INSTITUTE FOR DEFENSE ANALYSES

I-ITSEC DOE Tutorial

Breeana Anderson, Project Leader
Rebecca Medlin, Project Leader

John T. Haman
Kelly M. Avery
Keyla Pagan-Rivera

OED Draft

June 2023

IDA Document NS-D-33561

Log: H 2023-000231

INSTITUTE FOR DEFENSE ANALYSES
730 East Glebe Road
Alexandria, Virginia 22305

The Institute for Defense Analyses is a nonprofit corporation that operates three Federally Funded Research and Development Centers. Its mission is to answer the most challenging U.S. security and science policy questions with objective analysis, leveraging extraordinary scientific, technical, and analytic expertise.

About This Publication
This work was conducted by the Institute for Defense Analyses (IDA) under contract HQ0034-19-D-0001, Task BD-9-2299(98), "TestSci Training," and Task C9082 "Statistics and Data Science Working Group" for the Office of the Director, Operational Test and Evaluation. The views, opinions, and findings should not be construed as representing the official position of either the Department of Defense or the sponsoring organization.

For more information:
Dr. Breeana Anderson (primary); Dr. Rebecca Medlin, Project Leader
banderso@ida.org • (703) 845-6967

Dr. V. Bram Lillard, Director, Operational Evaluation Division
vlillard@ida.org • (703) 845-2230

Rigorous Analysis | Trusted Expertise | Service to the Nation

INSTITUTE FOR DEFENSE ANALYSES

IDA Document NS-D-33561

I-ITSEC DOE Tutorial

Breeana Anderson, Project Leader
Rebecca Medlin, Project Leader

John T. Haman
Kelly M. Avery
Keyla Pagan-Rivera

Executive Summary

Research staff members from the Institute for Defense Analyses (IDA) were invited to present a mini-tutorial on design of experiments (DOE) for the November 2023 Interservice/Industry Training, Simulation and Education Conference (I/ITSEC). I/ITSEC is the world's largest modeling, simulation, and training event. The conference is organized and sponsored by the National Training & Simulation Association (NTSA), which promotes international and interdisciplinary cooperation within the fields of modeling and simulation (M&S), training, education, analysis, and related disciplines at this annual meeting. The NTSA is an affiliate subsidiary of the National Defense Industrial Association (NDIA); hence, I/ITSEC also emphasizes themes related to defense and security.

The purpose of this training is to introduce attendees to DOE. DOE is a branch of applied statistics that deals with planning, conducting, analyzing, and interpreting controlled experiments. Specifically, DOE provides one with an analytical framework to determine whether a test is good enough for their purpose. A test strategy that employs DOE provides the most powerful allocation of test resources for a given number of events.

This training provides details regarding the use of design of experiments, from choosing proper response variables, to identifying factors that could affect such responses, to determining the amount of data necessary to collect. The training also explains the benefits of using a DOE approach to testing and provides an overview of commonly used designs (e.g., factorial, optimal, and space-filling). The briefing illustrates the concepts discussed using several case studies.

Lastly, all experiments are designed with an analysis methodology in mind: to reap the benefits, we need to follow through to the analysis. Rigorous test design and analysis techniques facilitate an efficient, objective, and credible evaluation. To that end, the training also covers proper analysis methods and reporting. Specifically, we discuss avoiding data averaging, reporting confidence intervals to convey uncertainty, and using graphical summaries to communicate impact.

After receiving this tutorial, attendees should be able to:

- Explain the importance of experimental design and its link to data analysis
- Identify essential elements of test design
- Recognize key differences between test designs
- Assess the adequacy of a test design
- Apply key tips to data analysis and reporting

Introduction to Design of Experiments

Presenter: John Haman, IDA

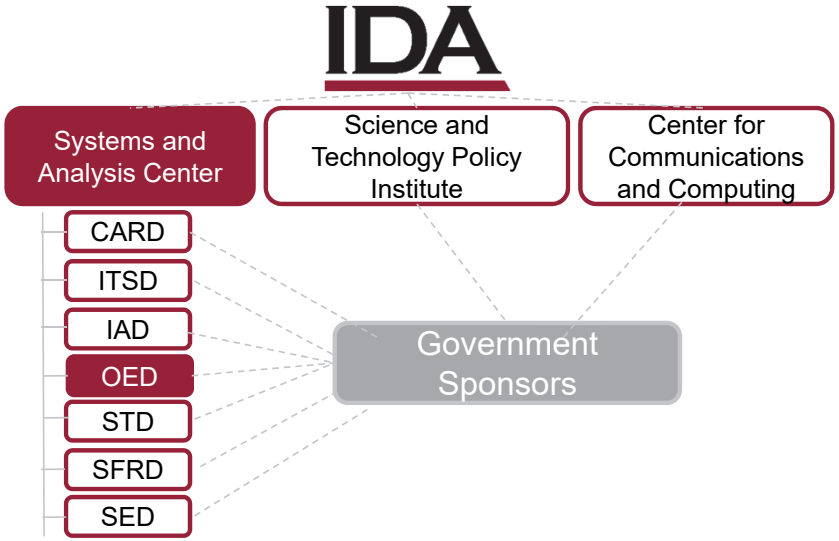
Co-Authors: Keyla Pagán-Rivera, Rebecca Medlin, and Kelly Avery, IDA

About Institute For Defense Analyses (IDA)

We answer challenging questions using scientific, technical, and analytic expertise.

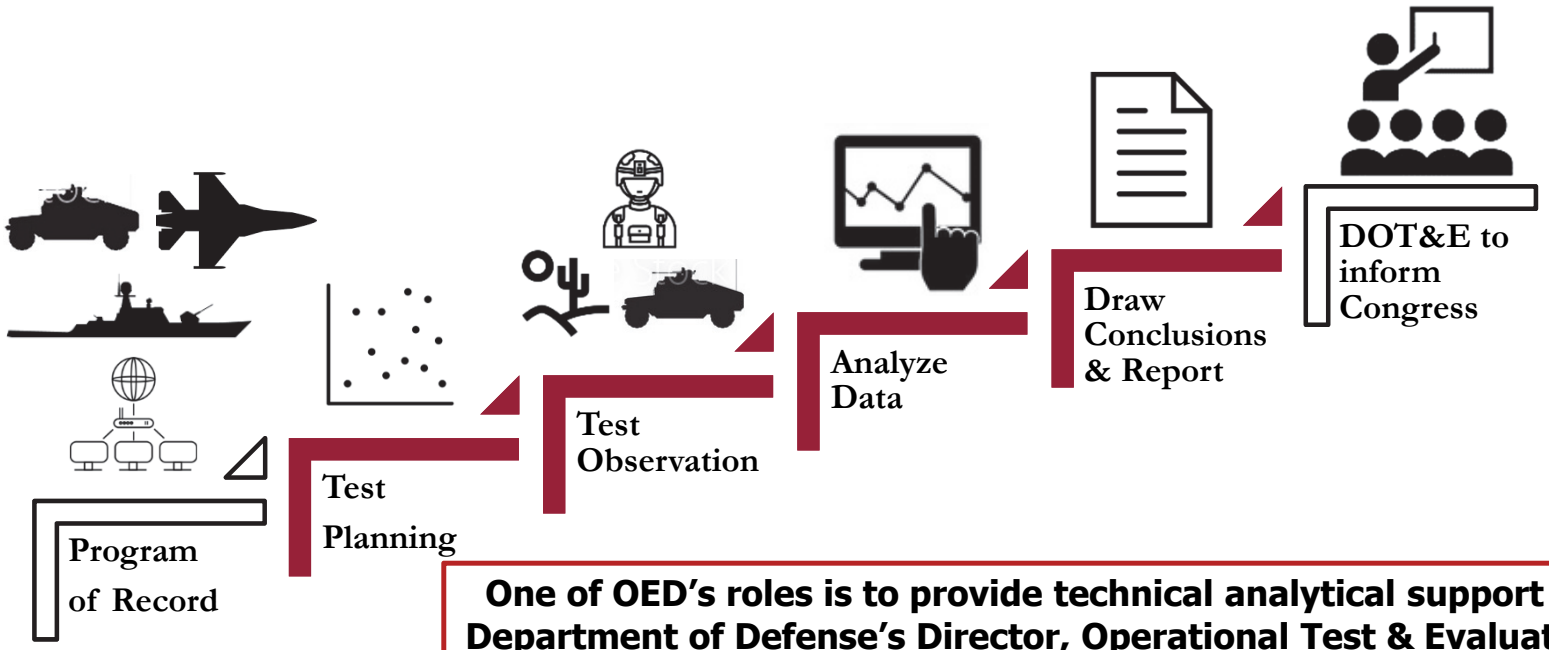
IDA is a private, nonprofit corporation headquartered in Alexandria, Virginia

IDA operates three Federally Funded Research and Development Centers



CARD – Cost Analysis and Research Division; ITSD – Information Technology and Systems Division; IAD – Intelligence Analyses Division; OED – Operational Evaluation Division; STD – Science and Technology Division; SFRD – Strategy, Forces and Resources Division; SED – System Evaluation Division

What Does The Operational Evaluation Division Do?



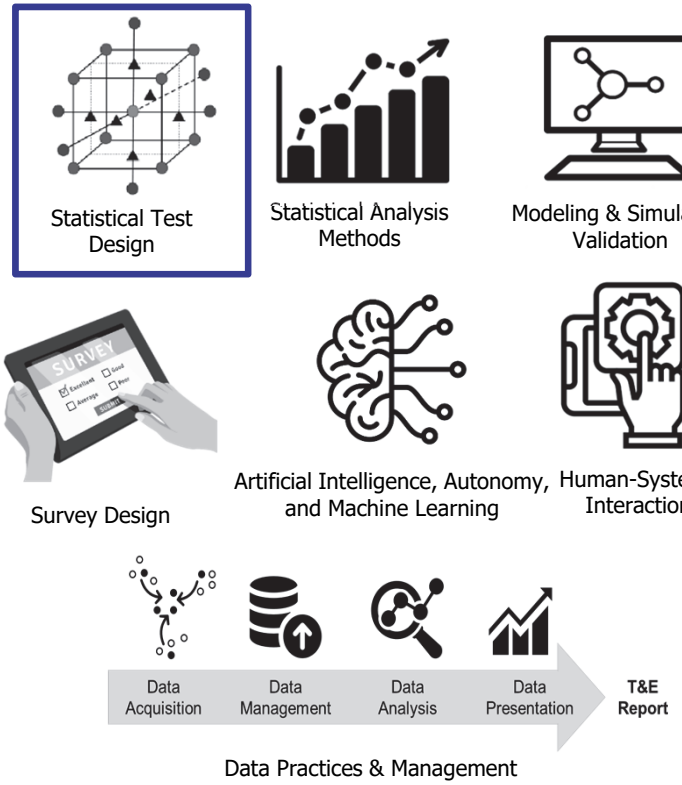
One of OED's roles is to provide technical analytical support to the Department of Defense's Director, Operational Test & Evaluation and its mission to conduct independent assessments of the military services' test and evaluation of new weapons systems.

- **Help design good tests** of military systems in realistic operational environments.
- **Evaluate** the operational **effectiveness, suitability, and survivability** of those systems from an objective, disinterested, and factual perspective.

The Test Science Team Provides Expertise To All Warfare Areas In OED

We develop, apply, and disseminate statistical, psychological, and data science methodologies

Core Areas of Expertise



What Will You Learn From This Tutorial?

After this tutorial, attendees should be able to:

- Explain the importance of experimental design and its link to data analysis
- Identify essential elements of test design
- Recognize key differences between test designs
- Assess the adequacy of a test design
- Apply key tips to data analysis and reporting

Introduction

General Framework For Designing A Test

Determine test objective
- Mission capability to be tested; questions to ask about the system

Identify appropriate metrics
- How system performance should be measured

Identify factors that affect performance
- Types of data to collect; operational envelope

Develop test design
- Quantity of data necessary; best resource allocation; objective plans

Conduct test
- Adjust test execution if necessary

Analyze data
- Structured mathematical data analysis plan appropriate for the design

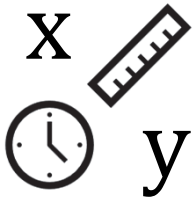
Draw conclusions
- Defensible risk assessments based on test results

Experimentation Requires Collaboration

Subject Matter Expertise

Analytical Expertise

Testing Framework: The Gold Standard Of Experimentation



- The design of experiments (DOE) process, at its core, is the scientific method.
- It is the gold standard for performing experiments and determining cause and effect.
- What we mean by “experiment” goes well beyond the ‘chemistry lab’ scenario it often evokes.
- An experiment includes any test where variables are purposely manipulated or inputs are purposely set.
- So, things like simulation studies certainly count as experiments.

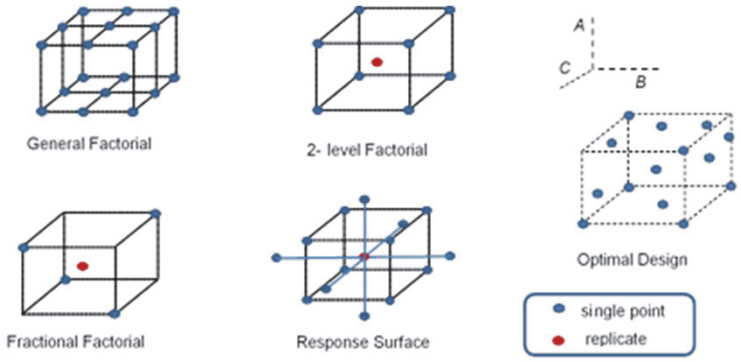
Rigorous Test Design And Analysis Techniques Facilitate Efficient, Objective, And Credible Evaluations

The test design process can be applied across multiple domains:

- System Performance and Effectiveness
- Modeling and Simulation Validation
- Human-Systems Interaction
- Cybersecurity
- Reliability
- Software
- Live Fire

What Is Design Of Experiments (DOE)?

