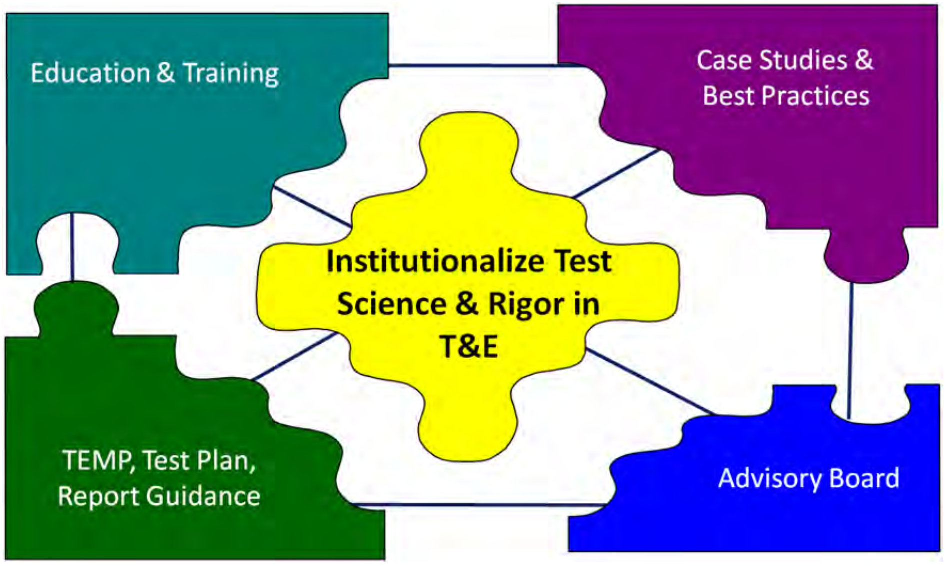
Appendices to the

Test Science Roadmap


Education & Training
Case Studies & Best Practices
Institutionalize Test Science & Rigor in T&E
TEMP, Test Plan, Report Guidance
Advisory Board

for the

Director, Operational Test and Evaluation (DOT&E)

July 2013

These appendices provide training, tutorials, case studies, and white papers that supplement the Test Science Roadmap Report.

This page intentionally left blank.

Appendices

Appendix 1: DOT&E Action Officer Training

Appendix 2: Tutorials

Appendix 3: Roadmap Case Studies

Appendix 4: IDA Background Case Studies

Appendix 5: White Papers

This page intentionally left blank.

Appendix 1
DOT&E Action Officer Training

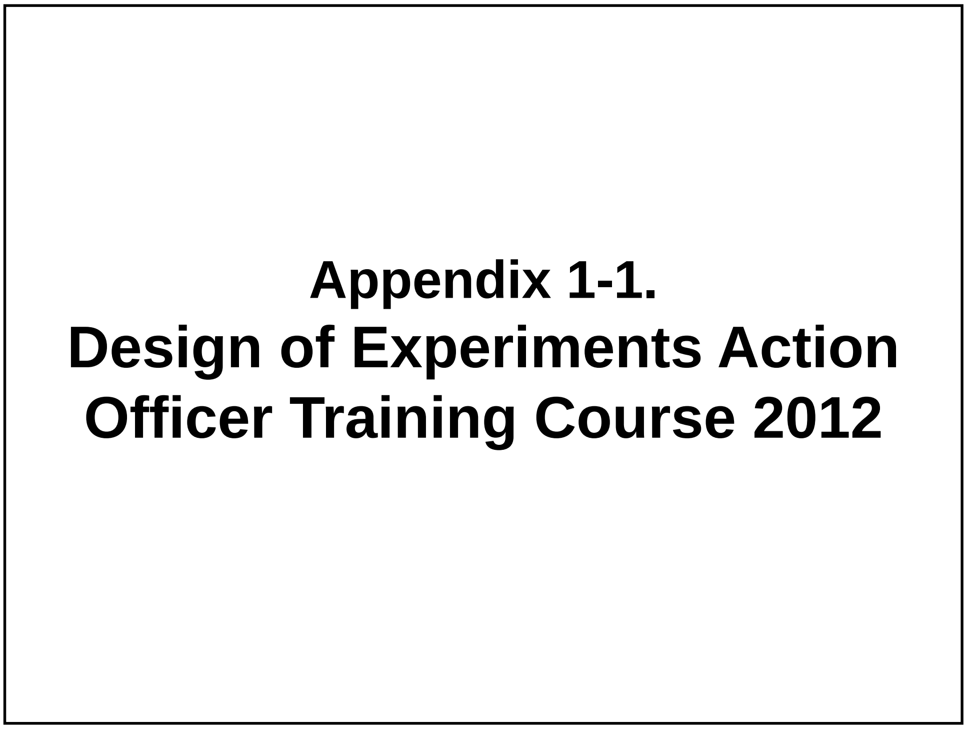
1-1. Design of Experiments Action Officer Training Course 2012

