a completely randomized layout and in a split-plot layout. Note that the split-plot layout requires at least one replicate of a whole-plot to get an estimation of whole-plot variance.

The fourth and last phase of the sequential strategy is the confirmation phase. The objective is to conduct a small number of runs to confirm the prediction capability of the response surface model. Good designs for verification or confirmation are resolution III fractional factorial designs and Plackett-Burman designs ([NIST Handbook Section 5.3.3.5](#)). Plackett-Burman designs are very economical resolution III designs where the number of runs is a multiple of four rather than of a power of two.

Building up a sequential ensemble, where the analysis results of one phase dictate the next set of runs, is the most efficient way to test. However, in a typical DoD environment many test events do not lend themselves to using a sequential ensemble approach. A real-time analysis of test data or a quick turn-around on analysis of data is not possible. Also, many test events require the allocation of resources and schedule well ahead of collecting data and test truncation and modification of the test design is not allowed. Due to these limitations, test designs tend to be one-shot, very large test designs to cover all potential outcomes. This can lead to a great amount of waste.

Clearly, there are compelling reasons for selecting a design with good features. Criteria to evaluate test designs has been refined over time. Table 17 provides a consensus of that evolution. Unfortunately, all the features listed in the criteria are rarely achieved in a design. Frequently the choice of the most useful design is driven by a compromise between statistical properties, cost, schedule, and performance considerations. Nonetheless, a rule that experienced test designers live by is that it is more important to select an approximate test design for the exact problem than to select an exact test design for an approximate problem.

There are situations in which the placement of test design points by the classical experiment designs discussed do not fit the situation of interest, when there is no a prior knowledge of how the model is, or when the placement of points has precedence over efficiencies gained thru sequential ensemble. That is, for instance, often the case for validating M&S where the objective is to validate that a representation or model of a system is consistent with reality or with another scientific model. For those situations, there is a class of designs that are generally referred to space filling designs that are adequate.

Table 17. General test design evaluation criteria

Design Feature	Implications
Test space coverage	Test points are reasonably distributed over the entire test space.
Easily interpretable model	The resulting empirical model is as simple as possible
Good prediction quality	The difference between predictions made by the empirical model and the true response value at the setting of factor levels are small. The test design has good prediction properties (i.e. good prediction variance over the design space).
Allows check of model assumptions	The test design provides information to investigate and fix issues with lack-of-fit, curvature, and violations of the basic assumptions.
Allows for blocking	The design allows for conducting the test in blocks, which isolate sources of variability (e.g. operational conditions, different operators, etc.).
Can be built up sequentially	A small initial design (in terms of number of runs) provides information to plan the next subsequent set of test runs to build a higher order model.
Estimation of pure error	The design includes enough runs to allow for estimating the natural variation between points at the same factor settings, especially if large experimental errors are expected.
Robust to outliers and missing data	The design is insensitive to outliers, missing observations, and errors in the settings and control of the factor levels.
Efficient and Effective	The test design is efficient in terms of the number of runs and maximizes the information obtained.

Space-filling designs, as the name suggests, are test designs that attempt to cover the test space as effectively as possible given a specified number of runs. These designs are commonly used for simulation experiments where a response is deterministic (repeated observations at the same factor inputs produces identical test results).

The factor space for these tests can be relatively broad, with more factors, larger ranges, and more levels than what would typically be seen for a physical experiment. The associated response surface has the potential to be very complex and highly non-linear (not easily modeled with a low order polynomial model) as illustrated in Figure 29.

Since these simulations are deterministic, these test designs do not need to consider randomization, replication, blocking, or whether factors are easy or hard-to-change. These test designs don't aim to optimize the same statistical properties as traditional test designs (e.g. power, confidence, prediction error), rather they try to maximize the minimum distance between points and avoid having replicates, or near replicates if the dimensions of factors space are reduced (uniformity). There are several methodologies to create these test designs, each with their own strengths and weaknesses.

Three common types of space-filling test designs are sphere sacking, Latin hypercube, and fast flexible space filling designs. Table 18 provides a brief description of each, a two-dimensional visual example for each, and a projection onto each individual dimension (on the left and on the bottom of each corresponding example).

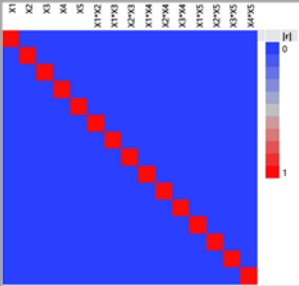
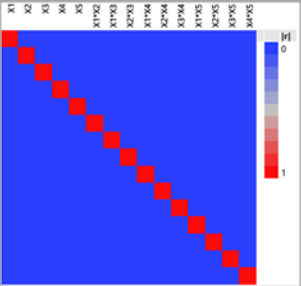
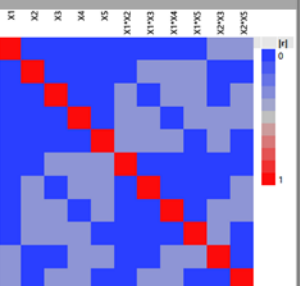
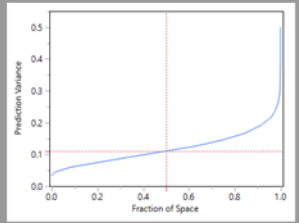
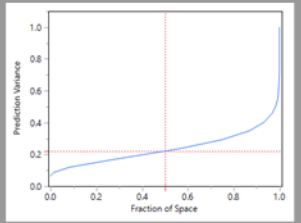
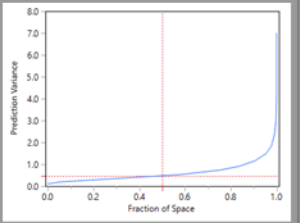
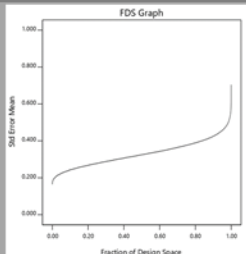
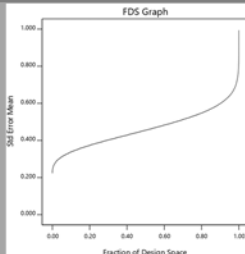
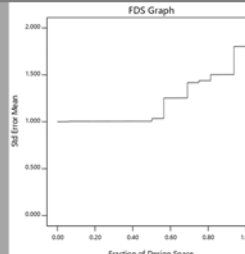
	Full Factorial + 5 Center Points	$\frac{1}{2}$ Fractional Factorial	D-Optimal
Runs	37	16	12
Estimable model terms	5 MEs 10 2FIs	5 MEs 10 2FIs	5 MEs 6 2FIs
Power	99% (all effects)	24.6% (all effects)	10.0 – 15.3%
dof for Model Lack-of-fit Pure error	15 17 4	15 0 0	11 0 0
VIF	1.0	1.0	1.3 – 1.9
Aliasing			
Fraction of design space			
Average prediction variance	0.114	0.236	0.222
Standard Error Mean vs. fraction of design space			
D-efficiency	87%	100%	100%
G-efficiency	87%	100%	100%

Figure 27. Partial list of test design metrics for some screening designs

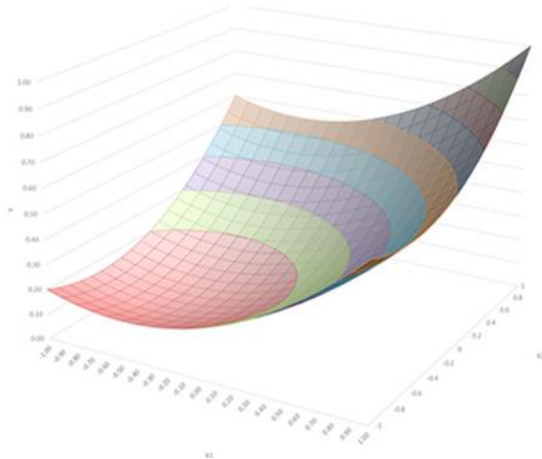
Std.	Run	x_1	x_2	x_3	x_4
7	1	-1	1	1	-1
1	2	-1	-1	-1	-1
4	3	1	1	-1	-1
8	4	1	1	1	-1
2	5	1	-1	-1	-1
6	6	1	-1	1	-1
12	7	1	1	-1	1
5	8	-1	-1	1	-1
3	9	-1	1	-1	-1
16	10	1	1	1	1
15	11	-1	1	1	1
14	12	1	-1	1	1
11	13	-1	1	-1	1
9	14	-1	-1	-1	1
13	15	-1	-1	1	1
10	16	1	-1	-1	1

(a) Completely randomized design

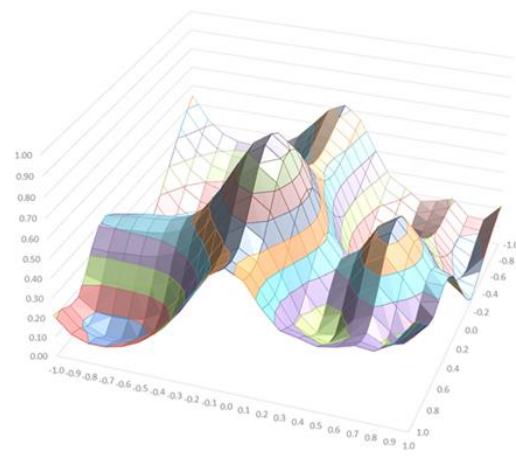
Std.	Run	WP	z_1	z_2	x_1	x_2
1	1	1	-1	-1	-1	-1
5	2	1	-1	-1	1	-1
9	3	1	-1	-1	-1	1
13	4	1	-1	-1	1	1
7	5	4	-1	1	1	-1
3	6	4	-1	1	-1	-1
15	7	4	-1	1	1	1
11	8	4	-1	1	-1	1
2	9	3	1	-1	-1	-1
6	10	3	1	-1	1	-1
14	11	3	1	-1	1	1
10	12	3	1	-1	-1	1
4	13	2	1	1	-1	-1
8	14	2	1	1	1	-1
12	15	2	1	1	-1	1
16	16	2	1	1	1	1
1	17	1	-1	-1	-1	-1
5	18	1	-1	-1	1	-1
9	19	1	-1	-1	-1	1
13	20	1	-1	-1	1	1

(b) Split-plot design

Figure 28. Example of 2^4 completely randomized and split-plot designs



(a) Low-order response surface



(b) Higher-order response surface

Figure 29. Low order vs. a higher order nonlinear response surface

Table 18. Common Types of Space Filling Designs

Design Type	Visual Example	Description
Sphere Packing		• Uses an algorithm to maximize the minimum distance between points. • Designs may not show good uniformity when projected into a smaller subspace. • All factors must be continuous.
Latin Hypercube		• Design that provide more of a balance between maximizing the minimum distance between points and exhibit good uniformity when projection into the test sub-space. • All factors must be continuous.
Fast Flexible Space Filling		• Method use a clustering algorithm for run selection. • Designs exhibit good space filling and uniformity. • Can be used with continuous and categorical factors.

Like with traditional test designs, the appropriate number runs for a space-filling design can be based on multiple factors. These factors include the time and resources needed to execute the runs as well as the expected complexity of the underlying model. The following equation provides a range of design sizes that have a good balance between efficiency and statistical power:

$$3 \left(K - L + \sum_{x \in L} (l_x - 1) \right) \leq n \leq 10 \left(K - L + \sum_{x \in L} (l_x - 1) \right)$$

Here, L is the number of categorical factors, with l_x levels, and K total factors. For example, if a test design has five total factors ($K = 5$), with two categorical factors ($L = 2$), one categorical factor has three levels ($l_1 = 3$) and the second has five levels ($l_2 = 5$), the recommended range of test runs is $27 \leq n \leq 90$. Note that if all five factors were continuous, $K = 5$ and the range would be smaller $15 \leq n \leq 50$.

The results from deterministic simulation test design are usually modeled using Gaussian Process (attempts to find the best match among a pool of candidate formulas that could generate the observations) as oppose to linear regression (attempts to model the relationship between the factors and the responses by fitting a linear equation over the data). This modeling method allows for an exact fit to data and is flexible enough for modeling complex and irregular surfaces. Figure 30 illustrates the difference in fit for the same test design data using both techniques.

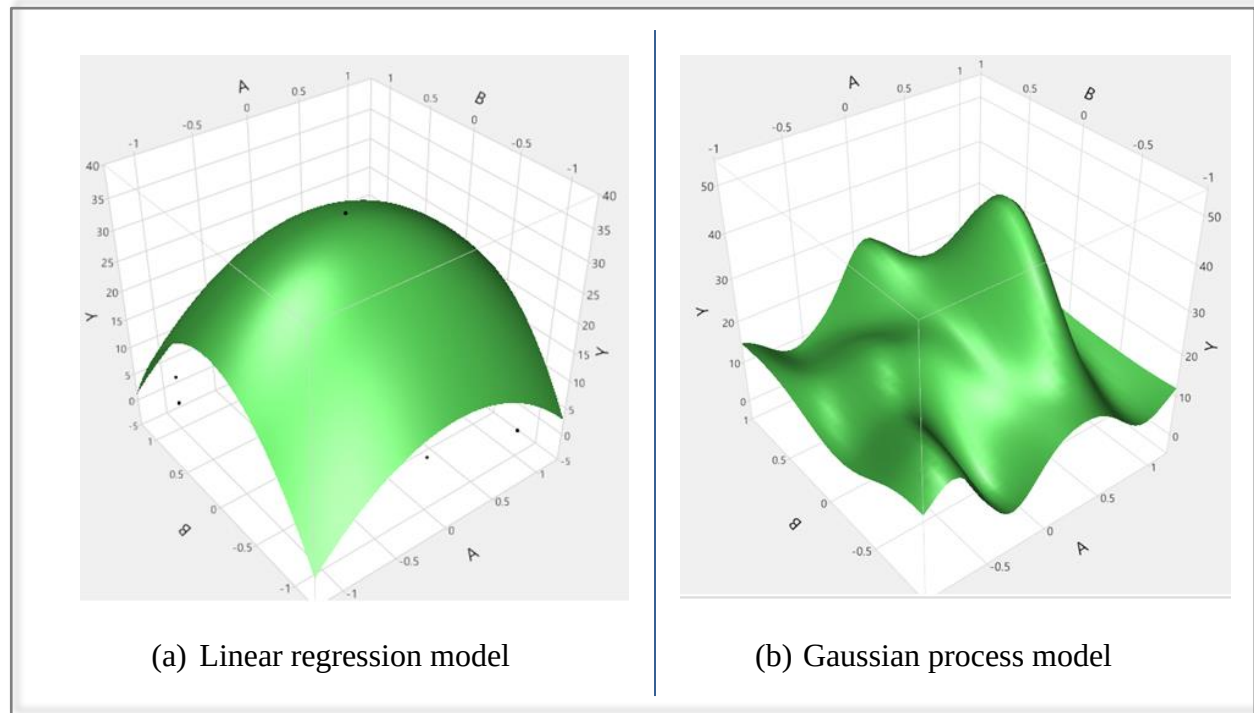


Figure 30. Linear regression vs. Gaussian process models

Other situations that involve response variables are deterministic and binary (pass/fail), meaning that running the same scenario repeatedly will not change the outcome, include software testing, network design, and cyber security testing. For these situations, combinatorial test designs and covering arrays help identify where faults are occurring due to some combination amongst factors, components, or subsystems. Due to the deterministic nature of the response variables, statistical

measures such as power, confidence, and collinearity are no longer useful to evaluate a proposed test design. Rather, designs are evaluated based on the concept of coverage.