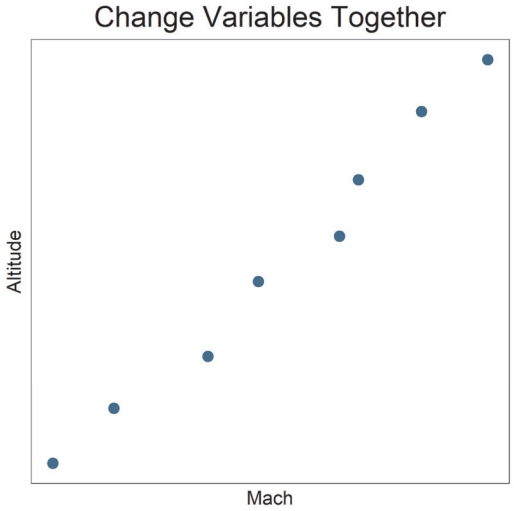
DOE is a branch of applied statistics that deals with planning, conducting, analyzing, and interpreting controlled tests.

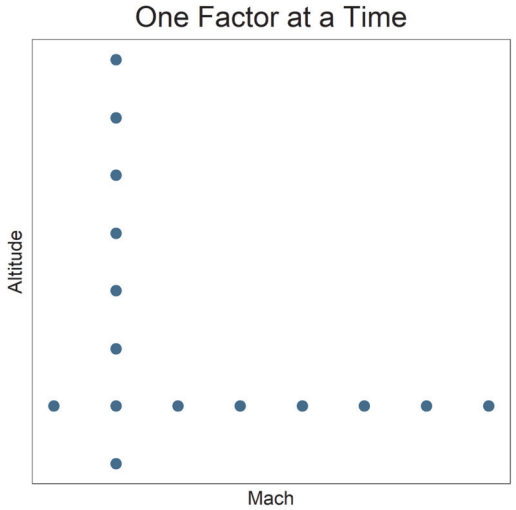


Benefits of DOE:

- Enables test designs that investigate multiple factors in one experiment
- Provides the tester with an analytical framework to determine whether a test is good enough for their purpose
- Provides the most powerful allocation of test resources for a given number of events

All Tests Are Designed, Many Poorly

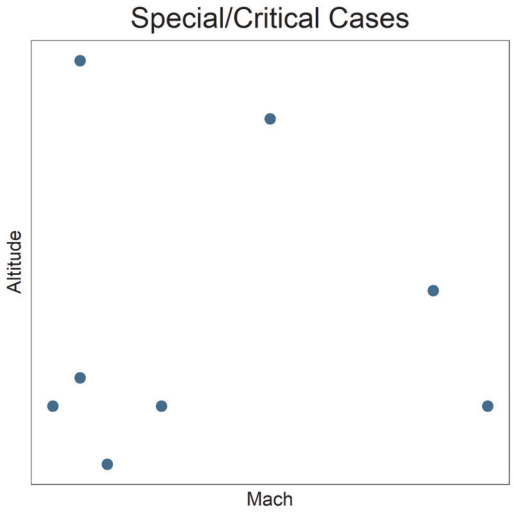


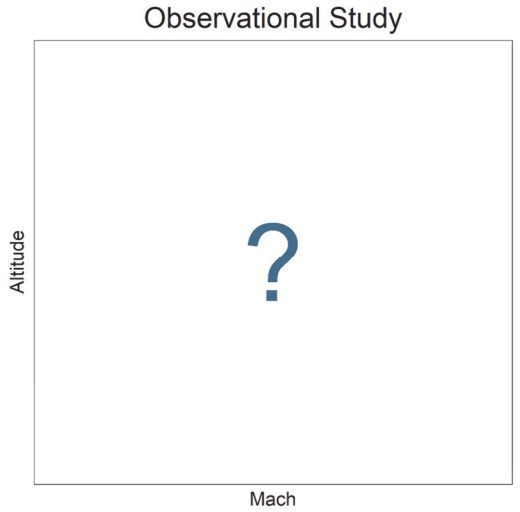


- Confounding variables!
- Loss of ability to determine cause and effect

- Interactions between conditions not examined
- Test is often overkill (unnecessarily large sample sizes)

All Tests Are Designed, Many Poorly

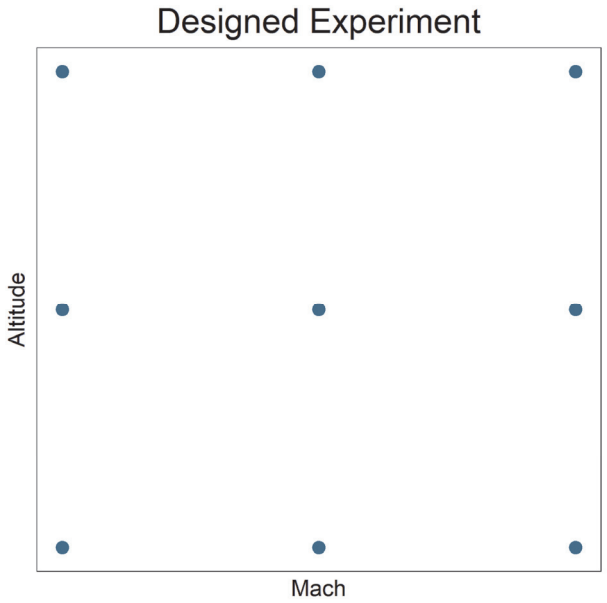




- Limited to the specific conditions selected; might miss important performance shortfalls
- Often poor statistical precision
- Possible loss of ability to determine cause and effect

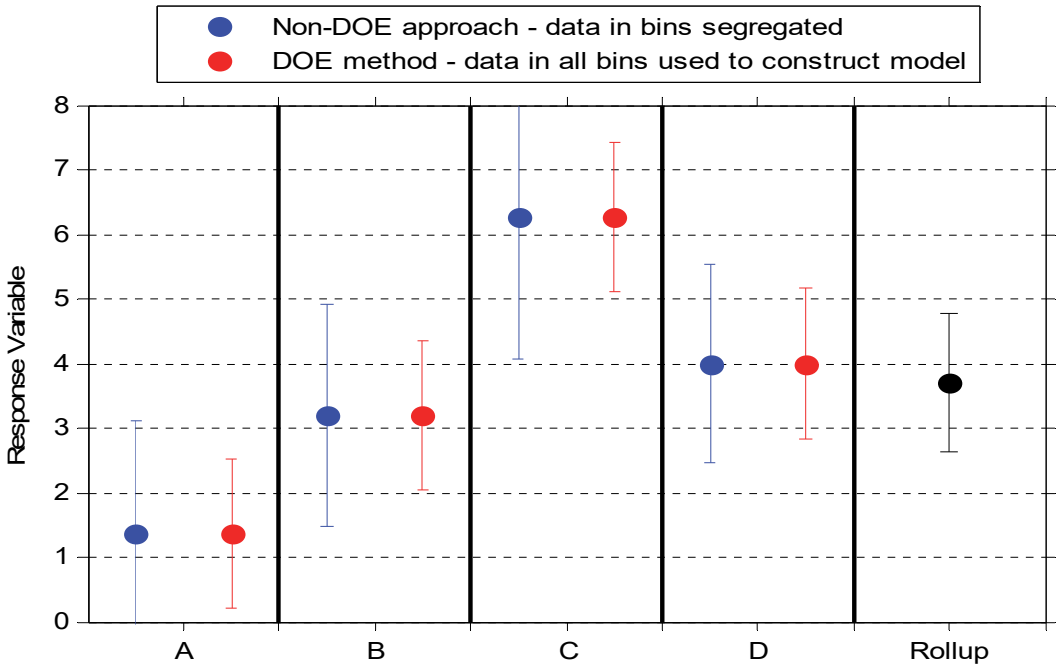
- No control over test conditions
- Data left up to chance
- No ability to determine cause and effect

Why DOE Over Other Data Collection Methods?



- Allows for inferences about causation
- Can directly understand and compare the effects of different treatments
- Minimizes bias/error
- Maximizes efficiency

DOE Supports A More Precise Characterization



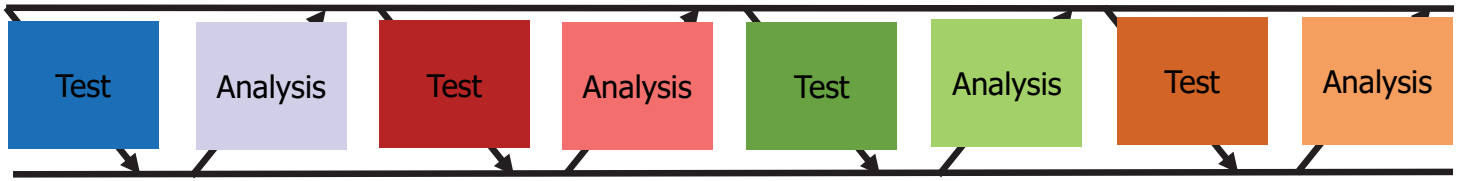
	Slow Speed Target	Fast Speed Target
With Countermeasures	A	B
No Countermeasures	C	D

Rollup (global mean) tells us little about system performance

- Non-DOE approach: Calculate confidence intervals using only data collected under each condition
- DOE approach: Construct a model (pool the data), use the model to estimate mean values in each condition
  - Note the reduction in confidence interval size!
    - In this case, intervals reduced by 25 to 50% compared to non-DOE approach
  - Can now tell significant differences in performance
    - E.g., system is **better** in C than in D conditions

DOE Encompasses A Large Body Of Techniques That Support Different Types Of Data, Use Cases, And Levels Of Complexity

- ***Classical designs*** (e.g. factorial, fractional factorial) are the gold standard in straightforward situations
- ***Optimal designs*** allow for constrained design spaces, disallowed combinations, and custom sample sizes
- ***Space-filling designs*** are specifically geared toward Modeling & Simulation experiments where outcomes are less stochastic and more complex
- ***Sequential design*** strategies build on information gained from the previous experiment in considering how to plan the next test



General Framework For Designing A Test

Outline for rest of tutorial

Determine test objective
- Mission capability to be tested; questions to ask about the system

Identify appropriate metrics
- How system performance should be measured

Identify factors that affect performance
- Types of data to collect; operational envelope

Develop test design
- Quantity of data necessary; best resource allocation; objective plans

Conduct test
- Adjust test execution if necessary

Analyze data
- Structured mathematical data analysis plan appropriate for the design

Draw conclusions
- Defensible risk assessments based on test results

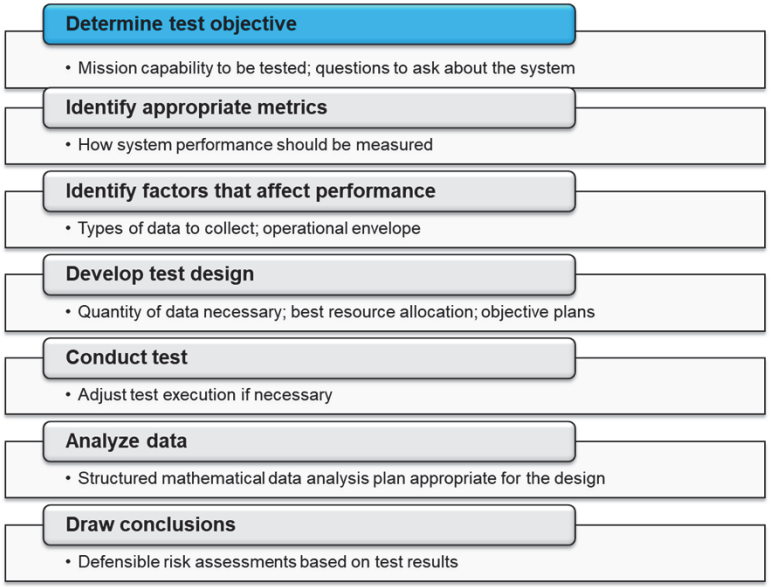
Experimentation Requires Collaboration

Subject Matter Expertise

Analytical Expertise

Data contained in presentation are notional

Test Objectives



Understanding *Why* We Are Testing Drives And *How* We Plan The Test

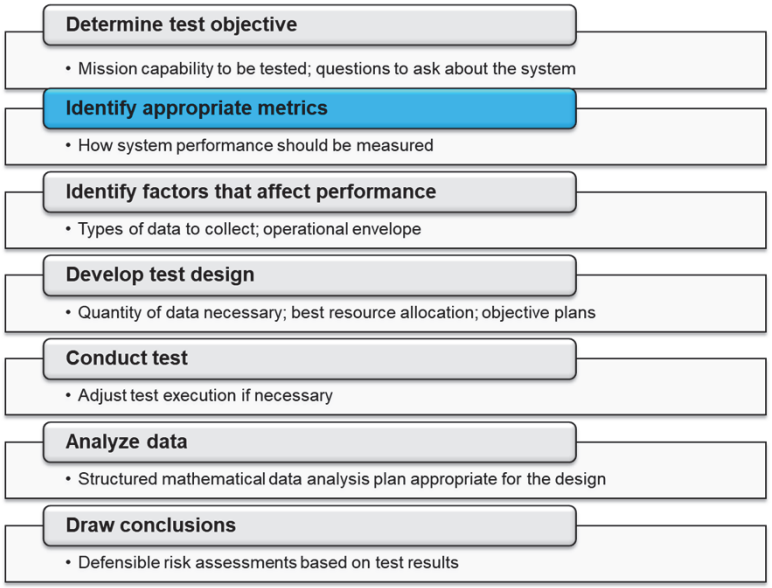
What questions must the study address?

What mission capabilities are we testing?

Common Test Objectives

- **Screen** for influential factors driving performance.
- **Characterize** performance across the relevant factor space.
- **Compare** two systems (or more) across a variety of operating conditions.
- **Identify problems** that degrade system performance.
- **Optimize** system performance with respect to a set of conditions.