1-2. Survey Action Officer Training Course

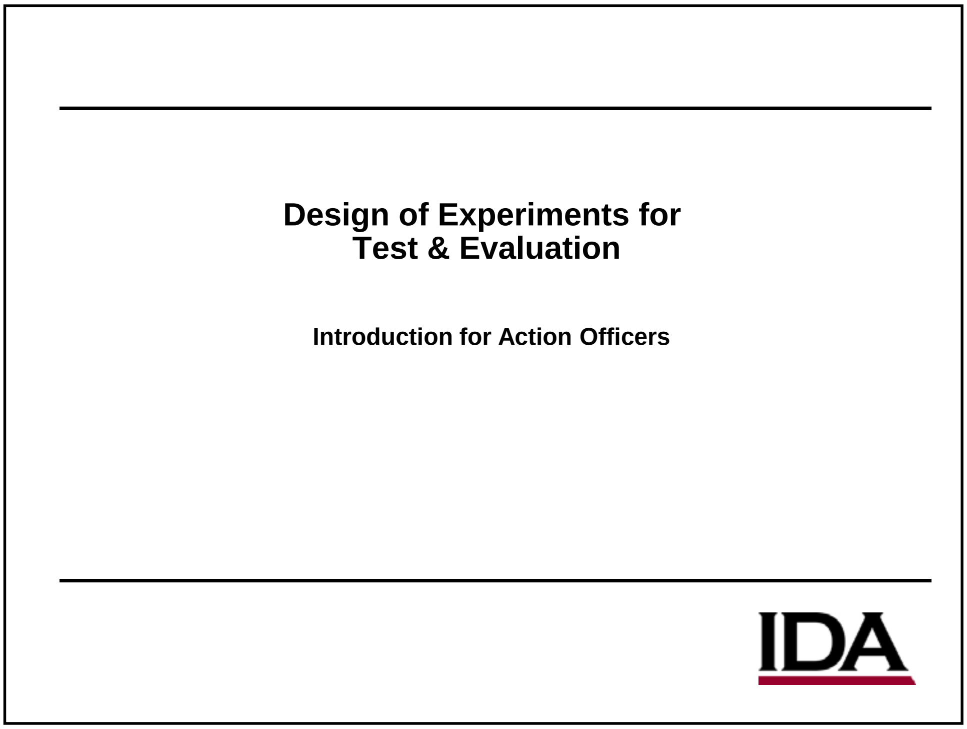
1-3. Reliability Action Officer Training Course

1-4. DOT&E Warfare Brownbag Examples

This page intentionally left blank.



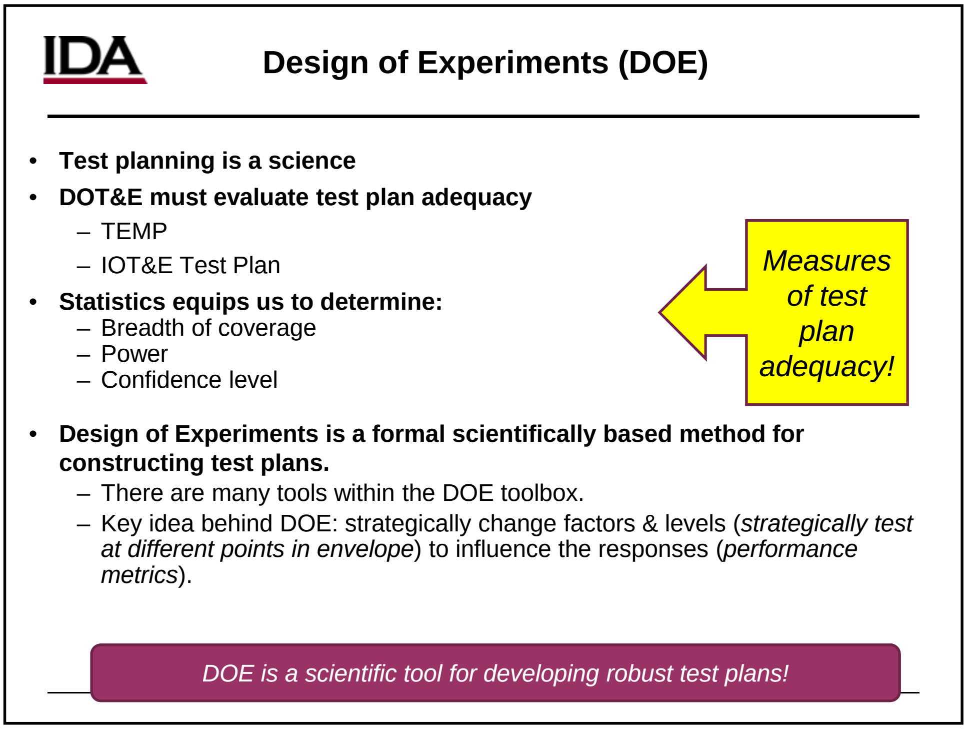
Appendix 1-1.
Design of Experiments Action Officer Training Course 2012