To illustrate this, consider testing a software application across different environments and factors as illustrated in Table 19. Assuming there are no disallowed combinations, it would take 3240 runs to look at all combinations of factor levels, which can be too laborious and expensive to conduct. However, a NIST study showed that most faults are caused by the interactions between a few factors. Those results mimic the idea behind the sparsity-of-effects concept in traditional DOE. Figure 31 shows the cumulative percentage of faults found versus the level of interactions between factors for various systems. The chart shows that more than 60% of all failures were found looking at 2-way interactions, 80% of failures were found by looking at just 3-way interactions, and over 90% with only 4-way interactions, regardless of the complexity of the system. Thus, test efficiencies can be obtained by focusing on examining all lower t-way interactions in the fewest number of runs as possible.

Table 19. Factor and level list for software testing application

Factor	Levels	No. Levels
Browser	IE, Edge, Firefox, Chrome, Opera, Safari	6
Connection Speed	0-5, 5-10, >10 Mbps	3
Database	MySQL, Sybase, Oracle	3
Operating System	Apple OS, Linux, Windows	3
Processor	Intel, AMD	2
Protocol Stack	IPv4, IPv6	2
RAM	4, 8, 16, 32, 64	5

Table 20 shows the efficiency of a combinatorial test design to cover t-way interactions for the seven-factor example of Table 19. Based on the NIST study, over 80% of failures could be found looking at 3-way interactions (100 runs) and about 95% of the failures could be found by looking at 4-way interactions (310 runs). Managing risk, combinatorial testing offers an efficient alternative for software testing, cyber security testing, etc. Software to build combinatorial designs are readily available from NIST as well as from various packages in R.

Once the capability-based test designs that stem from the XEF and the mission-based test designs that stem from the IEF are developed, the test designs are then combined to produce an integrated test design. The integrated test design identifies common data needs that can be met with a single test event—an integrated test. The integrated test provides the data for independent developmental and operational evaluations, in accordance with the XEF and IEF, realizing the vision (and fulfilling the legal obligation) of integrated test and independent evaluation.

The integrated test design produces a run matrix that is the foundation for identifying the LVC resources. The TEMP, or any other approved test planning documentation, will document resource requirements to include test articles, test sites and instrumentation, test targets and expendables, M&S, testbeds, distributed test networks, etc. Resources are mapped against the XEF to ensure they are adequate and available.

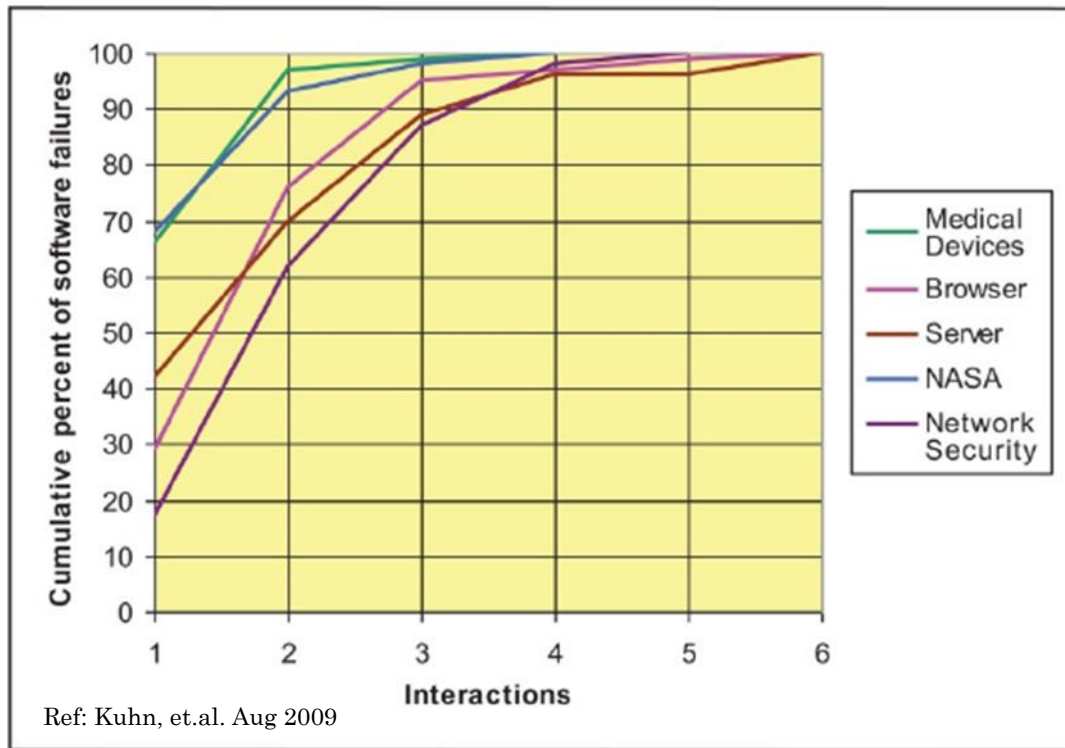


Figure 31. Coverage of faults due to t-way interactions

Table 20. Efficiency to cover t-way interactions (network security)

t-way	Runs	% Maximum No. Runs	% Faults Detected
2	30	1	70
3	100	3	90
4	310	10	95
5	817	25	99
6	1620	50	99
7	3240	100	100

The conclusion of this step is probably the point in the test strategy to publish a test run matrix, a data collection plan, and a data analysis and assessment plan to inform all stakeholders of the plan. The test plan and the data collection plan optimize test execution and data collection to fill data needs for the independent developmental and operational evaluations. At a minimum, the test plan and data collection plan detail developmental and operational test and evaluation objectives, developmental and operational measures, the required test conditions, the test design details specified in the DOT&E memos, and data collection requirements. Table 21 highlights some of the strategies for selecting designs.

Table 21. Strategies for selecting an experiment design

Strategy	Outcome
Allocate sufficient runs to estimate pure error, test for curvature, lack-of-fit, and most importantly for confirmation or validation, whenever possible	Provides information to completely characterize the system
Identify statistical figures of merit such as expected model terms, sample size, confidence, power, statistical signal-to-noise ratio, collinearity, prediction error, and coverage	Quantifies risk and gives decision makers insight into the amount of testing that is needed so they can make informed trade-offs and facilitates the comparison of designs
Evaluate the design (use historical data or simulations if possible)	Provides insight into the performance of the design and allows for the selection of a design with good properties
Choose an experiment design that fits the problem, a strategy that fits the resources, and that adheres to the principles of experiment design	Enables decision-quality information to answer the decision support objectives and to inform the evaluations and assessments
Apply a sequential strategy if possible (screening, characterization, optimization, validation)	Optimizes the allocation of resources throughout the execution of the test strategy
Estimate the difference to detect and standard deviation (signal-to-noise ratio) for each response variable	Forces the collection of information from subject matter experts, pilot studies, or historical data of similar systems under like conditions in order to estimate the change in a response variable that could result in substantial difference in system performance and an estimate of the inherent noise in the system, which helps right sizing the test

4.6 Coordinate and Execute the Test Strategy

The strategy for this phase includes to:

- Prepare evaluation plans, test plans, and data collection plans
- Ensure the physical set-up and the test execution matrix are consistent
- Execute the run matrix as planned
- Record and collect the data
- Perform a quick look analysis of the data before leaving the test stage

When conducting the test, it is of vital importance to ensure everything is being done according to the plan. It is extremely important to carry out the test following the fundamental principles of experiment design—factorization, replication, randomization, and blocking—as outlined by the test protocol or run matrix. The run order is important and should be as specified in the design matrix. Procedural errors could have a catastrophic effect on the response if they are not eliminated or at least significantly minimized. The data must be collected in a manner that is consistent with the way the test was conducted. Here is a list of common mistakes and their consequences:

- Missing runs – results in aliasing, loss of design resolution, loss of statistical power, and increased correlation between factors effects

- Finding a disallowed combination or a constrained region of space during testing – could be an indication of poor planning. It has the same consequences as missing runs. However, it is better to find it and not to execute the run than to executing a catastrophic test run
- Incorrect recording of results – may result in an incorrect model
- Inadvertently executing a completely randomized design as a split-plot - incorrect identification of significant factors and interactions
- Not resetting the factors between runs – extra variability could be induced in the experimental error, which could influence the tests of significance

DOE-based test design is applicable to the whole spectrum of LVC events. Adequate test venues and test resources may include hardware in the loop (HWIL) tests, M&S, live events, or real-world events. The *Test Plan* should identify the scenarios, the instrumentation plan, and the test execution plan. DOE-based test designs can be used in combination with test automation to maximize the amount of information while minimizing the number of tests to inform the evaluation objectives, which improves both the efficiency and effectiveness of test.

This is also the time to collect the data. Data collection refers to the gathering of test information from different sources, including runs generated by the test design with the objective of informing the evaluations. The data must be clearly organized and arranged in the order in which the test was conducted to facilitate its understanding. Record any observations associated with set-up and the outcome of the test. Before leaving the test stage, exploratory data analysis (EDA) and test design diagnostics should be used to identify potential areas of risk.

Table 22. Strategies for test execution

Strategy	Outcome
Develop an adequate test execution strategy	Maximizes the amount of decision-quality information collected while minimizing the use of test resources and allowing for risk assessment and trade-offs
Take advantage of test automation to stimulate the system and collect and analyze the data	Piles on efficiency and effectiveness gains achieved with DOE
Carry out the test as it was designed and documented in the run matrix	Ensures the use of proper scientific methods and statistical analysis techniques
Document the test protocol and deviations from it	Facilitates the data analysis process
Identify and manage unplanned or unexpected sources of variability, particularly those involving test operations and measurement	Enables a better understanding of how the factors affect the responses
Perform a hot wash-up and a quick-look analysis immediately after concluding the test runs	Identifies problems encountered during test and allows for taking immediate advantage of extra resources
Run statistical test diagnostics	Identifies outcomes or measurements that look suspect and those that need to be repeated
Re-run tests that have suspect test results, if possible	Improves the integrity of the test process and the quality of data to inform the evaluations
Confirm the results	Validates the test results, guarantees the integrity of the test process

4.7 Perform Exploratory Data Analysis

Exploratory data analysis (EDA) ([NIST Handbook Section 1.1.1](#)) is key in providing information to understand mission capability and system functionality, and is a pre-cursor to the evaluation of system performance and suitability. EDA is a philosophy that involves using a collection of statistical and graphical techniques to dissect a data set. EDA is used to gain insight into and uncover the structure of the system represented by the data. EDA also helps detect anomalies with the test and identify outliers.

Some critical EDA techniques are determined by the selection of the test design. So, if the test was performed according to the test design selected, critical analysis techniques are, for the most part, already identified. The data needed for analysis should be captured in the data collection plan and in the data management plan. The analysis and assessment methods and the evaluation criteria should be documented in the data analysis plan.

It is important to analyze the data immediately after the test to take advantage of any possible opportunities to identify bad data points and to re-run them, as necessary.

Typically, the steps in EDA are to:

- Use descriptive statistics to look for behavior and patterns in the data
- Identify factors and interactions that have a significant effect on the responses
- Fit a model to the data using numerical and graphical analysis methods
- Validate assumptions and calibrate the model
- Select the final model

So, the outcome would be identifying the significant factors and interactions and having a calibrated capability model that mathematically describes how the factors influence the responses across the entire test space (ideally the operational space). The model is then use as part of the MTM approach to update the authoritative source of truth.

There are a wide variety of statistical techniques and methods available for the analysis of test data. Those techniques and methods include one sample t-tests, two sample t-tests, analysis of variance (ANOVA), analysis of covariance (ANCOVA), linear regression, generalized linear models, Bayesian approaches, etc. Depending on the test objectives, other mathematical and statistical techniques such as gradient methods, Newton methods, or conjugate methods can be used along with the empirical model to search for the optimum response values. There are several off-the-shelf data analysis packages to help with the analysis. These software packages include graphical methods to aid with the visualization and interpretation of the data.

The first step in EDA is to perform descriptive statistics. Descriptive statistics is the process of using summary statistics and graphical techniques to help describe and understand data. Descriptive statistics is broken down into measures of central tendency ([NIST Handbook Section 1.3.5.1](#)), which describe the middle values of the data set, and measures of spread ([NIST Handbook Section 1.5.3.6](#)), which describe the distribution or dispersion of the data set. The most common measures of central tendency are the mean, median, and mode, and the most common measures of spread include range, quartiles, standard deviation, and variance. There are other graphical methods to present data that are useful and quite simple such as box plots, bar and pie charts, Pareto charts, steam-and-leaf plots, and histograms.

