
Science of Test: Improving the Efficiency and Effectiveness of DoD Test and Evaluation

Dr. Bram Lillard
Dr. Laura Freeman
Operational Evaluation Division
24 October 2014




- DOT&E was created by Congress in 1983.
- Responsible for all operational test & evaluation and to monitor and review live fire test & evaluation within DoD.
- Independent evaluation of the results of operational test and live fire test & evaluation.
- Objective reporting of these results to decision makers in DoD and Congress.
- DOT&E Focus
 - Is the system **operationally effective**?
 - Is the system **operationally suitable**?
 - Is the OT&E and/or LFT&E **adequate**?
 - Is the system **survivable** and **lethal**?



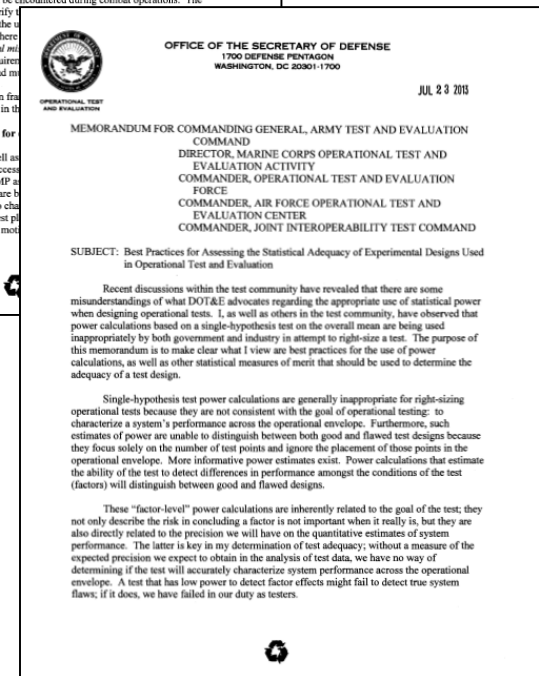
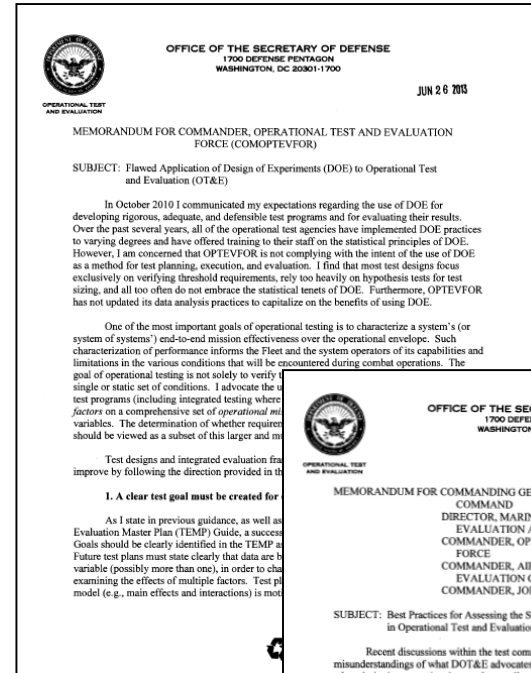
- **Test Planning**
 - Design of Experiments (DOE) – a structured and purposeful approach to test planning
 - » Ensures adequate coverage of the operational envelope
 - » Determines how much testing is enough – statistical power analysis
 - » Provides an analytical basis for assessing test adequacy
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

- **National Research Council Study (1998)**
 - “The current practice of statistics in defense testing design and evaluation does not take full advantage of the benefits available from the use of state-of-the-art statistical methodology.”
 - “The service test agencies should examine the applicability of state-of-the-art experimental design techniques and principles...”
- **Operational Test Agency Memorandum of Agreement (2009)**
 - “This group endorses the use of DOE as a discipline to improve the planning, execution, analysis, and reporting of integrated testing.”
- **DOT&E Initiatives (2009)**
 - “The DT&E and OT&E offices are working with the OTAs and Developmental Test Centers to **apply DOE across the whole development and operational test cycle** for a program.”
 - “Whenever possible, our evaluation of performance must include a rigorous **assessment of the confidence level of the test**, the **power of the test** and some measure of how well the **test spans the operational envelope** of the system.”