Design of Experiments for Test & Evaluation

Introduction for Action Officers

IDA



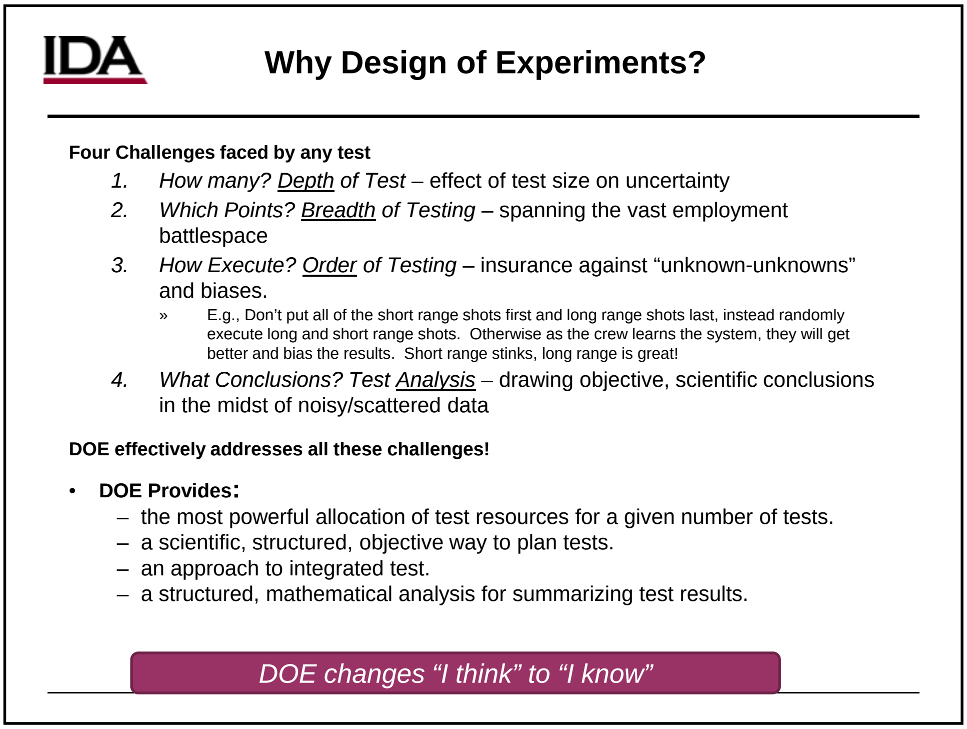
Design of Experiments (DOE)

- **Test planning is a science**
- **DOT&E must evaluate test plan adequacy**
  - TEMP
  - IOT&E Test Plan
- **Statistics equips us to determine:**
  - Breadth of coverage
  - Power
  - Confidence level

Measures of test plan adequacy!

- **Design of Experiments is a formal scientifically based method for constructing test plans.**
  - There are many tools within the DOE toolbox.
  - Key idea behind DOE: strategically change factors & levels (*strategically test at different points in envelope*) to influence the responses (*performance metrics*).

DOE is a scientific tool for developing robust test plans!



Why Design of Experiments?

Four Challenges faced by any test

1. *How many? <u>Depth</u> of Test* – effect of test size on uncertainty
2. *Which Points? <u>Breadth</u> of Testing* – spanning the vast employment battlespace
3. *How Execute? <u>Order</u> of Testing* – insurance against "unknown-unknowns" and biases.
   - » E.g., Don't put all of the short range shots first and long range shots last, instead randomly execute long and short range shots. Otherwise as the crew learns the system, they will get better and bias the results. Short range stinks, long range is great!
4. *What Conclusions? Test <u>Analysis</u>* – drawing objective, scientific conclusions in the midst of noisy/scattered data

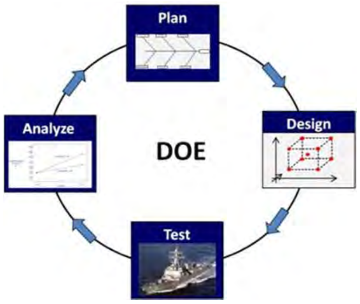
DOE effectively addresses all these challenges!

- **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.

DOE changes "I think" to "I know"

Steps in Testing and Evaluating a System (DOE Methodology)

1. Define the objective of the experiment
2. Select appropriate response variables
3. Choose factors, levels
4. Choose experimental design
5. Perform the test
6. Statistically analyze the data
7. Draw conclusions (a.k.a. Evaulation)

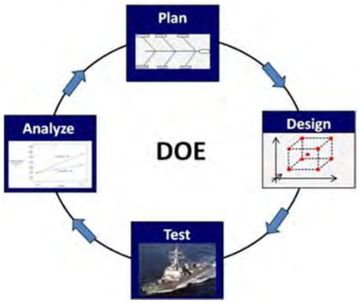


Steps are strategically linked into a defensible process!

Design of Experiments: Planning

Steps in Designing an Experiment

1. **Define the objective of the experiment**
2. **Select appropriate response variables**
3. **Choose factors, levels**
4. Choose experimental design
5. Perform the test
6. Statistically analyze the data
7. Draw conclusions

