Statistical Engineering Leadership Forum

The Honorable J. Michael Gilmore
Director Operational Test and Evaluation

December 9, 2014

Abstract available at https://engstat.ida.org

Statistical Engineering Partnership

- **DoD and NASA develop, acquire, evaluate, and operate some of the world's most advanced and sophisticated systems**
- **Infusing statistical engineering is promoting statistical thinking and the effective application of statistical methods**
  - **Generating better solutions to solve high-impact challenges**
- **To advance and broaden statistical engineering practice, a partnership was developed to share knowledge and experience**
- **In Feb. 2014, an interagency agreement between the Director, Operational Test & Evaluation (DOT&E) in the Office of the Secretary of Defense and the Chief Engineer of NASA was ratified**
- **Partnership is envisioned as catalyst to engage the broader statistical engineering community in industry, academia, and other government agencies.**

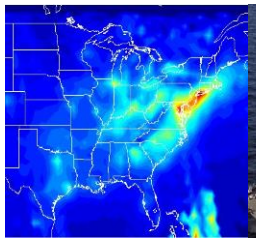
IDA

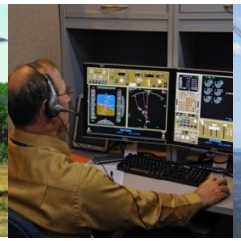
Statistical Engineering Partnership Operational Elements

Strategic framework to share leadership perspectives, foster practitioner collaboration, propagate training, and highlight accomplishments

Key Elements

- **Leadership Forum**
- **Tactical Site Visits**
- **Statistical Engineering Summit Meeting**
- **Collaboration on Documents/Handbooks**
- **Annual Report**





Rigorous, Defensible, Efficient Testing

Dr. J Michael Gilmore
Director, Operational Test and Evaluation (DOT&E)

Statistical Engineering Leadership Forum

9 December 2014

Overview

- **Introduction to DOT&E**
- **Science of Test**
  - The role of statistics in ensuring defense system testing is rigorous, defensible, and efficient
- **DOT&E Guidance on Design of Experiments (DOE)**
- **Better Test Planning Using DOE**
- **Benefits of Statistical Analyses**

Introduction to DOT&E

- **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?

What is the Science of Test?

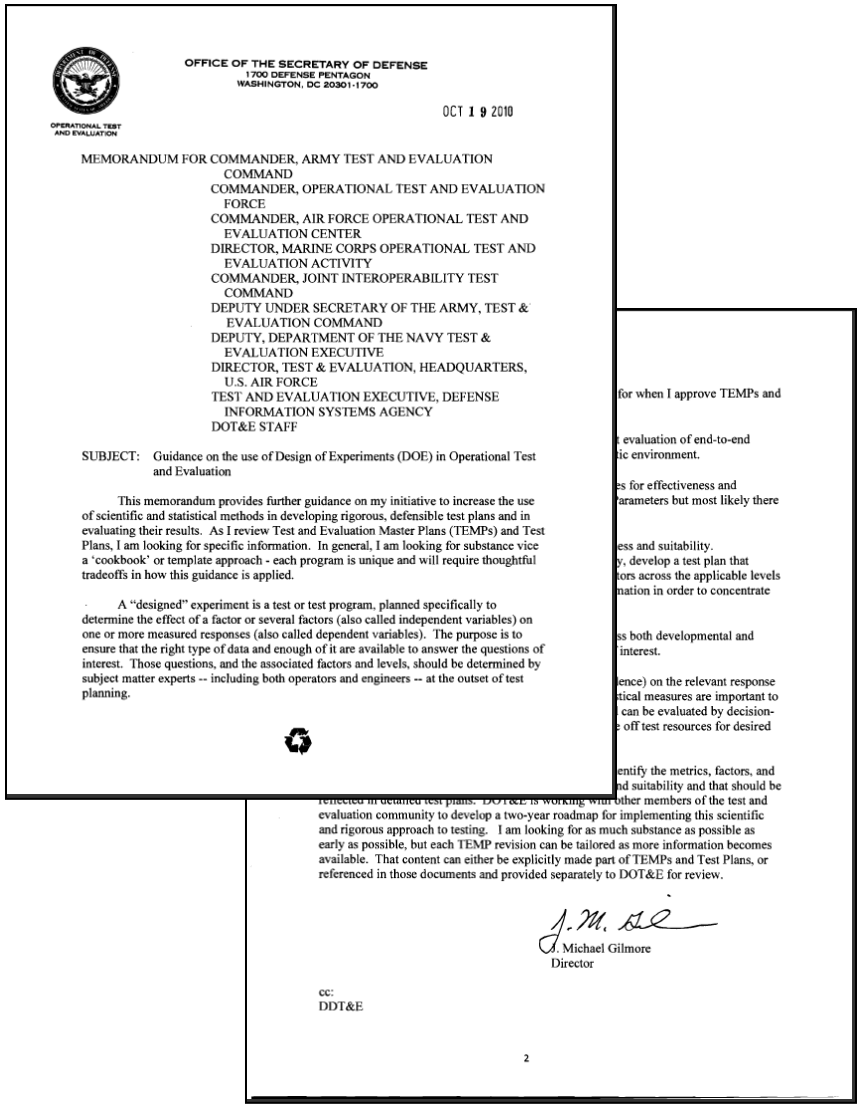
- **Test Planning**
  - Design of Experiments (DOE) – a structured and purposeful approach to test planning
    - » Ensures adequate coverage of the operational space
    - » Determines how much testing is enough – statistical power analysis
    - » Provides an analytical basis for assessing test adequacy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

The Science of Test enables Rigorous, Defensible, Efficient Testing

DOT&E Guidance

Design of Experiments in Operational Testing



OFFICE OF THE SECRETARY OF DEFENSE
1700 DEFENSE PENTAGON
WASHINGTON, DC 20301-1700

OPERATIONAL TEST AND EVALUATION

OCT 19 2010

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

SUBJECT: Guidance on the use of Design of Experiments (DOE) in Operational Test and Evaluation

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.

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.

for when I approve TEMPs and

evaluation of end-to-end
ic environment.

es for effectiveness and
arameters but most likely there

ess and suitability.
y, develop a test plan that
tors across the applicable levels
nation in order to concentrate

ss both developmental and
interest.

ence) on the relevant response
tical measures are important to
can be evaluated by decision-
e off test resources for desired

entify the metrics, factors, and
nd suitability and that should be
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.

J. Michael Gilmore
Director

cc:
DDT&E

2

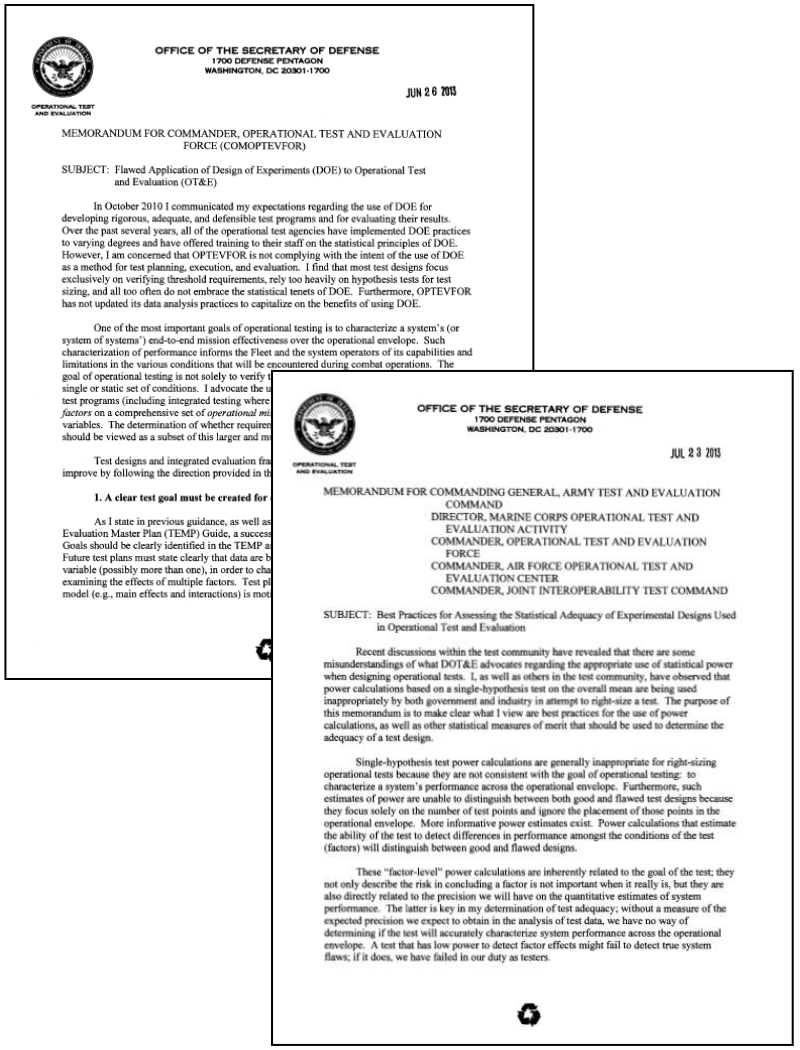
- **The goal of the experiment**. This should reflect evaluation of end-to-end mission effectiveness in an operationally realistic environment.
- Quantitative mission-oriented **response variables** for effectiveness and suitability. (These could be Key Performance Parameters but most likely there will 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.

http://www.dote.osd.mil/guidance.html

Additional DOT&E Guidance on DOE

- **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)



OFFICE OF THE SECRETARY OF DEFENSE
1700 DEFENSE PENTAGON
WASHINGTON, DC 20301-1700

JUN 26 2015

OPERATIONAL TEST AND EVALUATION

MEMORANDUM FOR COMMANDER, OPERATIONAL TEST AND EVALUATION FORCE (COMOPTEVFOR)

SUBJECT: Flawed Application of Design of Experiments (DOE) to Operational Test and Evaluation (OT&E)

In October 2010 I communicated my expectations regarding the use of DOE for developing rigorous, adequate, and defensible test programs and for evaluating their results. Over the past several years, all of the operational test agencies have implemented DOE practices to varying degrees and have offered training to their staff on the statistical principles of DOE. However, I am concerned that OPTEVFOR is not complying with the intent of the use of DOE as a method for test planning, execution, and evaluation. I find that most test designs focus exclusively on verifying threshold requirements, rely too heavily on hypothesis tests for test sizing, and all too often do not embrace the statistical tenets of DOE. Furthermore, OPTEVFOR has not updated its data analysis practices to capitalize on the benefits of using DOE.