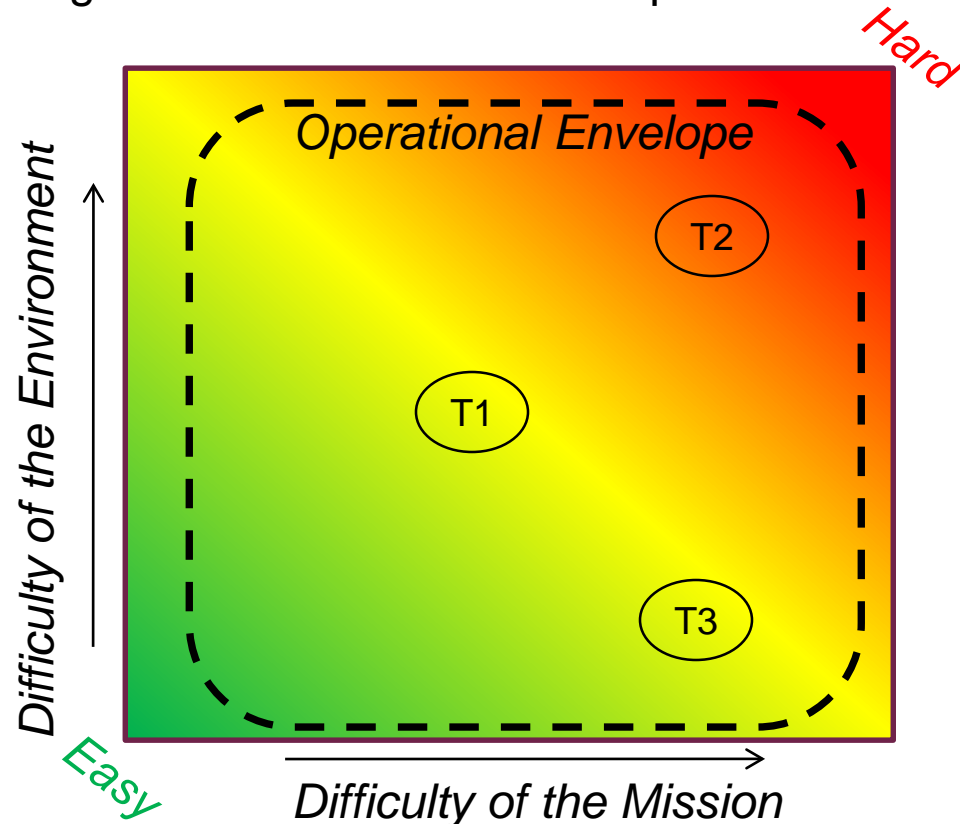
 <p>OFFICE OF THE SECRETARY OF DEFENSE 1700 DEFENSE PENTAGON WASHINGTON, DC 20301-1700</p> <p>OPERATIONAL TEST AND EVALUATION</p> <p>MEMORANDUM FOR COMMANDER, ARMY TEST AND EVALUATION COMMAND COMMANDER, OPERATIONAL TEST AND EVALUATION FORCE COMMANDER, AIR FORCE OPERATIONAL TEST AND EVALUATION CENTER DIRECTOR, MARINE CORPS OPERATIONAL TEST AND EVALUATION ACTIVITY COMMANDER, JOINT INTEROPERABILITY TEST COMMAND DEPUTY UNDER SECRETARY OF THE ARMY, TEST & EVALUATION COMMAND DEPUTY, DEPARTMENT OF THE NAVY TEST & EVALUATION EXECUTIVE DIRECTOR, TEST & EVALUATION, HEADQUARTERS, U.S. AIR FORCE TEST AND EVALUATION EXECUTIVE, DEFENSE INFORMATION SYSTEMS AGENCY DOT&E STAFF</p> <p>SUBJECT: Guidance on the use of Design of Experiments (DOE) in Operational Test and Evaluation</p> <p>This memorandum provides further guidance on my initiative to increase the use of scientific and statistical methods in developing rigorous, defensible test plans and in evaluating their results. As I review Test and Evaluation Master Plans (TEMPs) and Test Plans, I am looking for specific information. In general, I am looking for substance vice a 'cookbook' or template approach - each program is unique and will require thoughtful tradeoffs in how this guidance is applied.</p> <p>A "designed" experiment is a test or test program, planned specifically to determine the effect of a factor or several factors (also called independent variables) on one or more measured responses (also called dependent variables). The purpose is to ensure that the right type of data and enough of it are available to answer the questions of interest. Those questions, and the associated factors and levels, should be determined by subject matter experts -- including both operators and engineers -- at the outset of test planning.</p>	<p>for when I approve TEMPs and</p> <p>t evaluation of end-to-end tic environment.</p> <p>es for effectiveness and parameters but most likely there</p> <p>ess and suitability. y, develop a test plan that tors across the applicable levels nation in order to concentrate</p> <p>ss both developmental and interest.</p> <p>ence) on the relevant response tical measures are important to can be evaluated by decision- e off test resources for desired</p> <p>entify the metrics, factors, and nd suitability and that should be</p>
<p>reflected in detailed test plans. DOT&E is working with other members of the test and evaluation community to develop a two-year roadmap for implementing this scientific and rigorous approach to testing. I am looking for as much substance as possible as early as possible, but each TEMP revision can be tailored as more information becomes available. That content can either be explicitly made part of TEMPs and Test Plans, or referenced in those documents and provided separately to DOT&E for review.</p> <p><i>J. M. Gilmore</i> J. Michael Gilmore Director</p> <p>cc: DDT&E</p>	<p>2</p>

- ❑ **The goal of the experiment.** This should reflect evaluation of end-to-end mission effectiveness in an operationally realistic environment.
- ❑ Quantitative mission-oriented **response variables** for effectiveness and suitability. (These could be Key Performance Parameters but most likely there will be others.)
- ❑ **Factors** that affect those measures of effectiveness and suitability. Systematically, in a rigorous and structured way, develop a test plan that provides good breadth of coverage of those factors across the applicable levels of the factors, taking into account known information in order to concentrate on the factors of most interest.
- ❑ **A method for strategically varying factors** across both developmental and operational testing with respect to responses of interest.
- ❑ **Statistical measures of merit (power and confidence)** on the relevant response variables for which it makes sense. These statistical measures are important to understanding "how much testing is enough?" and can be evaluated by decision makers on a quantitative basis so they can trade off test resources for desired confidence in results.