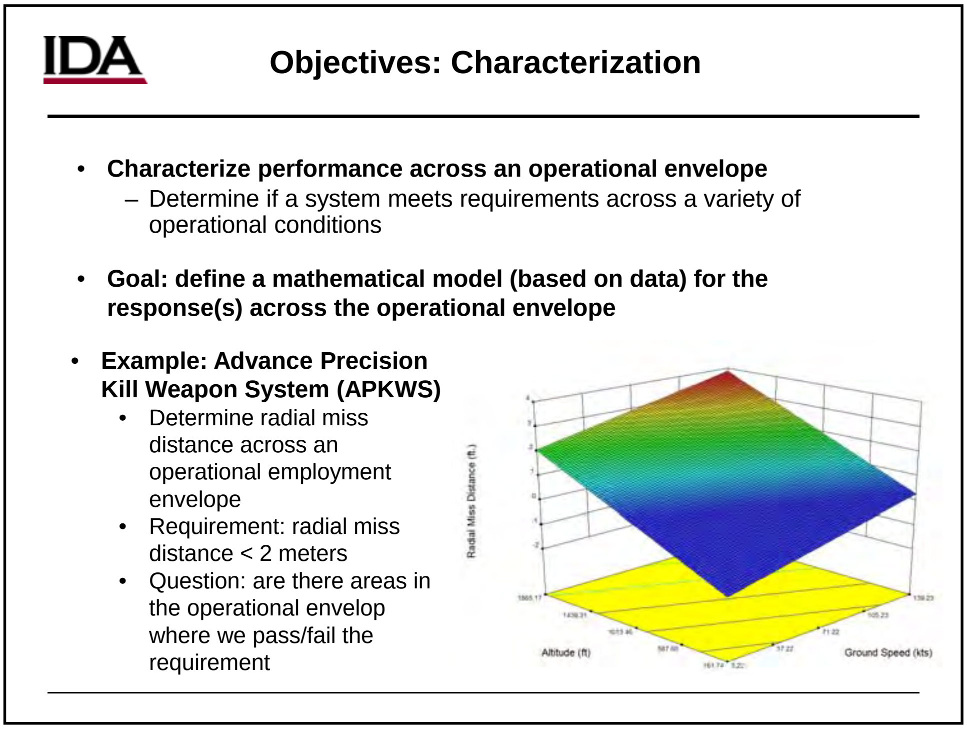


Steps are strategically linked into a defensible process!

Planning Essential Elements

- **Planning is essential for defensible test designs**
  - Poor planning → indefensible results
  - Proper planning → easily defendable results
- **Determine test objective(s)**
- **Determine response variables**
  - Definition: The response variable measures the outcome of interest for the test (a.k.a. Measures, dependent variables).
  - Objective, valid, informative, measureable - the gold standard
- **Determine factors and levels**
  - Definition: Factors are independent variables that are expected to impact the outcome of a test.
- **No math, no clever ideas here … just plain hard work**
  - Planning is a collaborative effort
  - Leverage operational experience of AO

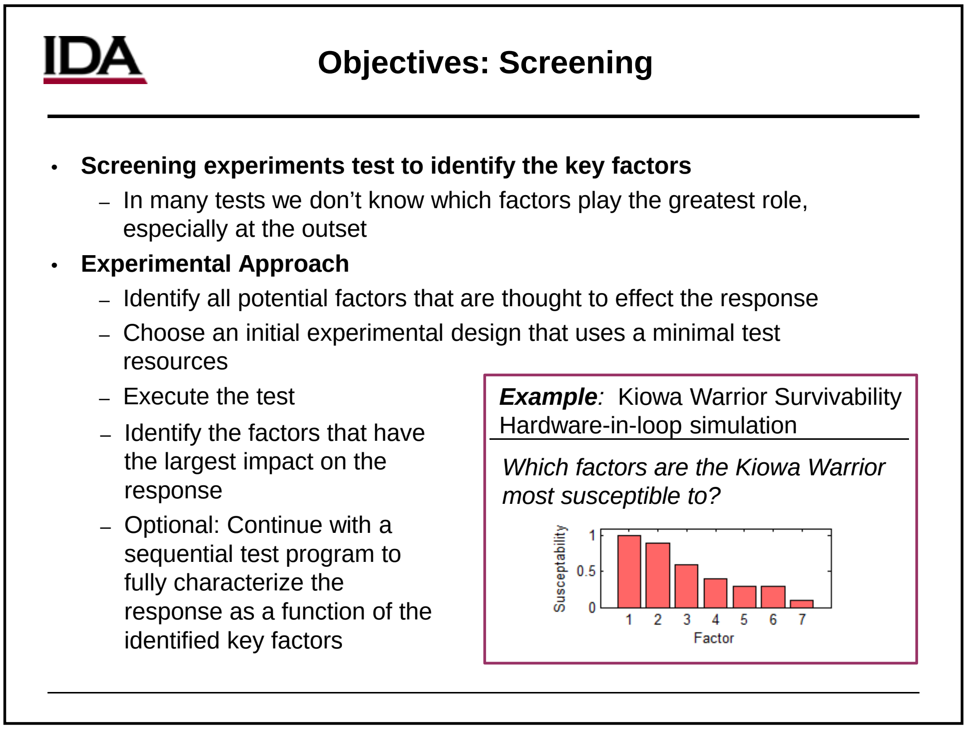
If the planning is wrong the design is meaningless!



Objectives: Characterization

- **Characterize performance across an operational envelope**
  – Determine if a system meets requirements across a variety of operational conditions
- **Goal: define a mathematical model (based on data) for the response(s) across the operational envelope**
- **Example: Advance Precision Kill Weapon System (APKWS)**
  - Determine radial miss distance across an operational employment envelope
  - Requirement: radial miss distance < 2 meters
  - Question: are there areas in the operational envelop where we pass/fail the requirement

Radial Miss Distance (ft.)
Altitude (ft)
Ground Speed (kts)