The second step in EDA is to identify significant effects and interactions. Use half normal probability plots (NIST Handbook Section 5.5.9.8) and Pareto charts (NIST Handbook Section 5.5.9.8) to determine the significant effects and interactions and what their rank is relative to other factors and interactions (those that are farther away from the line on the half-normal plot and over the Bonferroni and t-value limits in the Pareto chart as shown in Figure 32).

Once the significant factors have been identified, it is time to fit a model to the data by performing ANOVA to assess the sufficiency of the model terms and model misspecification. ANOVA (NIST Handbook Section 7.4.3) is a purely mathematical technique to analyze the differences between and within the means of groups of observations and to establish their significance by comparing the factor effects and interaction effects to the error term. Ideally, the error term is just pure experimental error, which is proportional to the variability between observations. Table 23 is an example of a typical ANOVA table (NIST Handbook Section 7.4.3.3), which in this case is consistent with the diagram in Figure 32 and the fictitious OASUW example from Table 11. Table 23 shows the source of variation, the variances for the treatments and the error (sum-of-squares and mean square), the test statistics (F-value and p-value) used in testing the equality of treatment means, and lack-of-fit test (NIST Handbook Section 4.4.4). A general rule of thumb in DoD testing is that p-values ≤ 0.05 , or 2σ , indicate the model terms are significant while p-values ≥ 0.1 indicate the model terms are not significant. p-values between 0.05 and 0.1 are evaluated carefully to determine the significance of model terms.

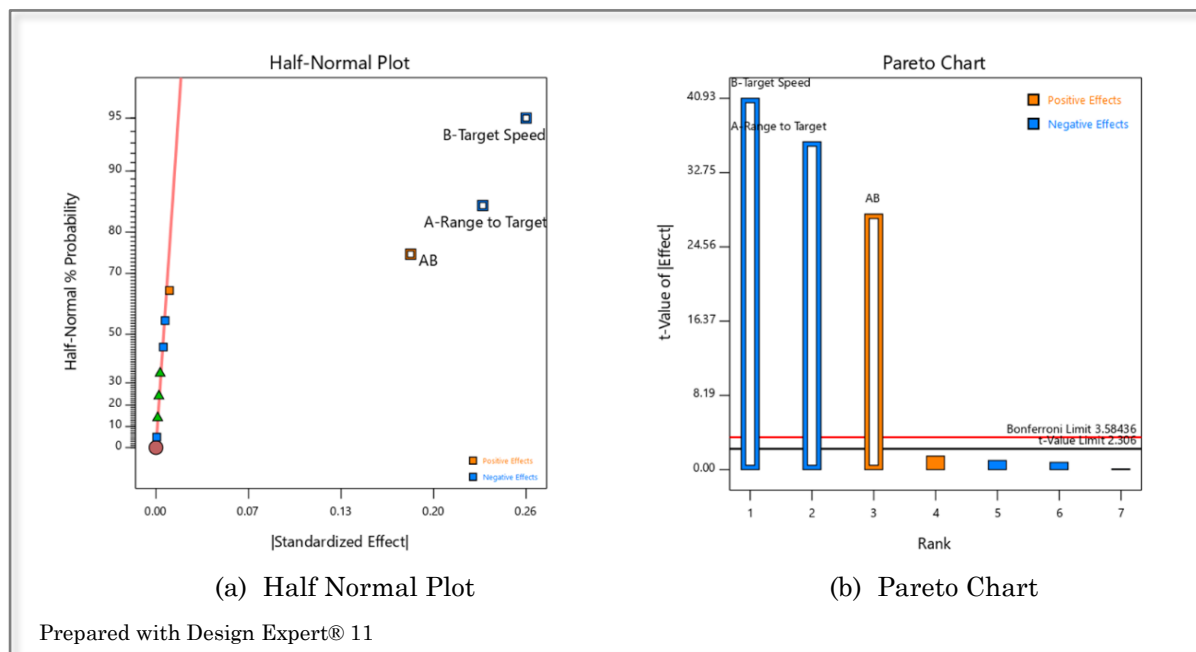


Figure 32. Graphical methods for identifying significant factors and interactions

The illustrative analysis in Table 23 (corresponding to the full factorial + center points in Figure 20) shows that main effects A and B as well as interaction A*B are significant to a p-value < 0.0001 (there is $< 0.01\%$ chance the result is due to the variability). The model is also significant to a p-value < 0.0001 . Curvature is significant to p-value < 0.0001 ; therefore, an augmentation to response surface design is needed to obtain an adequate model. Lack of fit is not significant relative to the pure error, which is good since the objective is to fit a model to the observations.

Table 23. ANOVA table for Response 1

Source of Variation	Sum of Squares	df	Mean Square	F-value	p-value	Significance
Model	0.5297	6	0.0883	743.53	< 0.0001	Significant
A - Range	0.2077	1	0.2077	1749.47	< 0.0001	Significant
B - Speed	0.1548	1	0.1548	1303.90	< 0.0001	Significant
C – Land Behind	0.0000	1	0.0000	0.0030	0.9587	Not significant
A*B	0.1671	1	0.1671	1407.58	< 0.0001	Significant
A*C	0.0000	1	0.0000	0.1976	0.6797	Not significant
B*C	0.0000	1	0.0000	0.0366	0.8575	Not significant
Curvature	0.0624	1	0.0624	545.49	<0.0001	Significant
Residual	0.0005	4	0.0001			
Lack of Fit	0.0002	1	0.0002	1.51	0.3062	Not significant
Pure Error	0.0003	3	0.0001			
Total	0.5926	11				

Once the factor analysis has been completed and a preliminary model has been defined, the next step is to determine how well the preliminary model fits the observations ([NIST Handbook Section 4.4.4](#)) and how well it can predict. Part of the process involves eliminating insignificant model terms (model reduction) as well as adjusting the model coefficients (model calibration) to improve the model. There are many statistical tools for this task, which include both numerical and graphical methods.

The primary numerical analysis tools are R^2 fit statistics, which focus on measuring the agreement between model and observations. Those include:

- R^2 – measures the proportion of the variance explained by the regression model (helps determining the capability of the regression model to fit the observations)
- Adjusted R^2 – adjusts R^2 based on the number of terms in the model (helps determining which terms do not improve the regression model)
- Predicted R^2 – measures how well the regression model predicts (helps determining the capability of the regression model to fit new observations vs. original observations)
- Adequate precision – measures the proportion of the variance explained by the model (how well the data fits the model)

Table 24 summarizes the fit statistics for the reduced model example. The value of R^2 indicates the model explains ~100% of the variability of the response data around its mean. A good agreement between Predicted R^2 and Adjusted R^2 (rule of thumb is a difference < 0.2) indicates the model is a good predictor. A high adequate precision (rule of thumb is > 4.0 points) indicates the model is adequate to navigate the design space.

Table 24. Fit statistics summary

Fit Statistic	Value	
	Full Model	Reduced Model
R ²	0.9991	0.9991
Adjusted R ²	0.9978	0.9986
Predicted R ²	0.9797	0.9975
Adequate Precision	69.10	111.75

The next steps in EDA are to validate the experiment design assumptions of normality, constancy of variance, independence of observations, and then to calibrate the model. For the ANOVA results to be valid, the basic assumptions of independence, normality, and constancy of variance must be valid. The violation of those assumptions and the adequacy of the model can be explored by graphical residual analysis. A residual ([NIST Handbook Section 4.4.4](#)) is the difference between the actual measurement and the estimate of the measurement predicted by the model for that point. There are several graphical methods that are useful for confirming the validity of the assumptions, adequacy of the model, and with diagnostics. Some of those are illustrated in Figure 33 and Figure 34:

- Normal probability plots ([NIST Handbook Section 1.3.3.21](#)) can help determine if the distribution of the observations is approximately normal (a check for normality assumption)
- Box-Cox transformation plots ([NIST Handbook Section 6.5.2](#)) and residual vs. factor plots ([NIST Handbook Section 4.4.4.2](#)) help detecting non-constant variation across the test space (check for constancy of variance assumption)
- Residuals vs. run order plots and actual residual plots vs. fitted values plot help assessing independency of the measurements, drifts in process, and distribution of errors (independence assumption)
- Cook distance plots, leverage vs run plots, difference in fits (DFFITS) graphs, and difference in beta coefficients (DFBETAS) graphs help detecting outliers ([NIST Handbook Section 1.3.5.17](#)) and informing the influence of any given run on a model term (diagnostics)

The final step is to select the calibrated model. The final model equation for the example is provided below, where x1 and x2 correspond to factors A and B, which were determined to be significant.

$$R1 = -2.37 + 0.09x_1 + 0.21x_2 - 0.003x_1x_2 - 0.0009x_1^2 - 0.003x_2^2 \text{ (in actual units)}$$

$$R1 = 0.66 - 0.16x_1 + 0.14x_2 - 0.14x_1x_2 - 0.09x_1^2 - 0.07x_2^2 \text{ (in coded units)}$$

Once the model has been calibrated and the response surface created, one can answer various questions regarding the system or capability. This is often done with the aid of M&S.

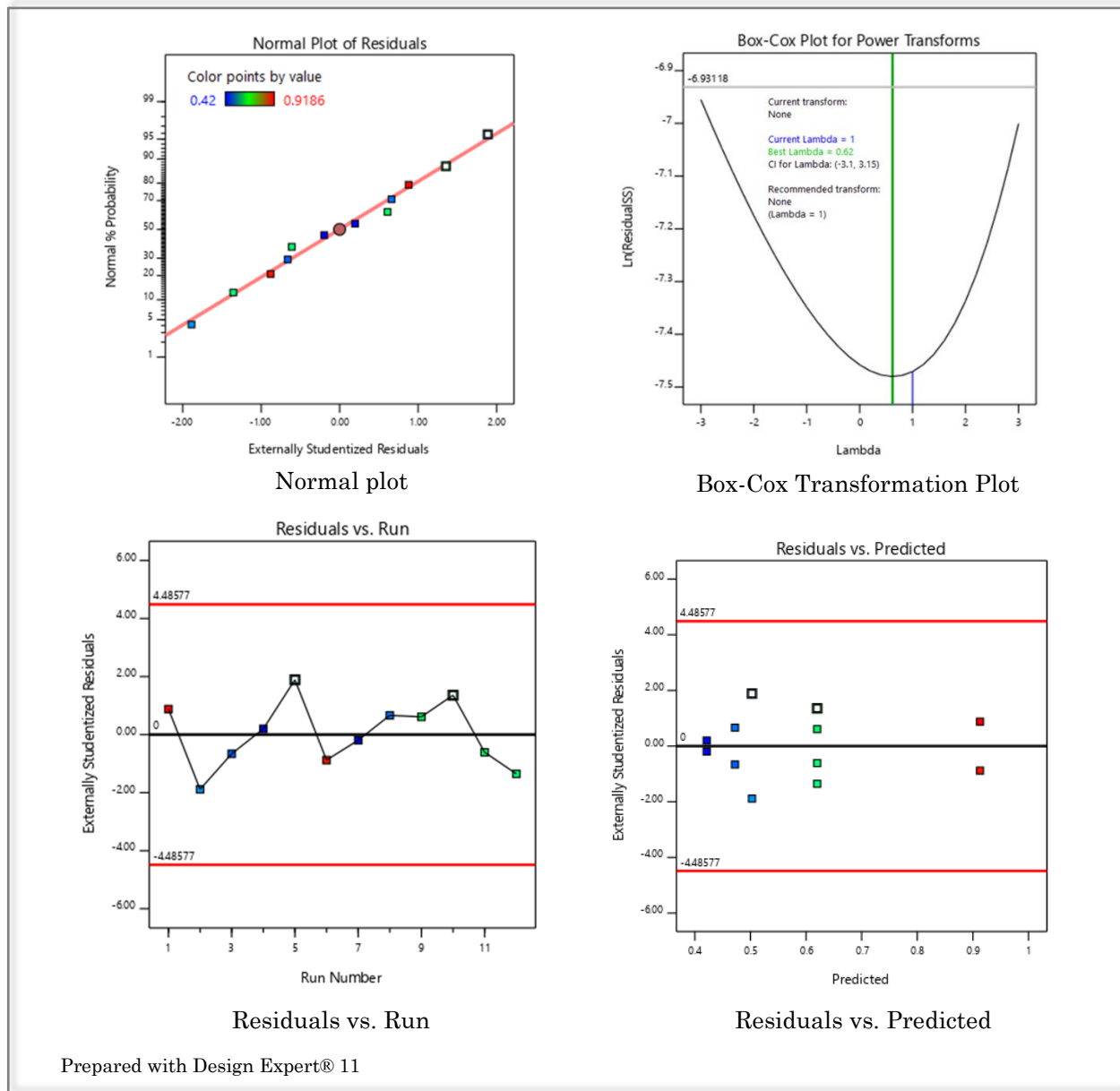


Figure 33. Graphical methods for validating DOE assumptions

For instance, consider a system that fails to fulfill its mission requirements if the response value is greater than 0.80. Aggregating the results from the 18 runs will produce an average of 0.55 and a 95% confidence interval around the mean between 0.43 and 0.66. In contrast, the response surface design model is more informative and provides an average of 0.66 that is adjusted by the factor settings. However, simulating the response surface with the information from the model allows for a more accurate representation of the space where the system will meet mission requirements. For instance, results from 50,000 Monte Carlo model runs (with $\sigma = 0.009$ in this case) shows a 4% lower probability of mission success than estimated by simply averaging the results from the 18 runs. While this number may seem small, it puts in perspective the difference between the estimates, but more importantly it may give warfighters a false sense of security regarding their

capability. Additionally, the descriptive statistics do not provide any information about the probability of mission success. Figure 35 shows the comparison between the response surface from the design and the response surface from the Monte Carlo simulation. The threshold for probability of mission success is illustrated by the contour line of 0.80.