- Flawed application of DOE memo emphasizes:**
 - Importance of clear test goals - Focus on characterization of performance, vice testing to specific requirements
 - Mission oriented metrics - Not rigidly adhering to requirements documents and using continuous metrics when possible
 - Not limiting factors to those in requirements documents
 - Avoiding single hypothesis tests
 - Considering all factors and Avoid confounding factors
- Best Practices for Assessing Statistical Adequacy memo emphasizes:**
 - Clearly identifying a test goal
 - Linking the design strategy to the test goal
 - Assessing the adequacy of the design in the context of the overarching goal
 - Re-emphasizes the importance of statistical power when used correctly.
 - Highlights quantitative measures of statistical test adequacy (power, correlation, prediction variance)



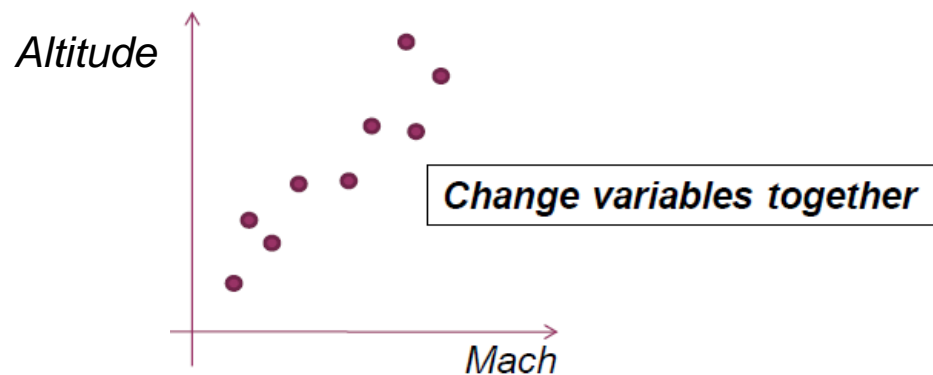
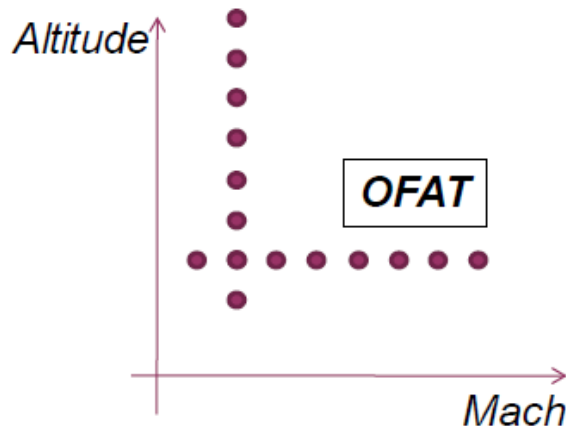
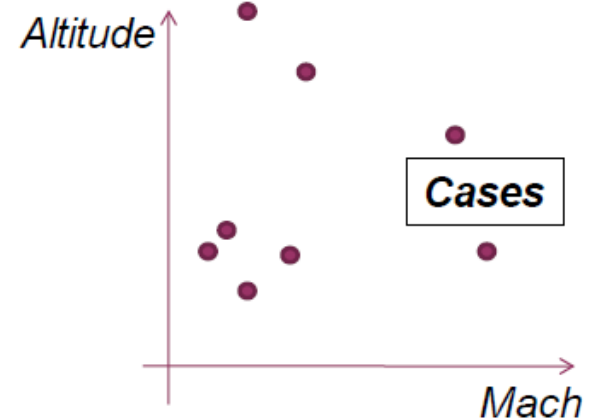
- The purpose of testing is to provide relevant, credible evidence with some degree of inferential weight to decision makers about the operational benefits of buying a system
 - DOE provides a framework for the argument and methods to help us do that systematically
- Statistical thinking/DOE provide:
 - a scientific, structured, objective test methodology answering the key questions of test:
 - How many points?
 - Which points?
 - In what order?
 - How to analyze?



DOE changes "I think" to "I know"

- **Types of data collection**

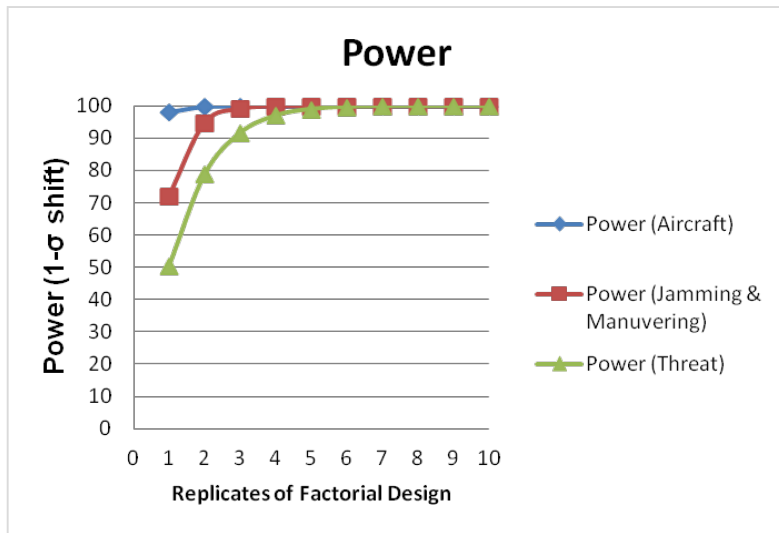
- DWWDLT – “Do what we did last time”
- Special/critical cases
- One-Factor-At-A-Time (OFAT)
- Historical data – data mining
- Observational studies
- Design of experiments
 - » Purposeful changing of test conditions



All tests are designed, many poorly!

