# Test Design, Statistical Analyses in TEMPs, Test Plans, and DOT&E Reports

Introduction and Course Overview

IDA

# Welcome and Overview

- **Administrative Details**
- **Introductions**
- **Test Planning**
  - Introduction & Overview
  - Test Planning Foundations
  - Experimental Designs
  - Best Practices
- **Statistical Analyses**

# Welcome and Overview

- **Administrative Details**
- **Introductions**
- **Test Planning**
- **Statistical Analyses**
  - Analysis Overview
  - Model Selection
  - Advanced Methodologies
    - » Generalized Linear Model
    - » Censored Data
    - » Bayesian Methodologies

# Course Objectives

- **Two primary goals:**
  - Provide a conceptual overview of Design and Analysis of Experiments
  - Educate Action Officers on the key elements of test adequacy when reviewing TEMPs and Test Plans
- **At the completion of the course Action Officers (AO) will:**
  - Understand the essential steps in designing an experiment
  - Be able to identify best practices in experimental design for operational testing
  - Be well equipped to review test designs in TEMPs and Test Plans for statistical test adequacy
  - Understand the basic statistical analysis methodologies that should be considered in analyzing OT data
  - Know the right level of analysis detail to provide in DOT&E reports

# Introduction

# What is the Science of Test?

- **Test Planning**
  - Design of Experiments (DOE) – a structured and purposeful approach to test planning
    - » Ensures adequate coverage of the operational envelope
    - » Determines how much testing is enough – statistical power analysis
    - » Provides an analytical basis for assessing test adequacy
  - Results:
    - » More information from constrained resources
    - » An analytical trade-space for test planning

- **Test Analysis and Evaluation**
  - Using statistical analysis methods to maximize information gained from test data
  - Incorporate all relevant information in analyses
  - Ensure conclusions are objective and robust

# Rational for DOE: Doing More with the Data

- **Goals: characterize initial detection range**
- **Hypothetical data**
  - IOT&E of a system: 12 runs, 3 in each condition

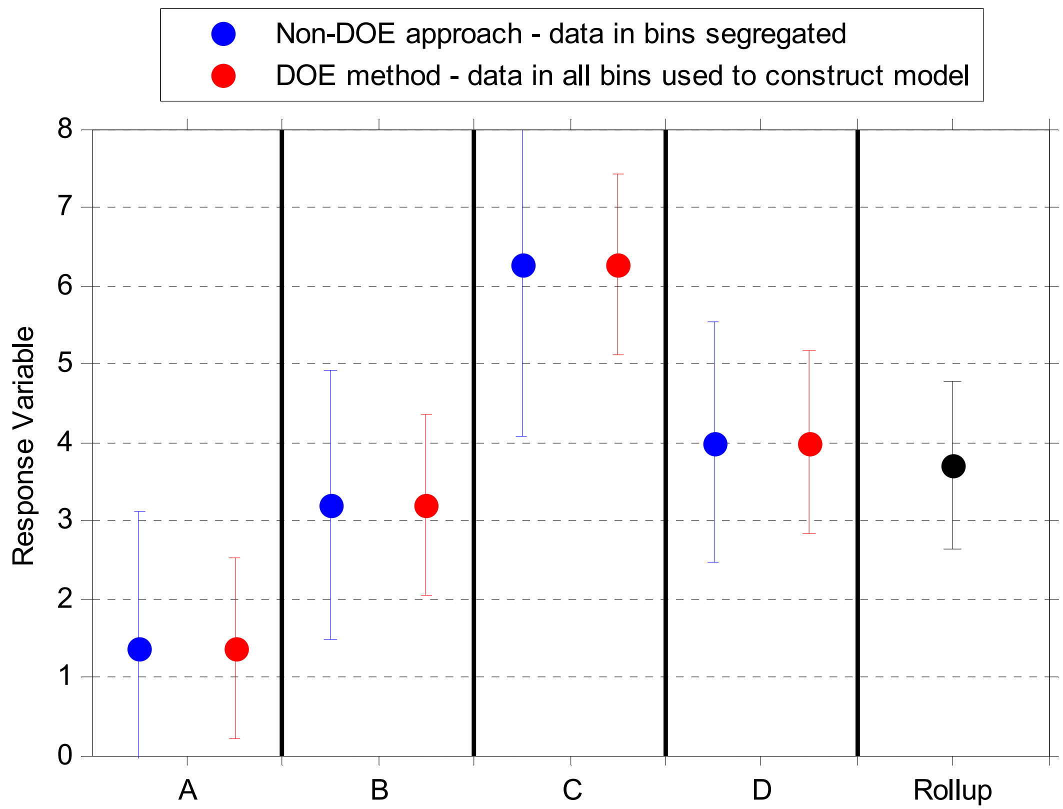
| | Slow Speed Target | Fast Speed Target |
|---|---|---|
| With Countermeasures | 0.2, 1.7, 2.2 | 2.1, 3.4, 4.1 |
| No Countermeasures | 4.9, 6.4, 7.5 | 3.2, 3.8, 5.0 |