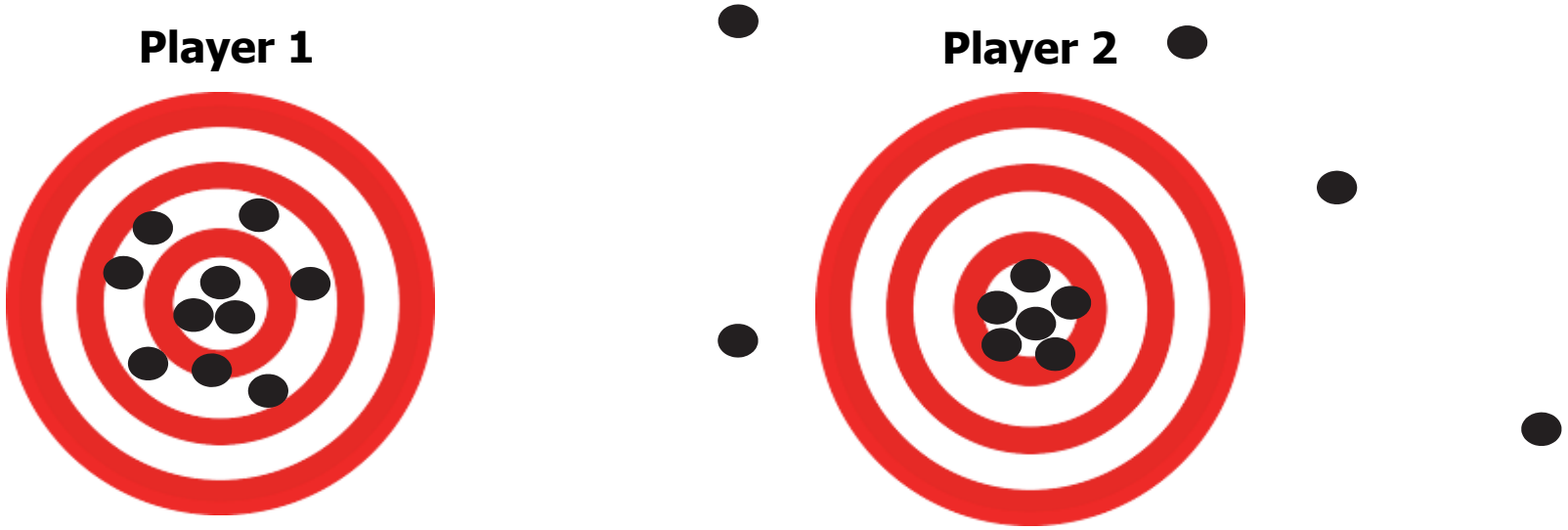
Identifying Metrics/Response Variables



The Response Variable(s) Measures The Outcome Of Interest For The Test

- Requirements often *inform* response variable selection but should not be the sole sort of metrics!
- Good response variables are:
  - **Measurable.** They can be measured at a reasonable cost and without affecting the test outcome.
  - **Valid.** They directly address the test objective.
  - **Informative.** Continuous responses always provide more information per test point than pass/fail metrics.
- Multiple response variables are almost always necessary!
- Must think about where and when to collect desired data.

Who Is The Better Darts Player?
(Hit = bullseye; Miss = all else)



Is Player 2 better because he hit the bullseye 5 of 10 times, whereas Player 1 hit the bullseye only 3 of 10 times?

We have a much better picture of the players' abilities by looking at the miss distances from bullseye compared to just a bullseye hit/miss.

Operational Examples Of Binary Responses Converted To Continuous

	Chemical Agent Detector	Submarine Mine Detection	Missile/Bomb	Radar Detection
Requirement	Probability of detection greater than 85% after one minute of exposure	Probability of detection greater than 80% outside 200 meters	Probability of hit at least 90%	Probability of detection greater than 90% at 300 kilometers
Original Test Measurement	Detect prior to 1 minute? (Yes/No)	Detect/Non-Detect	Hit/Miss	Detect/Non-Detect
Modified Test Measurement	Detection Time	Detection Range	Miss Distance	Detection Range

Ultimately, your primary metrics should always track with **the mission**

Many Elements And Decisions Influence The Size Of A DOE

Statistical

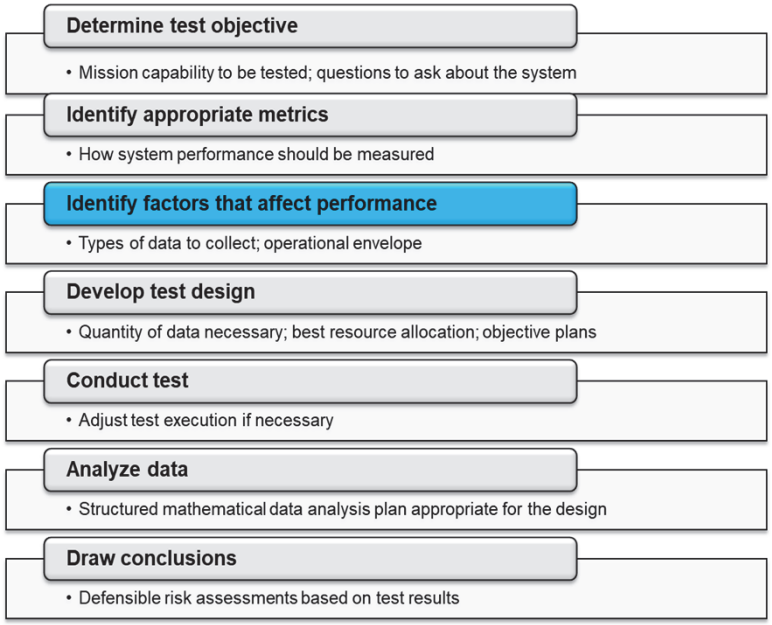
- How much power and confidence do we want?
- What is the type of response variable?
- Which model are we assuming? Do interactions and quadratic terms matter?
- Which signal-to-noise ratio can we assume?
  - Effect size
  - Variability
- Is the randomization restricted?

Practical

- Do we have enough resources to collect X number of samples?
- Are there any permissions we need for certain runs?
- Can we collect data on all conditions or are there political, geographical, or other restrictions?
- Are there disallowed combinations?

A pilot study takes out some of the guesswork

Determining Factors



Sources Of Variation

A source of variation is anything that might influence the performance of the process or system under test.

Brainstorm ALL the potential sources of variation that could affect test outcomes – then decide what to control/vary during test.

A fancy statistical design cannot redeem the quality of a test if we miss an important factor in the planning process!

Factor Management Process

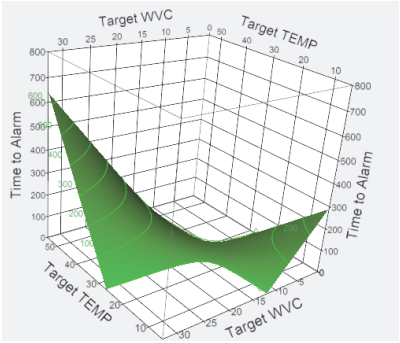
		Likelihood of Encountering Level During Operations		
		Multiple levels occur at balanced frequencies (e.g., 1/3, 1/3, 1/3)	Some levels are balanced, others are infrequent (e.g., 5/10, 4/10, 1/10)	One level dominates (e.g., 4/5, 1/10, 1/10)
Effect of Changing Level on Performance		**Balanced**	**Mixed**	**Dominant**
Significant Effect on Performance	**High**	Vary all	Vary balanced levels, Demonstrate infrequent levels	Fix dominant level, Demonstrate others
Moderate Effect on Performance	**Medium**	Vary all	Vary balanced levels, Demonstrate others	Fix dominant level, Demonstrate others
Low Effect on Performance	**Low**	Fix levels or record level used	Fix levels or record level used	Fix dominant level

Characteristics Of Good Experimental Factors

- **Important:** Factors are expected to have an effect on the test outcome.

- **Controllable:** Factors can be controlled (i.e., set to a specific level) at a reasonable cost.
  - Easy-to-change: a factor that is easy to randomize completely
  - Hard-to-change: a factor that is difficult to randomize completely, often due to time or cost constraints.
    - When you specify factors as Hard-to-change, your design should reflect these restrictions on randomization.
  - And sometimes the best we can do is simply record a setting….

- **Informative:** Continuous factors are preferred to categorical factors (e.g., if altitude is a factor, the preferable levels are 5,000, 10,000, and 15,000 as opposed to low, medium, and high).
  - Continuous factors allow for:
    - Interpolation
    - Explanation of change in performance
    - Higher power to detect significance of a factor
    - Strategic point placement



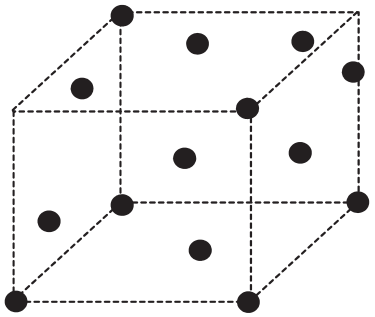
Data contained in presentation are notional

Developing the Test Design

- **Determine test objective**
  - Mission capability to be tested; questions to ask about the system
- **Identify appropriate metrics**
  - How system performance should be measured
- **Identify factors that affect performance**
  - Types of data to collect; operational envelope
- **Develop test design**
  - Quantity of data necessary; best resource allocation; objective plans
- **Conduct test**
  - Adjust test execution if necessary
- **Analyze data**
  - Structured mathematical data analysis plan appropriate for the design
- **Draw conclusions**
  - Defensible risk assessments based on test results

Which points?

How many points?



Testing A New Artillery Cannon

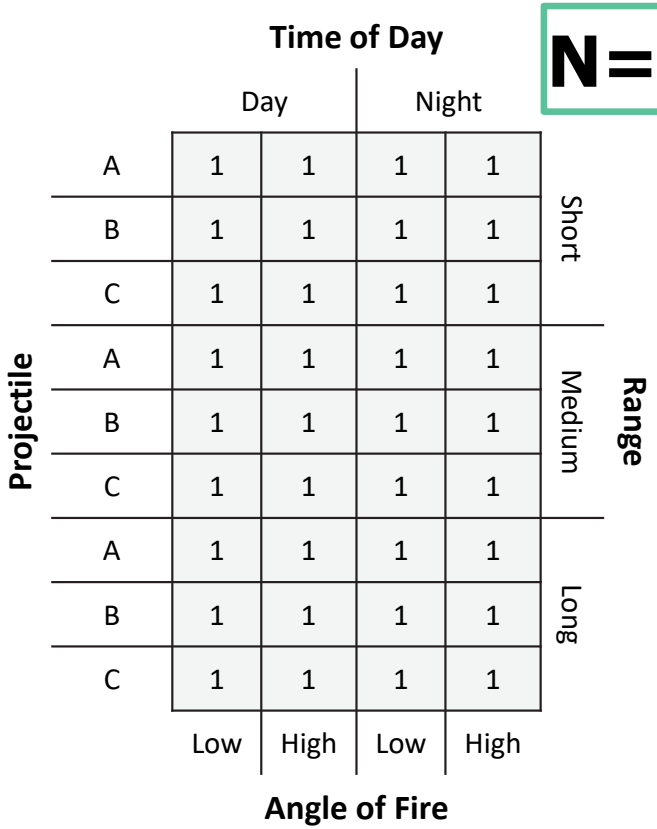
Response Variable

Accuracy of Fires - Miss Distance

Factor	Level
Time of Day	Day, Night
Range to Target	Short, Medium, Long
Projectile Type	A, B, C
Angle of Fire	Low, High

Data contained in presentation are notional

Factorial Designs - All Possible Combinations of Each Factor Level



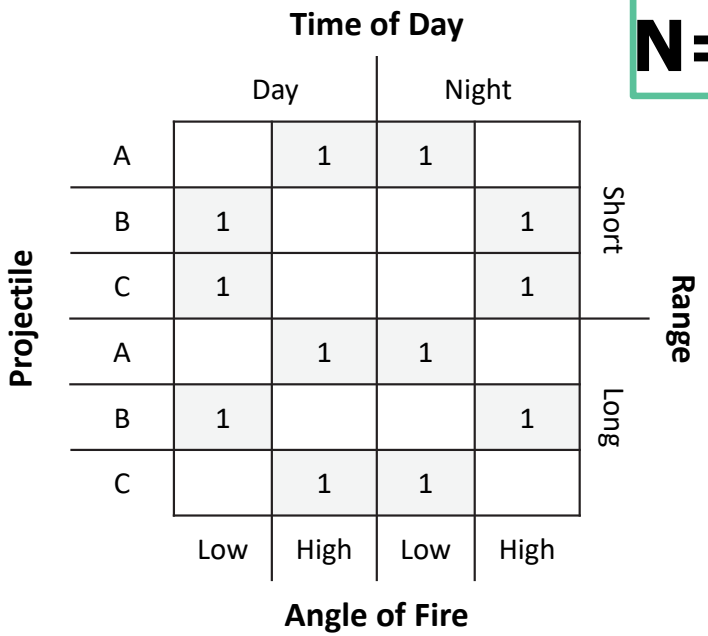
N=36

Tests All Possible Combinations.
The first fire mission is conducted during the Day at Short Range for Low Angle using Projectile A.

Full Factorial Designs:

- Support the test goals of characterize or compare.
- Examine every possible combination of each level.
- Allow for the estimation of all main effects and all possible interactions without aliasing.
- Are highly efficient and informative, though potentially prohibitively costly.

Typically, Screening Or Characterization Experiments Involve a Fractional Factorial Design

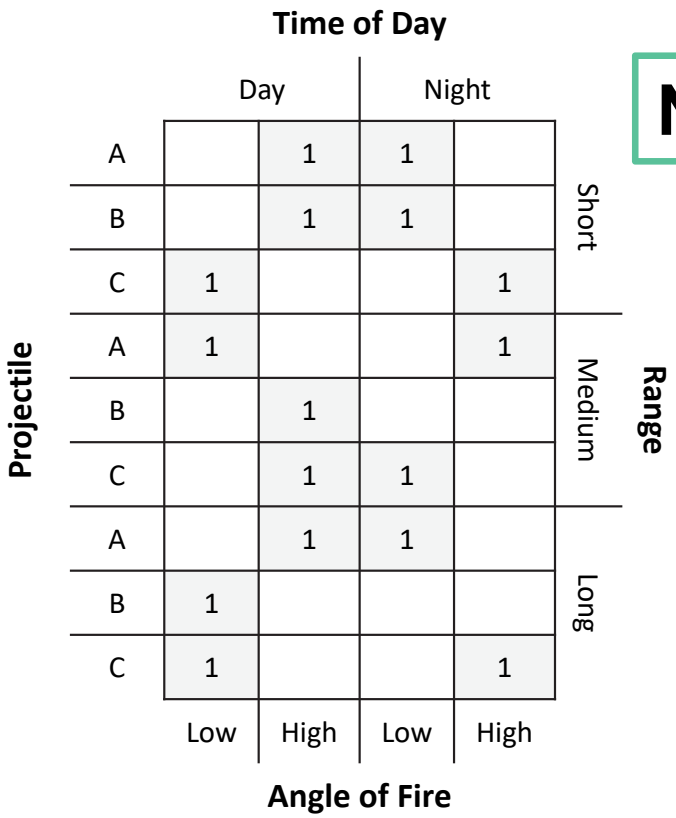


N=12 → Places test points at the right conditions to support a main effects model.