One of the most important goals of operational testing is to characterize a system's (or system of systems') end-to-end mission effectiveness over the operational envelope. Such characterization of performance informs the Fleet and the system operators of its capabilities and limitations in the various conditions that will be encountered during combat operations. The goal of operational testing is not solely to verify
single or static set of conditions. I advocate the u
test programs (including integrated testing where
factors on a comprehensive set of *operational mi*
variables. The determination of whether requirem
should be viewed as a subset of this larger and m

Test designs and integrated evaluation fra
improve by following the direction provided in th

1. A clear test goal must be created for

As I state in previous guidance, as well as
Evaluation Master Plan (TEMP) Guide, a success
Goals should be clearly identified in the TEMP a
Future test plans must state clearly that data are b
variable (possibly more than one), in order to cha
examining the effects of multiple factors. Test pl
model (e.g., main effects and interactions) is mot

OFFICE OF THE SECRETARY OF DEFENSE
1700 DEFENSE PENTAGON
WASHINGTON, DC 20301-1700

JUL 23 2015

OPERATIONAL TEST AND EVALUATION

MEMORANDUM FOR COMMANDING GENERAL, ARMY TEST AND EVALUATION COMMAND
DIRECTOR, MARINE CORPS OPERATIONAL TEST AND EVALUATION ACTIVITY
COMMANDER, OPERATIONAL TEST AND EVALUATION FORCE
COMMANDER, AIR FORCE OPERATIONAL TEST AND EVALUATION CENTER
COMMANDER, JOINT INTEROPERABILITY TEST COMMAND

SUBJECT: Best Practices for Assessing the Statistical Adequacy of Experimental Designs Used in Operational Test and Evaluation

Recent discussions within the test community have revealed that there are some misunderstandings of what DOT&E advocates regarding the appropriate use of statistical power when designing operational tests. I, as well as others in the test community, have observed that power calculations based on a single-hypothesis test on the overall mean are being used inappropriately by both government and industry in attempt to right-size a test. The purpose of this memorandum is to make clear what I view are best practices for the use of power calculations, as well as other statistical measures of merit that should be used to determine the adequacy of a test design.

Single-hypothesis test power calculations are generally inappropriate for right-sizing operational tests because they are not consistent with the goal of operational testing: to characterize a system's performance across the operational envelope. Furthermore, such estimates of power are unable to distinguish between both good and flawed test designs because they focus solely on the number of test points and ignore the placement of those points in the operational envelope. More informative power estimates exist. Power calculations that estimate the ability of the test to detect differences in performance amongst the conditions of the test (factors) will distinguish between good and flawed designs.

These "factor-level" power calculations are inherently related to the goal of the test; they not only describe the risk in concluding a factor is not important when it really is, but they are also directly related to the precision we will have on the quantitative estimates of system performance. The latter is key in my determination of test adequacy; without a measure of the expected precision we expect to obtain in the analysis of test data, we have no way of determining if the test will accurately characterize system performance across the operational envelope. A test that has low power to detect factor effects might fail to detect true system flaws; if it does, we have failed in our duty as testers.

http://www.dote.osd.mil/guidance.html

Current Efforts to Institutionalize Statistical Rigor in T&E

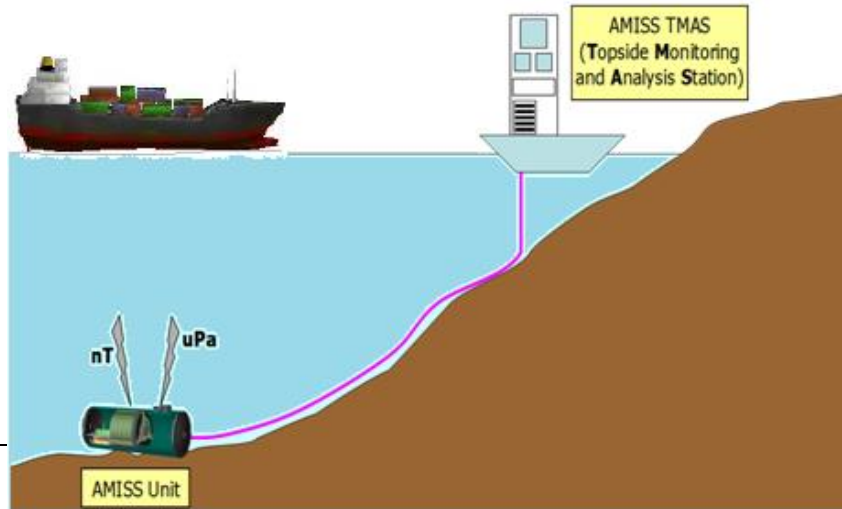
- **DOT&E/TRMC funded Research Consortium**
  - Air Force Institute for Technology, Arizona State University, Rochester Institute of Technology, Virginia Tech, University of Arkansas, North Carolina State
  - Graduate level research that focuses on DoD problems
    - » Potential pipeline of young researchers
    - » Expose academic world to DoD Test and Evaluation
- **DOT&E Test Science Roadmap – published June 2013**
  - Census of the OTAs, Training/Education opportunities
  - Best Practices and numerous Case Studies
- **DASD(DT&E) Scientific Test and Analysis Techniques (STAT) Implementation Plan**
  - Scientific Test and Analysis Techniques (STAT) Center of Excellence provides support to programs
- **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
  - DOT&E TEMP Guide
  - 5000.02 Updates
  - Defense Acquisition Guidebook

Better Test Planning with DOE

- **How much testing is enough?**
  - In test planning DOE allows us to determine how much testing is enough
  - Example: Mine susceptibility testing
- **DOE can provide an optimal allocation of test resources for spanning the operational envelope**
  - Example: Joint Strike Fighter
- **Smart application of DOE provides a better use of test resources**
  - Continuous metrics cut test resources or make limited resources stretch further
  - Example: Jammer example

How Much Testing is Enough?
Mine Susceptibility Testing

- **Goal:**
  - Develop an adequate test to assess the susceptibility of a cargo ship against a variety of mine types using the Advanced Mine Simulation System (AMISS).
- **Responses:**
  - Magnetic signature, acoustic signature, pressure
  - Slant range at simulated detonation
- **Factors:**
  - Speed, range, degaussing system status
- **Other considerations:**
  - Water depth
  - Ship direction



How much testing is enough? Power and Confidence

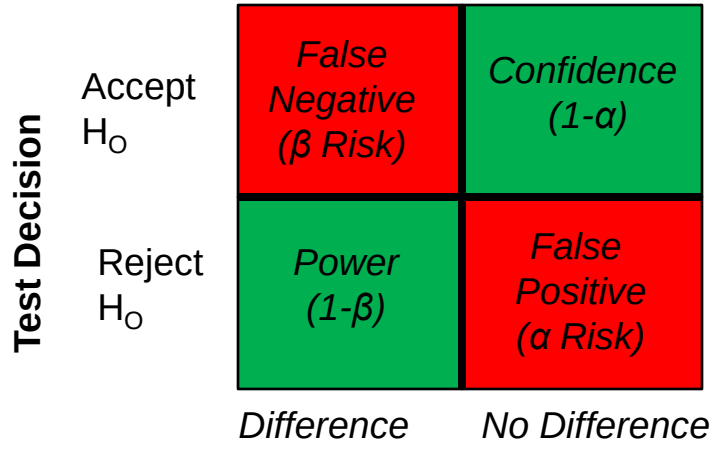
- **Power and confidence are only meaningful in the context of a hypothesis test!**
- **Statistical hypotheses:**