- **Traditional analysis:**
  - Condition by condition
  - Means (std. deviations)

| | Slow Speed Target | Fast Speed Target |
|---|---|---|
| With Countermeasures | 1.37 (1.04) | 3.20 (1.01) |
| No Countermeasures | 6.27 (1.31) | 4.00 (0.92) |

- **DOE based statistical analysis uses information across all conditions in one model**

# DOE versus Non-DOE Analysis




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

- **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
  - Now can tell significant differences in performance
    - » E.g., system is better in C than in D conditions

- ***Note: Rollup (global mean) tells us little about system performance***

# What test methods are available?

- **Types of data collection**
  - DWWDLT – "Do what we did last time"
  - Special/critical cases
  - One-Factor-At-A-Time (OFAT)
  - Observational studies
  - Design of experiments
    - » Purposeful changing of test conditions

Cases

| | Mission Type | With UAS Day | With UAS Night | Without UAS Day | Without UAS Night |
|---|---|---|---|---|---|
| AB2 | Recon | 1 | | 2 | 2 |
| | Attack | | | 1 | |
| AB3 | Recon | | 1 | 4 | 2 |
| | Attack | | 3 | 2 | 2 |

- **Design of Experiments is preferable**
  - Challenges with Case-based
    - » Little predictive ability; loss of ability to determine cause and effect
    - » Limited to the specific conditions selected – might miss important performance shortfalls
    - » Often poor statistical precision (demos)
  - Challenges with OFAT
    - » Often is over-kill, unnecessarily large test sizes
    - » Interactions between conditions often not examined
  - Challenges with observational studies
    - » Confounding data
    - » Loss of ability to determine cause & effect
    - » However, necessary in exercises

OFAT

| | Mission Type | With UAS Day | With UAS Night | Without UAS Day | Without UAS Night |
|---|---|---|---|---|---|
| AB2 | Recon | | | 1 | 1 |
| | Attack | 1 | 1 | | |
| AB3 | Recon | | | 1 | 1 |
| | Attack | 1 | 1 | | |

"All tests are designed, many poorly"

# DOE is an Industry Best Practice

**Design of Experiments has a long history of application across many fields.**

- **Agricultural**
  - Early 20th century
  - Blocked, split-plot and strip-plot designs
- **Medical**
  - Control versus treatment experiments
- **Chemical and Process Industry**
  - Mixture experiments
  - Response surface methodology
- **Manufacturing and Quality Control**
  - Response surface methodology
  - DOE is a key element of Lean Six-Sigma
- **Psychology and Social Science Research**
  - Controls for order effects (e.g., learning, fatigue, etc.)
- **Software Testing**
  - Combinatorial designs test for problems

**There are many tools within the DOE toolbox!**
**New fields employing DOE tend to lead to new tools**

- **Pratt and Whitney Example**
  - Design for Variation process DOE
  - Turbine Engine Development
- **Key Steps**
  - Define requirements (probabilistic)
  - Analyze
    - Design experiment in key factors (heat transfer coefficients, load, geometric features, etc.)
    - Run experiment through finite element model
  - Solve for optimal design solution
    - Parametric statistical models
  - Verify/Validate
  - Sustain
- **Results**
  - Risk Quantification
  - Cost savings
  - Improved reliability