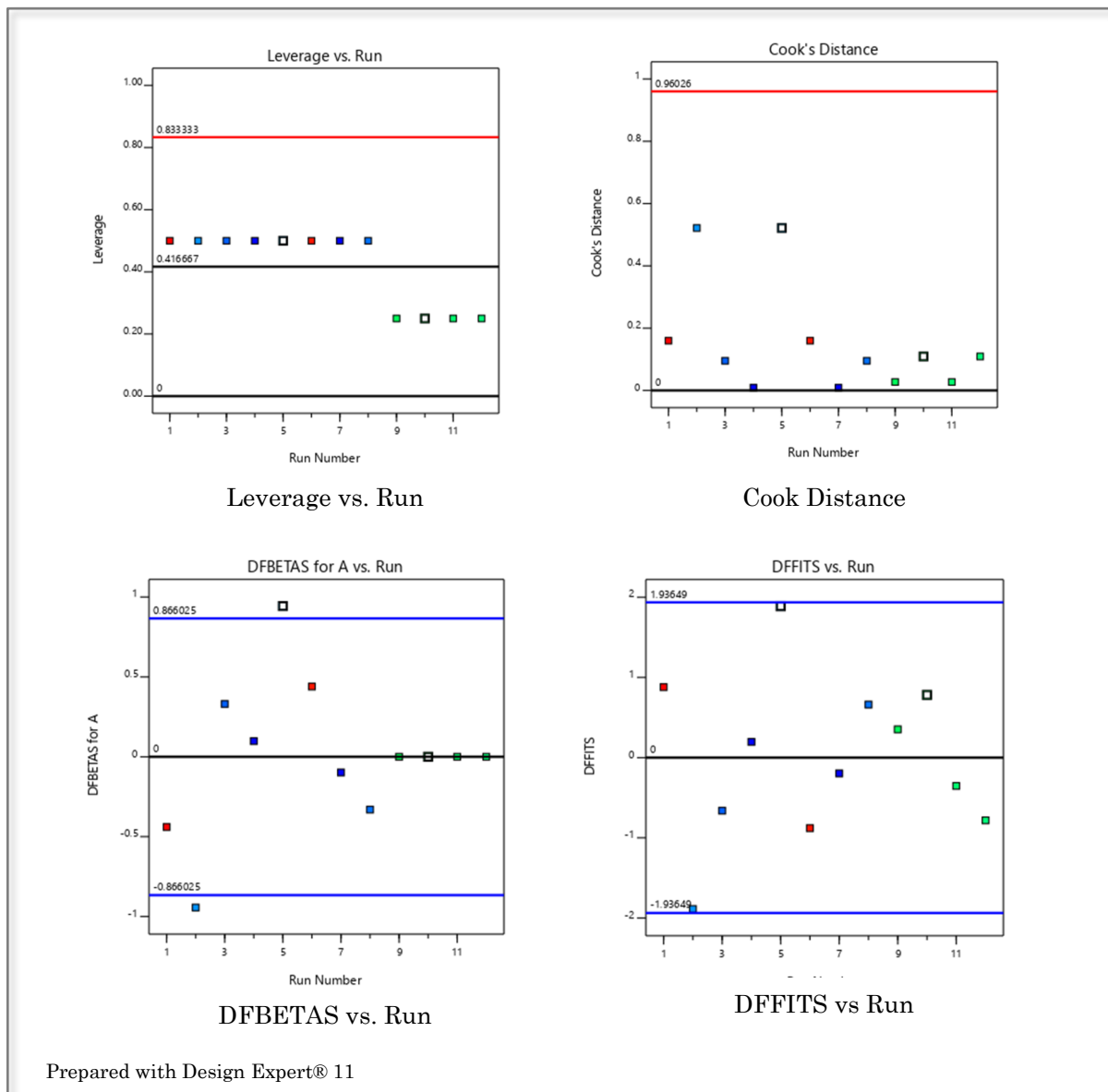


Figure 34. Graphical methods for diagnostics

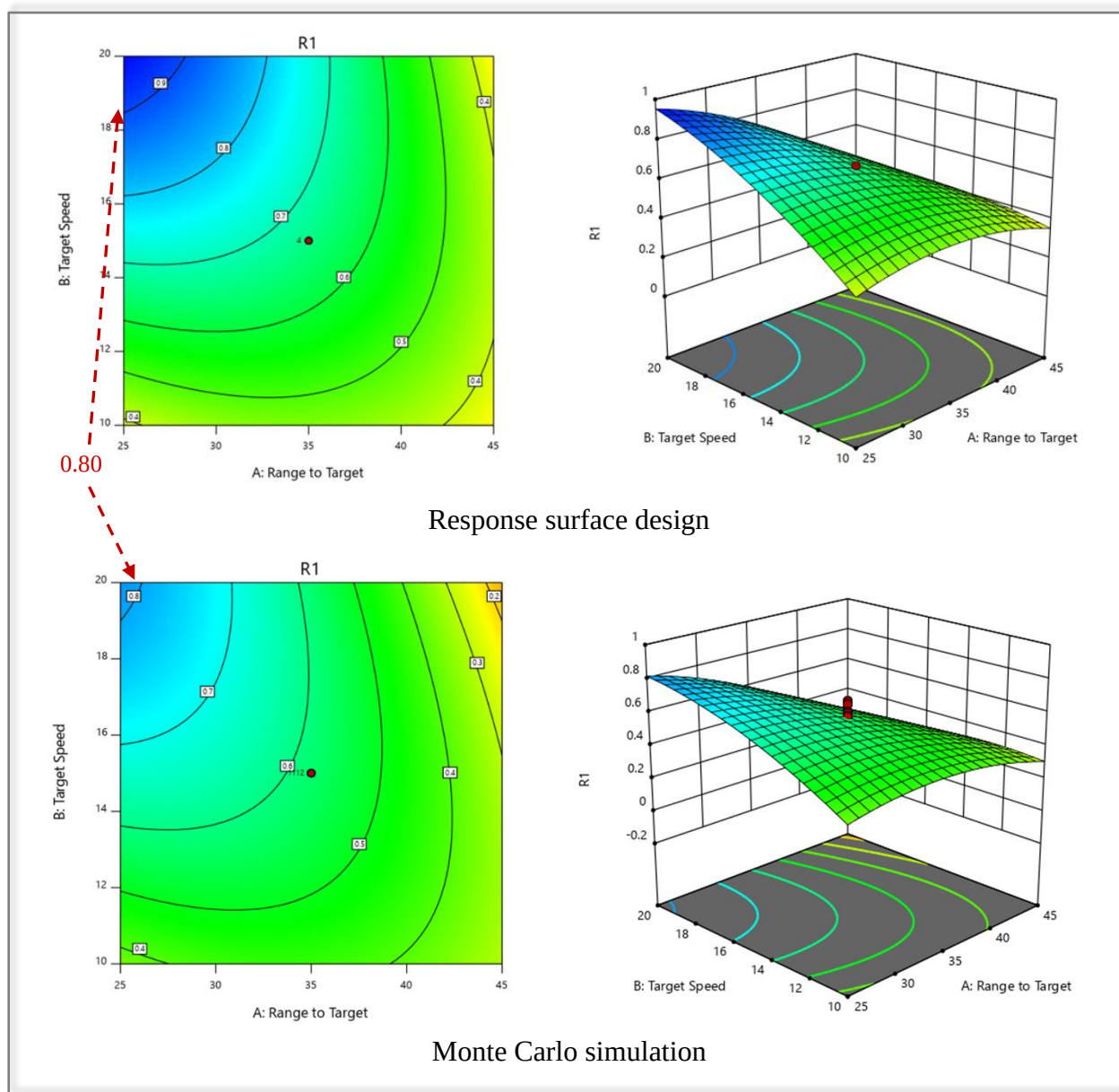


Figure 35. Contour maps and response surfaces comparisons

DOT&E Memo of 26 June 2013 criticized the practice of reporting a single average of a system's performance across all test conditions. Instead, the analysis should provide estimates of system performance anywhere within its operational space. These estimates should be accompanied with some statistical interval that accounts for uncertainty due to limited sample data.

Three commonly calculated intervals are the confidence, prediction, and tolerance intervals. A confidence interval provides a calculated range of values based on sampled data where the true population parameter (e.g. mean) likely resides. Confidence intervals consider the variation in the estimation. A prediction interval provides a range of values where the response is expected to fall in within the factor space in the future. Prediction intervals are wider than confidence

intervals due to not only considering the variation in the estimation, but also the variation in the response. A tolerance interval provides a range of values where a user specified X% of the population should fall. Tolerance intervals are not widely used as prediction and confidence intervals and are thus underutilized in DoD testing.

Prediction and tolerance intervals are highly sensitive to the assumption that the data collected follows the normal distribution. Diagnostic plots and tests for normality should be conducted to ensure this assumption is not violated. If there is a violation of normality, a transformation of the response could be employed such that the transformed response is normal. Interpretation of the transformed results can be difficult since estimates are in a transformed scaled and not in the real-world scale. Another alternative would be to employ a distribution free calculation of the interval.

DOE combined with regression analysis and statistical intervals allows a program to assess a system's performance with greater rigor than the practice of reporting a single average of a system's performance across all test conditions. DOE helps define the test space and helps determine the specific test scenarios that would best aid the analysis in the most efficient number of runs. Regression analysis helps create an empirical model that can be used to predict future performance. Statistical intervals quantify the level of uncertainty in the performance estimates from the empirical model.

Table 25. Strategies for data analysis

Strategy	Outcome
Outline how the analysis will address the objectives	Focuses data analysis on informing assessments of performance, reliability, interoperability, and cybersecurity
Quantify uncertainty	Quantifies statistical uncertainty
Analyze the data with rigorous and appropriate analysis techniques, including inferential statistics, that are consistent with test protocols (completely randomized designs, randomized block designs, or split-plot designs)	Produces the correct characterization of system performance
Analyze the data immediately after the test to identify outliers and test protocol problems	Affords the opportunity to recommend re-running tests that have suspect results
Verify that the fundamental DOE assumptions have not been violated – independence, normality, and constant variance	Ensures validity of test results and model verification
Use the appropriate form of regression techniques (based on the response variables, factors, and test protocols) to build an empirical model	Produces an empirical model that characterizes system performance across the entire test space and that can aid in assessment, training, and warfighting analysis
Validate results from previous tests, when appropriate	Produces information about the maturity of the systems as it progresses through the lifecycle

4.8 Inform Key Knowledge Points and Decisions

T&E produces knowledge about the capability of the system by comparing the analysis of empirical observations obtained from stimulating the system to measure its technical and operational attributes. The goal is to transform the knowledge gained from testing into “decision-quality information” to inform the system engineering process, key acquisition decisions, and determinations of operational effectiveness, operational suitability, survivability, and safety for intended use. The use of DOE-based test design can simultaneously help generate test efficiencies, improve the fidelity of test results, illuminate risks, enable better-informed decisions, and ultimately, enable fielding a more effective, suitable, survivable, lethal, and safe capability.

The strategy for this phase is to:

- Package information to support decisions related to test strategy activities
- Organize data analysis results to help draw informed decisions about the capability of the system

The positive outcomes of DOE based test design are far reaching and provide substantial contributions to the test strategy. Those contributions are listed in Table 26.

Table 26. Contributions of the test design framework to the test strategy

Deliverables	Contributions
Requirements Traceability Matrix	Testable and measurable response variables and a holistic factor space
Scenario Certification	Refence to authoritative models for predictive analytics
Test and Evaluation Master Plan, Test Strategy Briefing, and Test Concept Brief	Test design details such as methodology, design properties, figures of merit, overall test design strategy, resources, and objectives
Test Readiness Review (TRR), Analysis Structure (Evaluation Framework with supporting Data Source Matrix)	Consistent, defensible plans that can be incorporated into TRR entrance criteria
Test Plan, Data Collection Plan, Data Management Plan, and Instrumentation Plan	Protocols for detailed test procedures and their associated data collection, management, and instrumentation plans coded from the test design run matrix
Data Authentication Group	Appropriate statistical data analysis techniques, assessment methods, and evaluation criteria plus a suite of diagnostic methods that permit the validation of test design assumptions and the identification of outliers
Reports and Briefings (Hot Wash, DAG Briefings, Emerging Results Briefing, Quick Look, Final Test Report)	Briefing templates and reports that contain data analysis that reflect the rigorous, disciplined approach embedded in the DOE-based test design framework
Risk Assessments	A framework that illuminates the tradeoff between risk, cost, and utility of information

DOE methodologies can also help produce defensible artifacts to inform decisions in activities and processes such as:

- *Performance assessment* - DOE has a large repertoire of **screening** designs that are practical for identifying significant factors and interactions that influence system performance while requiring a minimum number of trials
- *Tactical decision making* - DOE can produce empirical models that are useful for **characterizing** the performance of a system over the entire test space
- *Analysis of alternatives* - DOE can be used to gather information for **comparing** competing alternatives tested in common background conditions purposely and systematically varied
- *System or mission area qualifications* – DOE could be used for quantifying mission effectiveness, personnel, and readiness by **comparing** results of test events over time
- *Product improvement* - DOE can be used for **comparing** the effectiveness of the system prior to and after corrections of deficiencies or any other engineering changes made to systems under development as well as fielded, such as ordnance alterations (ORDALT)
- *Regression testing* – DOE can help **confirming** that a system is robust to recent changes
- *Trade-off analysis* - DOE allows for defining and **optimizing** system attributes based on multi-criteria objective function relating cost, schedule, and performance that require a compromise to achieve a desirable outcome
- *Sensitivity analysis* - DOE has protocols to identify and minimize the influence of operating conditions found in the field on capability attributes

It is important to organize the results to help draw appropriate conclusions and practical recommendations about the test process, test structure, test designs, and more importantly, about the performance of the system. Potential knowledge points may require evaluating:

- Mission readiness and combat capability
- System design and specifications
- Manufacturing processes
- Correction of deficiencies (design, implementation, readiness, training)
- Performance limiters
- Models and simulations