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 that’s what got approved last time
 - 10 – because that sounds like enough
 - 30 – because something good happens at 30!
- DOE methods provide the tools to calculate



Loosely speaking:

“Plot of Likelihood of Finding Problems vs N ”

Or

“Plot of Likelihood of Seeing a Performance Degrade in Certain Conditions vs. N ”

Analytical trade space for test planning – balancing risk and resources



2. Which Points in a 12-D Battlespace?

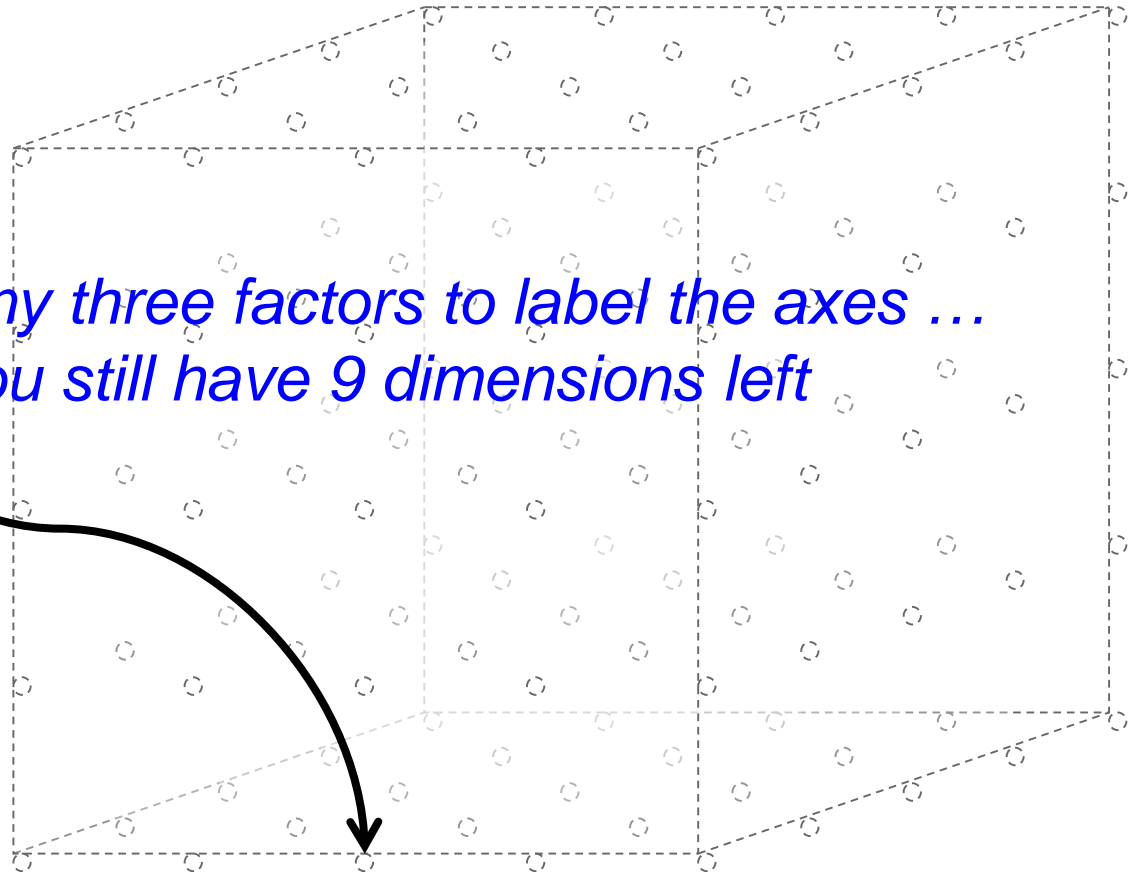
Test Condition
Target Type:
Num Weapons
Target Angle on Nose
Release Altitude
Release Velocity
Release Heading
Target Downrange
Target Crossrange
Impact Azimuth (°)
Fuze Point
Fuze Delay
Impact Angle (°)

*Pick any three factors to label the axes ...
And you still have 9 dimensions left*

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

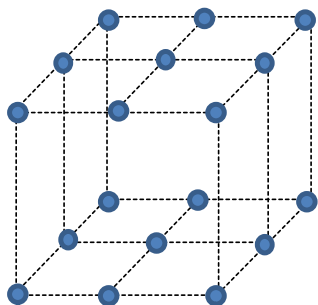
$$2^{12} = 4096$$

... lattice points!

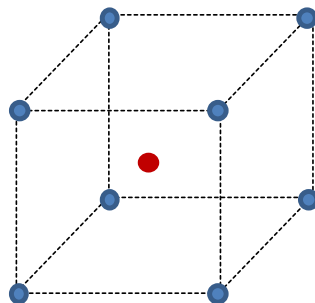


IDA A Structured Approach to Picking Test Points

(Tied to Test Objectives and Connected to the Anticipated Analysis!)

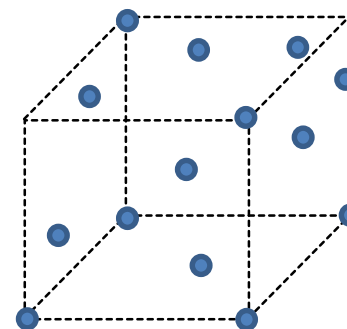


General Factorial
3x3x2 design

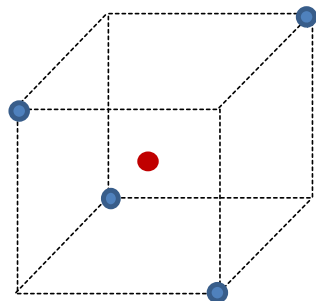


2-level Factorial
 2^3 design

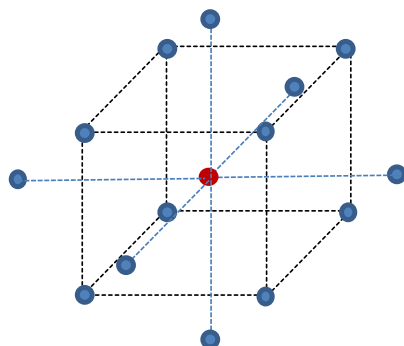
*“Just Enough”
test points:
– most efficient*



Optimal Design
IV-optimal



Fractional Factorial
 2^{3-1} design



Response Surface
Central Composite design



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

- Operational Envelope Defined – 128 possible cases
- Test team identified factors and their interactions and refined them to identify the most important aspects of the test design

Background Complexity	Threat	Formation Size	Location Confidence	Time of Day	Variant	Weapon
Black	Yellow	Red	Green	Green	Green	Yellow
Black	Red	Red	Green	Red	Red	Red
Black	Green	Green	Red	Green	Green	Red
Black	Green	Green	Red	Green	Green	Yellow
Black	Green	Green	Red	Green	Green	Red
Black	Green	Green	Red	Green	Green	Yellow

Green	No significant interaction expected
Yellow	Significant interaction in one response
Red	Significant interaction in multiple responses

			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																

- 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 *important* factor interactions will be estimable

- **Note the significant reduction to 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											

- **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?