# Experiments… A Misunderstood Term

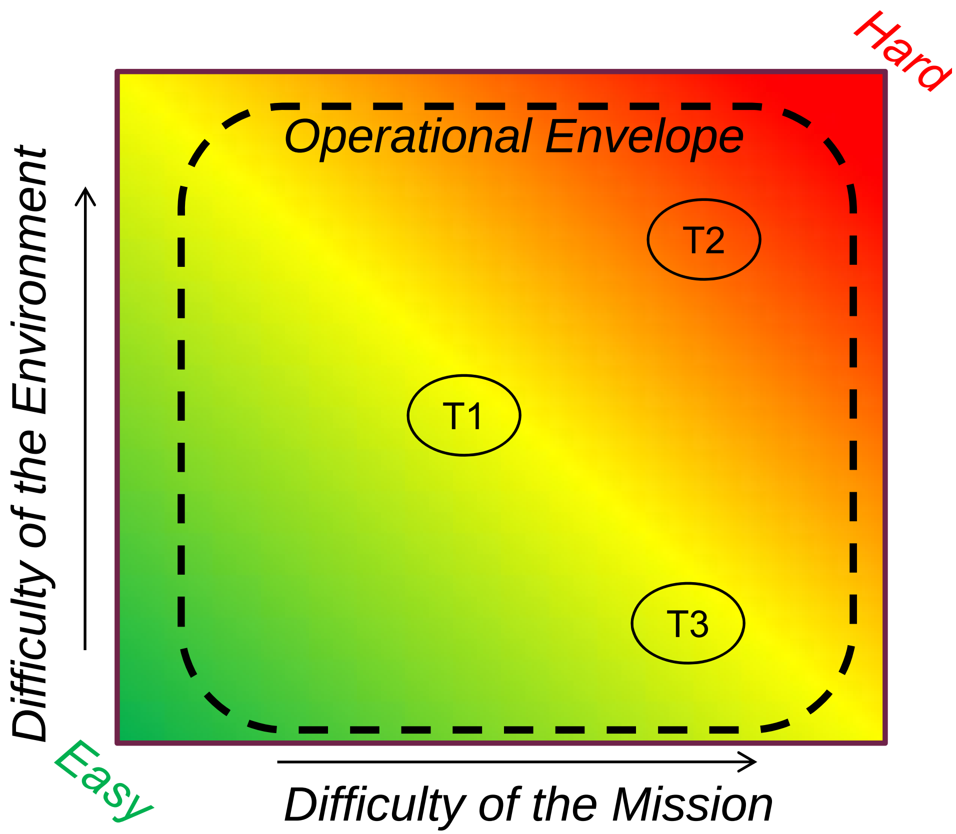
- **Definition:** a test or series of tests in which **purposeful** changes are made to the input variables in order to observe an outcome, which will be determined by a statistical **analysis**.
  - Note: an experiment is a **test or a series of tests**
- **Experiments are used widely throughout the engineering world**
  - Process characterization & optimization
  - Evaluation of material properties
  - Product design & development
  - Component & system tolerance determination
- **Experiments can be conducted in highly variable situations**
  - Agriculture
  - Human factors
  - However, they require **purposeful** changes to the **factors of interest**
- **"All experiments are designed experiments, some are poorly designed, some are well-designed" G.E.P. Box**

# Rationale for DOE in Test and Evaluation

- **The purpose of testing is to provide relevant, credible evidence with some degree of inferential weight to decision makers about the operational benefits of buying a system**
  - DOE provides a framework for the argument and methods to help us do that systematically

**Four Challenges faced by any test**

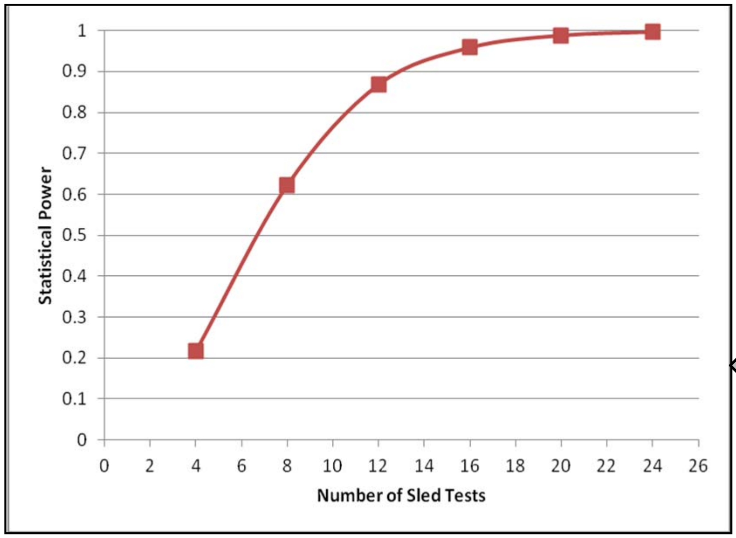
1. How many? Depth of Testing
2. Which Points? Breadth of Testing – spanning the operational envelope
3. How to Execute? Order of Testing
4. What Conclusions? Test Analysis – drawing objective, robust conclusions while controlling noise