Table 27. Strategies for informing the evaluation

Strategy	Outcome
Draw rational and practical conclusions about system performance, the overall test process, and the efficiency of the test design	Provides unambiguous, decision-quality information to decision makers
Validate the empirical results and the model	Enables the model to support tactical decision making
Use DOE results to inform the assessments - performance, interoperability, reliability, and cybersecurity	Facilitates tradeoffs in risks, cost, and utility of information

5 Summary and Recommendations

T&E provides essential information to decision makers regarding the capabilities and operational conditions of a system. There are key strategies that should be employed to ensure a successful evaluation and a successful test design process. When correctly employed, these strategies will produce better interpretation of test data and a better understanding of the state of the system capabilities and the risks associated with the decisions to be made. Clearly, the DOE framework meets the expectations set forth by DoDI 5000.UF, the DOT&E Memos, and the NRC report. The outcomes of DOE demonstrate the exceptional benefits that can be achieved when key stakeholders – spanning program management, systems engineering, and test equities – work together under a standardized, rigorous approach for test planning. Some of the practical benefits afforded by injecting the rigor of *design and analysis of experiments* into the fabric of T&E include:

- Means for identifying factors and conditions that affect the response variables, and consequently the mission
- Ability to identify interactions between test factors
- Methods for strategically varying factors and for controlling sources of unwanted variability that could influence performance
- Methods to make informed trade-offs between the cost of test and the quantity and quality of information obtained
- Optimized run matrices that optimize the number of tests needed to obtain the required information for the evaluation
- Statistical figures of merit (such as confidence, power, and aliasing) to assess proposed test design, which are key in understanding the number and placement of test runs and the likelihood of successfully identifying performance drivers
- Changes in the resources required due to adding or removing factors or changing their levels
- Analysis techniques that go beyond single hypothesis tests
- Graphical methods for assisting in the statistical analysis of test data in the presentation of test results to inform the decisions

The place of design and analysis of experiments in T&E is illustrated in Figure 36. *Design and analysis of experiments* adds rigor and discipline to T&E and facilitates a comprehensive understanding of the tradeoffs in the techno-programmatic domains of risks, cost, and utility of information. DOE can help increase mission impact, reduce test assets, shorten the test schedule, and providing decision-quality information to the warfighter and decision makers. *Design and analysis of experiments* leverages statistical engineering methods and know-how into statistically driven strategies for planning, executing, analyzing, and reporting tests that produce:

- Well-defined mission-oriented objectives within the scope of the decision to be informed
- Testable mission-oriented response variables linked to capability evaluations

- Complete coverage of the test space mapped to capability attributes and a comprehensive list of factors and conditions that influence the attributes
- Efficient and effective test designs with good statistical properties
- Disciplined test execution protocols and well-developed run matrices
- Inferential statistical techniques and standardized evaluation criteria to analyze the data
- Defensible and reliable decision-quality products to inform decisions

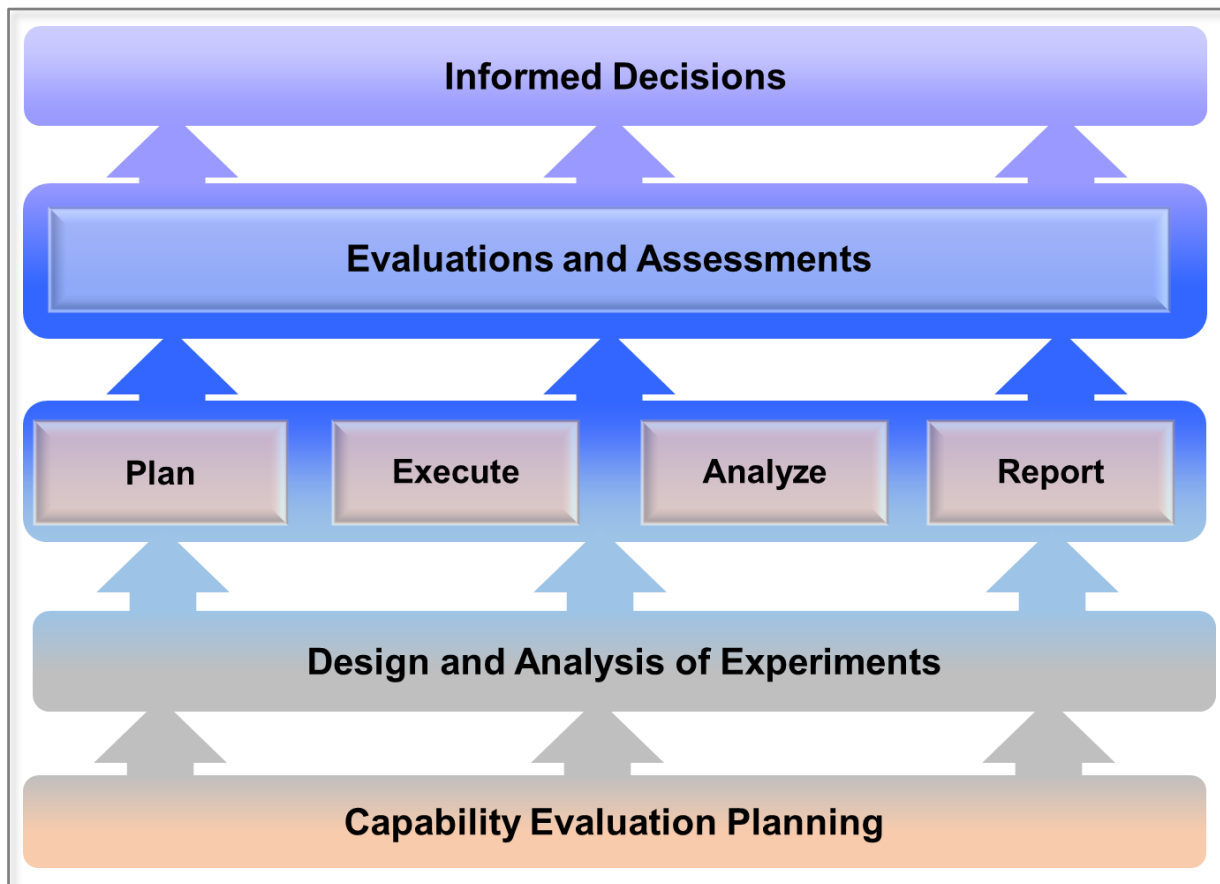


Figure 36. Evaluation and test process flow

DOE allows a program to assess a system's performance with greater rigor than the practice of reporting a single average of a system's performance across all test conditions. DOE helps define the test space and helps determine the specific test scenarios that would best aid the analysis in the most efficient number of runs. Regression analysis helps create an empirical model that can be used to predict future performance. Statistical intervals quantify the level of uncertainty in the performance estimates from the empirical model.

MITRE recommends NSWC PHD: (1) adopts and starts implementing the CBT&E philosophy; (2) adopts and starts implementing the test design framework outlined below; and (3) incorporates the test design process into the digital engineering transformation initiative, particularly into an MBSE strategy to calibrate the authoritative capability model with empirical data.



- Identify key knowledge points and decisions to be informed by the evaluation of the system capabilities
 - ✓ Charter a DOE WG to facilitate communications and foster a common understanding of the test strategy
 - ✓ Identify knowledge points and decisions across the lifecycle
 - ✓ Identify technical and operational capabilities to be evaluated to gain knowledge required to inform decisions
 - ✓ Identify mission area measures and technical parameters (response variables) to quantify the systems capabilities
 - ✓ Identify LVC events required to generate the data for estimating the response variables
 - ✓ Document all the information above in the XEF and use it to guide evaluation planning and test design
- Identify key operational measures and technical parameters (i.e. response variables) that serve as means to quantify system capabilities
 - ✓ Identify the response variables associated with the capabilities to be evaluated
 - ✓ Identify the data type (e.g. continuous, discrete, ordinal nominal) for each response variable
 - ✓ Draw functional flow block diagrams and mission decomposition diagrams
 - ✓ Assess the metrology system that will be used to measure each response
 - ✓ Baseline sources and type of variability that could influence each response
 - ✓ Identify the expected range of values for each response



- Develop practical evaluation objectives
 - ✓ Formulate clear, testable, measurable, and unbiased evaluation objectives of practical consequences
 - ✓ Develop a sequential evaluation and test framework
 - ✓ Identify the purpose of each objective: screening, characterization, etc.
- Identify the test space
 - ✓ Identify all potential design factors and all nuisance factors that could influence the response variables, their type, and how to control them
 - ✓ Use factors that make sense and that are consistent with the evaluation objectives
 - ✓ Identify potential interactions between factors
 - ✓ Select factor levels that are practical and that can produce the maximum amount of information with the minimum number of runs
 - ✓ Identify combinations of factors that are not feasible (constraints and disallowed combinations)
 - ✓ Establish if blocking or other randomization restrictions are involved

-
- Construct a design that is adequate for the type of event, objectives, and technical attributes of interest
 - ✓ Design a sequential ensemble of test designs, if practicable, for each phase of test that is adequate for the event and for the response variable of interest
 - ✓ Decide on the desired confidence ($1 - \alpha$)
 - ✓ Decide on the change (δ) that is important to detect for each response
 - ✓ Estimate the standard deviation (σ) for each response variable
 - ✓ Choose a model that describes the relationship between factors and response variables
 - ✓ Select and set-up a set of candidate test designs while considering situations such as hard-to-change factors, constraints, disallowed combinations, and number of runs per test period (blocking)
 - ✓ Evaluate the candidate designs and compare them based on figures of merit such power ($1 - \beta$), sample size, aliasing among effects, prediction variance, and design efficiency
 - ✓ Select the design that fits the situation and determine if the sample size is adequate to meet program objectives (may have to negotiate for additional samples, test time, or additional budget)
 - ✓ Generate a run matrix specifying the number of replicates and blocks



- Coordinate and execute the test design strategy
 - ✓ Prepare evaluation plans, test plans, and data collection plans
 - ✓ Ensure the physical set-up and the test execution matrix are consistent
 - ✓ Execute the run matrix as planned
 - ✓ Record and collect the data
 - ✓ Perform a quick look analysis of the data before leaving the test stage



- Perform exploratory data analysis
 - ✓ Use descriptive statistics to look for behavior and patterns in the data and to describe system capabilities. Use inferential statistics to draw conclusions about those capabilities
 - ✓ Identify factors and interactions that have a significant effect on the responses
 - ✓ Fit a model to the data using numerical and graphical analysis methods
 - ✓ Validate assumptions and calibrate the model
 - ✓ Select the final model



- Inform key knowledge points and evaluations
 - ✓ Package information to support decisions related to test strategy activities
 - ✓ Organize test data analysis results to help draw informed decisions about the capability of the system

This page intentionally left blank.

References

(Policy and Guidance)

1. DoD Test and Evaluation Management Guide (6th ed., 2012). Defense Acquisition University Press. Retrieved from <http://www.acqnotes.com>
2. Defense Acquisition Guidebook (DAG), Chapter 8, Test & Evaluation. Defense Acquisition University. Retrieved from <https://dag.dau.mil>
3. Office of the Under Secretary of Defense for Research and Engineering Website <https://www.cto.mil/>
4. OSD Memorandum Definition of Integrated Testing dated 25 April 2008
5. DoD Directive 5000.01, The Defense Acquisition System (Change 2). 31 August 2018. Retrieved from AcqNotes, [https://www. <http://acqnotes.com/acqnote/acquisitions/dodd-5000>](https://www.acqnotes.com/acqnote/acquisitions/dodd-5000)
6. DoD Instruction 5000.02, Operation of the Adaptive Acquisition Framework, 18 August 2020
7. *Implementing Capabilities-Based Test and Evaluation (CBT&E): A Guide for Test Professionals* (Draft Revision 4). US Navy Cross-SYSCOM CBT&E Integrated Product Team. 5 August 2020.
8. NIST/SEMATECH e-Handbook of Statistical Methods, <http://www.itl.nist.gov/div898/handbook/>, 30 October 2013

This page intentionally left blank.

References

(Design and Analysis of Experiments)

1. Anderson-Cook, Interval Training. *Quality Progress* 2009; **42**: 58-60.
2. Anderson-Cook, C. M., Borror, C. M., and Montgomery, D. C. Response surface design evaluation and comparison (with discussion). *Journal of Statistical Planning and Inference* 2009; **139**: 629-674.
3. Bisgaard, S. The design and analysis of $2^{k-p} \times 2^{q-r}$ split-plot experiments. *Journal of Quality Technology* 2000; **32**: 39-56.
4. Box, G. E. P. Signal-to-noise ratios, performance criteria, and transformation. *Technometrics* 1988; **30**: 1-40.
5. Box, G. E. P. Sequential experimentation and sequential assembly of designs. *Quality Engineering* 1992; **5**: 321-330.
6. Box, G. E. P. Statistics as a catalyst to learning by the scientific method Part II—A discussion. *Journal of Quality Technology*; **31**: 16-29.
7. Box, G. E. P., and Behnken, D. W. Some new three-level designs for the study of quantitative variables. *Technometrics* 1960; **2**: 455-476.
8. Box, G. E. P., and Cox, D. R. An analysis of transformations. *Journal of the Royal Statistical Society, B* 1964; **26**: 211-243.
9. Box, G. E. P., and Draper, N. R. On minimum-point second-order designs. *Technometrics* 1974; **16**: 613-616.
10. Box, G. E. P., and Hunter, J. S. The 2^{k-p} fractional factorial designs, Part I. *Technometrics* 1961; **3**: 311-352.
11. Box, G. E. P., and Hunter, J. S. The 2^{k-p} fractional factorial designs, Part II. *Technometrics* 1961; **3**: 449-458.
12. Box, G. E. P., and Wilson, K. G. On the experimental attainment of optimum conditions (with discussion). *Journal of the Royal Statistical Society B* 1951; **13**: 1-45.
13. Cochran, W. G. Some consequences when the assumptions for the analysis of variance are not satisfied. *Biometrics* 1947; **3**: 22-38.
14. Cochran, W. G. Analysis of covariance; Its nature and uses. *Biometrics* 1957; **13**: 261-281.
15. Cohen, J. The cost of dichotomization. *Applied Psychological Measurement* 1983; **7**: 249-253.