Fractional Factorial Designs:

- Support the test goals of screen, characterize, or compare.
- Require a subset of the runs required for a full factorial.
- Achieve a large reduction in test points by trading off the ability to estimate high-order interaction effects.

Optimal Designs Are Most Useful When The Number Of Test Points Is Constrained To Preclude A Factorial Design



N=16 → Places test points at the right conditions to estimate specified model terms.

Optimal Designs

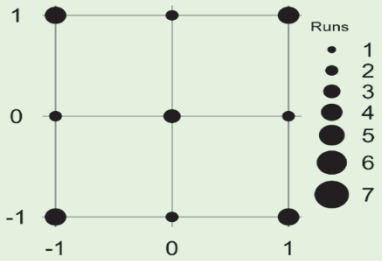
- Support the test goals of characterize, optimize, predict.
- Useful when the number of test points is constrained to preclude a factorial design.
- Requires a researcher-specified model and a fixed sample size (a subset of the runs required for a full factorial).

Relevant Optimality Criteria Overview

D-Optimal

- D-optimal designs **minimize the variance of the parameter estimates**
- Most useful for characterizing performance
- Generally, provides the "best" power across all parameter terms
- Tends to place most points on the edges of the design space

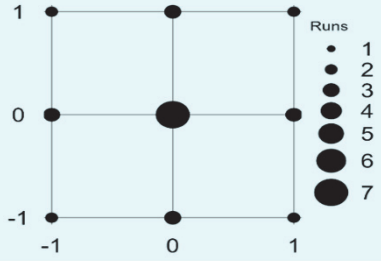
Example: 2-factor model with quadratic terms and all interactions



I-Optimal

- I-optimal designs **minimize the average prediction variance**
- Most useful when you want good predictions across your design space, particularly if you want to model curvature
- Tends to allocate more points on the interior

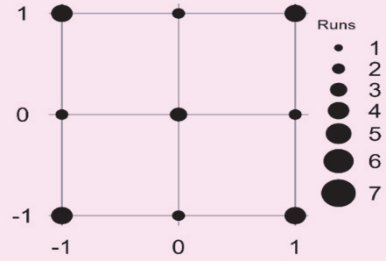
Only difference here in inputs is design is I-optimal—resulting in more points allocated in the interior.



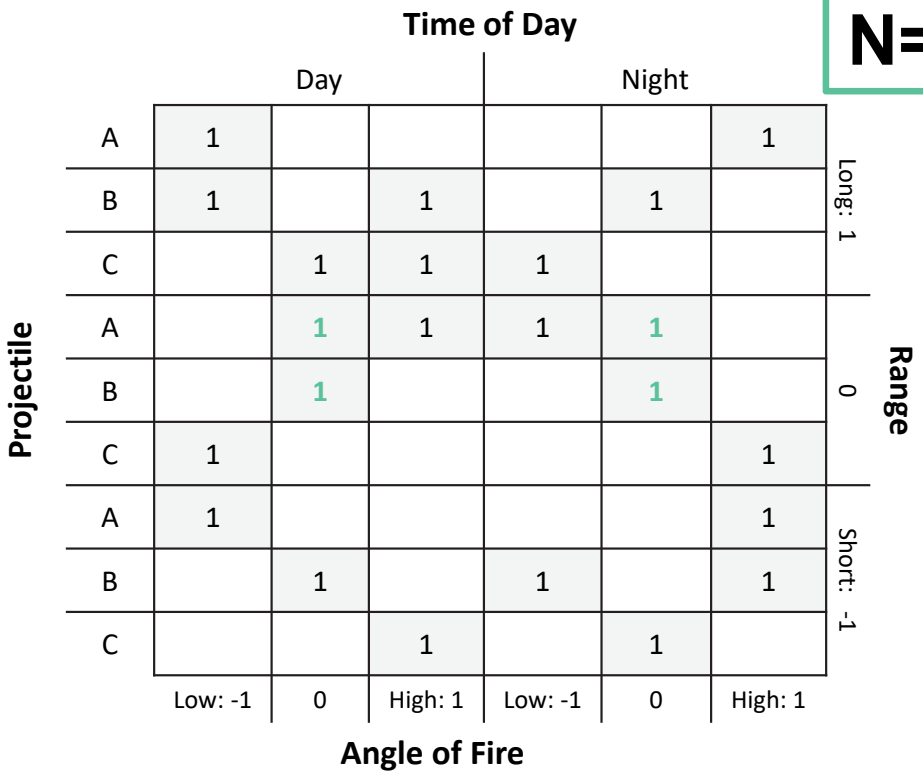
Alias-Optimal

- Alias-optimal designs minimize correlation between main effect and interaction terms
- Most useful when performing screening experiments
- Usually less power to estimate effect sizes, but can reduce overall resource requirements with sequential testing

Note: Here, the Alias-optimal design is the same as D-optimal design!



Response Surface Designs Spread Test Points Throughout The Experimental Region To Support A Detailed Model Of The Response



N=24 → Includes center points to check for curvature.

Response Surface Designs

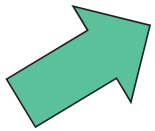
- Support the test goals of characterize, optimize, predict, improve.
- One of the best designs if the researcher needs to minimize or maximize a response.

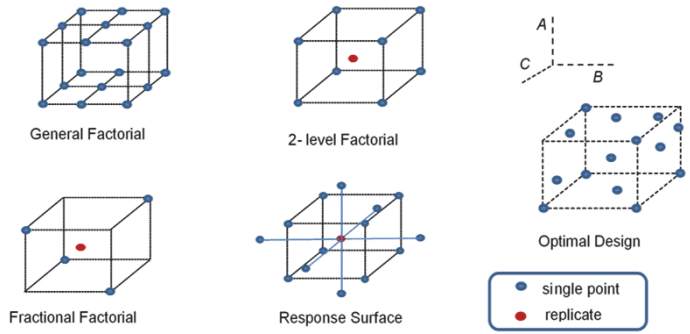
So Far All The Designs Presented Have Been Geared Toward Stochastic Outcomes

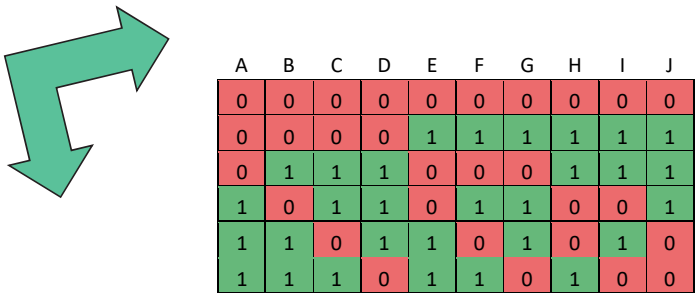
The field of DOE includes two broad classes of designs:

Classical – geared toward <u>stochastic</u> test outcomes

Computer – assume a (near) <u>deterministic</u> outcome

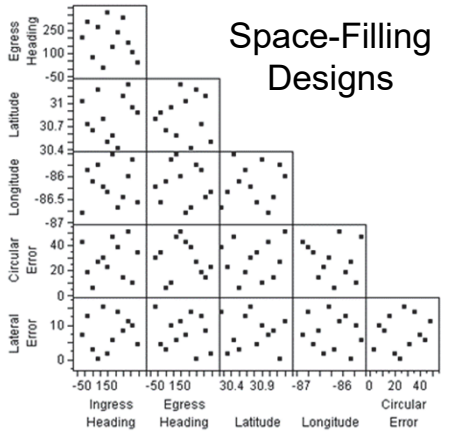






A	B	C	D	E	F	G	H	I	J
0	0	0	0	0	0	0	0	0	0
0	0	0	0	1	1	1	1	1	1
0	1	1	1	0	0	0	1	1	1
1	0	1	1	0	1	1	0	0	1
1	1	0	1	1	0	1	0	1	0
1	1	1	0	1	1	0	1	0	0

Covering Arrays



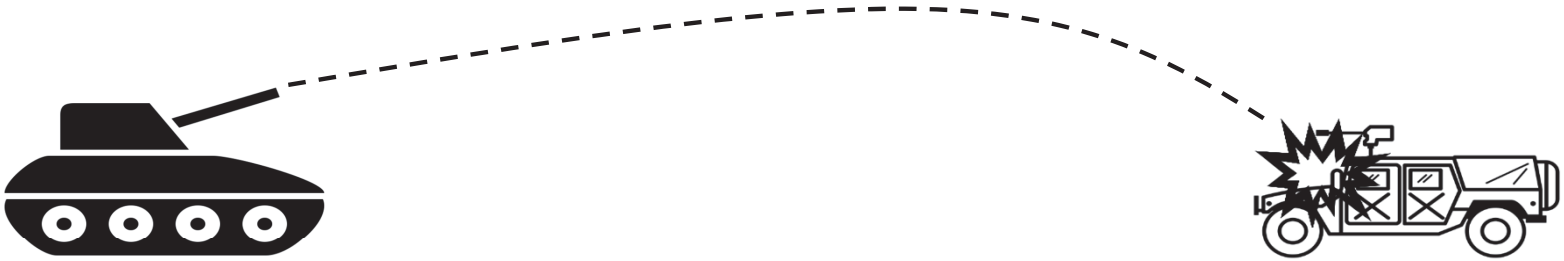
Space-Filling Designs

Choosing the right class of design for your specific outcome is critical!

Leveraging M&S To Expand Testing For New Artillery Cannon

Response Variable

Accuracy of Fires - Miss Distance



Factor	Level
Time of Day	Day, Night
Range to Target	500 m, 4000 m
Projectile Type	A, B, C
Angle of Fire	10, 45 degrees

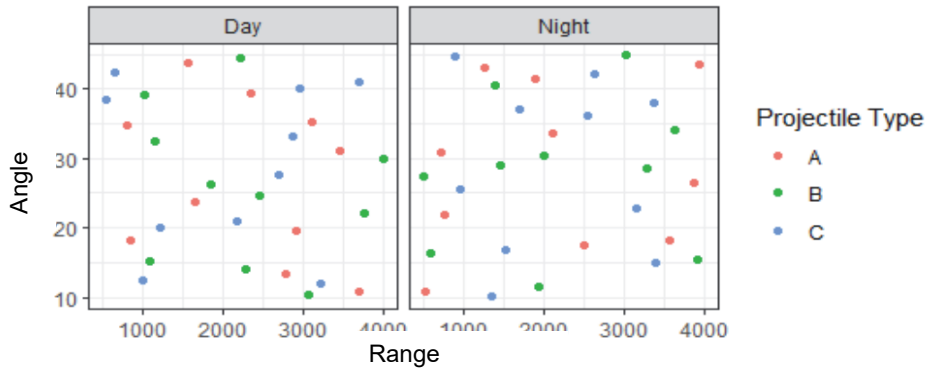
Data contained in presentation are notional

Space-Filling Designs Can Help Recover More Trends In Simulation Outputs

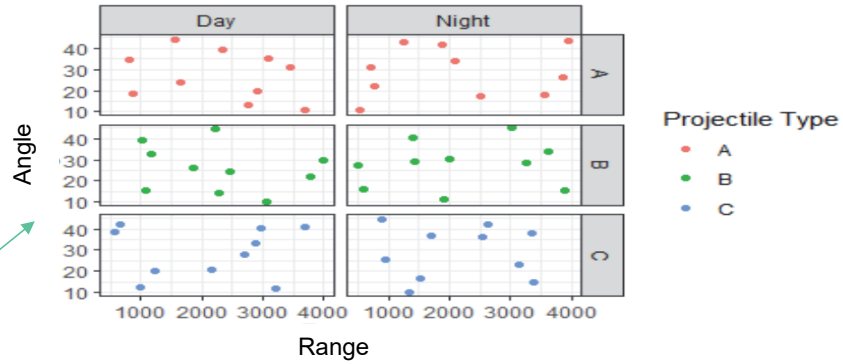
Space-Filling Designs

- When designing a test that involves M&S, we should consider using space-filling designs (SFDs) to better spread out points across the factor space

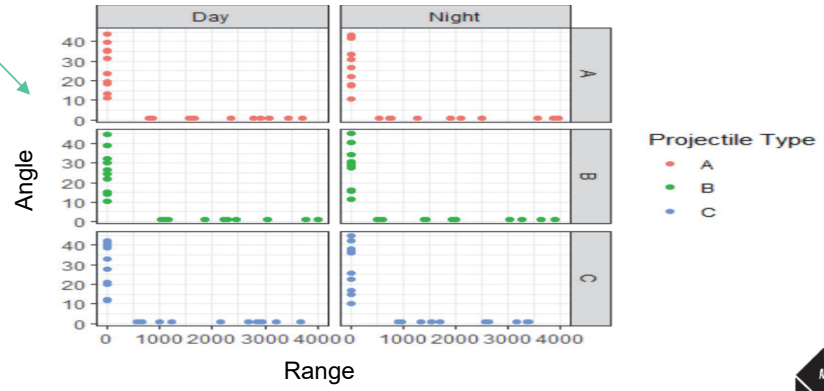
60-run fast flexible space-filling design for 2 categorical and 2 continuous factors



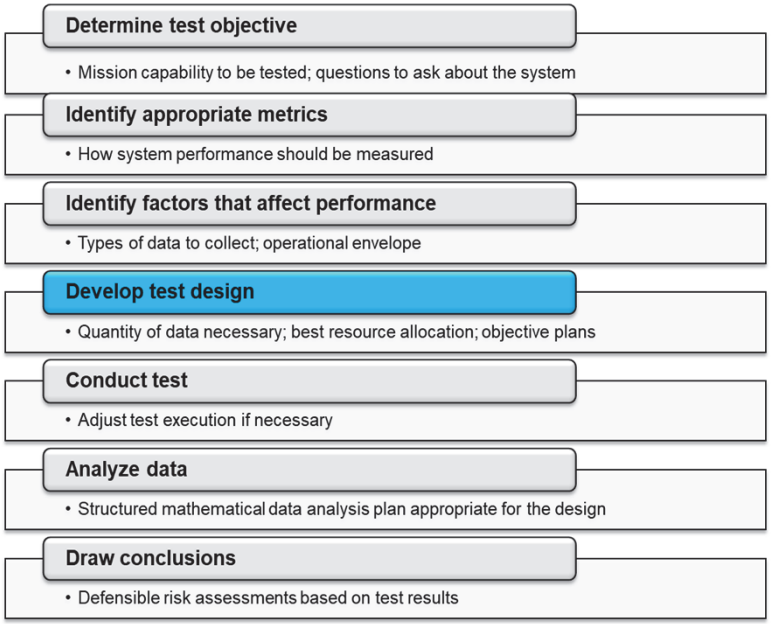
Smaller SFD within factor combinations



With good projection properties



Evaluating the Test Design



Defensible Test Designs Ensure That You Get The Data You Need To Evaluate The System