DOE provides the analytical basis for test planning tradeoffs

# 1. How Many?

- **Need to execute a sample of _*n*_ drops/events/shots/measurements**
- **How many is enough to get it *right*?**
  - 3 – because that's how much $/time we have
  - 8 – because I'm an 8-guy
  - 10 – because I'm challenged by fractions
  - 30 – because something good happens at 30!
- **DOE methods provide the tools to calculate statistical power**

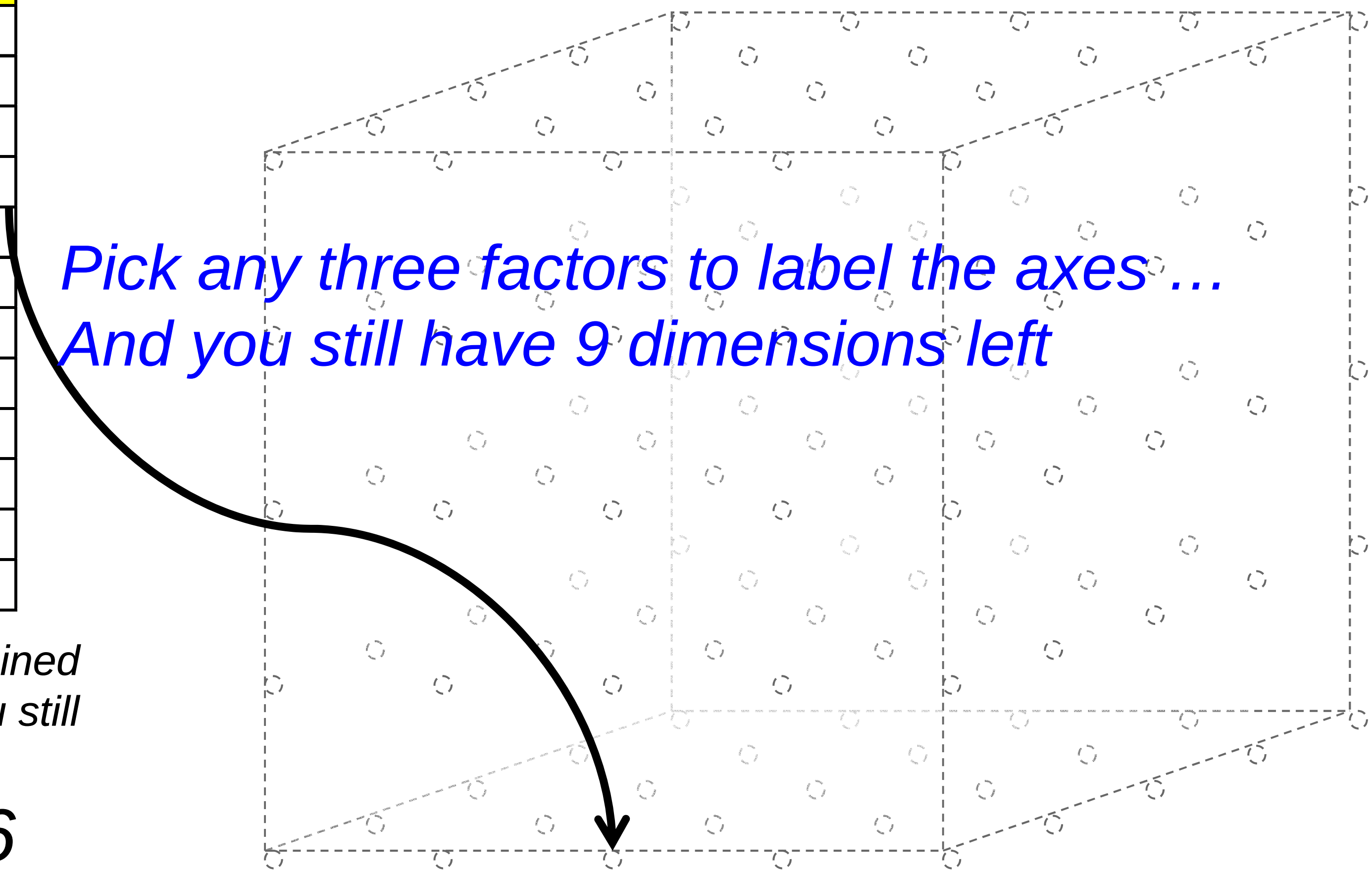



*Loosely speaking:*
*"Plot of Likelihood of Finding Problems vs N"*

*Analytical trade space for test planning – balancing risk and resources*

# 2. Which Points in a 12-D Battlespace?

| Test Condition |
|---|
| Target Type: |
| Num Weapons |
| Target Angle on Nose |
| Release Altitude |
| Release Velocity |
| Release Heading |
| Target Downrange |
| Target Crossrange |
| Impact Azimuth (°) |
| Fuze Point |
| Fuze Delay |
| Impact Angle (°) |