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**

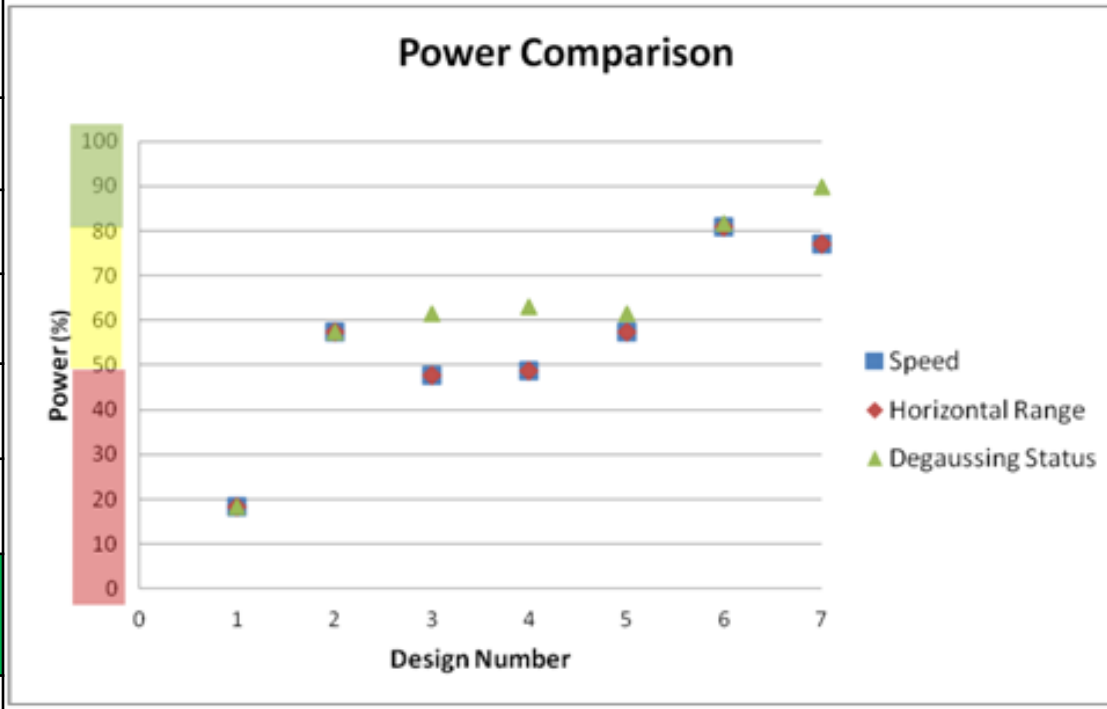


Power and confidence allow us to understand risk

Test Design Comparison: Statistical Power

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

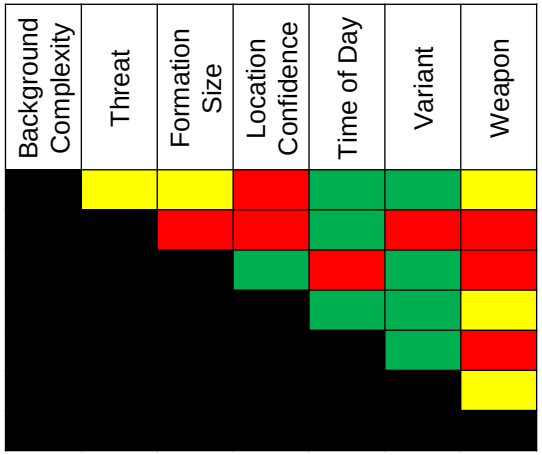
	Design Type	Number of Runs
1	Full Factorial (2-level)	8
2	Full Factorial (2-level) replicated	16
3	General Factorial (3x3x2)	18
4	Central Composite Design	18
5	Central Composite Design (replicated center point)	20
6	Central composite Design with replicated factorial points (Large CCD)	28
7	Replicated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



Statistical power provides an object measure of how much testing is enough

Optimal Allocation of Test Points: 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**



	No significant interaction expected
	Significant interaction in one response
	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																

JSF: 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 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											

JSF Air-to-Ground DOE Summary

- **Original 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?
- **Similar approaches to test planning were used for all JSF mission areas**
  - Close Air Support, Suppression of Enemy Air Defenses, Destruction of Enemy Air Defenses, Reconnaissance, etc.
  - Allowed objective comparison to legacy systems in important mission areas

Making the Most of Limited Test Resources:
Notional Jammer Test Design

- **Goals of the Test**
  - Characterize performance of a new jammer

- **Response variables**
  - Probability of hit (pass/fail requirement)
  - Miss distance of missile shots

- **Factors and levels (48 conditions)**
  - Aircraft variant: 2 variants (A, B)
  - Threat: 6 different type of threats (T1, T2, T3, T4, T5, T6)
  - Counter Measures: dry(no jamming), wet(jamming) non-maneuvering, or wet with one of three maneuvers

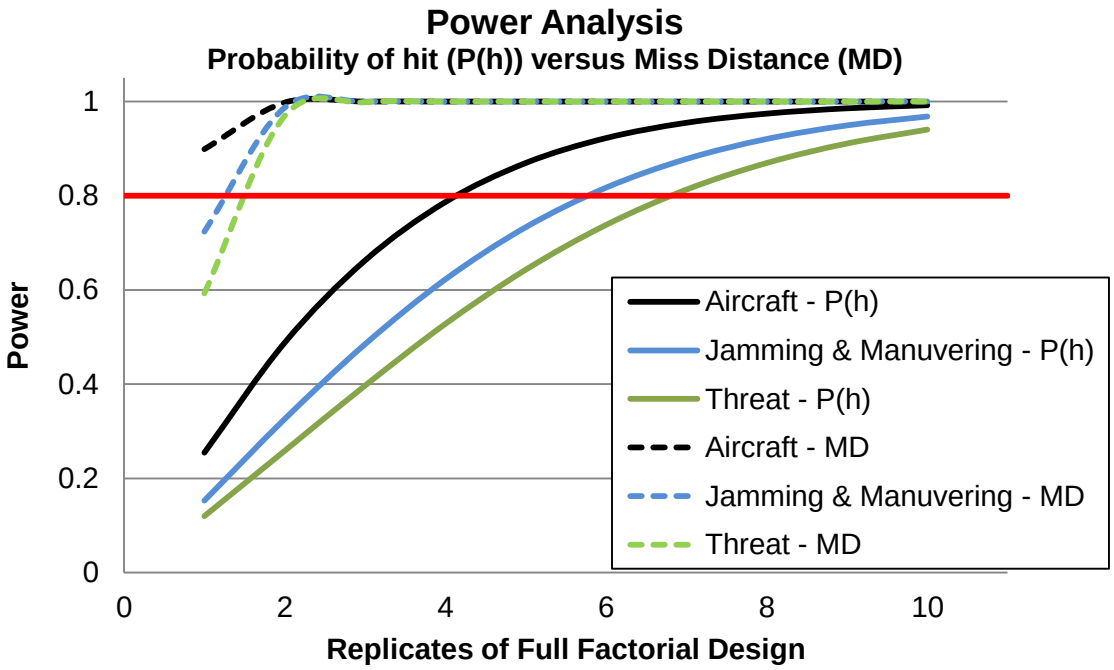
Jammer DOE Matrix

A/C	Threat	Dry Run	Wet Runs			A/C	Threat	Dry Run	Wet Runs		
		NM	M0	M1	M2			NM	M0	M1	M2
A	T1	?	?	?	?	B	T1	?	?	?	?
	T2	?	?	?	?		T2	?	?	?	?
	T3	?	?	?	?		T3	?	?	?	?
	T4	?	?	?	?		T4	?	?	?	?
	T5	?	?	?	?		T5	?	?	?	?
	T6	?	?	?	?		T6	?	?	?	?
Totals											

Making the Most of Limited Test Resources:
Notional Jammer Test Design

- **Original design proposal considers each condition separately**
  - Primary response variable is reduction in lethality
  - Result:
    - » 10 replicates in each condition, based on single hypothesis test in each condition
    - » A test that allows zero failures concludes that the probability of hit is at most 0.20 at the 90% confidence level

- **Statistical Design**
  - Capitalizes on relationships between conditions
  - Regression analysis of Miss Distance:
    - » 2 replicates in each condition
  - Logistic regression of Probability of hit:
    - » 6 replicated in each condition



Continuous measure and experimental design reduces test costs by more than half!

Benefits of Statistical Analyses for DoD T&E

- **Provides objective summary of test results**
  - Allows the data to speak for itself
    - » As opposed to traditional analyses that have focused on one-sample “roll-up” averages or selective binning of the data
  - Quantifies what conditions and groups are important
  - Allows for prediction of future performance
  - Example: Apache Link-16 Addition

- **Ensures that areas of poor performance are identified**
  - Statistical model selection
  - Example: Tracking of boat targets

- **Advanced methods provide more information than traditional analyses**
  - Censored data analysis coupled with cumulative distribution estimation
  - Example: Submarine detection time

- **Supports Integrated Testing**
  - Incorporate relevant information across multiple tests without biasing results
  - Example: Stryker reliability analysis

Objective Summary of Test Results: Apache Link-16 Addition

- **Test Goal: How does Link 16 contribute to the AH-64E crew's ability to find targets?**
- **Response Variable:**
  - Time to find initial target (minutes)
  - Measured once per mission
- **Factors:**
  - Battlefield density (low, high)
  - Link 16 targeting data (yes, no)
  - Light (day, night)
- **Test Design Executed**

	Battlefield Density	Low		High	
	Light	Day	Night	Day	Night
L16 Targeting Data	no	3	1	2	2
	yes	6	2	3	3

Cells indicate missions executed per condition

Battlefield density describes the amount of targets (enemy and friendly) in the objective area (>5 vehicles is high density).

L16 targeting data indicates whether an Air Force C2 platform was sending Link 16 surveillance tracks to the Apache Team during the mission.

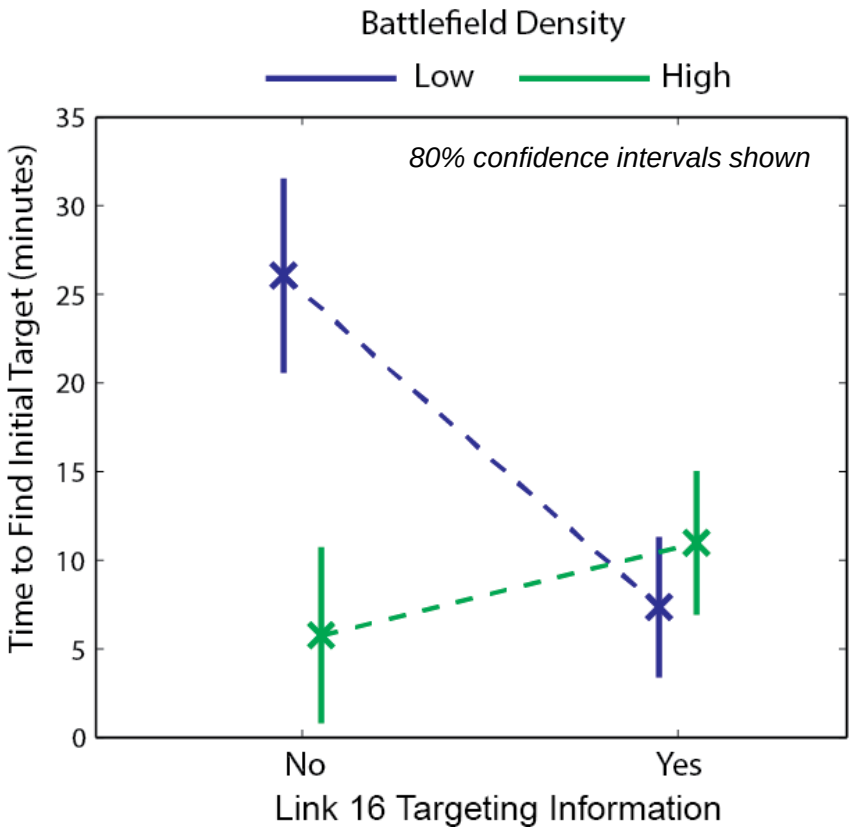
Apache Results

- **Two-Factor Interaction Regression Model**

Term	Estimate (Minutes)	Prob>|t|
Intercept	12.54	
L16 Targeting[no]	3.38	0.07
Battlefield Density[low]	4.17	0.03
Time[day]	0.96	0.60
L16 Targeting[no]*Battlefield Density[low]	5.98	0.00
L16 Targeting[no]*Time[day]	-0.36	0.84
Battlefield Density[low]*Time[day]	1.11	0.53