4. What Conclusions? (Traditional Analysis)

- Cases or scenario settings and findings

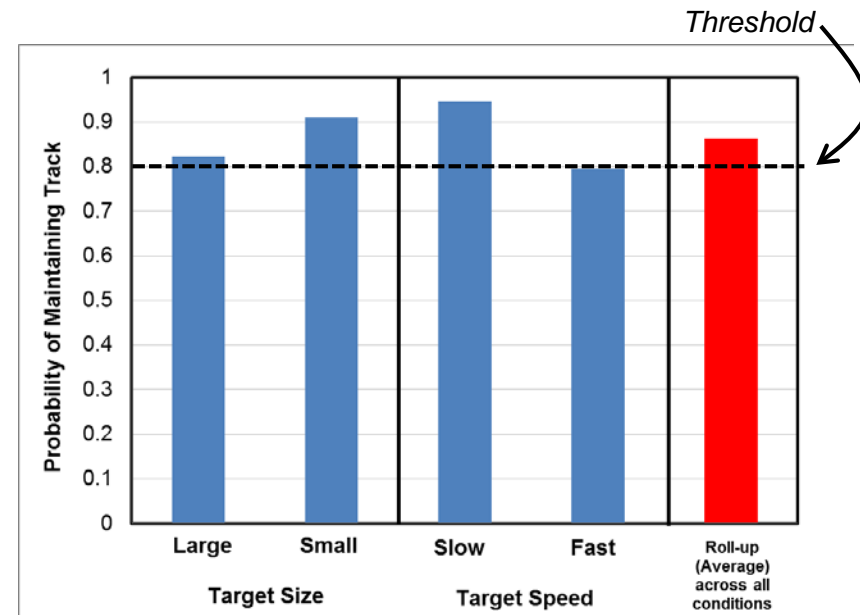
Sortie	Alt	Mach	MDS	Range	Tgt Aspect	OBA	Tgt Velocity	Target Type	Result
1	10K	0.7	F-16	4	0	0	0	truck	Hit
1	10K	0.9	F-16	7	180	0	0	bldg	Hit
2	20K	1.1	F-15	3	180	0	10	tank	Miss

- Run summaries

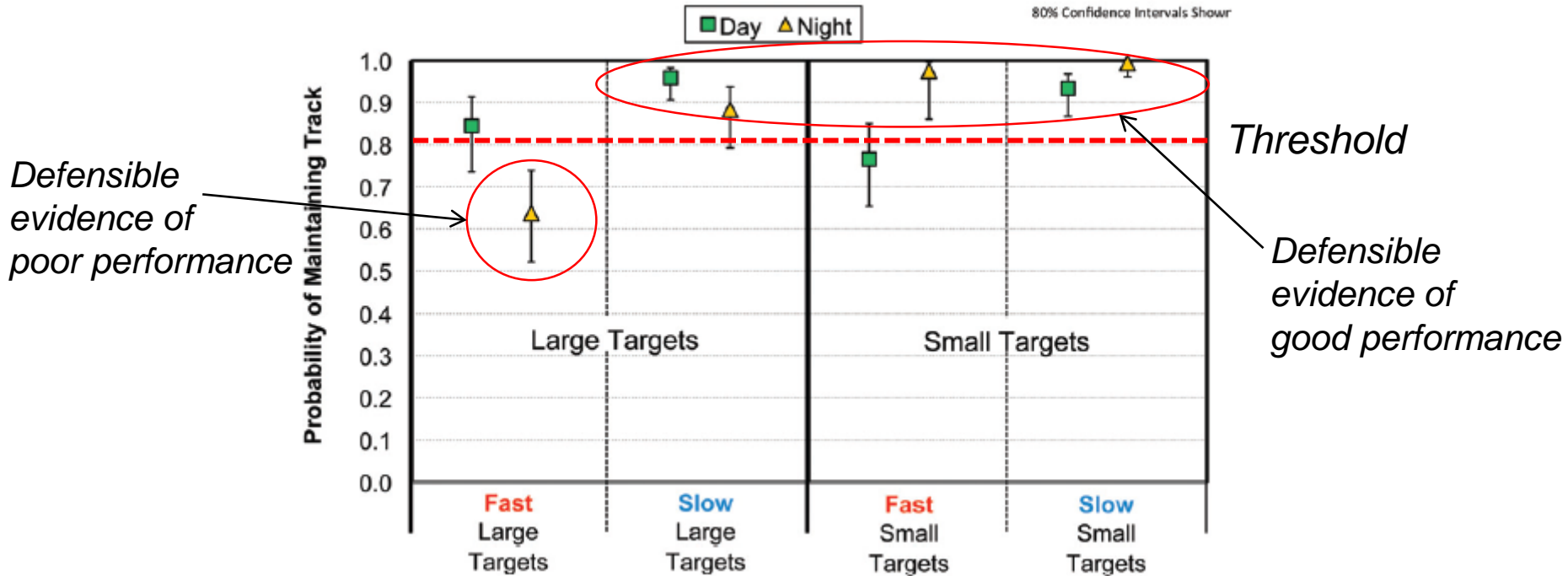
- Subject to removing “anomalies” if they don’t support expected trend
- No link to cause and effect