*If each factor constrained to just two levels, you still have ...*

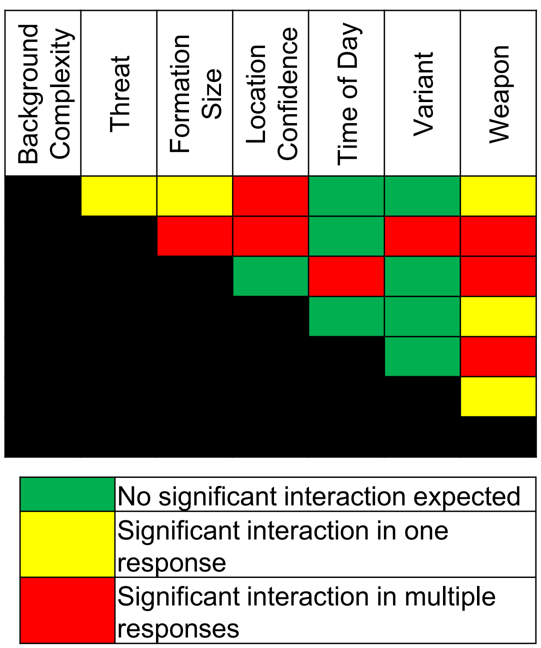
$$2^{12} = 4096$$

DOE provides efficient test design techniques to identify an adequate and significantly smaller test design than 4096 runs!

IDA

# Picking Test Points Case Study: JSF Air-to-Ground Missions

- **Operational Envelope Defined – DOE used to find reasonable subset of 128 cases**
- **Test team identified factors and their interactions and refined them to identify the most important aspects of the test design**




| | | | Variant - B | | | | | | | | Variant - A | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Category-B Threat | | | | Category-C Threat | | | | Category-B Threat | | | | Category-C Threat | | | |
| | | | Low TLC | | High TLC | | Low TLC | | High TLC | | Low TLC | | High TLC | | Low TLC | | High TLC | |
| | | | L | H | L | H | L | H | L | H | L | H | L | H | L | H | L | H |
| 2-Ship | Day | JDAM | | | | | | | | | | | | | | | | |
| | | LGB | | | | | | | | | | | | | | | | |
| | Night | JDAM | | | | | | | | | | | | | | | | |
| | | LGB | | | | | | | | | | | | | | | | |
| 4-Ship | Day | JDAM | | | | | | | | | | | | | | | | |
| | | LGB | | | | | | | | | | | | | | | | |
| | Night | JDAM | | | | | | | | | | | | | | | | |
| | | LGB | | | | | | | | | | | | | | | | |

# Proposed Design

- **Test team used combination of subject matter expertise, and test planning knowledge to efficiently cover the most important aspects of the operational envelope**

- **Provided the data are used together in a statistical model approach, plan is adequate to evaluate JSF performance across the full operational envelope.**

- ***Determined that 21 trials was the minimum test size to adequately cover the operational space***
  - *Ensures <u>important</u> factor interactions will be estimable*

- ***Note the significant reduction from the 128 possible conditions identified.***