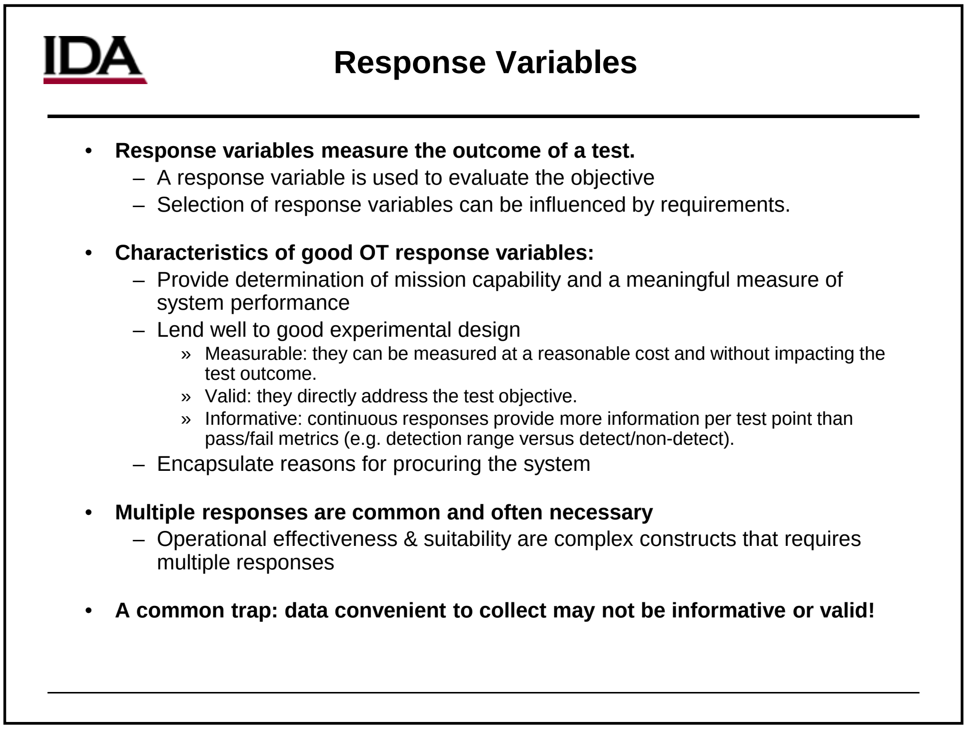


Objectives: Screening

- **Screening experiments test to identify the key factors**
  – In many tests we don’t know which factors play the greatest role, especially at the outset
- **Experimental Approach**
  – Identify all potential factors that are thought to effect the response
  – Choose an initial experimental design that uses a minimal test resources
  – Execute the test
  – Identify the factors that have the largest impact on the response
  – Optional: Continue with a sequential test program to fully characterize the response as a function of the identified key factors

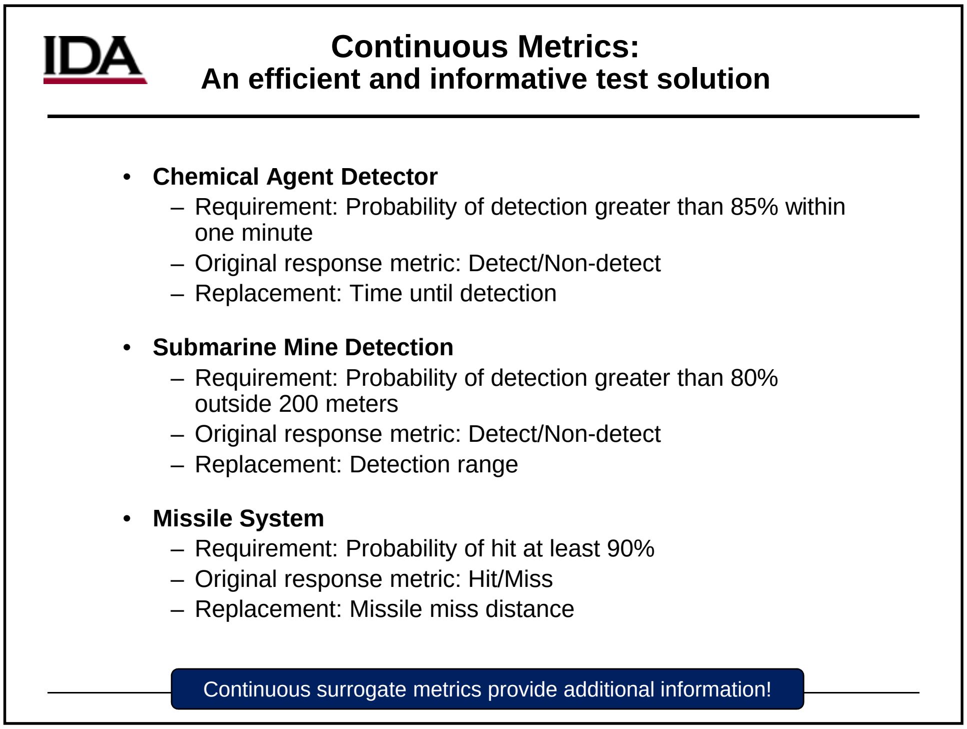
***Example**:* Kiowa Warrior Survivability Hardware-in-loop simulation

Which factors are the Kiowa Warrior most susceptible to?