-
16. Coleman, D. E., and Montgomery, D. C. A systematic approach to planning for a designed industrial experiment. *Technometrics* 1993; **35**: 1-27.
 17. Cook, D. R. Detection of influential observations in linear regression. *Technometrics* 1977; **19**: 15-18.
 18. Cook, D. R. Influential observations in linear regression. *Journal of the American Statistical Association* 1979; **74**: 169-174.
 19. Cortes, L. A. *Best practices for highly effective test design; Part 2 – Beginners guide to design of experiments in T&E*. Scientific Test and Analysis Techniques Center of Excellence (STAT COE) [White Paper], 2014.
 20. Cortes, L. A., Simpson, J. R., and Parker, P. A. Response surface split-plot designs; A literature review. *Quality and Reliability Engineering International* 2018; **34**: 1374-1389.
 21. Daniel, C. *Applications of statistics to industrial experimentation*. Wiley & Sons: New York, NY, 1976.
 22. Doehlert, G. H. Uniform shell designs. *Journal of the Royal Statistical Society, Series C* 1970; **19**: 231-239.
 23. Draper, N. R. Small composite designs. *Technometrics* 1985; **27**: 173-180.
 24. Fisher, R. A. The arrangement of field experiments. *Journal of the Ministry of Agriculture of Great Britain* 1926; **33**: 503-513.
 25. Fisher, R. A. *The design of experiments*. (8th ed.) Hafner Publishing Company, New York, NY, 1966.
 26. Goos, P. *The optimal design of blocked and split-plot experiments*. Springer: New York, NY, 2002.
 27. Gryze, et al. A tutorial on tolerance intervals of ordinary least-squares regression, *Chemometrics and Intelligent Laboratory Systems* 2007; **87**: 147-154.
 28. Hamada, M. The advantages of continuous measurements over pass/fail data. *Quality Engineering* 2002; **15**, 253-258.
 29. Harman, M. *Understanding requirements for effective test planning*. Scientific Test and Analysis Techniques Center of Excellence (STAT COE) [White Paper], 2018.
 30. Harman, M., Cortes, L. A., Hill, R. *Quantifying test risk using design of experiments*. Scientific Test and Analysis Techniques Center of Excellence (STAT COE) [White Paper], 2015.
 31. Hartley, H. O. Smallest composite designs for quadratic response surfaces. *Biometrics* 1959; **15**: 611-624.

-
32. Hoke, A. T. Economical second-order designs based on irregular fractions of the 3^n factorial. *Technometrics* 1974; **16**: 375-384.
 33. John, J. A., and Prescott, P. Critical values of a test to detect outliers in factorial experiments. *Applied Statistics* 1975; **24**: 56-59.
 34. Jones, B., and Nachtsheim, C. J. Split-plot designs: What, why, how. *Journal of Quality Technology* 2009; **41**: 340-361.
 35. Jones, B., and Nachtsheim, C. J. A class of three level designs for definitive screening in the presence of second-order effects. *Journal of Quality Technology* 2011; **43**: 1-15.
 36. Jones, B. 21st century screening experiments: What, why, and how? (With discussion). *Quality Engineering* 2016; **28**: 98-126
 37. Kiefer, J. C. Optimum experimental design (with discussion). *Journal of the Royal Statistical Society B* 1959; **21**: 272-319.
 38. Kiefer, J. C. Optimum designs in regression problems. *Annals of Mathematical Statistics* 1961; **32**: 298-325.
 39. Lucas, J. M. Optimum composite designs. *Technometrics* 1974; **16**: 561-567.
 40. Meeker, W. Q., Hahn, G. J., and Escobar, L. A. *Statistical Intervals: A Guide for Practitioners*, Gilmore, 2012
 41. Montgomery, D. C. *Design and analysis of experiments*. (9th ed.). Wiley & Sons: New York, NY, 2017.
 42. Morris, M. A class of three-level experimental designs for response surface modeling. *Technometrics* 2000; **42**: 111-121.
 43. Myers, R. H., Montgomery, D. C., and Anderson-Cook, C. M. *Response surface methodology*. (3rd ed.). Wiley & Sons: New York, NY, 2016.
 44. Natrella, M. G. *Experimental Statistics, NBS Handbook 91*, US Department of Commerce, 1963.
 45. Notz, W. Minimal point second-order designs. *Journal of Statistical Planning and Inference* 1982; **6**: 47-58.
 46. Oehlert, G. W., and Whitcomb, P. Small, efficient, equ-replicated resolution V fractions of 2^k designs and their application to central composite designs. *46th Annual Fall Technical Conference* 2002, ASQ and ASA: Valley Forge, PA.
 47. Ortiz, F. *Categorical data in a designed experiment Part 1; Avoiding categorical data*. Scientific Test and Analysis Techniques Center of Excellence (STAT COE) [White Paper], 2018.

-
48. Ortiz, F. *Categorical data in a designed experiment Part 2; Sizing with a binary response*. Scientific Test and Analysis Techniques Center of Excellence (STAT COE) [White Paper], 2018.
 49. Plackett, R. L, and Burman, J. P. The design of optimum multifactorial experiments. *Biometrika* 1946; **33**: 305-325, 328-332.
 50. Perry, M. A prediction interval estimator for the original response when using Box-Cox transformations. *Journal of Quality Technology* 2015; **47**: 278-297.
 51. Pesotchinsky, L. L. D-optimum and quasi D-optimum second-order designs on a cube. *Biometrika* 1975; **62**: 335-340.
 52. Rechtschaffner, R. I. Saturated fractions of 2^n and 3^n factorial designs. *Technometrics* 1967; **9**: 569-575.
 53. Roquemoire, K. G. Hybrid designs for quadratic response surfaces. *Technometrics* 1976; **4**: 419-423.
 54. Rucker, A. Improving statistical rigor in defense test and evaluation: Use of tolerance intervals in designed experiments. *Defense Acquisition Research Journal* 2014; **21**: 804-825.
 55. Taguchi, G. *System of experimental design, Vols. 1 and 2*. Kraus International: White Plains, NY, 1987.
 56. Vieira H., Sanchez, S. M., Kienitz, K. H., and Belderrain, M. C. N. Efficient, nearly orthogonal-and-balanced, mixed designs: an effective way to conduct trade-off analyses via simulation. *Journal of Simulation* 2013; **7**: 264-275.
 57. Westlake, W. J. Composite designs based on irregular fractions of factorials. *Biometrics* 1965; **21**: 324-336.
 58. Wu, C. F.F., and Hamada, M. *Experiments; Planning, analysis, and optimization*. (2nd ed.). Wiley & Sons: Hoboken, NJ, 2009.

Appendix A

Top T&E Policy and Guidance in DoD and DoN

This page intentionally left blank.

Top T&E Policy and Guidance in DoD and DoN

This appendix outlines top T&E policy within DoD and DoN that governs the application of DOE in T&E.

A1 Top DoD T&E Policy

DoDD 5000.01

The Defense Acquisition System (DoDD 5000.01) (Change 2) of 31 August 2018 outlines the overarching policy, operating principles, and management process that govern the system by which the DoD provides products to the users. DoDD 5000.01 assigns responsibilities to key positions within OSD beyond those stipulated in Title 10 of the United States Code (USC) and other DoD directives. DoDD 5000.01 applies to OSD, the Military Departments, the Office of the Chairman of the Joint Chiefs of Staff (CJCS) and the Joint Staff, the Combatant Commands, the Office of the Inspector General (IG) of the DoD, DoD Agencies, DoD Field Activities, and all other DoD Components. For this document, two policy statements are key:

- Use an adaptive acquisition framework employing tailored management pathways to emphasize these principles
- Employ capability portfolio management, mission engineering, and integration analysis using an effects/kill chain framework to assess the integration and interoperability of the SoS required to execute critical mission requirements

DoDI 5000.02

Operation of the Adaptive Acquisition Framework (DoD Instruction 5000.02) of 23 January 2020 provides policies to support the Defense Acquisition System with the objective of delivering effective, suitable, survivable, sustainable, and affordable solutions to the end user in a timely manner. To achieve those objectives, MDAs, other Decision Authorities (DAs), and PMs have broad authority to plan and manage their programs consistent with sound business practice. The AAF, illustrated in Figure A-1, outlines the acquisition pathways that affords acquisition authorities and program managers opportunities to develop acquisition strategies and employ acquisition processes that match the characteristics of the capability being acquired. In addition to a DoD instruction for each pathway, there is a DoD instruction for each of ten functional areas as illustrated in Figure A-2. DoDI 5000.UF is the *Test and Evaluation Instruction*.

DoDI 5000.UF

DoD Test and Evaluation Instruction (DoDI 5000.UF), a new instruction currently in coordination, establishes the T&E policy and procedures across the AAF. It encapsulates Enclosure 4 (DT&E) and Enclosure 5 (OT&E and LFT&E) from DoDI 5000.02T. DoDI 5000.UF requires DoD to:

- Establish and implement T&E policy and procedures across all pathways of the AAF
- Conduct IT&E, whenever feasible, as part of an adequate T&E program
- Conduct DT&E, OT&E, and LFT&E as part of an adequate T&E program

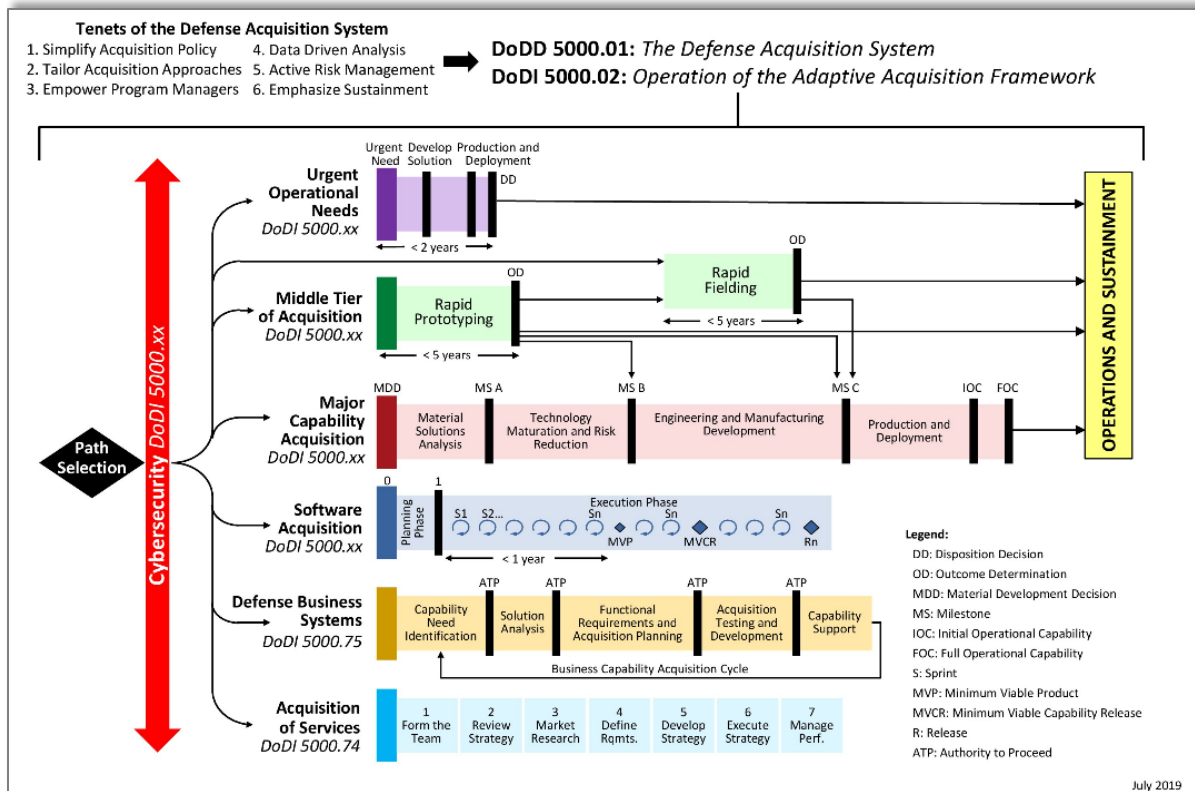


Figure A-1. The Adaptive Acquisition Framework (AAF)

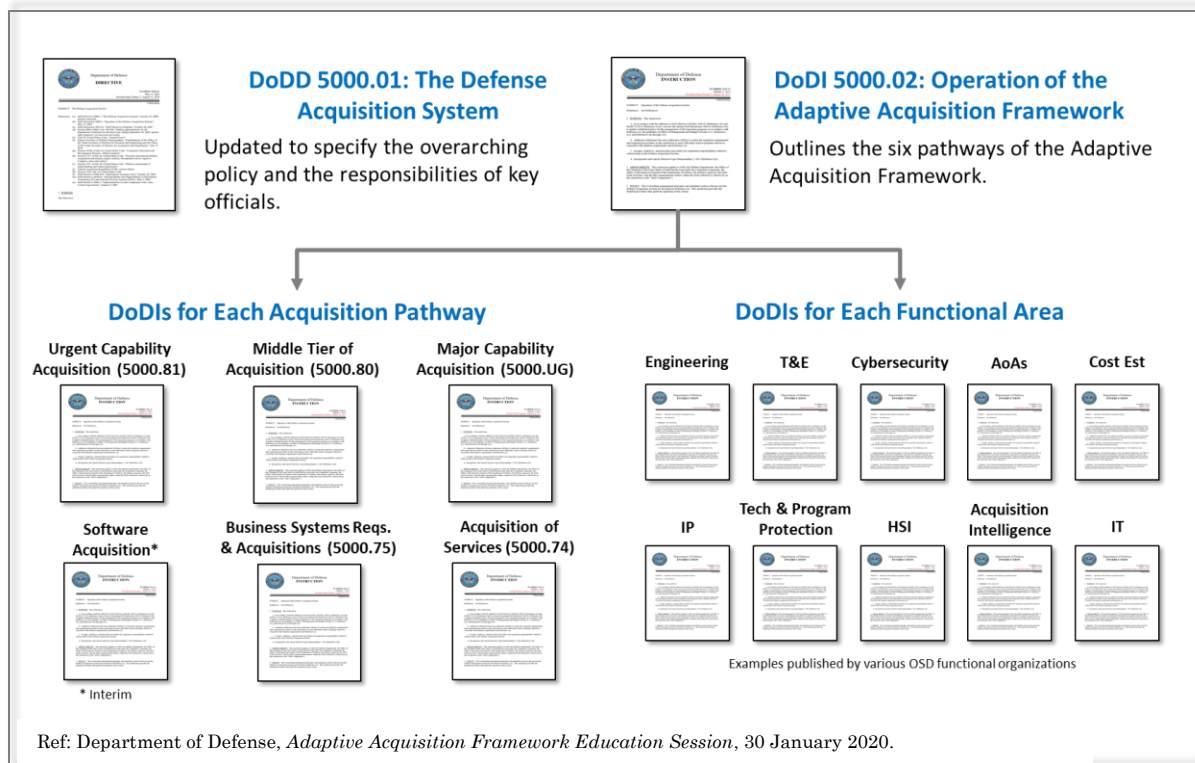


Figure A-2. Defense Acquisition System policies

DoDI 5000.UF:

- Outlines all T&E responsibilities and mutually supportive oversight roles of Senior OSD T&E stakeholders, and the responsibilities of the PM and the T&E Working-level Integrated Product Team (WIPT)
- Outlines the procedures and statutory requirements for each AFF pathways
- Defines DT&E, OT&E, and LFT&E activities
- Requires the Programs to identify how design and analysis of experiments “will be used to design and effective and efficient test program that will produce the required data to characterize system behavior and combat mission capability across an appropriately selected set of factors and conditions”

DoDI 5000.UF requires the TEMP includes a developmental evaluation methodology starting at MS-B that provides essential information on programmatic and technical risks as well as information for major decisions. The strategy should include:

- Decisions to be informed by an evaluation of the system’s capabilities
- Capability evaluation to inform those decisions
- Test and M&S events to be conducted to generate the data for the evaluation
- Resources to be used and schedule to be followed to execute the test events

DoDI 5000.UF adds new language to promote the use of IT&E. DoDI 5000.UF and OSD Memorandum *Definition of Integrated Testing* dated 25 April 2008 define IT&E as the collaborative planning and simultaneous execution of test phases and events to provide shared data in support of independent analysis, evaluation, and reporting by all stakeholders. IT&E focuses on designing, developing, and implementing a comprehensive and economical test strategy that produces information that can be used collaboratively and unobtrusively by the organizations participating in the test event to provide evaluation results from different points of views to the decision makers. Figure A-3 illustrates the comparison between a traditional test program and an integrated test program.

DOT&E Memos

Operational Test Agency (OTA) Directors Memorandum of Agreement (MOA) *Using design of experiments for operational test and evaluation* of 20 February 2009 endorsed the use of DOE for OT&E:

- Experimental design provides a valuable tool to identify and mitigate risk in all test activities
- It offers a framework from which test agencies may make well-informed decisions on resource allocation and scope for adequate testing
- Using DOE in all phases of program DT&E and OT&E affords the opportunity for rigorous systematic improvement in test processes

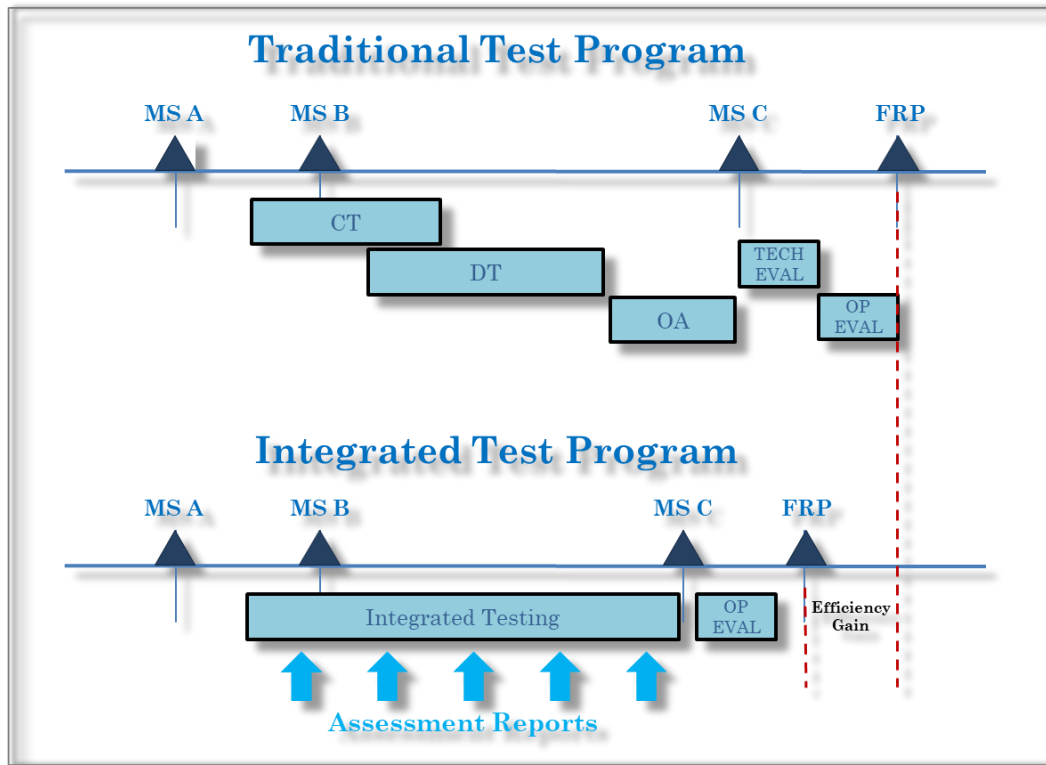


Figure A-3. Traditional vs. integrated test programs

Guidance on the use of design of experiments (DOE) in Operational Test and Evaluation of 19 October 2010 sets expectations on elements of DOE that must be addressed in the TEMP:

- The goal of the experiment should reflect evaluation of end-to-end mission effectiveness in an operationally realistic environment
- Quantitative mission-oriented response variables for effectiveness and suitability
- Factors that affect those response variables and a good breadth of coverage of their associated levels
- A method for strategically varying factors across both developmental and operational testing with respect to responses of interest
- Statistical measures of merit on the response variables for which it makes sense

Case studies for the use of design of experiments (DOE) in Developmental Test and Evaluation (DT&E) of 21 January 2011 examines the applicability of DOE to DT&E activities, from early engineering analysis, through incremental development, to final verification of requirements:

- Industry has a long history of using DOE, and it is considered a best practice
- DOE has many benefits, including coverage of the test space, high probability of detecting faults, potential efficiencies, and the ability to quantify risks to test programs
- DOE is applicable in many areas of DT&E

Flawed application of design of experiments (DOE) to Operational Test and Evaluation (OT&E) of 26 June 2013 reinforces the expectations set forth in the October 2010 memo and addresses areas for improving the use of DOE:

- Create clear test goals for each phase of test
- Use relevant and informative mission oriented metrics
- Evaluate the system across the conditions under which the system will be employed
- Strategically vary and control the factors
- Avoid confounding factors
- Avoid rolling up results in the analysis of data
- Avoid single hypothesis tests to size the design
- Understand the consequences of adding or removing factors
- Statistical tests of significance

Best practices for assessing the statistical adequacy of experimental designs used in Operational Test and Evaluation of 23 July 2013 identifies statistical figures of merit that should be used to determine the adequacy of a test design and:

- Establishes the importance of statistical power in test design
- Links the importance of a clear test goal and the appropriate tests designs for screening, characterization, optimization, and comparison
- Advocates different statistical figures of merit for assessing test design, such as empirical model strength, collinearity, power, and confidence, scaled prediction variance, and optimality criteria

Inadequacy of recently proposed test design for the P-8A Increment 2 and Multistatic Active Coherent (MAC) Programs of 24 September 2013 identifies deficiencies in the test design concept for evaluating the P-8A Increment 2 and MAC systems:

- The design is insufficient to assess the operational performance of the P-8A aircraft crew conducting Anti-Submarine Warfare (ASW) search
- The design is insufficient to assess the operational performance of the MAC system
- The design does not account for the critical roles both the operator and environment play in ASW search

Guidance on the validation of models and simulation used in Operational Test and Live Fire Assessments of 14 March 2016 sets expectations for the TEMP and Test Plan to describe the M&S capability, its intended use, and its validation and accreditation approach, including a discussion covering:

- Response variables and/or mission-level metrics of interest
- Range of conditions over which the M&S will be validated
- Plan for collecting the necessary live and simulation data for M&S validation

-
- Analysis of statistical risk
 - Validation methodology

Clarification on guidance on the validation of models and simulation used in Operational Test and Live Fire Assessments of 17 January 2017 re-emphasizes the expectations for the M&S validation strategy, including the quantitative comparison of live and M&S outcomes using statistical methods and a comprehensive strategy to assess M&S output across the operational domain for which the M&S will be accredited. At a minimum, the following information should be included in the validation and accreditation plan:

- A strategically selected set of test points designed to compare the M&S and live data
- Analysis methodologies to empirically model the live and M&S outcomes and test for statistical differences between the two outcomes
- A robust design for the M&S that systematically covers the range of operationally realistic inputs over which the model will be accredited—space-filling design methodologies are preferred
- A quantification of uncertainties associated with the M&S to the maximum extent possible

A2 TOP DoN T&E Policy

SECNAVINST 5000.2F

Defense Acquisition System and Joint Capabilities Integration and Development System Implementation (SECNAVINST 5000.2F of 26 March 2019) prescribes specific policies and procedures for the operation of the JCIDS and Defense Acquisition system within DoN. Enclosure (5), Test and Evaluation, supplements DoDI 5000.02 with policies for DT&E, OT&E, and LFT&E.

SECNAV-M 3960.1

Department of the Navy (DoN) Test and Evaluation Guidebook (SECNAV-M 3960.1) supplements chapter 5 of SECNAVINST 5000.2F with policy, guidance, process, and best practices for DT&E, OT&E, and LFT&E. It consolidates existing T&E policy in *DoD Test and Evaluation Instruction* (DoDI 5000.UF), *Business Systems Requirements and Acquisition* (DoDI 5000.75), *Middle Tier of Acquisition* (DoDI 5000.80), *Major Capability Acquisition* (DoDI 5000.81), and *Urgent Capability Acquisition* (5000.81).

Detailed policy, guidance, and processes are captured in the following twelve enclosures: References; DoN T&E Organizational Roles and Responsibilities; T&E Planning; Phases of T&E; CBT&E; Cybersecurity; T&E Funding Responsibilities; Scheduling RDT&E Fleet Support; M&S, Environmental, Safety, and Occupational Health in T&E; Templates and Enclosures; and Acronyms. This document touches on T&E Planning, Phases of T&E, CBT&E, and M&S.

NAVSEAINST 3960.2E

Test and Evaluation Policy (NAVSEAINST 3960.2E) has not been promulgated and will replace NAVSEAINST 3960.2D of 22 April 1988. NAVSEAINST 3960.2E provides policy and procedures for acquisition related T&E activities of programs and systems managed by NAVSEA. It implements policy for NAVSEA, affiliated PEOs, and NAVSEA Warfare Center Divisions.

Detailed policy, guidance, and processes are captured in a ten-chapter Technical Manual as follows: Introduction; Enterprise T&E; Planning, Execution, Analysis, and Reporting (PEA&R process); CBT&E; T&E and Digital Transformation; T&E Tools and Enablers; Cybersecurity T&E; LFT&E; DT Sufficiency; and T&E for Open Architecture and Interoperability. This document touches on the PEA&R process and CBT&E.

PEOIWSINST 3960.2

Planning, Execution, Analysis, and Reporting of Integrated Combat/Weapon System Test and Evaluation Events (PEOIWSINST 3960.2) of 27 April 2015 provides a standardized process for the planning, execution, analysis, and reporting (PEA&R) of PEO IWS integrated combat and weapon system T&E events. The IWS T&E process is illustrated in Figure A-4.

Planning a test event starts with the development of test objectives (step 1 in Figure A-4). Those test objectives are mapped to the TEMP, CDD, Naval Capabilities Document (NCD), systems specifications, M&S VV&A plans, or Fleet Concept of Operations (CONOPS). Subsequently, the test objectives are reviewed to allow for scenario development (step 2 in Figure A-4) and to incorporate analysis requirements and its associated key performance parameters (KPP), measures of effectiveness (MOE), and measures of performance (MOP). Downstream in the process, the test is conducted (step 6 in Figure A-4) and the data is analyzed (step 8 in Figure A-4) based on the test objectives, KPPs, MOEs, and MOPs. Finally, comprehensive reviews are performed and a report (step 9 in Figure A-4) is generated. Steps 1, 2, 6, 8, and 9 are integral steps of the DOE process and are discussed in detail in the subsequent sections.

NAVSEAINST 9093.1D

Combat System Ship Qualification Trials for Surface Ships of 15 August 2019 provides policy and identifies responsibilities for Combat System Ship Qualification Trials (CSSQT). As it relates to this document, the policy requires:

- A closed-loop system engineering approach that provides for long range and near-term planning, execution, reporting, and feedback
- Repeatable metrics-based method to help determining the significance of the combat systems changes made during availability or cumulative availabilities
- A hierarchical approach to identify test requirements from element and system level capabilities to platform, ship class, and inter-platform capabilities
- Test objectives that provide resource requirements, MOPs, MOEs, data collection requirements, data analysis methods, and determination of when M&S is required as a prerequisite
- The identification of test resources to support CSSQT events

TP-60

Surface Ship Weapon/Combat Systems Test and Evaluation Guide (NSWC PHD TP-60) of 31 July 2017 is a resource for the organization's workforce to support T&E throughout the lifecycle, including Combat System Ship Qualification Trials (CSSQT). The document provides consolidated and standardized T&E guidance and serves as a training guide. It contains examples, templates, checklists, step-by-step procedures, and references to policy, standards, and tutorials to aid test engineers in organizing the T&E. The document is organized in fourteen sections, each of

which links completely or partially to portions of the T&E process. Figure A-5 illustrates the organization of TP-60.