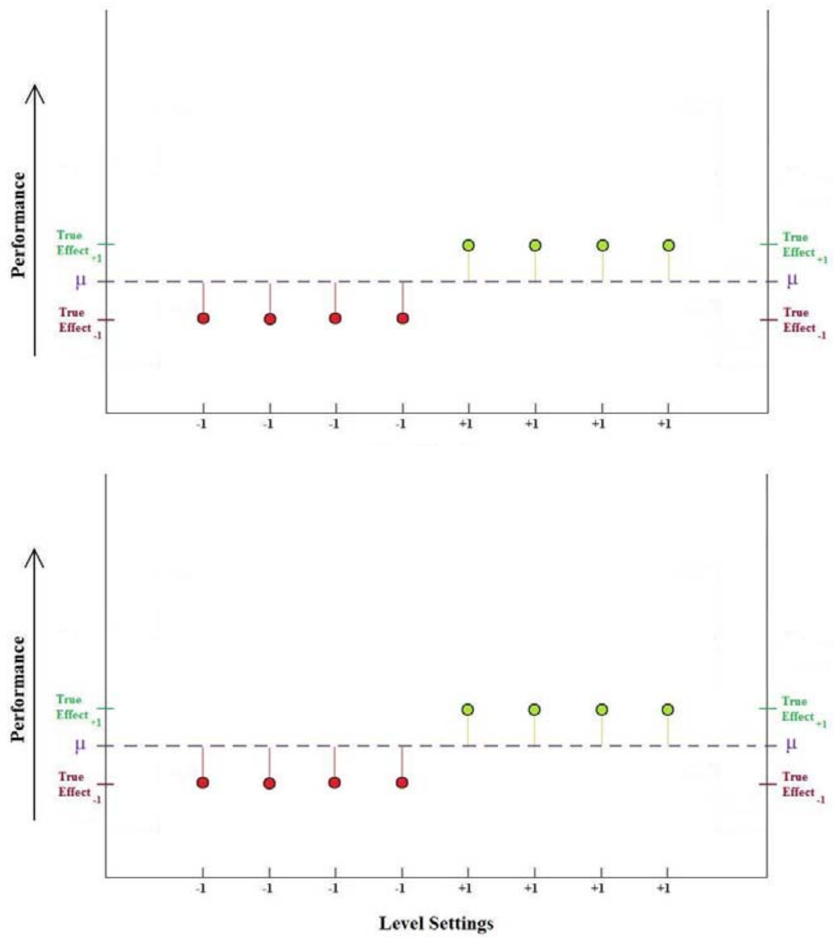
| | | | Variant - A | | | | | | | | Variant - B | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Category-B Threat | | | | Category-C Threat | | | | Category-B Threat | | | | Category-C Threat | | | |
| | | | Low TLC | | High TLC | | Low TLC | | High TLC | | Low TLC | | High TLC | | Low TLC | | High TLC | |
| | | | L | H | L | H | L | H | L | H | L | H | L | H | L | H | L | H |
| 2-Ship | Day | JDAM | | | 1 | | | | | | | 1 | | | | | | |
| | | LGB | | | | | | | | 1 | 1 | | | 1 | | | | |
| | Night | JDAM | 1 | | | | | | | 1 | | | | | 1 | | | |
| | | LGB | | 1 | | | | | | | | | 1 | | | 1 | | |
| 4-Ship | Day | JDAM | | | | | 1 | | | | | | | 1 | | | | |
| | | LGB | | | 1 | | | 1 | | | | | | | | | 1 | |
| | Night | JDAM | | 1 | | | | | | | | | 1 | | | | | 1 |
| | | LGB | | 1 | | | 1 | | | | | | | | | | | |

# JSF Air-to-Ground DOE

- **TEMP test design required 16 trials**
  - Would have been insufficient to examine performance in some conditions
- **Updated test design requires 21 trials but provides full characterization of JSF Pre-planned Air-to-Ground capabilities.**
- **New test design answers additional questions with the addition of only 5 trials:**
  - Is there a performance difference between the JSF variants?
    - » Do those differences only manifest themselves only under certain conditions?
  - Can JSF employ both primary weapons with comparable performance?