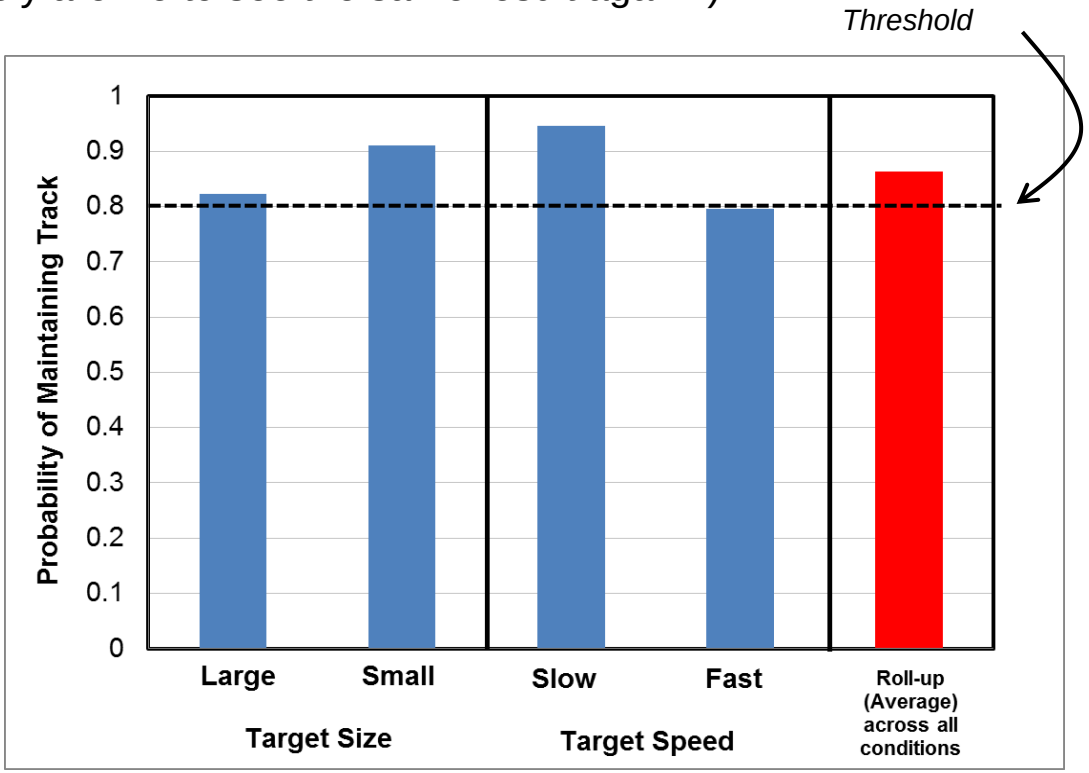
- **Statistical Result**
  - L16 targeting data, battlefield density were statistically significant; light was not.
  - Two factor interaction between BF density and L16 targeting data was significant
- **Bottom Line Result**
  - L16 targeting data significantly reduces Apache Team's ability to find targets
  - Effect is especially strong in a low density environment

- **Graph shows interaction between factors**



Ensures Areas of Poor Performance are Identified: Boat Tracking

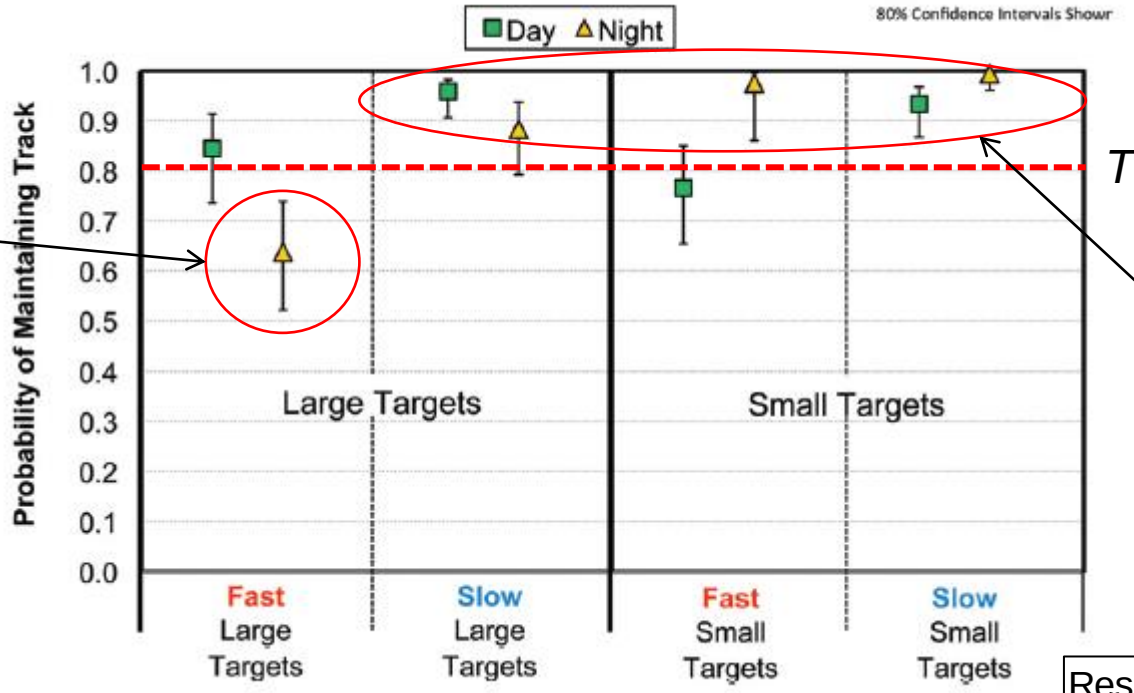
- **Traditional Analysis:**
  - Report *average performance* in common conditions or global average
  - Compare point estimate to threshold
  - No estimate of precision/uncertainty
    (how likely are we to see the same result again?)



Ensures Areas of Poor Performance are Identified: Boat Tracking

- **Regression Analysis**
  - Model selection revealed a significant 3rd order interaction
  - 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 sure am I that it failed threshold in Condition X?**



Defensible evidence of poor performance

Threshold

Defensible evidence of good performance

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

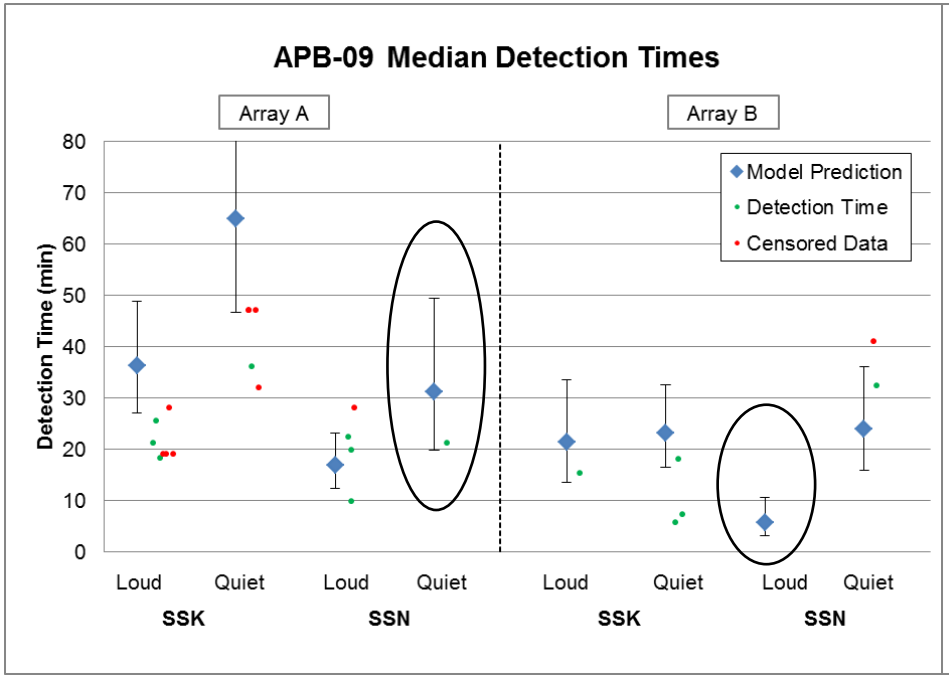
** Many model choices: regression, ANOVA, mixed models, Censored Data, Gen Linear Model, etc. etc.*

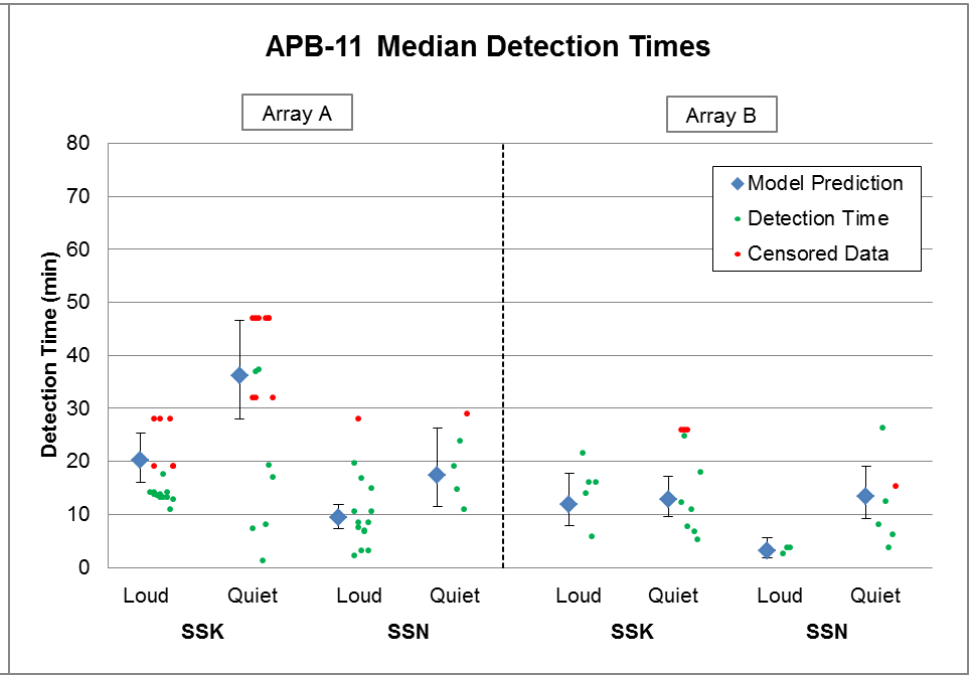
Advanced Methods Provide More Information: Submarine Detection Time