Response Variables

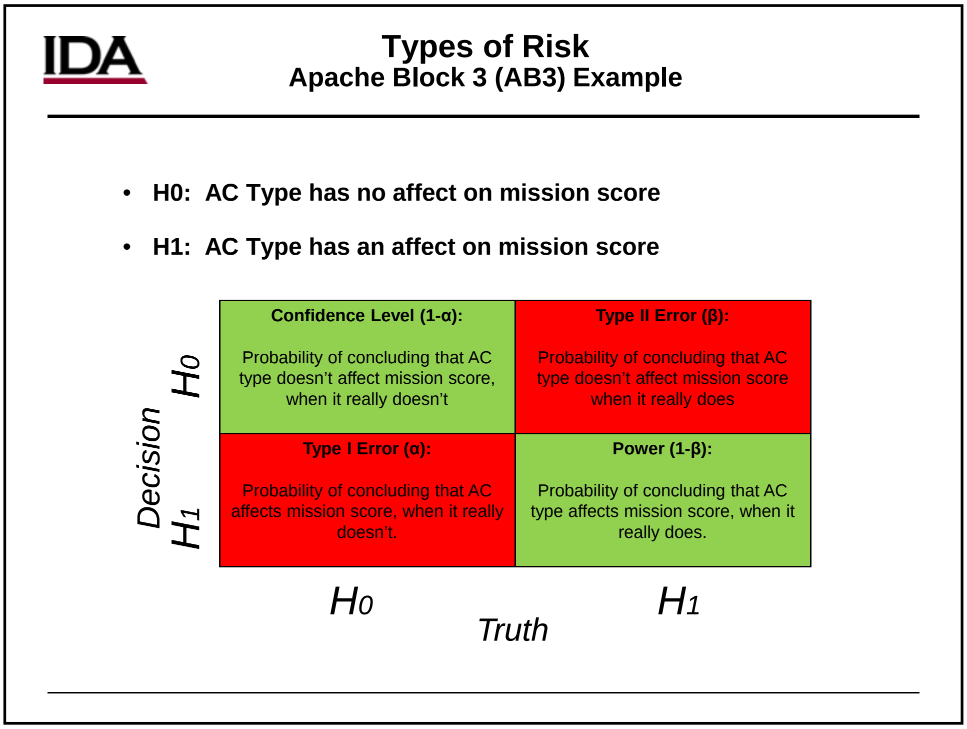
- **Response variables measure the outcome of a test.**
  - A response variable is used to evaluate the objective
  - Selection of response variables can be influenced by requirements.
- **Characteristics of good OT response variables:**
  - Provide determination of mission capability and a meaningful measure of system performance
  - Lend well to good experimental design
    - Measurable: they can be measured at a reasonable cost and without impacting the test outcome.
    - Valid: they directly address the test objective.
    - Informative: continuous responses provide more information per test point than pass/fail metrics (e.g. detection range versus detect/non-detect).
  - Encapsulate reasons for procuring the system
- **Multiple responses are common and often necessary**
  - Operational effectiveness & suitability are complex constructs that requires multiple responses
- **A common trap: data convenient to collect may not be informative or valid!**



Continuous Metrics:
An efficient and informative test solution

- **Chemical Agent Detector**
  - Requirement: Probability of detection greater than 85% within one minute
  - Original response metric: Detect/Non-detect
  - Replacement: Time until detection
- **Submarine Mine Detection**
  - Requirement: Probability of detection greater than 80% outside 200 meters
  - Original response metric: Detect/Non-detect
  - Replacement: Detection range
- **Missile System**
  - Requirement: Probability of hit at least 90%
  - Original response metric: Hit/Miss
  - Replacement: Missile miss distance

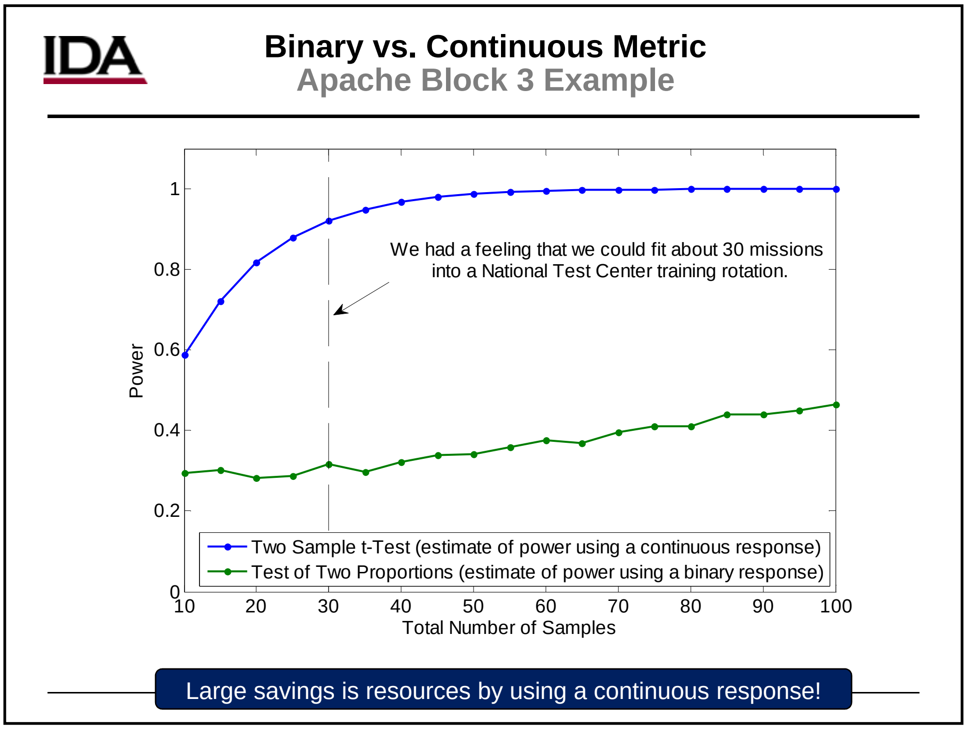
Continuous surrogate metrics provide additional information!