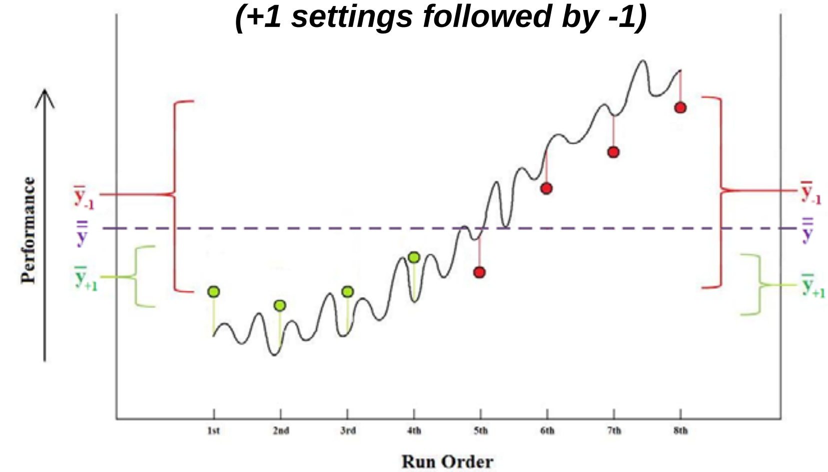
# 3. How to Execute - Randomizing

### *Truth Model*

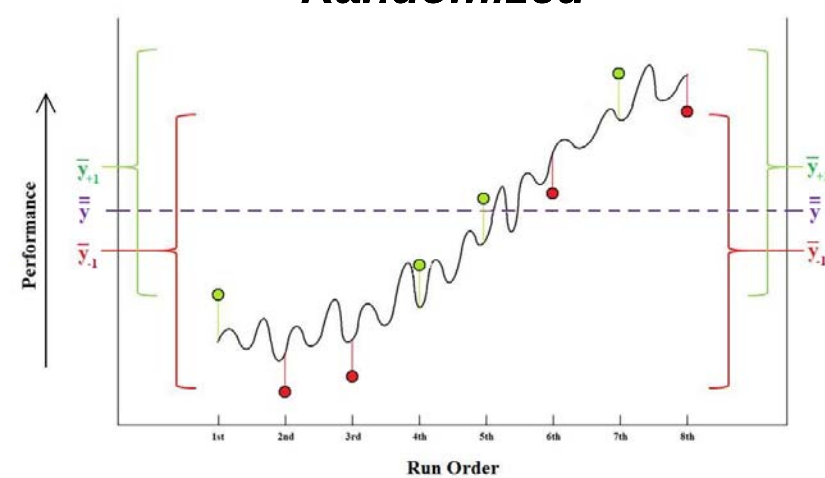



### *Non-Randomized*

***(+1 settings followed by -1)***




### *Randomized*




*Randomization avoids confounding test conditions with underlying uncontrollable noise (e.g. weather conditions).*

# Operational Test Implications of Randomization

- **Complete randomization is often not possible in operational testing**
- **Implications**
  - Action Officers should ensure that execution avoids systematic confounding of any factor with test execution
  - For example, avoid conducting all night missions first and day missions second
  - Design techniques exist such as split-plot and blocking which can provide executable designs

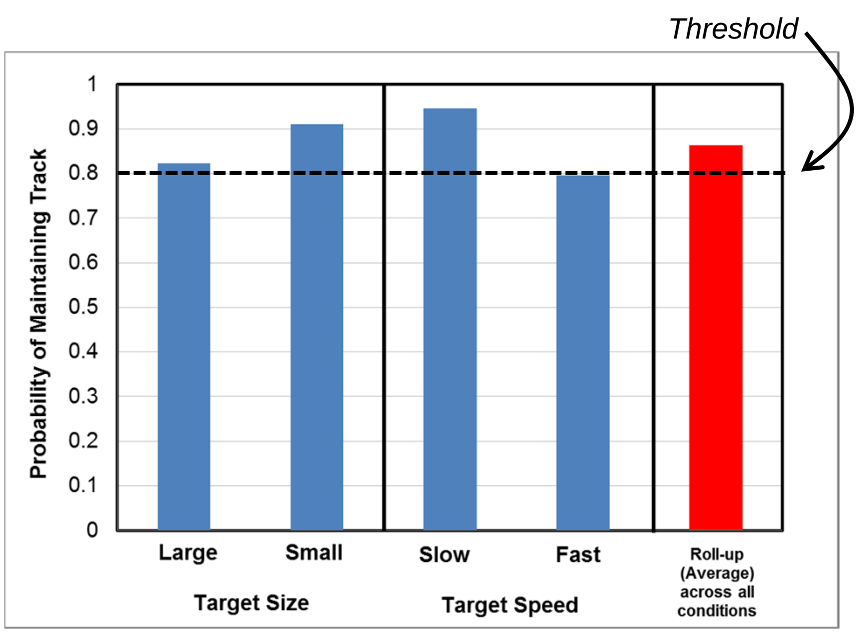
# 4. What Conclusions?
## (Traditional Analysis – Misses Important Information)