- **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)

- **Test 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)

Submarine Detection Time: Results





- **Right censored data analysis allowed for analysis of median detection times and probability of detection (often requirements are specified in probabilities)**
  - Median detection times show a clear advantage of APB-11 over the legacy APB – Result was statistically significant
- **Statistical model provides insights in areas with limited data**

Supports Integrated Testing: Stryker Reliability Analysis

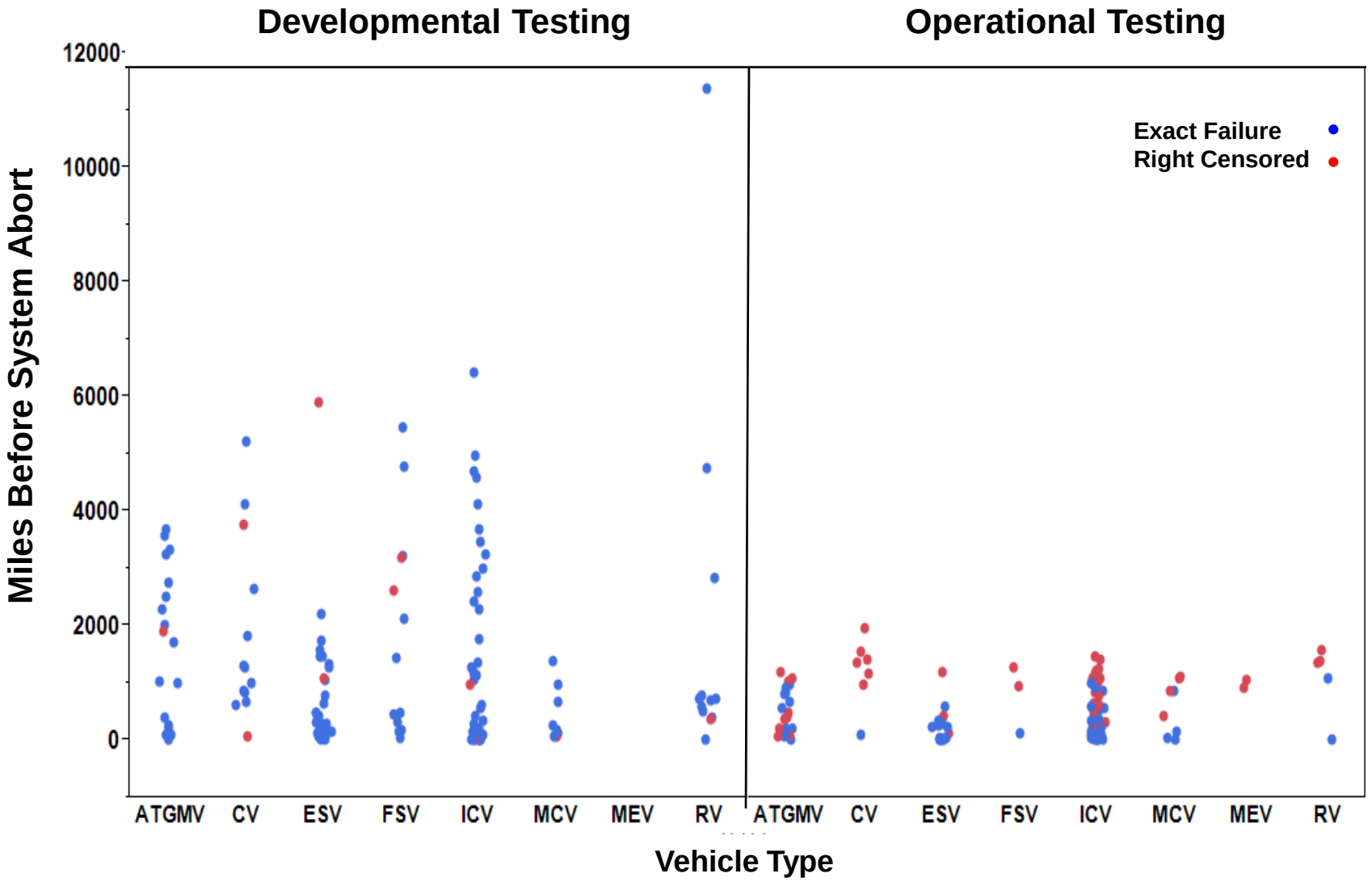
- **Statistical methods (including DOE) apply to reliability data as well as performance data**

- **Stryker Retrospective Case Study**
  - Infantry Carrier Vehicle (ICV) - the infantry/mission-vehicle type
  - Base vehicle for eight separate configurations
  - IOT&E Results:

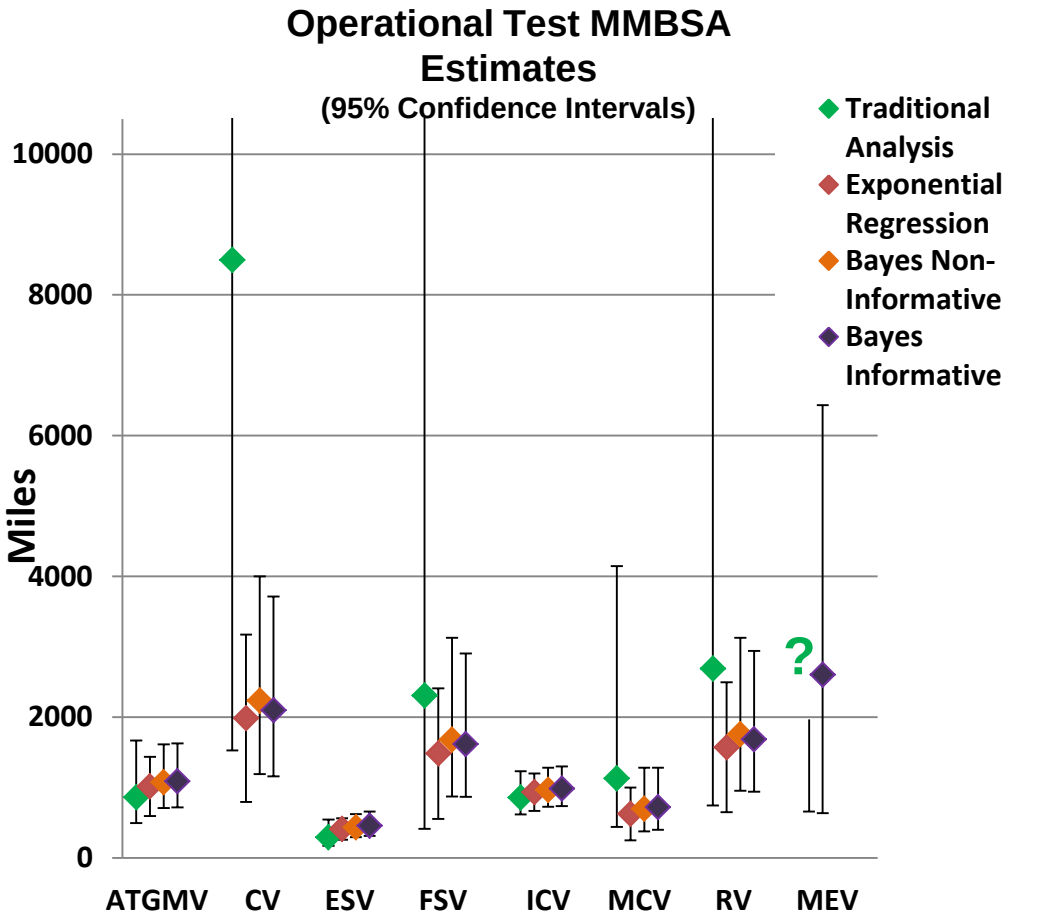
Stryker Reliability by Variant using Operational Test Data					
Vehicle Variant	**Total Miles Driven**	**System Aborts**	**MMBSA**	**MMBSA 95% LCL**	**MMBSA 95% UCL**
Antitank Guided Missile Vehicle (**ATGMV**)	10334	12	861	493	1667
Commander's Vehicle (**CV**)	8494	1	8494	1525	335495
Engineer Squad Vehicle (**ESV**)	3771	13	290	170	545
Fire Support Vehicle (**FSV**)	2306	1	2306	414	91082
Infantry Carrier Vehicle (**ICV**)	29982	35	857	616	1230
Mortar Carrier Vehicle (**MCV**)	4521	4	1130	441	4148
Medical Evacuation Vehicle (**MEV**)	1967	0	-	657	-
Reconnaissance Vehicle (**RV**)	5374	2	2687	744	22187
Total	**66749**	**68**	**982**	774	1264

- **Results do not leverage DT data or relationships between vehicles**

The Stryker Reliability Data Set



Stryker Reliability Results



- **Traditional Approach:**

$$MMBSA = \frac{Miles}{\#\,Failures}$$

  - Extremely wide confidence intervals
  - Results in unrealistic estimates for the Commander's Vehicle

- **Exponential Regression Approach & Bayesian Approaches**

$$MMBSA = f\left(TestPhase \quad Variant\right)$$

  - Allows for a degradation in MMBSA from DT to OT (increases could occur as well).
  - Leverages all information
    - » Better estimates of MMBSA
    - » Tighter confidence intervals

Statistical models support using all data in evaluations

Conclusions

- **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**
  - Ensure objectivity in the data analysis
  - Do more with the data you have
  - Incorporate all relevant information in evaluations

- **Areas for improvement**
  - Incorporating these methods earlier in defense testing
  - Testing sequentially
  - Improving access to education and training for the T&E workforce
  - Improving access to statistics experts across the T&E workforce
  - Use of statistical analysis techniques
  - Sharing case studies more widely

- **Statistical Engineering Agreement supports all of these areas for improvement!**

Backups

What is the Envisioned "End State"?