When evaluating a test design, ask:

1. Does this design ensure that **adequate data** is collected? (power/confidence)
2. If the system doesn't perform well, will I be able to **determine why**? (factor identifiability)
3. Does the DOE reflect **the way the test will be conducted**? (restricted randomization)
4. Will the data described in this test let me do the **analysis** I want to do? (characterize performance across the relevant factor space)

When evaluating a test design, ask:

1. Does this design ensure that **adequate data** is collected? (power/confidence)
2. If the system doesn't perform well, will I be able to **determine why**? (factor identifiability)
3. Does the DOE reflect **the way the test will be conducted**? (restricted randomization)
4. Will the data described in this test let me do the **analysis** I want to do? (characterize performance across the relevant factor space)

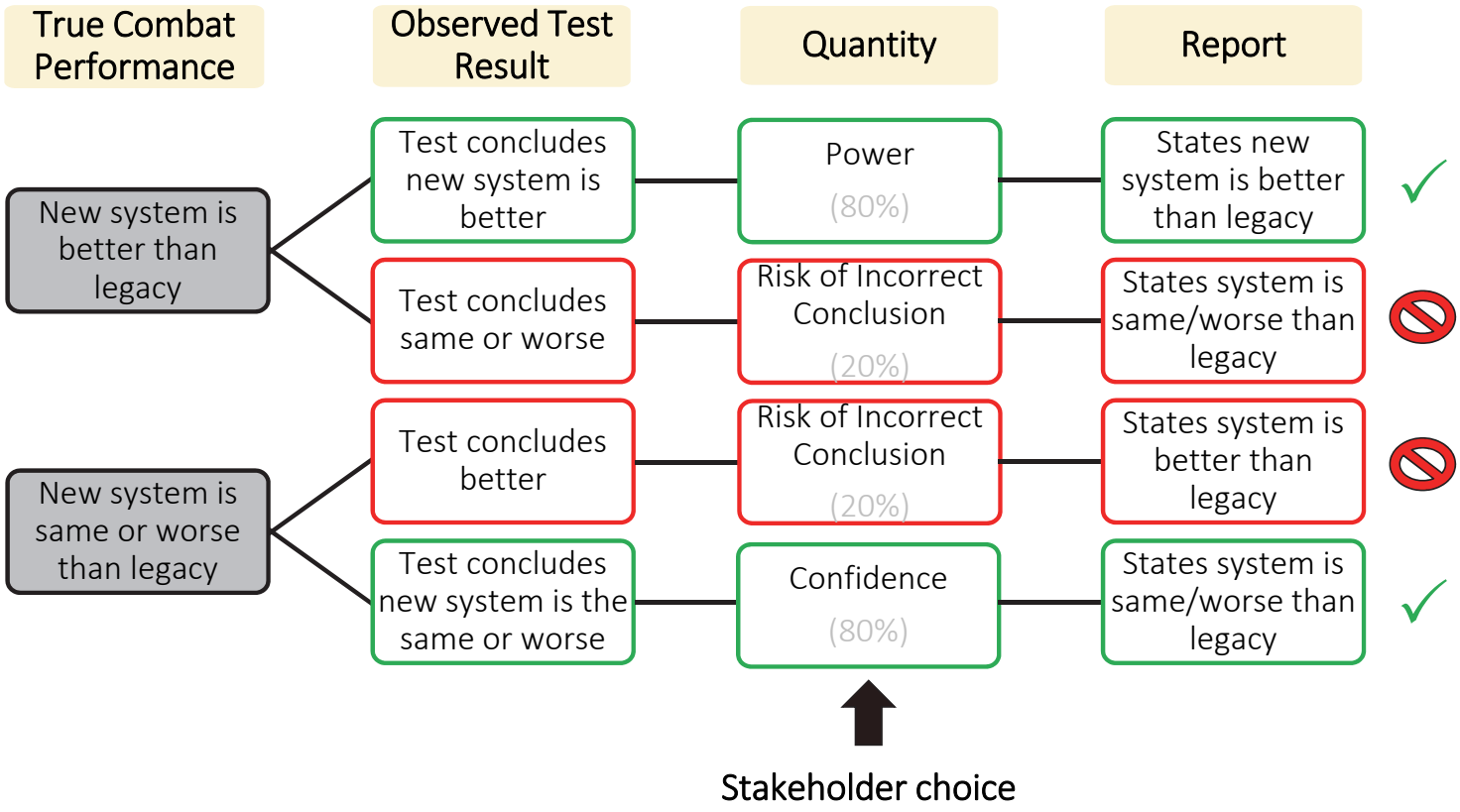
One Of The Main Reasons To Use DOE Is To Identify The Number Of Observations Required To Satisfactorily Address The Critical Questions

To know whether you have enough data, you first need to clarify the **specific questions** you want to address.

Questions of common interest:

1. Is the new system **better than the legacy** version?
2. What is the system's performance [according to some operationally relevant response variable] **across the operational space**?
3. Does the system's performance meet its **requirements**?

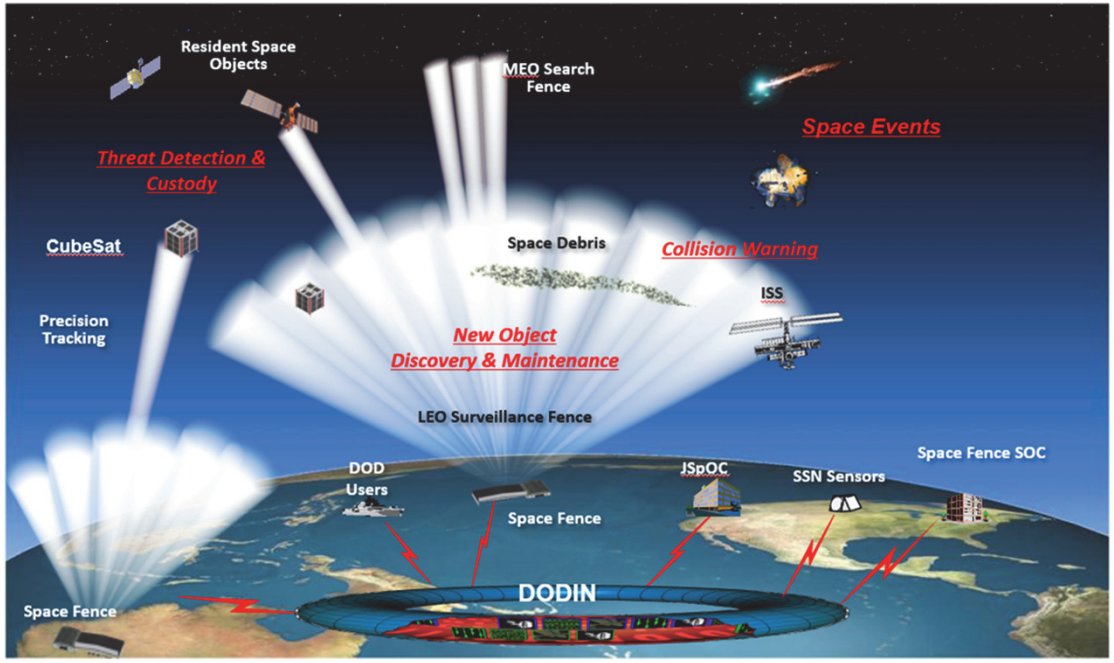
Statistical Power Is A Tool For Telling You If You Have Enough Data To Answer Critical Questions



Type-I Error α (probability of false positive)	Type-II Error β (probability of false negative)
	
“The new radar indicates incoming balloons!”	“The new radar doesn’t detect any balloons entering our airspace.”

Space Fence Is A Terrestrial Space-Directed S-band (2-4 Ghz) Radar System Designed To Detect, Track, And Catalog Space Objects

DOE question: **How many days** do we need to test to ensure that we observe enough space objects of each type to evaluate the system?



Mission

- Populate and maintain the Space Catalog (SATCAT)
- Routinely detect and track smaller objects
- Provide event notifications
- Support safety of flight

Test Objectives

- Determine probability of track for Space Fence
- Determine accuracy of Space Fence's position estimates

Space Fence Test Design Considerations

Challenges

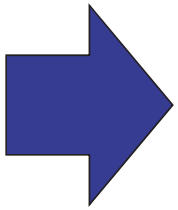
- Need to **characterize** system performance across full operational space, not just for the types of space objects most frequently observed
- **Cannot control** which space objects are observable during a given test day
- Satellite catalog analysis can provide information on the **frequency** with which objects meeting certain characteristics are observed

Test Design 1: Full factorial design for **Probability of Track**

Key Factors: Object Altitude and Inclination

Using Power To Design A Test Of The Right Size

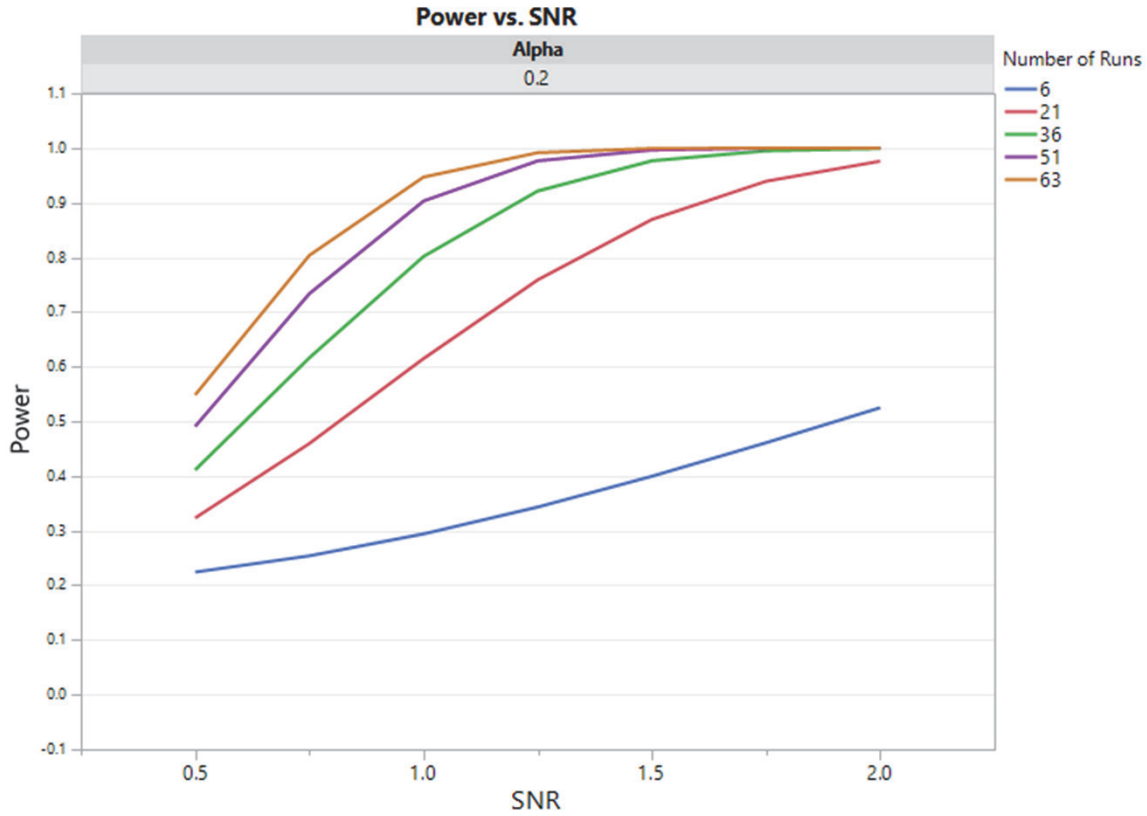
Inclination (degrees)	Altitude (kilometers)	Number
9 < θ ≤ 45	250-550	22
	550-800	60
	800-3,000	37
45 < θ ≤ 80	250-550	67
	550-800	1,094
	800-3,000	1,536
80 < θ ≤ 171	250-550	156
	550-800	1,356
	800-3,000	4,039
Total		8,367



Factor	Power for different test lengths (days)			
	6	7	8	10
Inclination	70.4%	85.0%	96.3%	99.9%
Altitude	62.5%	78.1%	92.8%	99.5%
I × A	46.0%	60.7%	79.7%	96.1%

Comparing different test sizes, we can see that in 8 days of testing, we expect to see enough objects to **fully characterize** Space Fence's Probability of Track

Power Curves Help Visualize And Compare Power For Different Test Designs



Power estimated by JEDIS v3.0

How Much Power? It's A Balancing Act

- Not enough power: **Knowledge risk!**
  - An experiment that is so small that we are unlikely to detect effects of interest is **a waste of resources.**

- Too much power: **Financial risk!**
  - A too-large (and therefore too-powerful) experiment **is also wasteful** because it uses more resources than are likely to be needed for detecting the alternative.

- Is there a sweet spot?
  - We *often* aim for at least 80% power and 80% confidence for screening and characterization, but there is nothing intrinsically correct about 80%! Sometimes, you may want more power if a comparison or factor is especially important.

When evaluating a test design, ask:

1. Does this design ensure that **adequate data** is collected? (power/confidence)
2. If the system doesn't perform well, will I be able to **determine why**? (factor identifiability)
3. Does the DOE reflect **the way the test will be conducted**? (restricted randomization)
4. Will the data described in this test let me do the **analysis** I want to do? (characterize performance across the relevant factor space)

Even The Best Systems Have Missions Or Situations That Aren't Their Strength

Defense Medical Information Exchange test report:

Operationally effective for queries of DoD and VA data, but not for queries of external partner data.