- Report **average performance** in common conditions or global average alone

- Compare point estimate to threshold
- No estimate of precision/uncertainty



4. What Conclusions? (DOE Analysis)



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

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

Case Study

System Description

- Sonar system replica in a laboratory on which hydrophone-level data, recorded during real-world interactions can be played back in real-time.
- System can process the raw hydrophone-level data with any desired version of the sonar software.
- Upgrade every two years; test to determine new version is better
- Advanced Processor Build (APB) 2011 contains a potential advancement over APB 2009 (new detection method capability)



- **Response Variable: Detection Time**
 - Time from first appearance in recordings until operator detection
 - » Failed operator detections resulted in *right censored data*
- **Factors:**
 - Operator proficiency (quantified score based on experience, time since last deployment, etc.)
 - Submarine Type (SSN, SSK)
 - System Software Version (APB 2009, APB 2011)
 - Array Type (A, B)
 - Target Loudness (Quiet, Loud)

		SSK		SSN	
		Quiet	Loud	Quiet	Loud
APB-11	Array A	12	12	6	12
	Array B	6	6	6	6
APB-09	Array A	8	8	4	8
	Array B	4	4	4	4

- **A full-factorial design across the controllable factors provided coverage of the operational space**
- **Replication was used strategically:**
 - Allowed for characterization across different operator skill levels (randomly assigned)
 - Provided the ability to support multiple test objectives
 - Skewed to the current version of the system under evaluation (APB-11)
- **Power analysis was used to determine an adequate test**
 - Power was 89% detecting a 1σ difference between APB versions – primary goal of the test
 - Power was > 99% for all other factor differences
 - Power was lower for APB due to blocking by day

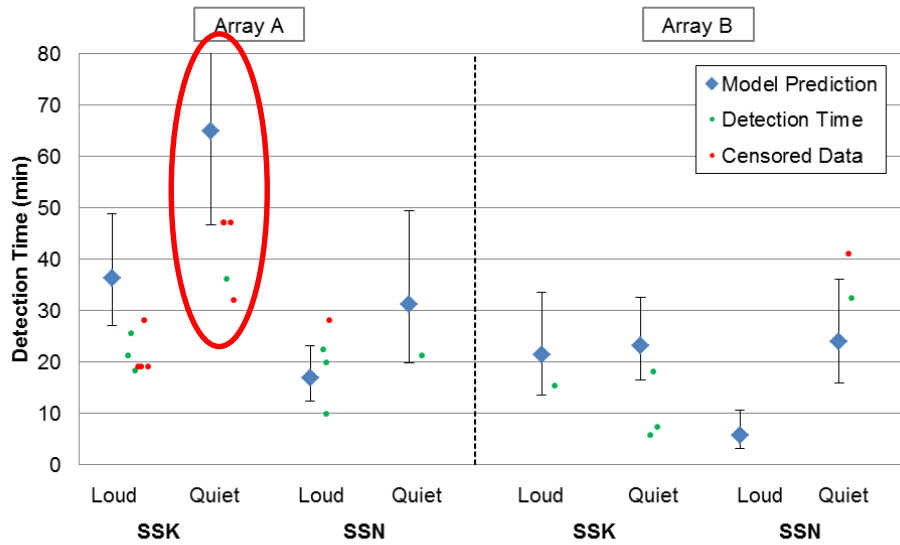
		SSK		SSN	
		Quiet	Loud	Quiet	Loud
APB-11	Array A	16	18	5	14
	Array B	10	5	6	3
APB-09	Array A	5	7	1	4
	Array B	3	1	2	0

- **Execution did not match the planned test design**
- **Test team used the DOE matrix at the end of the first round of testing to determine the most important points to collect next**
 - Real time statistical analyses revealed that there was only limited utility in executing the remainder of the planned test
 - Analysis revealed that there was a significant difference in APB versions
 - Additionally all other factors considered were statistically significant due to larger effects than anticipated

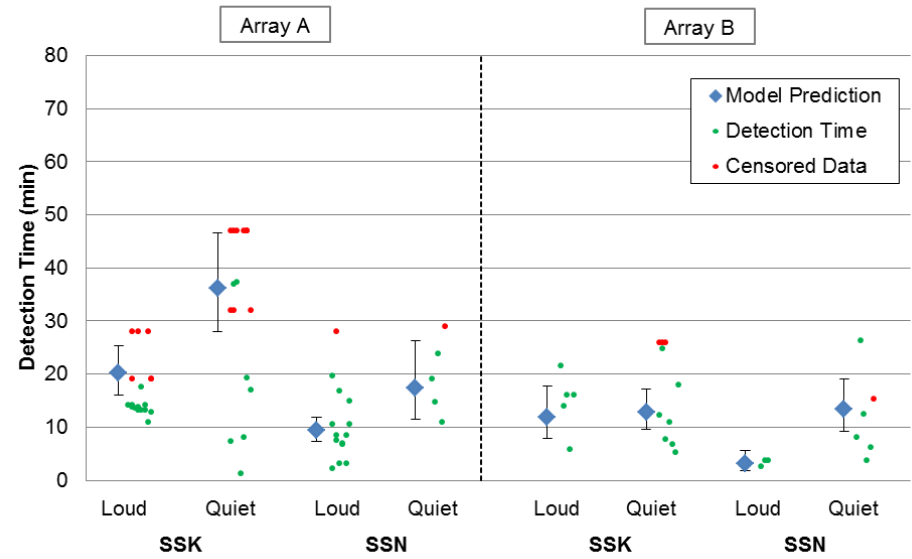
- **Advanced statistical modeling techniques incorporated all of the information across the operational space.**
 - Generalized linear model with log-normal detection times
 - Censored data analysis accounts for non-detects
- **All factors were significant predictors of the detection time**