(What does the institutionalization of these methods look like?)

- **GOAL: DoD (including both government and industry) is the premier test organization, leading the way in efficient and effective test design, data analysis, and objective decision making that relies on robust test conclusions**
- **A workforce that is well-trained and competently employs these methods as their standard operating procedure**
  - Note that ORSAs will need better education/training and the support from the military community (i.e. better career trajectories)
- **Access to practitioners – all test organizations, program offices, and test teams (gov't and industry) have at least one statistics/data analysis/test design SME, and have access to PhD level statisticians to tackle harder problems**
  - STAT COE is a potential model to build upon; each Service/SYSCOM needs built in capability with a community hub to go to for advice, tools, etc.
- **Community of excellence must expand – practitioners get together discuss best practices, lessons learned, formally, informally and at conferences**

What is the Envisioned "End State"? (cont.)

(What does the institutionalization of these methods look like?)

- **Elimination of the mindset that DOE/STAT is just a part of the test program (e.g., the DOE portion of the test plan versus the other tests that we need to get done), and elimination of the mindset that DOE is just a DOT&E requirement**
  - These methods need to become the **cornerstone** of our business and the methods that drive the entirety of the test program and test analysis
- **Eliminate Test Designs that unnecessarily require thousands of test points, and eliminate test designs limited to demos, selected cases, etc.**
- **Test Designs that balance control, operational realism, and the need to make defensible conclusions in OT**
- **We learn from DT and use that information to size OT (sequential test design, realistic estimates of the SNR)**
- **We employ statistical analysis methods commensurate with the test design, and developers rely on these analyses to determine future development needs, correction of deficiencies, etc.**
- **We employ advanced techniques without sales pitches or misuse (censored data, Bayesian statistics)**

Rationale for DOE in Test & Evaluation

- **The purpose of testing is to provide relevant, *credible evidence* of system performance with some degree of inferential weight to decision makers**
  - 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 Test*
  2. *Which Points? Breadth of Testing* – spanning the operational envelope
  3. *How to Execute? Order of Testing*
  4. *What Conclusions? Data Analysis* – drawing objective, robust conclusions while accounting for noise

- **DOE Provides:**
  - the most powerful allocation of test resources for a given number of tests.
  - a scientific, structured, objective way to plan tests.
  - an approach to integrated test.
  - a structured, mathematical analysis for summarizing test results and discovering performance problems and proving system capability.
  - enables defensible conclusions about system performance.

DOE changes “I think” to “I know”

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)
  - Historical data – data mining
  - Observational studies
  - Design of experiments
    - » Purposeful changing of test conditions
- **Design of Experiments is preferable**
  - Challenges with Case-based
    - » Little predictive ability; loss of ability to determine cause & 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 historical data
    - » Missing data
    - » Data often collected for a different purpose
    - » Unclear documentation of test conditions
  - Challenges with observational studies
    - » Confounding data
    - » Loss of ability to determine cause & effect
    - » However, sometimes necessary for safety and/or ethical reasons

Cases

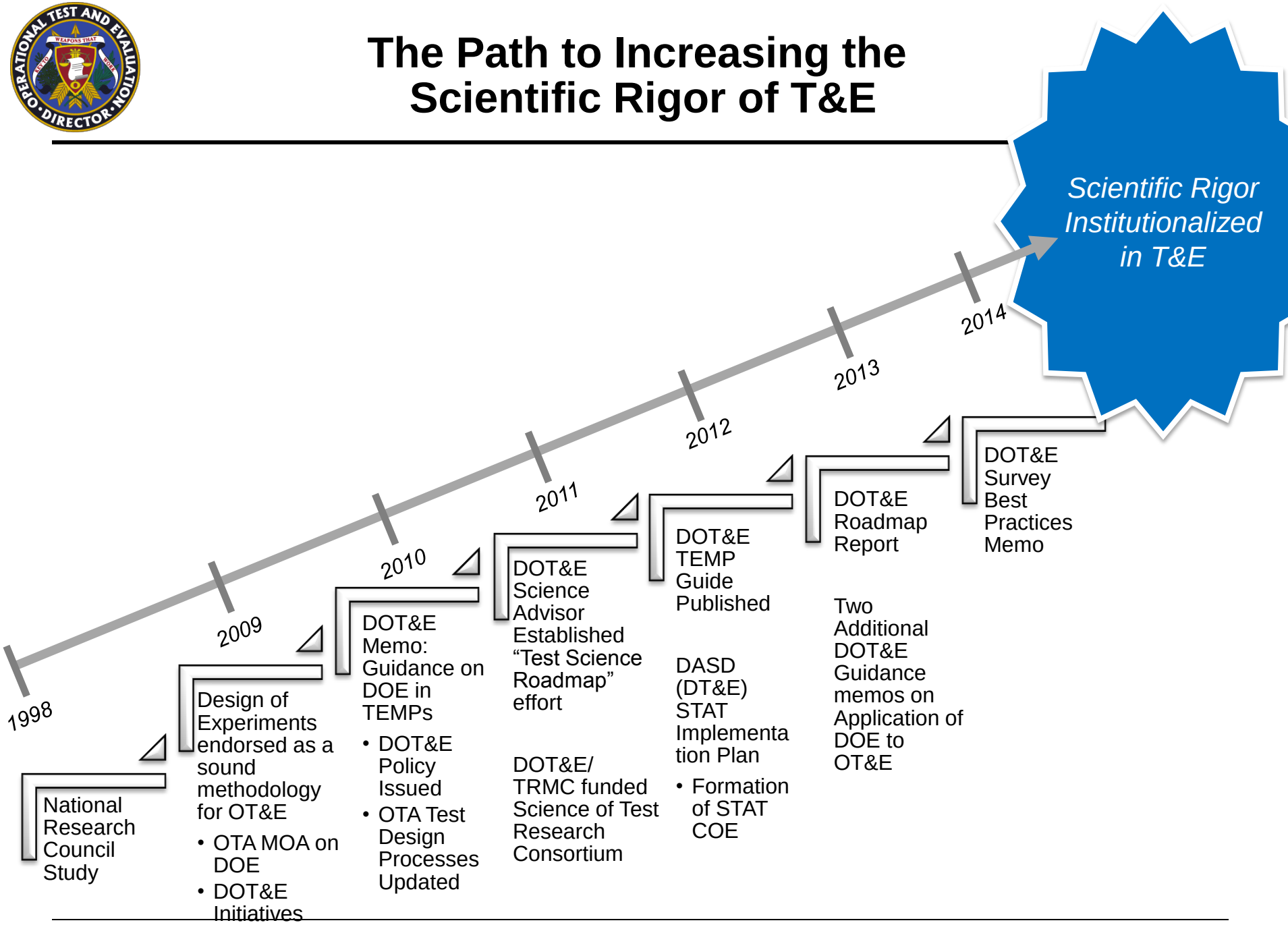
		With UAS		Without UAS	
	Mission Type	Day	**Night**	Day	**Night**
AB2	Recon	1		2	2
	Attack			1	
AB3	Recon		1	4	2
	Attack		3	2	2

OFAT

		With UAS		Without UAS	
	Mission Type	Day	**Night**	Day	**Night**
AB2	Recon			2	2
	Attack	2	2		
AB3	Recon			2	2
	Attack	2	2		

"All tests are designed, many poorly"

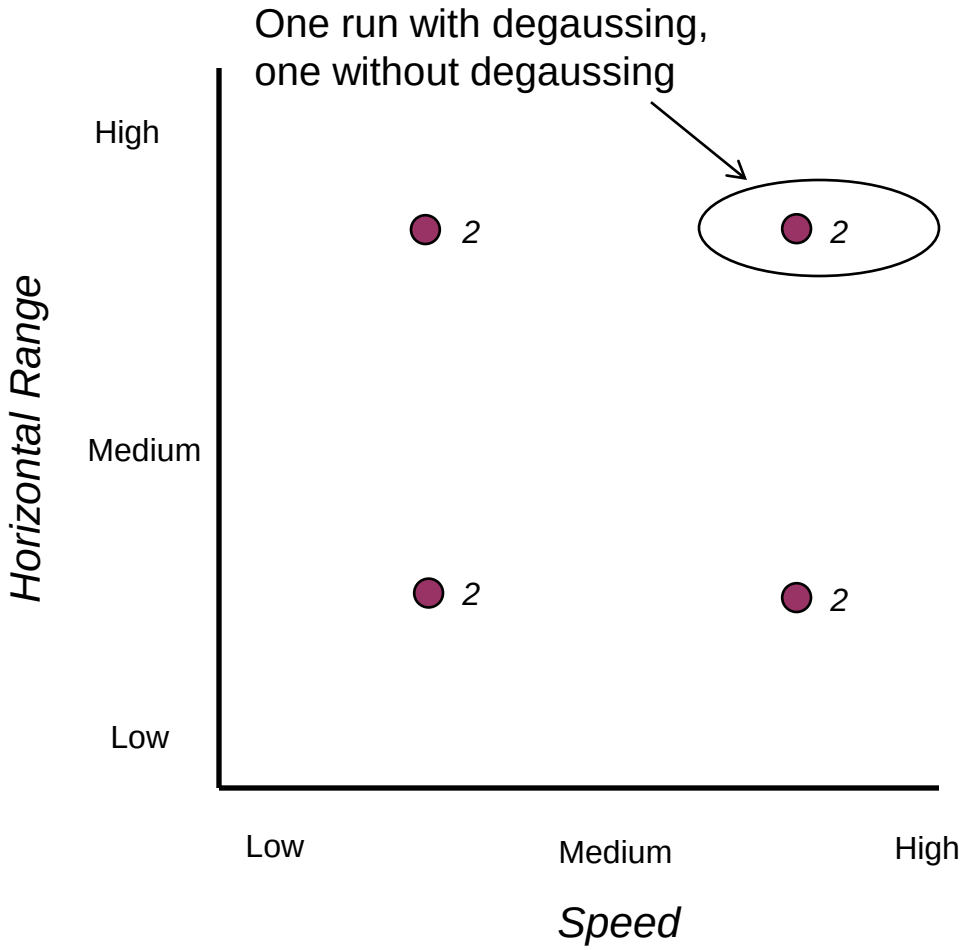
The Path to Increasing the Scientific Rigor of T&E