Types of Risk
Apache Block 3 (AB3) Example

- **H0: AC Type has no affect on mission score**
- **H1: AC Type has an affect on mission score**

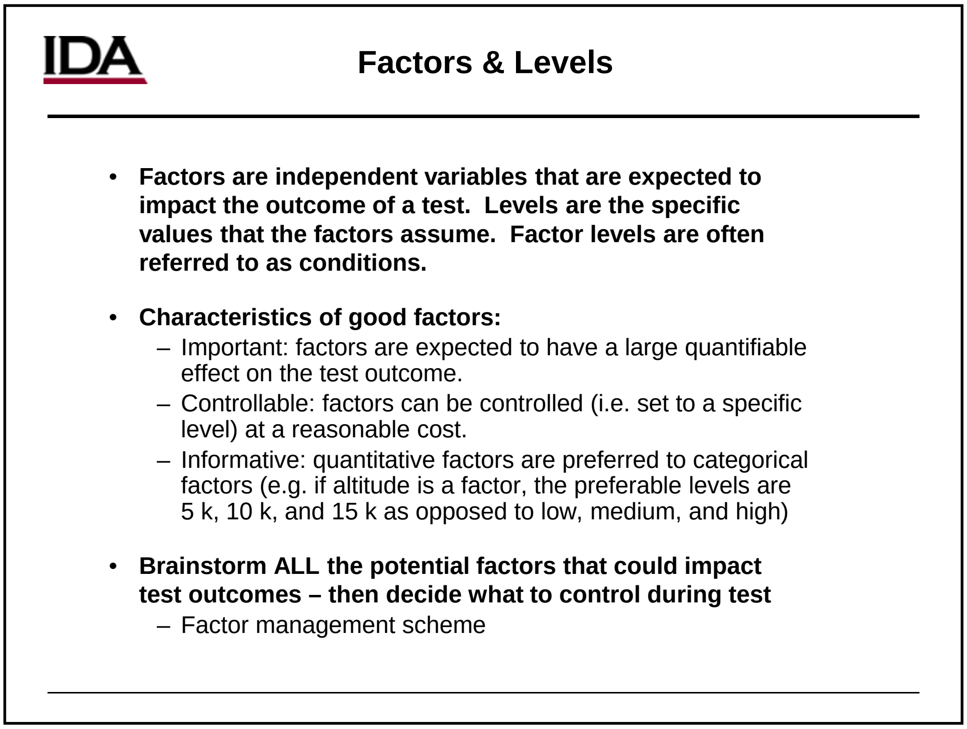
Decision	Truth: *H0*	Truth: *H1*
H0	**Confidence Level (1-α):** Probability of concluding that AC type doesn't affect mission score, when it really doesn't	**Type II Error (β):** Probability of concluding that AC type doesn't affect mission score when it really does
H1	**Type I Error (α):** Probability of concluding that AC affects mission score, when it really doesn't.	**Power (1-β):** Probability of concluding that AC type affects mission score, when it really does.



Binary vs. Continuous Metric
Apache Block 3 Example

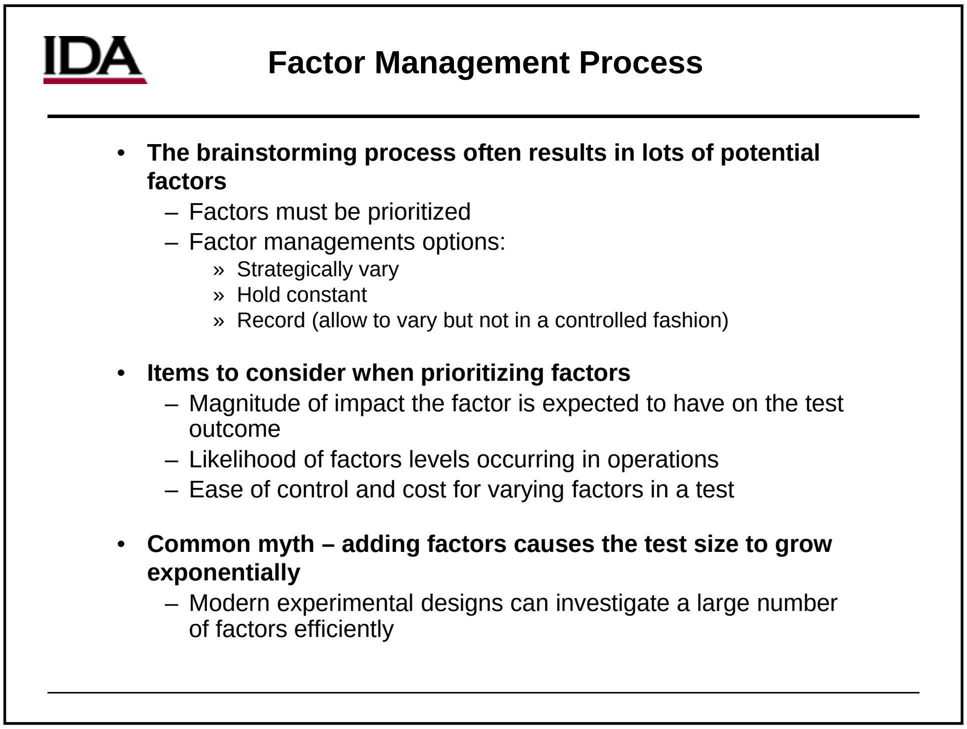
We had a feeling that we could fit about 30 missions into a National Test Center training rotation.

Large savings is resources by using a continuous response!