Q-53 Counterfire Radar test report:

Not effective for volley-fired mortars, but effective for other types of artillery and rockets and single-fire mortars

Tests should be designed so that we can identify the situations or missions where performance drops off

If You Only Ever Observe The Same Combination Of Factors, You Can't Determine Which One Is Driving Performance

Due to concerns like **convenience** and **resource availability**, data points may be arranged in a way that precludes differentiating factor effects

Small Diameter Bomb II

12 Run Main Effects Only Design – smaller than the real design!

Time of Day	Target Speed	Target Type	Update Rate	Clutter
Day	Fast	Wheeled	30	Y
Day	Fast	Wheeled	12	N
Day	Fixed	Wheeled	12	N
Day	Fixed	Wheeled	30	N
Day	Slow	Wheeled	12	Y
Day	Slow	Wheeled	30	Y
Night	Fast	Tracked	30	N
Night	Fast	Tracked	12	Y
Night	Fixed	Tracked	30	Y
Night	Fixed	Tracked	12	Y
Night	Slow	Tracked	12	N
Night	Slow	Tracked	30	N

Poor performance during the day?

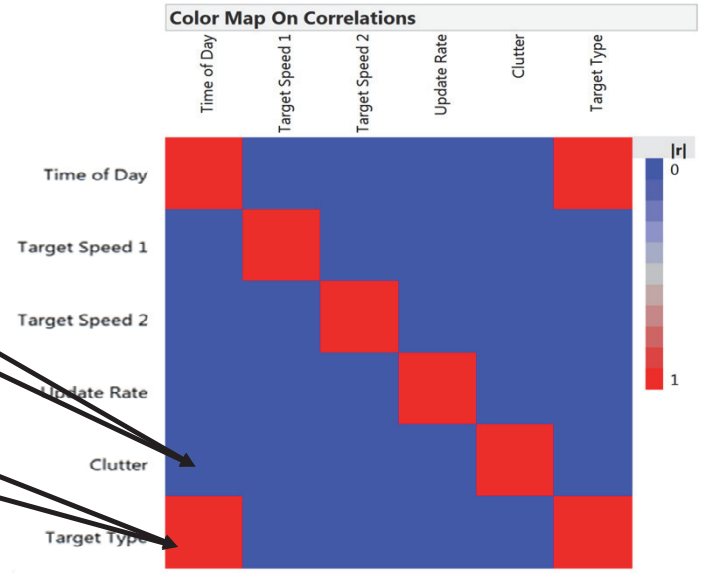
Poor performance against wheeled targets?

Poor performance against wheeled targets during the day?

Data contained in presentation are notional

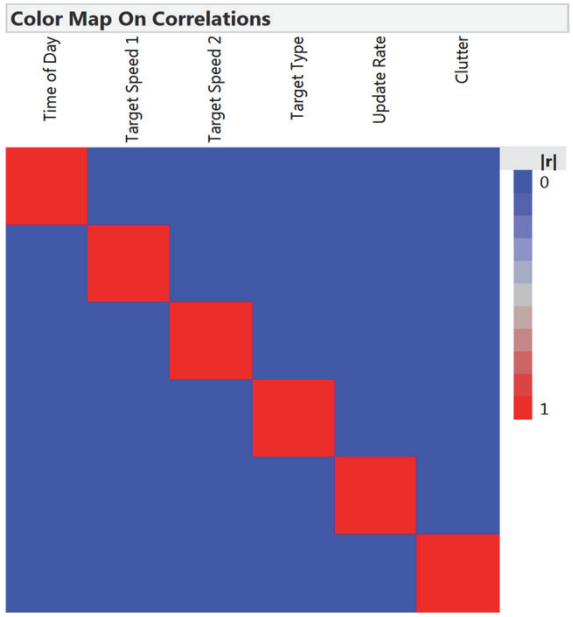
Use Correlation Maps To Quickly Visualize How Hard It Will Be To Distinguish Between Factors

Time of Day	Target Speed	Target Type	Update Rate	Clutter
Day	Fast	Wheeled	30	Y
Day	Fast	Wheeled	12	N
Day	Fixed	Wheeled	12	N
Day	Fixed	Wheeled	30	N
Day	Slow	Wheeled	12	Y
Day	Slow	Wheeled	30	Y
Night	Fast	Tracked	30	N
Night	Fast	Tracked	12	Y
Night	Fixed	Tracked	30	Y
Night	Fixed	Tracked	12	Y
Night	Slow	Tracked	12	N
Night	Slow	Tracked	30	N



- Blue means perfectly uncorrelated (No problems!)
- Red is perfectly (100%) correlated (Cannot differentiate)
- Colors in between indicate some degree of correlation (Can differentiate but estimates may be biased/incorrect)

Ideally, Your Test Design Will Have Factors Completely Independent To Ensure Easy Estimation



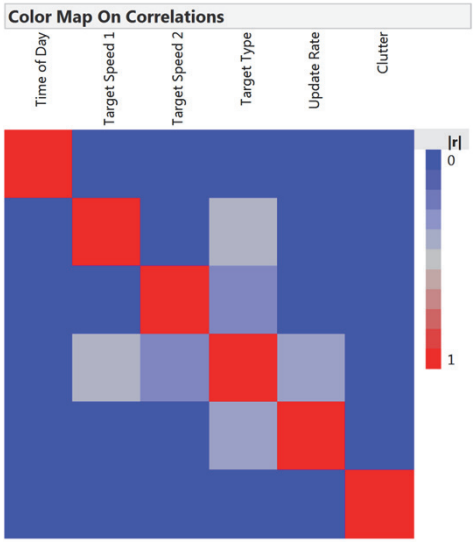
Time of Day	Target Speed	Target Type	Update Rate	Clutter
Day	Fast	Tracked	30	Y
Day	Fast	Wheeled	12	N
Day	Fixed	Tracked	12	N
Day	Fixed	Wheeled	30	N
Day	Slow	Tracked	12	Y
Day	Slow	Wheeled	30	Y
Night	Fast	Tracked	30	N
Night	Fast	Wheeled	12	Y
Night	Fixed	Tracked	30	Y
Night	Fixed	Wheeled	12	Y
Night	Slow	Tracked	12	N
Night	Slow	Wheeled	30	N

Depending on the type of design, this ideal might be achievable

- Full and fractional factorial designs
- Some optimal designs

Sometimes, Practical Constraints And Operational Realism Mean That Factors Will Be Partially Correlated

This is okay! – But it means that we cannot easily separate all factor effects on the response



Time of Day	Target Speed	Target Type	Update Rate	Clutter
Day	Fast	Wheeled	30	Y
Day	Fast	Wheeled	12	N
Day	Fixed	Tracked	12	N
Day	Fixed	Wheeled	30	N
Day	Slow	Tracked	12	Y
Day	Slow	Wheeled	30	Y
Night	Fast	Wheeled	30	N
Night	Fast	Wheeled	12	Y
Night	Fixed	Tracked	30	Y
Night	Fixed	Wheeled	12	Y
Night	Slow	Tracked	12	N
Night	Slow	Wheeled	30	N

Tracked vehicles don't have the same max speed as wheeled vehicles, so there will be no tracked targets moving at the "fast" speed.

When evaluating a test design, ask:

1. Does this design ensure that **adequate data** is collected? (power/confidence)
2. If the system doesn't perform well, will I be able to **determine why**? (factor identifiability)
3. Does the DOE reflect **the way the test will be conducted**? (restricted randomization)
4. Will the data described in this test let me do the **analysis** I want to do? (characterize performance across the relevant factor space)

Range Limitations And System Operations May Affect The Way Factors Are Varied In An Operational Test

Artillery Cannon

Survivability Moves:

- Goal is to evade enemy counterfire
- Occur frequently
- Cover short distances

Tactical Moves:

- Goal is to reposition as battle lines move
- Occur less frequently
- May cover longer distances

Operations:

Fire → Move → Fire

Response Variable

Miss Distance

Operations

- Fire platoons execute survivability moves between fire missions. These are generally short, so the range to target won't change much.
- Units will perform tactical moves once or twice every 12 hours. These are longer, meaning that the unit may move from one firing area to another.
- Between tactical moves, both illumination and range to target will be more or less constant, while other factors may be randomized.

Our test of the new cannon includes factors that are difficult to completely randomize and have disallowed combinations

Factor	Level	Difficult to Randomize	Disallowed Combinations
Time of Day	Day, Night	Hard	Night and Short Range
Range to Target	Short, Medium, Long	Hard	
Projectile Type	A, B, C	Easy	A and Long Range
Angle of Fire	Low, High	Easy	

Time of Day	Range	Angle of Fire	Projectile Type
Day	Long	Low	B
		Low	C
		High	B
		High	C
Tactical Move			
Day	Short	Low	A
		Low	B
		High	C
		High	B
Tactical Move			
Night	Medium	Low	C
		High	B
		Low	B
		Low	A
Tactical Move			
Day	Medium	High	A
		High	B
		Low	C
		Low	A
Tactical Move			
Night	Medium	High	A
		Low	B
		Low	A
		High	C
Tactical Move			
Night	Long	High	B
		Low	B
		High	C
		Low	C
Tactical Move			
Night	Long	Low	B
		Low	C
		High	C
		High	C
Tactical Move			
Day	Medium	High	C
		High	B
		High	A
		Low	B

Whole-Plot

Sub-Plots

Split-plot designs allow your DOE to reflect your test execution

Failing to correctly design and account for execution order could lead to wrong conclusions.

Lighting and **Range** are ***hard*** to change between Tactical Moves.

Angle of Fire and **Projectile Type** are ***easy*** to vary from one fire mission to the next.

Data contained in presentation are notional

Power And Restricted Randomization

- If you ignore the structure of your test, the power numbers you get will look higher, but this is fool's gold.
- If you cannot vary some of your factors completely at random, your power estimates should reflect this.
- Use a realistic estimate for how much group-to-group variation you can expect relative to run-to-run variation.

When evaluating a test design, ask:

1. Does this design ensure that **adequate data** is collected? (power/confidence)
2. If the system doesn't perform well, will I be able to **determine why**? (factor identifiability)
3. Does the DOE reflect **the way the test will be conducted**? (restricted randomization)
4. Will the data described in this test let me do the **analysis I** want to do? (characterize performance across the relevant factor space)

Design Your Test To Get The Information You Need

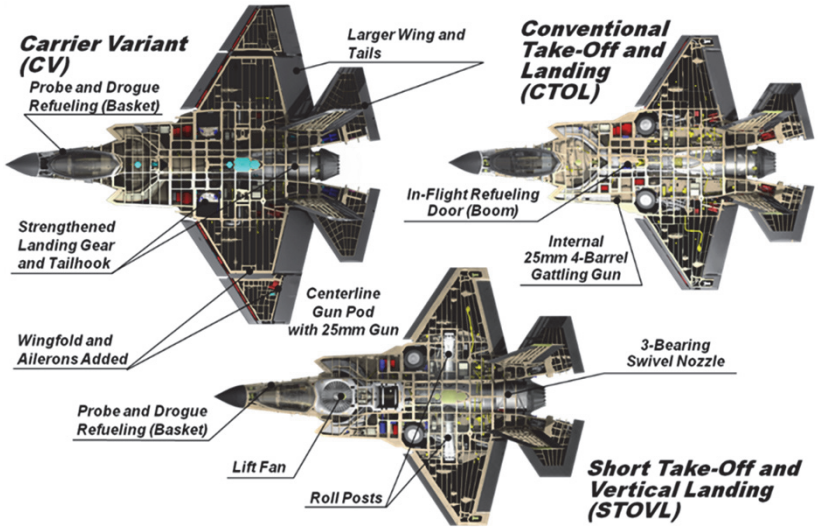
When resources are constrained, carefully describe the conditions the system will be employed in and make sure to get the data points required to fully describe that space.

Operational Test Should Cover The Full Spectrum Of Conditions In Which Systems Will Be Employed

F-35 Joint Strike Fighter test

Goal: Evaluate effectiveness of Block 3F focusing on the core mission areas designated in the JSF Operational Requirements Document (ORD). F-35 IOT&E will evaluate the aircraft's effectiveness in several core mission areas:

- Air-to-Surface (A/S) Attack
- Aerial Reconnaissance (AR) / Strike Coordination & Reconnaissance (SCAR)
- Close Air Support (CAS)
- Offensive Counter Air (OCA)
- Defensive Counter Air (DCA)
- Destruction – Suppression of Enemy Air Defenses (D-SEAD)
- Combat Search and Rescue (CSAR)
- Forward Air Controller – Airborne (FAC-A)



Differences between variants mean that the test must provide adequate information for all **three variants** in the **core mission areas**.

Choose Appropriate Factor Levels As Part Of Mission-Based Test Design

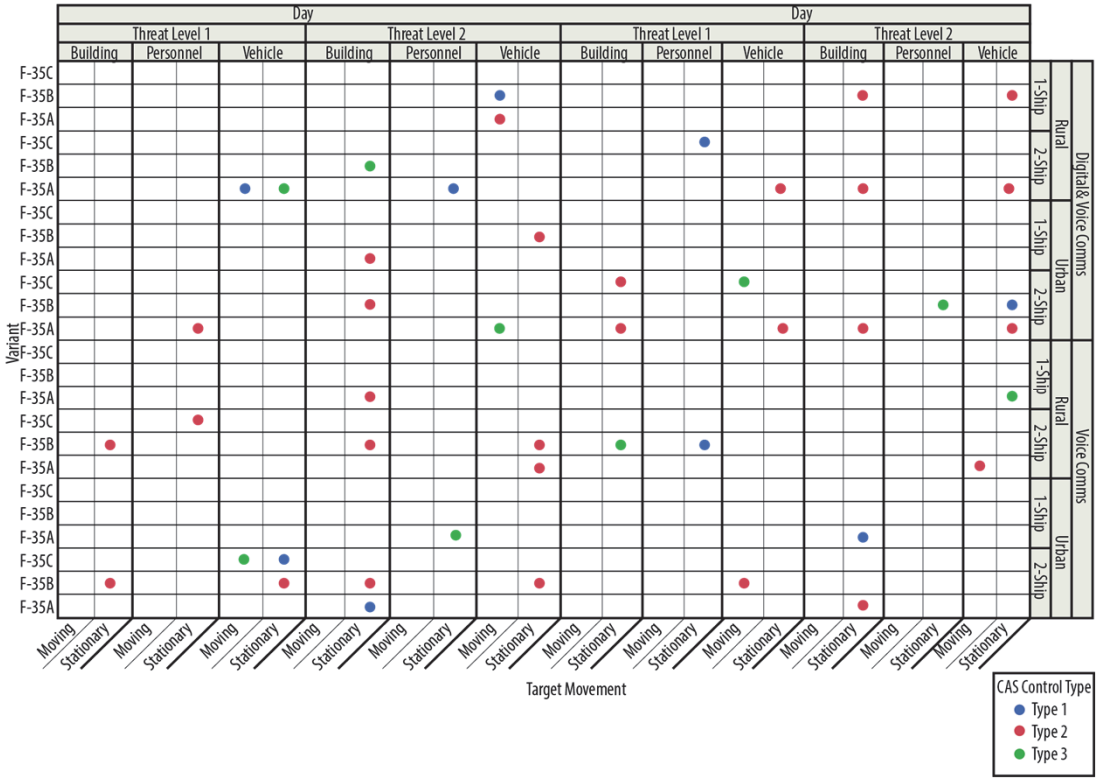
- Test design approach leverages DOE principles to span multiple factors across missions:
  - F-35 Variant – All Missions
  - Time of Day (Day/Night) – All Missions
  - Ground Threat – Covered Across Missions
  - Air Threat Level – Covered Across Missions
- Mission-specific factors allow for specific capability characterization within mission areas
  - E.g., close air support mission-specific factors include environment (urban/rural), target type, control type, target movement, formation size