Test Designs – Mine Susceptibility

- **DOE provides a vast library of test design types**

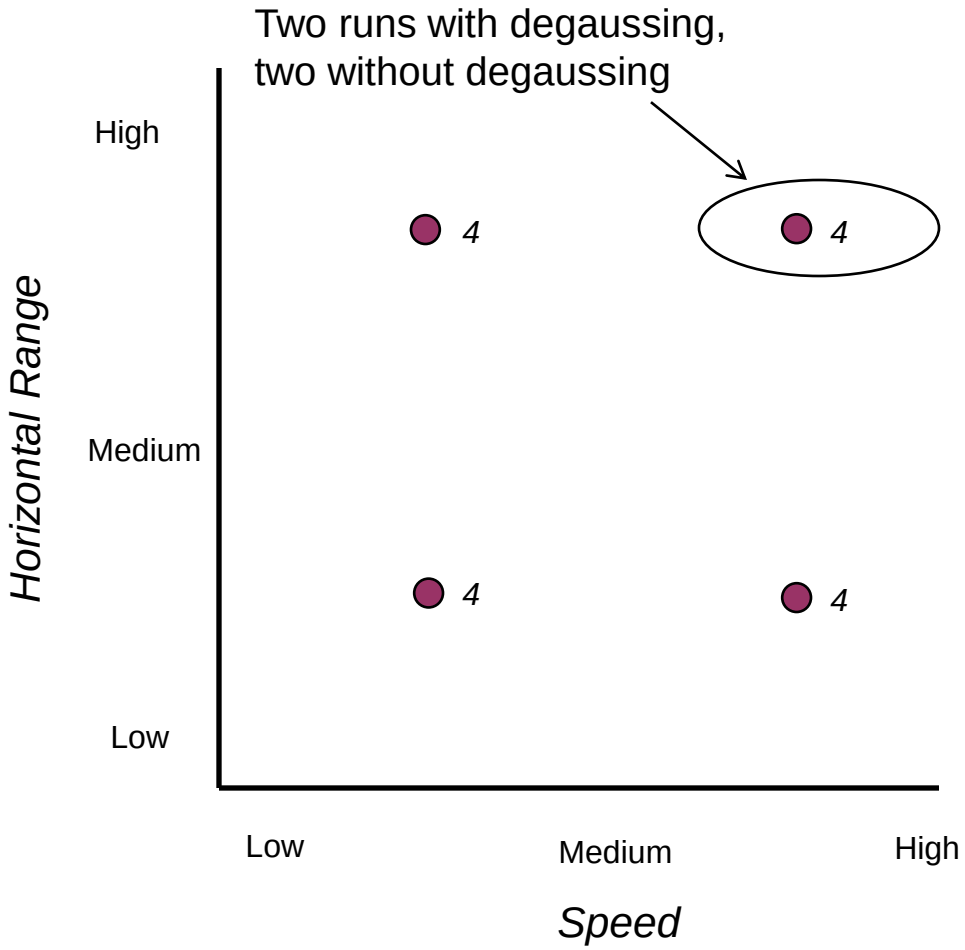
	Design Type	Number of Runs
1	Full Factorial (2-level)	8
2	Full Factorial (2-level) replicated	16
3	General Factorial (3x3x2)	18
4	Central Composite Design	18
5	Central Composite Design (replicated center point)	20
6	Central composite Design with replicated factorial points (Large CCD)	28
7	Replicated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



Test Designs – Mine Susceptibility

- **DOE provides a vast library of test design types**

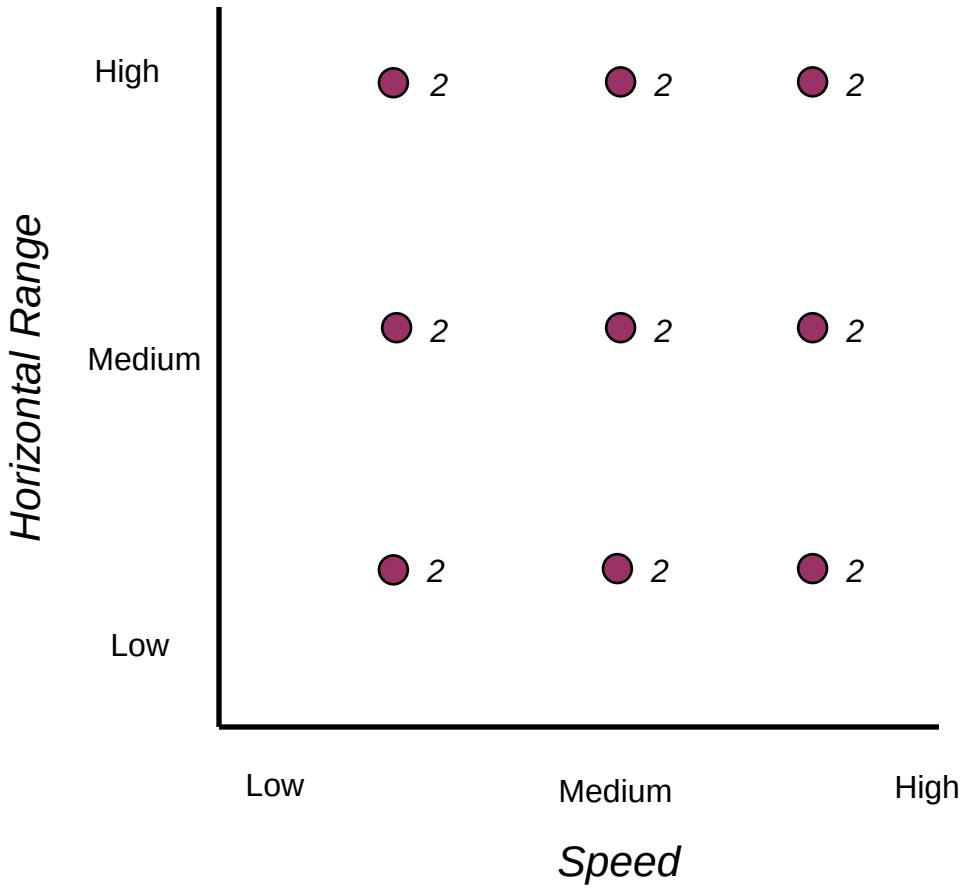
	Design Type	Number of Runs
1	Full Factorial (2-level)	8
2	Full Factorial (2-level) replicated	16
3	General Factorial (3x3x2)	18
4	Central Composite Design	18
5	Central Composite Design (replicated center point)	20
6	Central composite Design with replicated factorial points (Large CCD)	28
7	Replicated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



Test Designs – Mine Susceptibility

- **DOE provides a vast library of test design types**

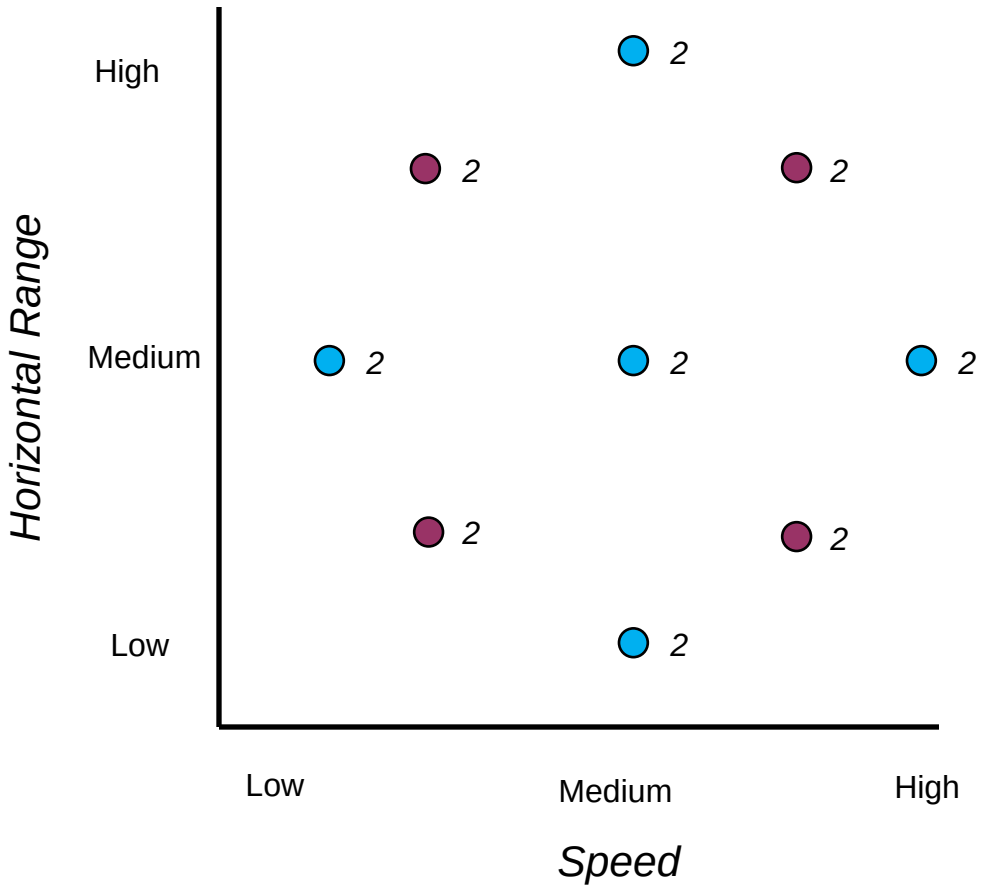
	Design Type	Number of Runs
1	Full Factorial (2-level)	8
2	Full Factorial (2-level) replicated	16
3	General Factorial (3x3x2)	18
4	Central Composite Design	18
5	Central Composite Design (replicated center point)	20
6	Central composite Design with replicated factorial points (Large CCD)	28
7	Replicated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