Factor/Model Term	Description of Effect	P-Value
Recognition Factor	Increased recognition factors resulted in shortened detection times	0.0227
APB	Detection time is shorter for APB-11	0.0025
Target Type	Detection time is shorter for SSN targets	0.0004
Target Noise Level	Detection time is shorter for loud targets	0.0012
Array Type	Detection time is shorter for Array B	0.0006
Type* Noise		0.0628
Type* Array	Additional model terms improve predictions. Third order interaction is marginally significant, therefore all second order terms are retained.	0.9091
Noise*Array		0.8292
Type* Noise*Array		0.0675

APB-09 Median Detection Times



APB-11 Median Detection Times



- **Median detection times show a clear advantage of APB-11 over the legacy APB**
- **Confidence interval widths reflect weighting of data towards APB-11**
- **Statistical model provides insights in areas with limited data**

- **DOE provides us the *Science of Test***
 - We understand sys-engineering, guidance, aero, mechanics, materials, physics, electromagnetics, ...
- **Design of Experiments (DOE) – a structured and purposeful approach to test planning**
 - Ensures adequate coverage of the operational test space
 - Determines how much testing is enough
 - Quantifies test risks
 - Results:
 - » More information from constrained resources
 - » An analytical trade-space for test planning
- **Statistical analysis methods**
 - Do more with the data you have
 - Incorporate all relevant information in evaluations
 - » Supports integrated testing
- **DOT&E Guidance Memos**
 - Guidance on Design of Experiments
 - Flawed Application of DOE to OT&E
 - Assessing Statistical Adequacy of Experimental Designs in OT&E

- **DOT&E Test Science Roadmap** – published June 2013
- **DDT&E Scientific Test and Analysis Techniques (STAT) Implementation Plan**
- **Scientific Test and Analysis Techniques (STAT) Center of Excellence** provides support to programs
- **Research Consortium**
 - Navel Post Graduate School, Air Force Institute for Technology, Arizona State University, Virginia Tech
 - Research areas:
 - » Case studies applying experimental design in T&E.
 - » Experimental Design methods that account for T&E challenges.
 - » Improved reliability analysis.
- **Current Training and Education Opportunities**
 - DOT&E AO Training: Design, Analysis, and Survey Design
 - Air Force sponsored short courses on DOE
 - Army sponsored short courses on reliability
 - AFIT T&E Certificate Program
- **Policy & guidance**
 - DOT&E Guidance Memos
 - DOD 5000
 - Defense Acquisition Guidebook