Mission Areas	Air Threat	Ground Threat
Air-Surface (SCAR & AR)		
Strike		
Destruction/Suppression of Enemy Air Defenses		
Defensive counter air		
Offensive counter air		
Close air support		
Search and rescue		

Determine What You Want To Learn From The Test And Build The Design Accordingly

- All main effects
- Two-way interactions involving aircraft type
- Certain other two-way interactions

Close Air Support D-Optimal Design



Data contained in presentation are notional

Beginning With Good DOEs Allows To Produce More Accurate And Meaningful Results

A well-designed test will…

ensure enough data are collected

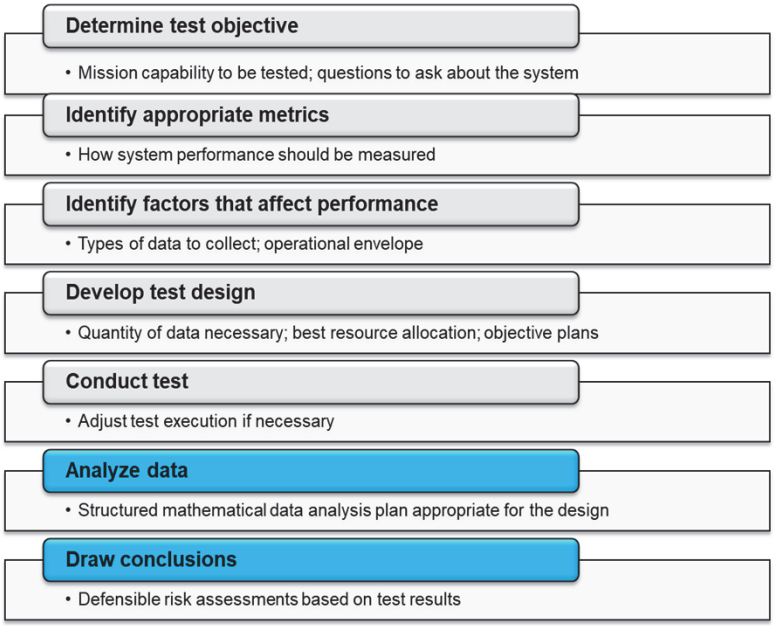
be able to identify the drivers of mission performance

reflect the way the test will be conducted

support a credible and defensible analysis

In Most Cases (If Not All) Test Designs Are Generated And Evaluated Using Statistical Software

Analysis and Reporting



Motivation

- A successful analysis should follow the structure of the test design.
- Building a statistical model using the factors specified in our test design allows us to quantify differences across relevant conditions and draw conclusions with confidence.

All experiments are designed with an analysis methodology in mind: to reap the benefits, we need to follow through to the analysis!

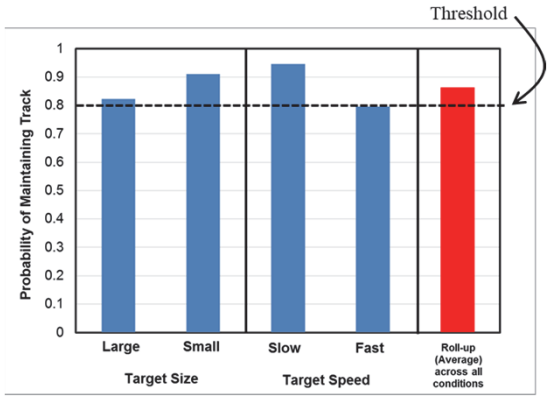
Key Tip #1: Avoid Data "Roll-Ups"

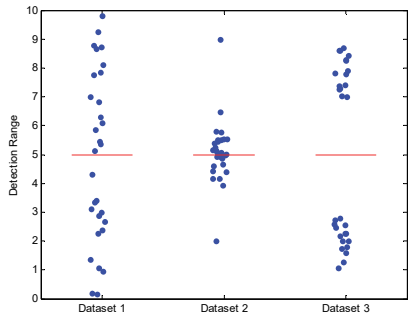
- A "roll-up" is the term for averaging or summarizing across test conditions to arrive at a smaller set of values, for example:
  - Reporting a mean across all conditions
  - Reporting a mean across all levels of one factor, ignoring "interaction" with other factors

- "Roll-ups" appear simple and straightforward, but they obscure true performance and reduce information.

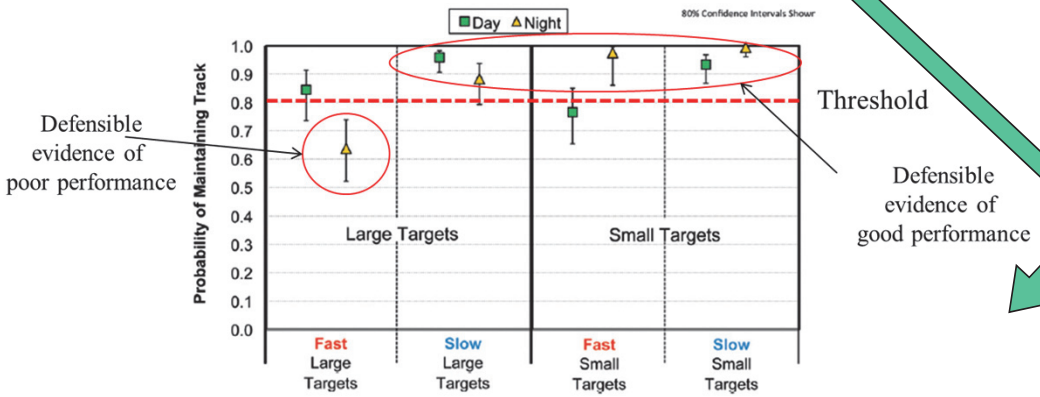
- We want to avoid data "roll-ups" because they can:
  - Arrive at a summary value that represents no actual test condition
  - Miss potentially important conditions where performance is poor

Statistical Analyses Can Summarize And Characterize Information Better Than Simple Averages And Roll-Ups





Same mean in every case, but very different distributions!



Area of poor performance would not have been uncovered without statistical modeling.

Data contained in presentation are notional

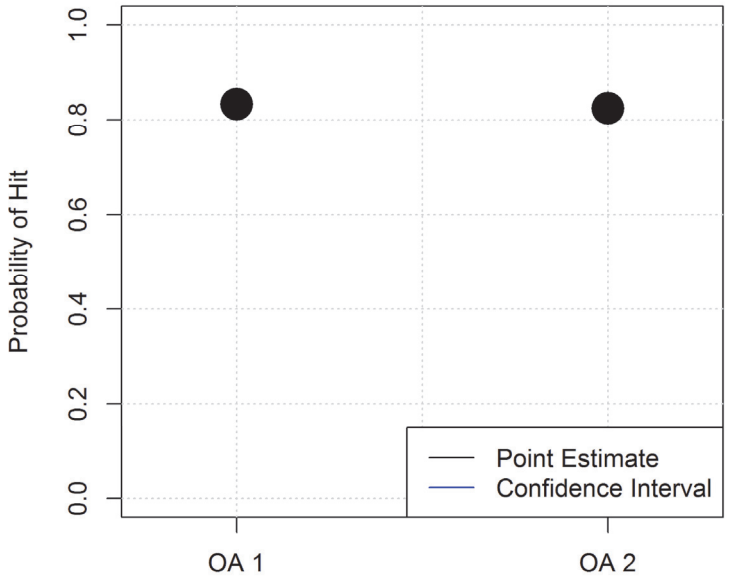
Key Tip #2: Interval Estimates Are Important!

Results from a single test event cannot predict exactly how a system will perform in the field.

- Confidence intervals tell us how precise our test results are.
- More data → tighter confidence bounds.
- Tighter confidence bounds → better estimate of our system.

Example: New turret for LAV Anti-Tank variant (notional data)

- Shoots TOW missiles
- OA 1: 12 shots
  - 10 hits
- OA 2: 40 shots
  - 33 hits



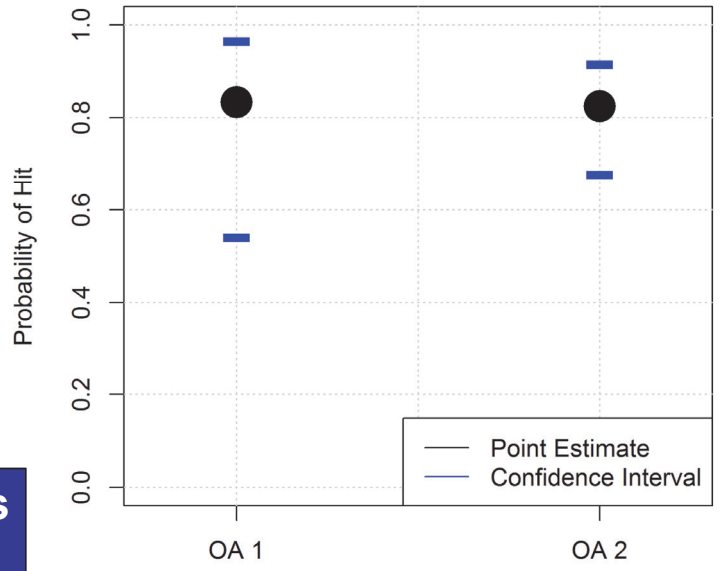
Key Tip #2: Interval Estimates Are Important!

Results from a single test event cannot predict exactly how a system will perform in the field.

- Confidence intervals tell us how precise our test results are.
- More data → tighter confidence bounds.
- Tighter confidence bounds → better estimate of our system.

Example: New turret for LAV Anti-Tank variant (notional data)

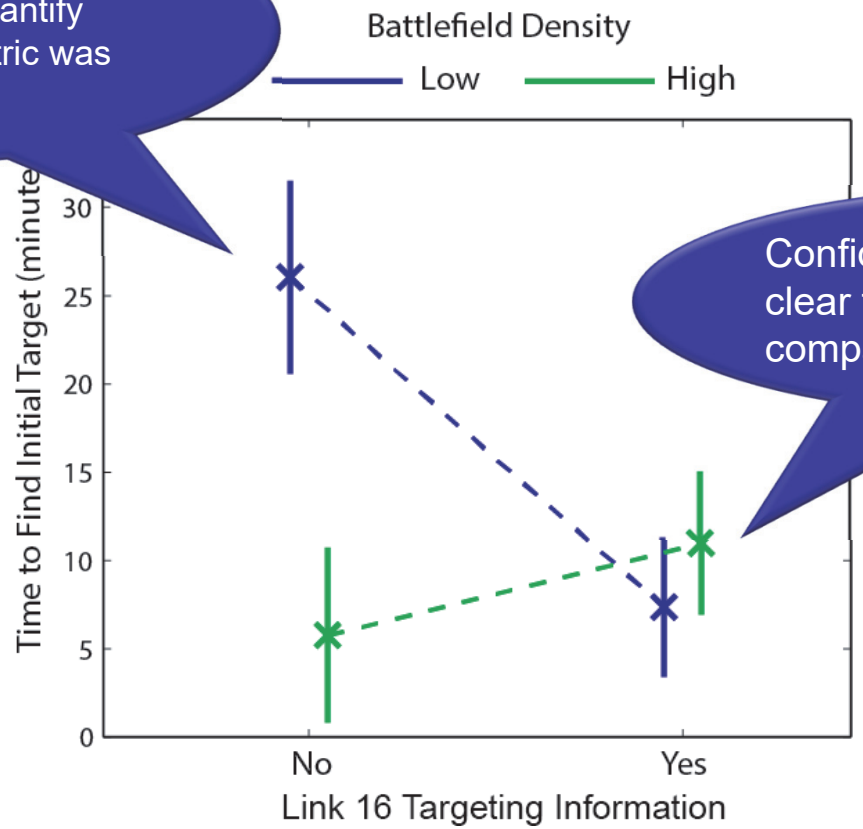
- Shoots TOW missiles
- OA 1: 12 shots
  - Interval Width: 42.5%
- OA 2: 40 shots
  - Interval Width: 23.8%



Interval estimates show the range of values for the system's lifetime performance under similar conditions.