Test Designs – Mine Susceptibility

- **DOE provides a vast library of test design types**

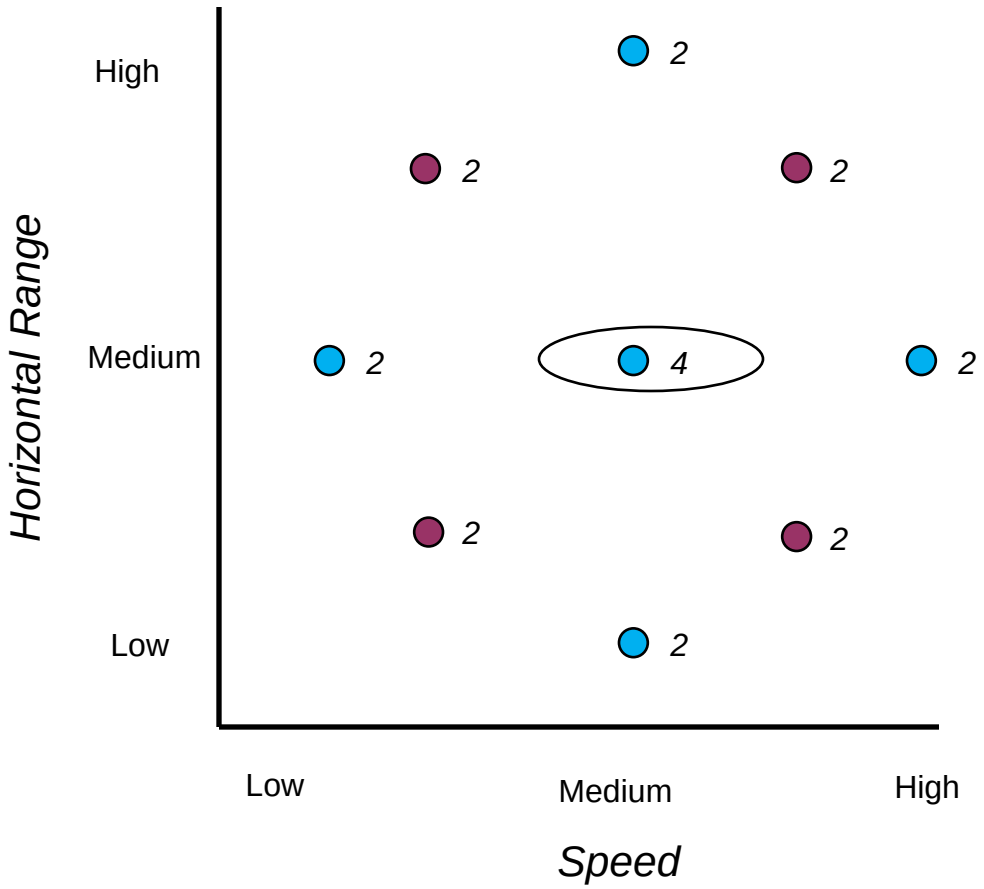
	Design Type	Number of Runs
1	Full Factorial (2-level)	8
2	Full Factorial (2-level) replicated	16
3	General Factorial (3x3x2)	18
4	Central Composite Design	18
5	Central Composite Design (replicated center point)	20
6	Central composite Design with replicated factorial points (Large CCD)	28
7	Replicated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



Test Designs – Mine Susceptibility

- **DOE provides a vast library of test design types**

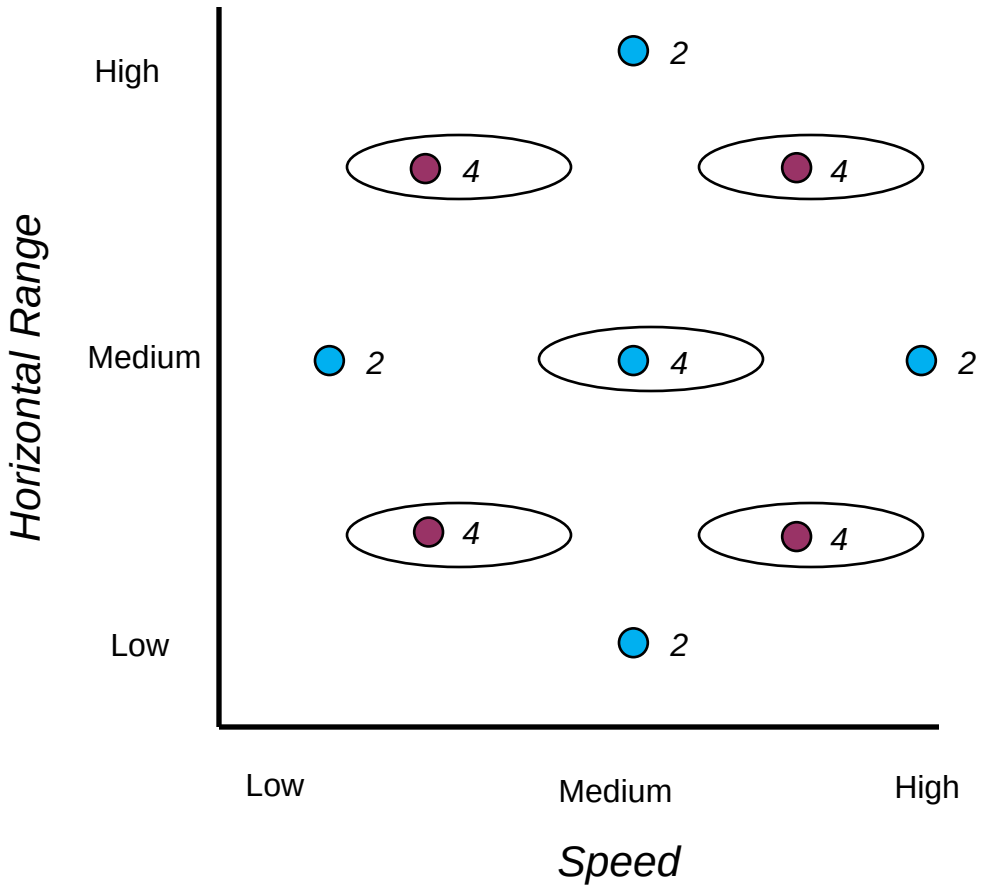
	Design Type	Number of Runs
1	Full Factorial (2-level)	8
2	Full Factorial (2-level) replicated	16
3	General Factorial (3x3x2)	18
4	Central Composite Design	18
5	Central Composite Design (replicated center point)	20
6	Central composite Design with replicated factorial points (Large CCD)	28
7	Replicated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



Test Designs – Mine Susceptibility

- **DOE provides a vast library of test design types**

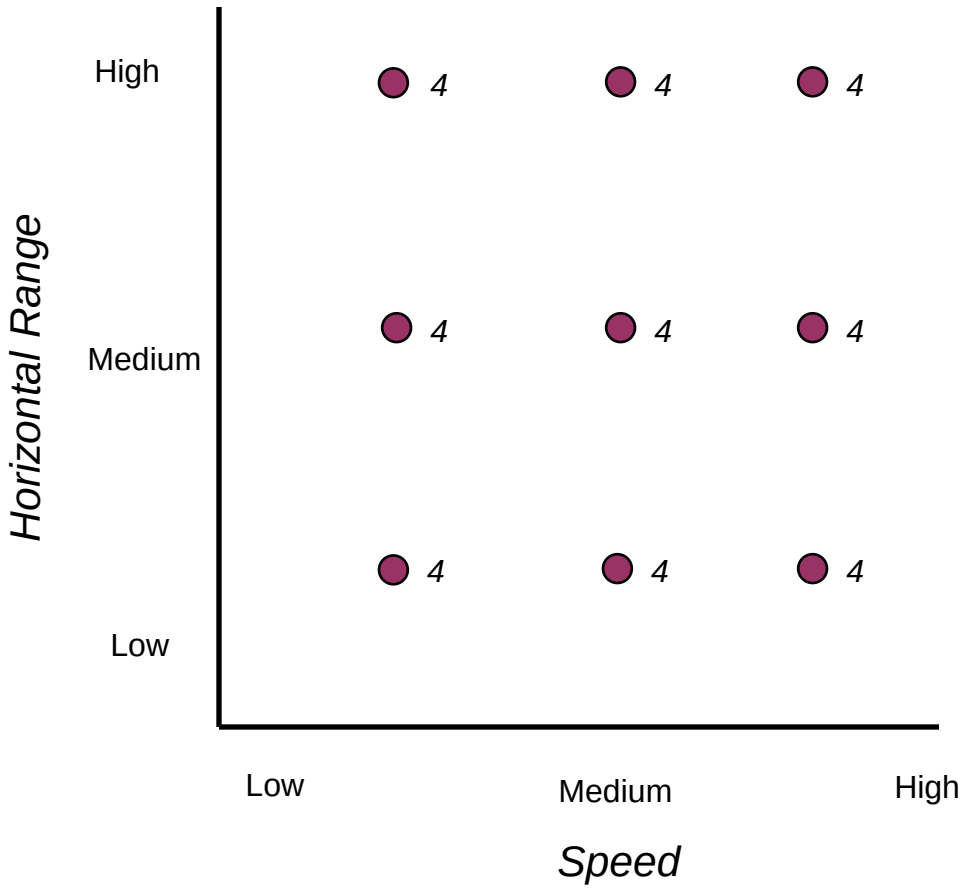
	Design Type	Number of Runs
1	Full Factorial (2-level)	8
2	Full Factorial (2-level) replicated	16
3	General Factorial (3x3x2)	18
4	Central Composite Design	18
5	Central Composite Design (replicated center point)	20
6	Central composite Design with replicated factorial points (Large CCD)	28
7	Replicated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



Test Designs – Mine Susceptibility

- **DOE provides a vast library of test design types**

	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



Submarine Detection Time: DOE Plan

		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

Submarine Detection Time: Data Collected

		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

Submarine Detection Time: Analysis

- **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	0.0154
APB	Detection time is shorter for APB-11	0.0020
Target Type	Detection time is shorter for SSN targets	0.0003
Target Noise Level	Detection time is shorter for loud targets	0.0017
Array Type	Detection time is shorter for Array B	0.0004
Type* Noise	Additional model terms improve predictions. Third order interaction is marginally significant, therefore all second order terms are retained.	0.0601
Type* Array		0.9286
Noise*Array		0.8547
Type* Noise*Array		0.0643