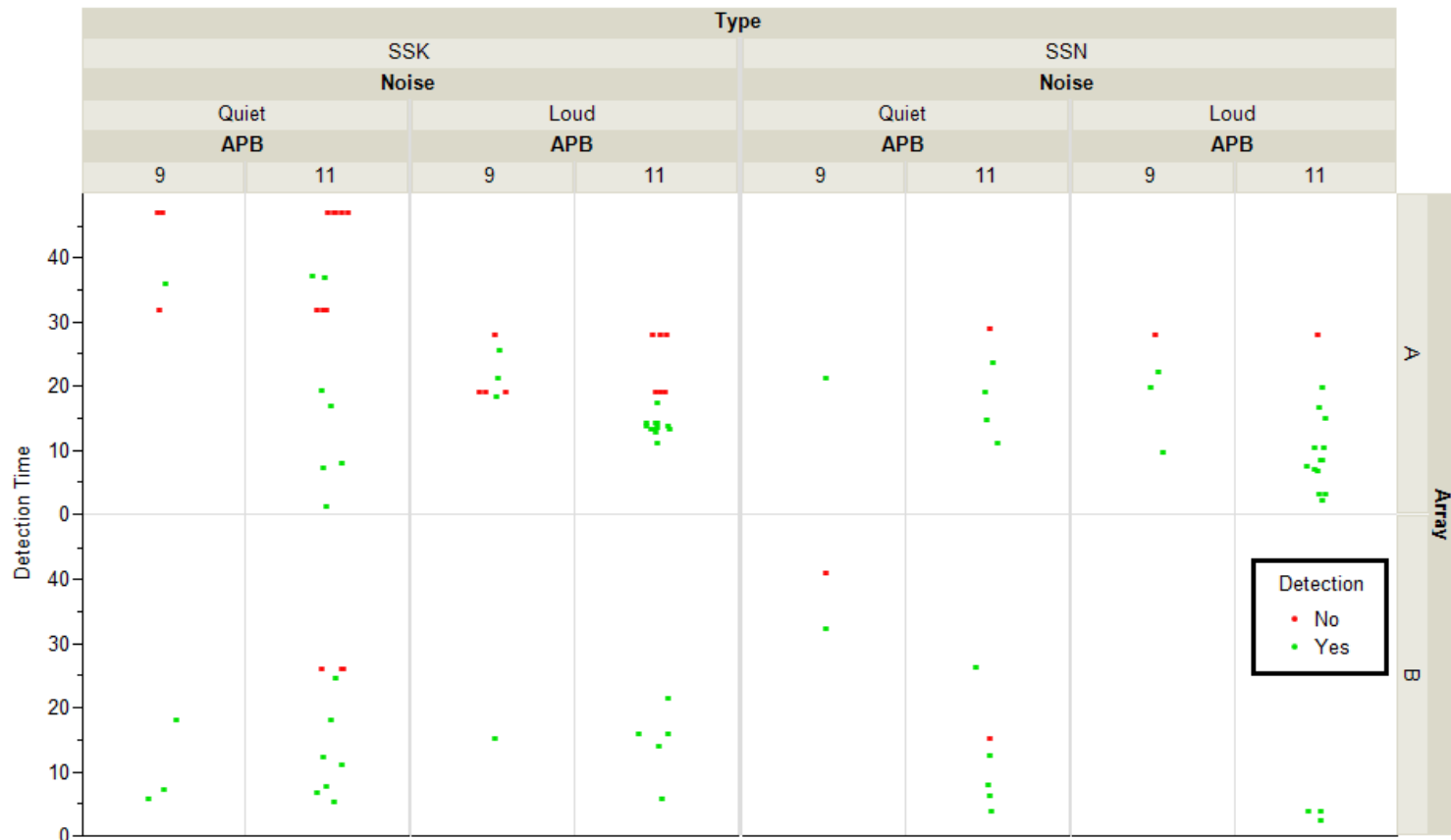
Backups

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
- **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



- A closer look at the data



- Power = Prob(Detect problem if problem exists)
- Power and confidence are only meaningful in the context of a hypothesis test! Example:

H_0 : Detonation slant range is the same with and without degaussing

H_1 : Detonation slant range differs when degaussing is employed

$$H_0: \mu_D = \mu_{ND}$$

$$H_1: \mu_D \neq \mu_{ND}$$

- Power is the probability that we conclude that the degaussing system makes a difference when it truly does have an effect.
- Similarly, power can be calculated for any other factor or model term

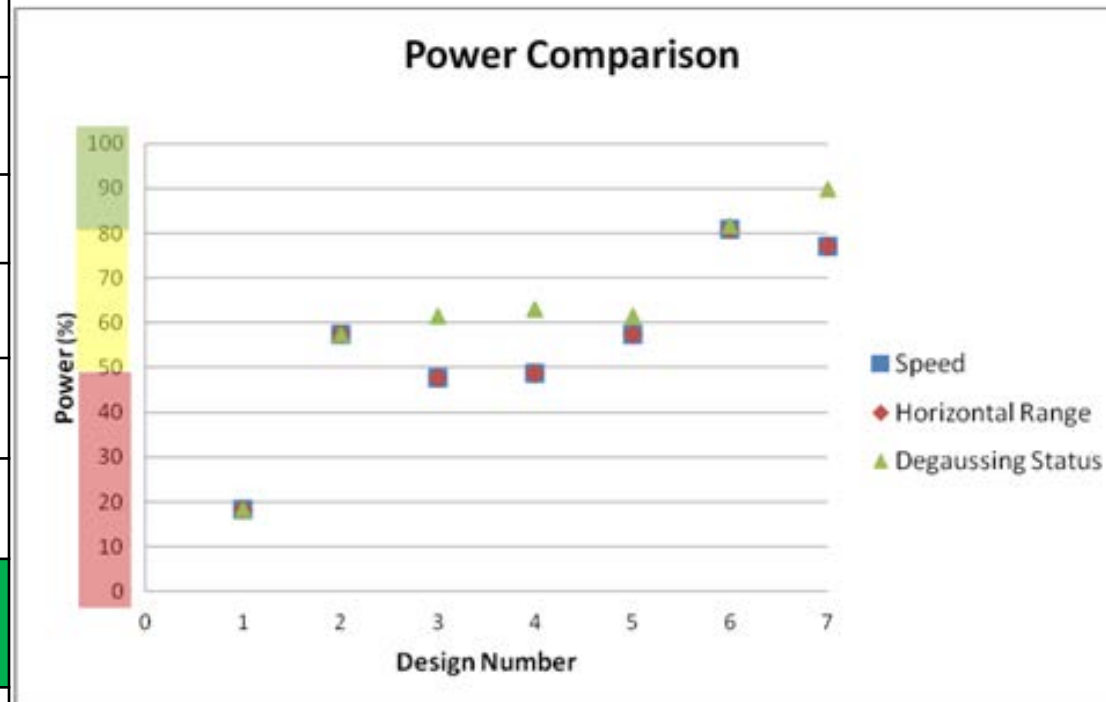
Test Decision	Accept H_0	False Negative (β Risk)	Confidence ($1-\alpha$)
	Reject H_0	Power ($1-\beta$)	False Positive (α Risk)
		Difference	No Difference

Real World

We need to understand risk!

- Compared several statistical designs
 - Selected a replicated central composite design with 28 runs
 - Power calculations are for effects of one standard deviation at the 90% confidence level

	Design Type	Number of Runs
1	Full Factorial (2-level)	8
2	Full Factorial (2-level) replicated	16
3	General Factorial (3x3x2)	18
4	Central Composite Design	18
5	Central Composite Design (replicated center point)	20
6	Central composite Design with replicated factorial points (Large CCD)	28
7	Replicated General Factorial	36



- **Best Practices**
 - Continuous Metrics where
 - Power calculations consistent with test goal (rarely use single hypothesis test)
 - Power curves to show tradeoffs
 - Include all relevant factors (cast as continuous where possible!) in design
 - Test goals not limited to verifying requirements under limited set of conditions
 - Use of statistical measures of merit to judge designs
- **Areas to Emphasize/Improve Upon**
 - Analysis of data commensurate with DOE design
 - » Employ regression techniques (linear regression, logit for binomial)
 - » Include “recordable” variables as covariates
 - » Model terms included based on factors/levels varied
 - Model verification methods and model reduction methods
 - Employment of advanced methods
 - » Bayesian approaches to reliability (data from multiple test phases)
 - » Censored data analysis for continuous measures
 - » Regression models not limited to the normal-distribution assumption
 - » Regression models flexible to all effects in the data (e.g., variance terms)
 - Power calculations for more advanced model approaches
 - Survey Design and Use