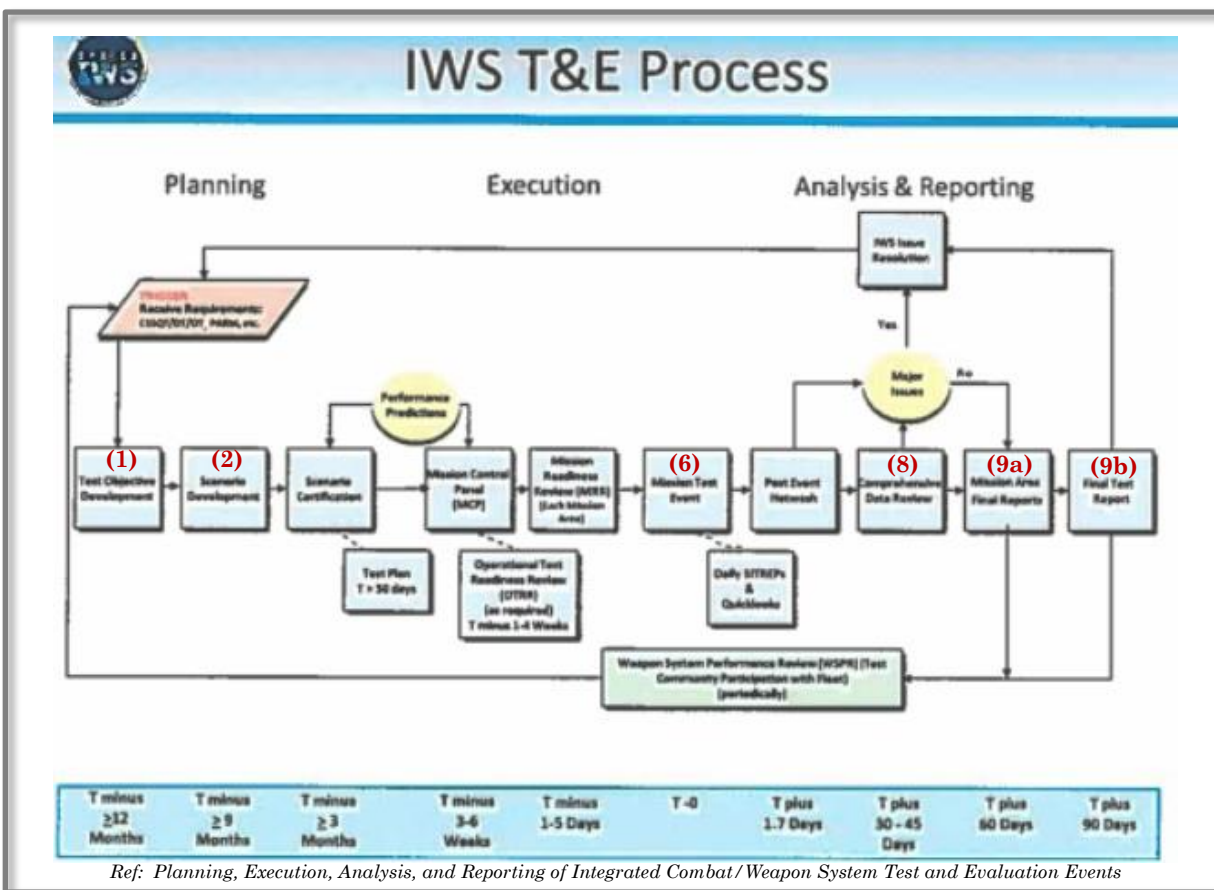


Figure A-4. IWS T&E PEA&R Process

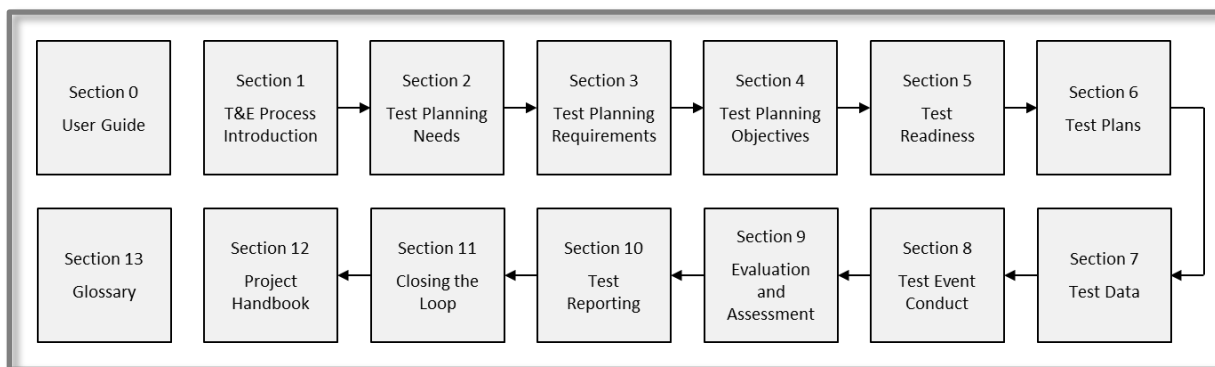


Figure A-5. TP-60 Document Organization

Appendix B

Catalog of Popular Experiment Designs

Catalog of Popular Experiment Designs

This appendix outlines common classes of DOE designs.

B1 Screening Designs

Full factorial design – this class of design consists of two or more factors with a discrete number of levels that produces runs for all possible combinations of those levels across all factors. Sometimes these designs are also referred to as 2^k , where k represents the number of factors, and 2 represents the number of levels. For example, consider a full factorial design that consists of three factors with two levels each. The number of possible combinations in this type of design is thirty-two (2^5).

Fractional factorial design – this design is constructed by systematically choosing a fraction of runs from a full factorial design to start the process of efficiently determining the relationship between the factors and responses. Fractional factorial designs are described mathematically as “ 2^{k-p}_{Res} ” where k refers to the number of factors, $1/2^p$ refers to the fraction, 2 represents the number of levels, and Res refers to the ability of the design to reveal main effects and two-factor interactions (2FI). For example, a 2^{5-1} is a $\frac{1}{2}$ fraction of the 2^5 full factorial that requires only sixteen runs.

Optimal design – this class of design is a good option whenever it is inadequate to use classical designs. The design is generated from computer algorithms that could be optimal with respect to a single criterion for a specified statistical model, but they could be sub-optimal according to another criterion. The designs are model dependent and may require a model that the user may not have. The efficiency of these designs depends on the number of factors, the number of points, and the maximum standard error for prediction over the design space. Typically, the best design for an application is the design with the highest optimality efficiency. The designs have designations corresponding to the letters of the alphabet. The most popular are the D-optimal design, which is mostly used for screening, and the I- (or Q-) optimal design, which is used for response surface. Other designs include the A-, V-, and E- optimal designs:

- D-optimal - Minimizes the generalized variance of the model regression coefficients
- G-optimal - Minimizes the maximum scaled prediction variance over the design region
- I-optimal - Minimizes the average scaled prediction variance over the design region
- A-optimal - Minimizes the average variance of the regression coefficients
- V-optimal - Minimizes the average prediction variance over a set of specific m points of interest in the design region
- E-optimal - Maximizes the minimum eigenvalue of the information matrix

Definitive screening design (DSD) – DSD is a class of three-level screening designs for numeric $k > 5$ that provide estimates of the main effects that are uncorrelated with two-factor interactions and pure quadratic terms. Because they are three-level designs, the quadratic effects are estimable (more on that later in this section). DSDs require $2k+1$ runs. Two-factor interactions (2FI) are only partially confounded with other 2FI as opposed to other screening designs in which 2FI are

completely confounded with other 2FI. Pure quadratic effects are not completely confounded with interactions.

Oehlert and Whitcomb minimum run resolution IV (MR Res IV) –MR Res IV design is a class of economical designs that offers savings over the 2^{k-p} fractional factorial designs. Like all minimum run designs, MR Res IV designs are extremely sensitive to missing data.

General factorial design – is a class of design where there are a levels of factor A, b levels of factor B, and c levels of factor C to produce a total number of $abc...n$ combinations.

B2 Response Surface Designs

Central composite design (CCD) – CCD is a five-level design that consists of a combination of 2^k factorial design or $2\sqrt{k-p}$ fractional factorial design, center points, and axial points located a distance α from the center of the design. The 2^k or $2\sqrt{k-p}$ designs provide for estimating first-order effects, two-factor interactions, and lack of fit if there is replication. The center points provide for estimating pure error, determining the presence of curvature, and for a more uniform estimation of the prediction variance. The axial points provide for estimating second-order effects. The position of the axial points produces designs for spherical regions or designs for cuboidal regions of space. The CCD is the workhorse of response surface methodology.

Box-Behnken design (BBD) – BBD is a three-level design for fitting response surfaces. The construction technique relies on using balanced incomplete block designs and 2^k factorial designs. The design avoids the corners of the design space in favor of edge points located at the mid-level ($x_i = 0$) of the factor levels, which results in poor estimation at the factorial point locations. Thus, a BBD is more useful for situations in which there is no interest in predicting at the factorial points of the cube. Like for a CCD, replicated runs at the center points permit a more uniform estimation of the prediction variance over the design space. Some BBDs with four or less factors are more economical than their equivalent CCDs. Practitioners often associate the BBD with cuboidal regions because of its cubic appearance; however, the BBD is a spherical design. BBDs require sufficient center points to improve their prediction accuracy.

3^k general factorial design – this is a three-level factorial design that allows for the estimation of curvature. However, this class of design is generally inadequate since it requires a large sample size.

Space-filling design –this is a class of design that aims at filling the test space with as few gaps as possible. These designs are suitable for computer simulations and other situations in which the response is deterministic.

B3 Small Response Surface Designs

There are situations in which scarce resources—funds, time, material, manpower, equipment—make it impractical to allow the use of traditional designs for fitting second-order models, especially when the number of factors k is high. For those situations, small or saturated response surface designs could be attractive since they are more economical. Some of the most popular small or saturated response surface designs and their features are listed below.

Draper-Lin design – a minimal point design for $k = 3, 5, 6, 7$, and 10 . These designs share a common structure with other small composite designs.

Koshal design – a class of design with the same number of runs as the number of model coefficients to be estimated.

Hoke design - saturated, non-orthogonal, second-order design based on irregular fractions of partially balanced 3^k factorial designs (for $k > 2$). Hoke designs are suitable for cuboidal regions. For a small number of factors ($k < 7$), some of the designs are near-saturated and permit estimating pure error and lack-of-fit. Except for the quadratic terms, the best Hoke designs (for $2 < k < 10$) are comparable to BBDs and Hartley small composite designs based on the determinant and the trace of the information matrix. Hoke designs have better prediction performance and similar sample size than Hartley small composite designs.

Doehlert uniform shell design - uniform shell designs (for $k < 11$) have an equally spaced distribution of points lying on concentric spherical shells.

Roquemore hybrid design - a set of saturated or near saturated second-order rotatable or near-rotatable designs for $k = 3, 4, 5$, and 6 . Roquemore hybrid designs are competitive with CCDs based on the scaled prediction variance criteria.

Box and Draper minimum point design – this class of design fits main effects plus interaction models that are optimal for $k = 2$ and 3 but not optimal for $k > 3$.

Hartley small composite second-order design – this design is based on the idea that the cube portion of the composite design can be as low as a resolution III fraction if the 2FIs are not aliased with other two-factor interactions.

Westlake cuboidal design - second-order cuboidal designs for $k = 5, 7$, and 9 in $22, 40$, and 62 runs that expands the idea from Hartley.

Rechtschaffner dummy factor design – adds a dummy factor, whose main effect and 2FIs are assumed to be zero, to saturated fractions of 2^k and 3^k factorial resolution V designs to increase the degrees-of-freedom for estimating pure error and preserving the balanced structure of the design.

Pesotchinsky minimum point D-optimal design – designs for $k < 8$.

Lucas saturated D-Optimum composite design - saturated D-optimum composite design using a subset of points from the saturated resolution V of Rechtschaffner's.

Mitchell and Bayne exchange algorithm design – uses an algorithm to find a k -run design that maximizes the information matrix $|X'X|$ given the number of factors k , a specified model, and a set of candidate points.

Notz minimal point asymptotic D-efficiency design - a class of 3^k (for $k < 7$) minimal point second-order design with asymptotic D-efficiency of 1 relative to the number of factors k and the number of minimal points q .

Draper near saturated design - cuboidal, near-saturated design using the Plackett-Burman design for the factorial portion of Hartley's small composite design, which are a compromise between a saturated small composite design and a CDD allowing degrees-of-freedom for estimating lack-of-fit.

Morris augmented-pairs design - a class of three-level design constructed by combining the levels of every pair of points in a two-level first-order design to form the third level.

Oehlert and Whitcomb minimum run resolution V (MR Res V) design – a class of equ-replicated, irregular fractions of 2^k designs constructed using D-optimality criterion algorithm. This design provides resolution V designs in fewer runs when regular fraction designs contain significantly more degrees-of-freedom than are needed to estimate the model up to 2FIs. Judged only on a D-optimality criterion, these designs are more efficient than many other types.

Haines' San Cristobal design – fits a quadratic response surface for k factors that are restricted to non-negative levels.

Gilmour subset design – a class of three-level response surface design obtained by using subsets of 2^k factorial designs at levels of -1 and 1 for each combination of k factors while holding the other $q - k$ factors at their middle level.