- **Cases or scenario settings and findings**

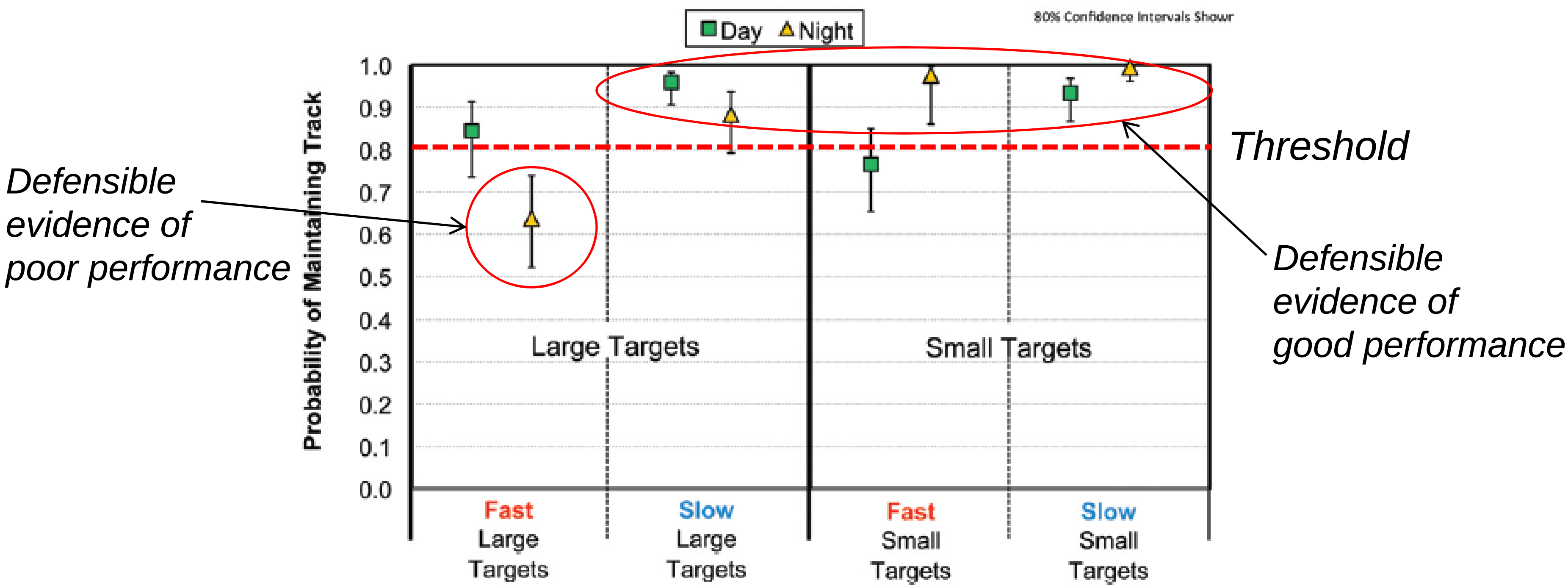
| Mission | Target Size | Target Speed | Time of Day | Result |
|---|---|---|---|---|
| 1 | Small | Fast | Night | Success |
| 1 | Large | Fast | Night | Failure |
| 2 | Large | Slow | Day | Success |

- **Run summaries**
  - Subject to removing "anomalies" if they don't support expected trend
  - No link to cause and effect

- **Report *average performance* in common conditions or global average alone**
  - Compare point estimate to threshold
  - No estimate of precision/uncertainty (how likely are we to see the same result again?)



# 4. What Conclusions?
## (DOE Analysis – Correctly Identifies Performance Changes)




- **DOE enables tester to build math-models* of input/output relations, quantifying noise, controlling error**

$$\boxed{\text{Responses} = f\left(\text{Factors}\right) + \varepsilon}$$

- **Enables performance characterization across multiple conditions**
  - Find problems with associated causes to enable system improvement
  - Find combinations of conditions that enhance/degrade performance (lost by averaging)
- **Rigorous determination of uncertainty in results – how confident am I that it failed threshold in Condition X?**



* *Many model choices: regression, ANOVA, mixed models, Censored Data, Gen Linear Model, etc. etc.*

# Overview Summary

- **Take-away: we already have good science in our system development**
  - We understand sys-engineering, guidance, aero, mechanics, materials, physics, electromagnetics …
  - DOE provides us the *Science of Test*
- **Design of Experiments (DOE) – a structured and purposeful approach to test planning**
  - Ensures adequate coverage of the operational envelope
  - Determines how much testing is enough
  - Quantifies test risks
  - Results:
    - » More information from constrained resources
    - » An analytical trade-space for test planning
- **Statistical *measures of merit* provide the tools needed to understand the quality of any test design to support statistical analysis**
- **Statistical analysis methods**
  - Do more with the data you have
  - Incorporate all relevant information in evaluations
    - » Supports integrated testing
- **DOT&E Memos provide expectations and outline best practices**
  - Flawed Application of DOE to OT&E
  - Assessing Statistical Adequacy of Experimental Designs in OT&E