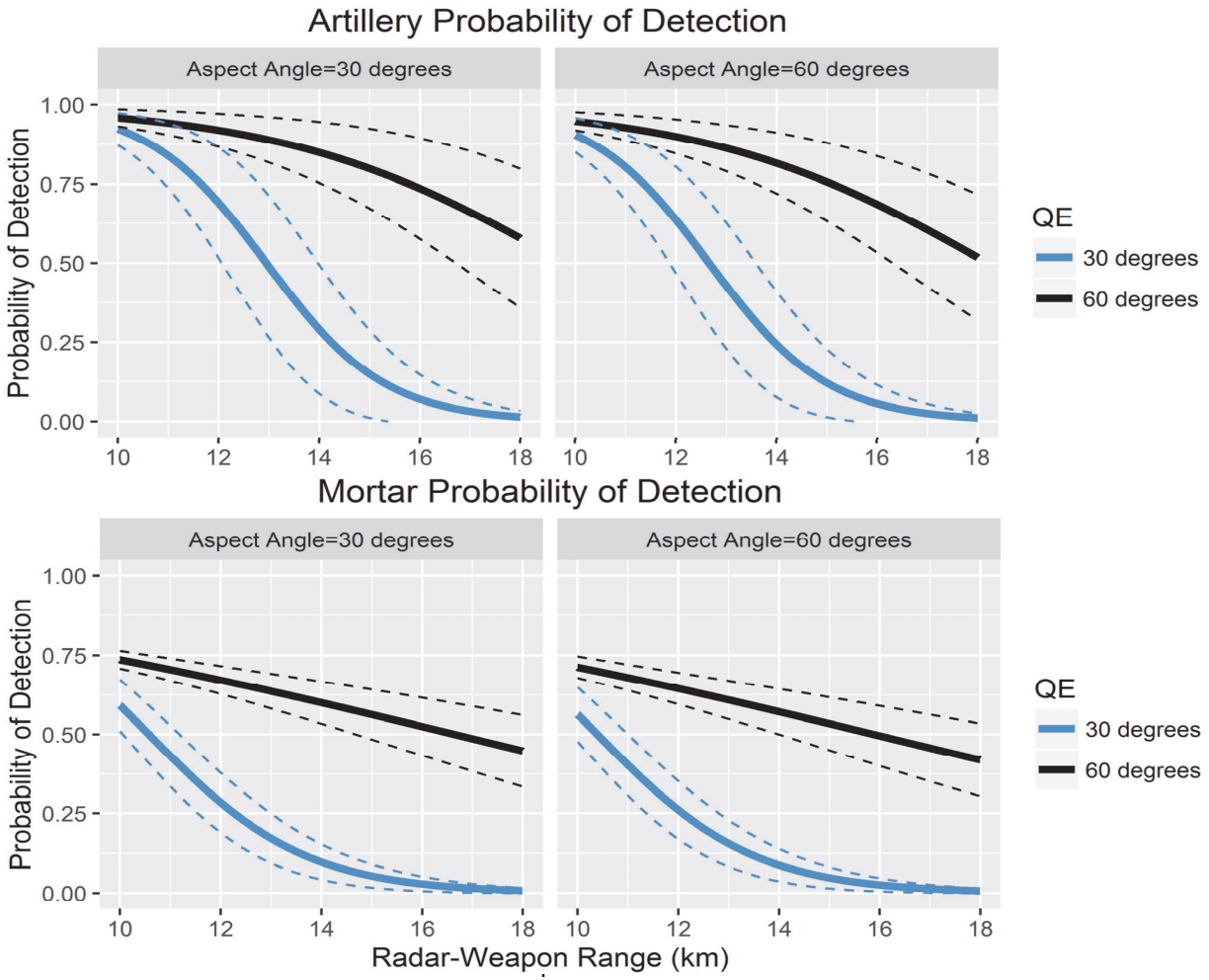
Conveying Level Of (Un)certainty In The Results Is Critical

Even without a threshold, confidence intervals quantify how accurately the metric was measured.



Confidence intervals make it clear that performance is comparable.

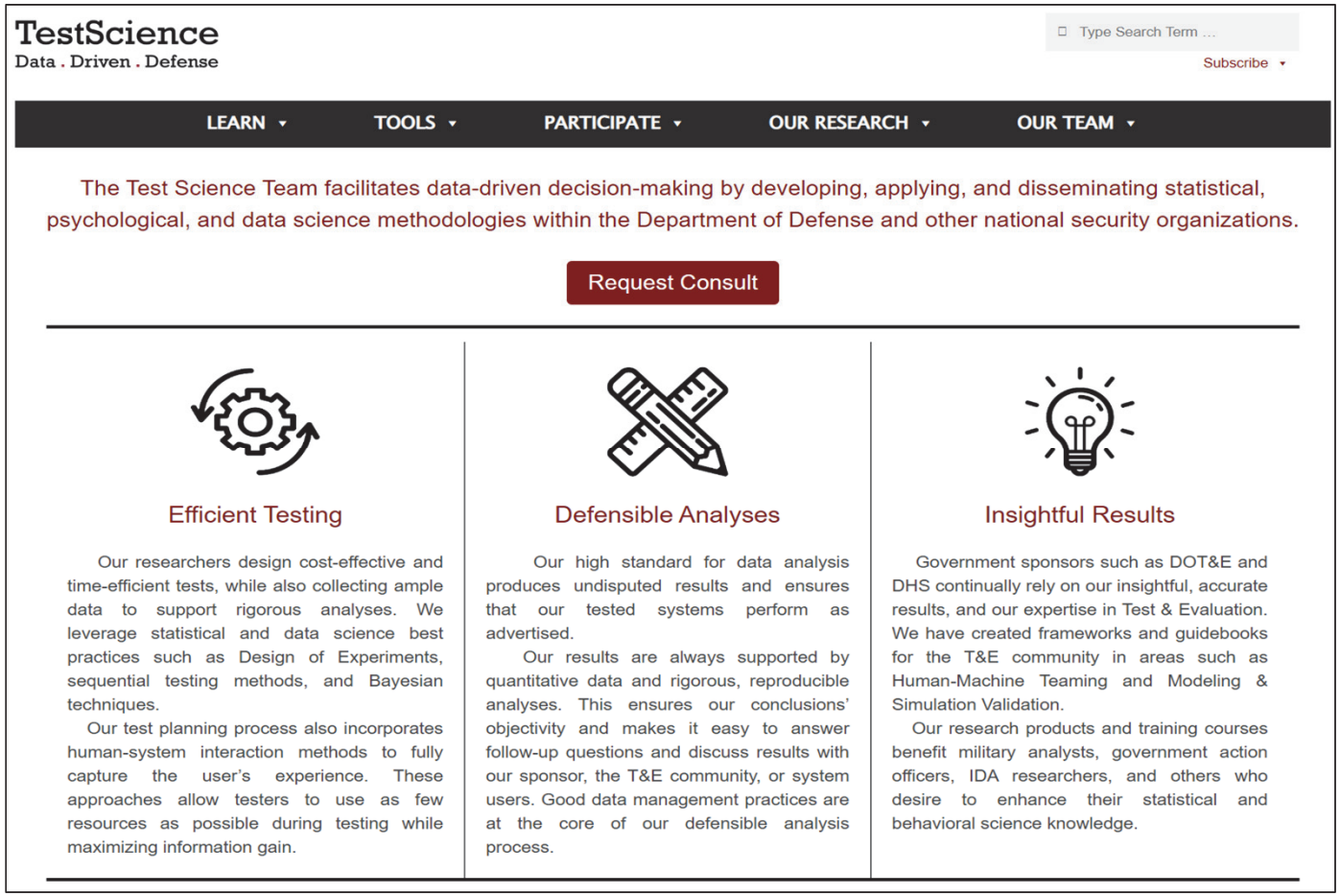
Key Tip #3: Use High-Level Graphical Summaries To Communicate Impact Across The Operational Space



Some Final Thoughts

- DOE encompasses a large body of techniques that support different types of data, use cases, and levels of complexity
- DOE enables you to design tests that investigate multiple factors in one experiment
- DOE provides the tester with an analytical framework to determine whether a test is good enough for their purpose
- A test strategy that employs DOE will provide the most powerful allocation of test resources for a given number of events
- Rigorous test design and analysis techniques facilitate an efficient, objective, and credible evaluation

Want To Learn More? Visit Us At Testscience.org!



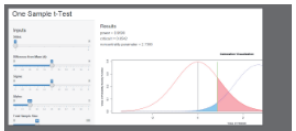
Test Science Website Tools

https://testscience.org/

Design, Analysis |code|
One Sample Proportion Test and Power
Interactive Shiny App

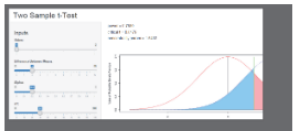
Contributed by: IDA Staff
Dec-03-2020
Power

Design, Analysis
One Sample t-test and Power
Interactive Shiny App



Contributed by: IDA Staff
Dec-03-2020
Power

Design |code|
Two Sample t-test Power
Interactive Shiny App



Contributed by: IDA Staff
Dec-03-2020
Power

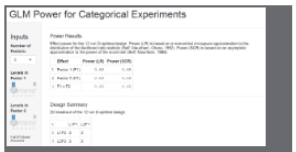
Design
JEDIS JMP AddIn
JMP Add In

Contributed by: IDA Staff
Jan-12-2022
Power

Design |code|
Categorical Analysis Power
Interactive Shiny App

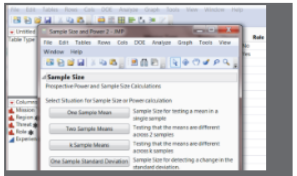
Contributed by: IDA Staff
Dec-03-2020
Power

Design
GLM Power for Categorical Factors
Interactive Shiny App



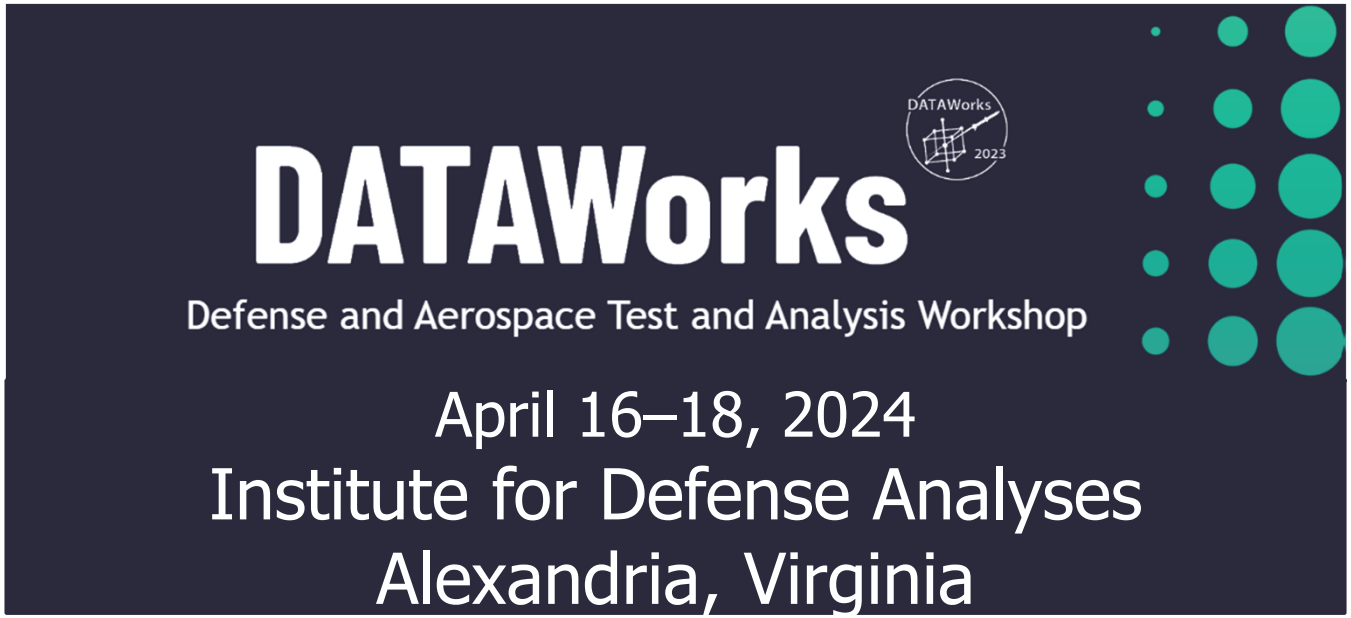
Contributed by: IDA Staff
Dec-03-2020
Power

Design
Power in JMP Tutorial
Document-PDF



Contributed by: IDA Staff
Dec-03-2020
Power

Want To Learn More About Statistics For T&E? Join Us At Dataworks 2024!



http://dataworks.testscience.org/

Additional Resources

- Test Science website: https://testscience.org/
- Freeman, L. J., Johnson, T., Avery, M., Lillard, V. B., & Clutter, J. (2018). Testing Defense Systems. Analytic Methods in Systems and Software Testing, 441.
- Montgomery, D. C. (2017). Design and analysis of experiments (Ninth ed.): Hoboken, NJ: John Wiley & Sons, Inc.
- Director, Operational Test & Evaluation (2010). Guidance on the use of Design of Experiments (DOE) in Operational Test and Evaluation.
- Director, Operational Test & Evaluation (2013). Best Practices for Assessing the Statistical Adequacy of Experimental Designs Used in Operational Test and Evaluation.
- Ahrens, M., Medlin, R., Pagán-Rivera, K., & Dennis, J. W. (2022). Case study on applying sequential analyses in operational testing. Quality Engineering, 1–12.
- Avery, K. M., Freeman, L. J., Parry, S. H., Whittier, G. S., Johnson, T. H., Flack, A. C., & Wojton, H. (2019). Handbook on Statistical Design and Analysis Techniques for Modeling and Simulation Validation. Institute for Defense Analyses D-10455.

BACKUP

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
XX-06-2023	Final	XXXX-XXXX - Jun 2023

4. TITLE AND SUBTITLE

I-ITSEC DOE Tutorial

5a. CONTRACT NUMBER

HQ0034-19-D-0001

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

John T. Haman
Kelly M. Avery
Keyla Pagan-Rivera
Rebecca M. Medlin

5d. PROJECT NUMBER

BD-09-2299(98)

5e. TASK NUMBER

2299(98)

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Institute for Defense Analyses
730 East Glebe Road
Alexandria, Virginia 22305

8. PERFORMING ORGANIZATION REPORT NUMBER

D-33561

H 2023-000231

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Director, Operational Test and Evaluation
1700 Defense Pentagon
Washington, DC 20301

10. SPONSOR/MONITOR'S ACRONYM(S)

DOT&E

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

This OED Draft has not been approved by the sponsor for distribution and release. Reproduction or use of this material is not authorized without prior permission from the responsible IDA Division Director.

13. SUPPLEMENTARY NOTES

Project Leader: Dr. Breeana Anderson

14. ABSTRACT

Understanding the experimental design process is fundamental to conducting efficient and effective tests and model and simulation experiments. This tutorial aims to provide a comprehensive understanding of the test design process, with practical examples and demonstrations. Our goal is to provide participants with the skills to create a good test design and effectively communicate statistical results in reports.
The tutorial begins with an introduction that explores the test design framework and explains the concept and purpose of experimental design. We then delve into the various aspects of planning, design selection and evaluation, and the analysis of an experimental design. Within the planning phase, we cover the essential elements such as identifying test objectives, response variables, and factor selection. We then provide an overview of different types of experimental designs and their purpose. In evaluating a test design, we discuss important metrics like power and confidence that are

15. SUBJECT TERMS

Design of Experiments; Operational Testing; Statistical Analysis

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Breeana Anderson
Unclassified	Unclassified	Unclassified	SAR	88	19b. TELEPHONE NUMBER *(Include area code)* (703) 845-6967