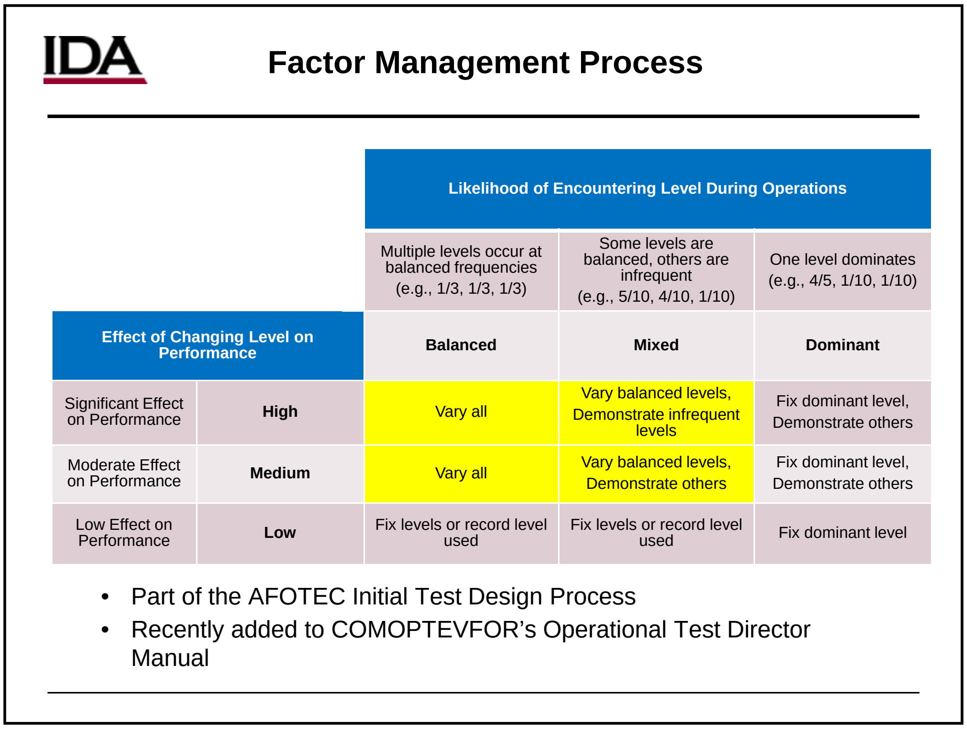
Factors & Levels

- **Factors are independent variables that are expected to impact the outcome of a test. Levels are the specific values that the factors assume. Factor levels are often referred to as conditions.**
- **Characteristics of good factors:**
  - Important: factors are expected to have a large quantifiable effect on the test outcome.
  - Controllable: factors can be controlled (i.e. set to a specific level) at a reasonable cost.
  - Informative: quantitative factors are preferred to categorical factors (e.g. if altitude is a factor, the preferable levels are 5 k, 10 k, and 15 k as opposed to low, medium, and high)
- **Brainstorm ALL the potential factors that could impact test outcomes – then decide what to control during test**
  - Factor management scheme



Factor Management Process

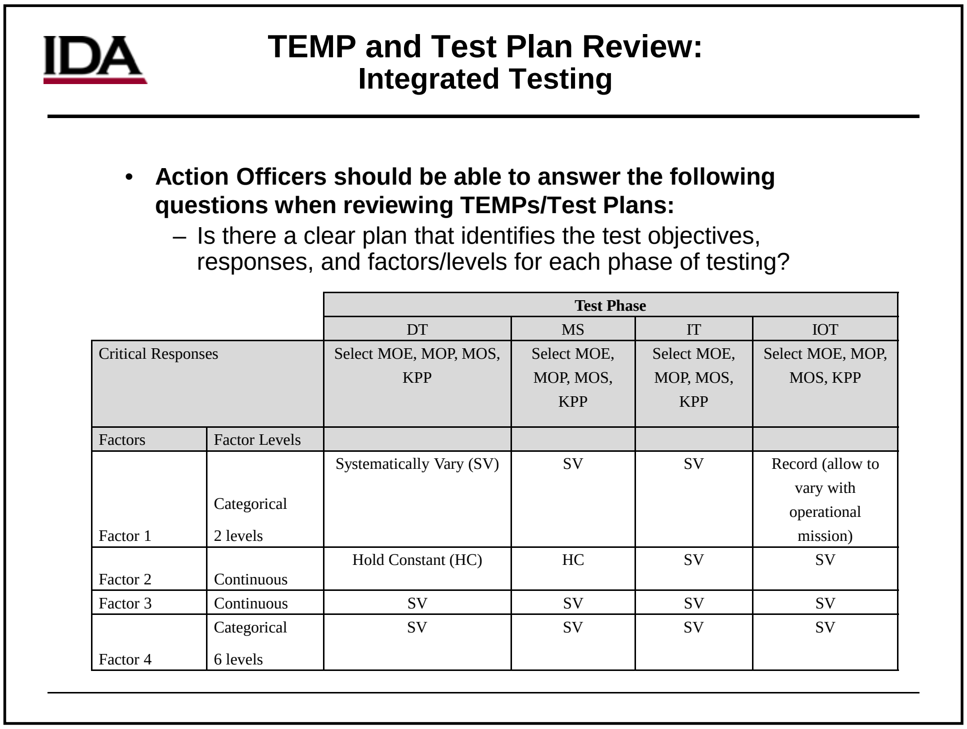
- **The brainstorming process often results in lots of potential factors**
  - Factors must be prioritized
  - Factor managements options:
    - » Strategically vary
    - » Hold constant
    - » Record (allow to vary but not in a controlled fashion)
- **Items to consider when prioritizing factors**
  - Magnitude of impact the factor is expected to have on the test outcome
  - Likelihood of factors levels occurring in operations
  - Ease of control and cost for varying factors in a test
- **Common myth – adding factors causes the test size to grow exponentially**
  - Modern experimental designs can investigate a large number of factors efficiently



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

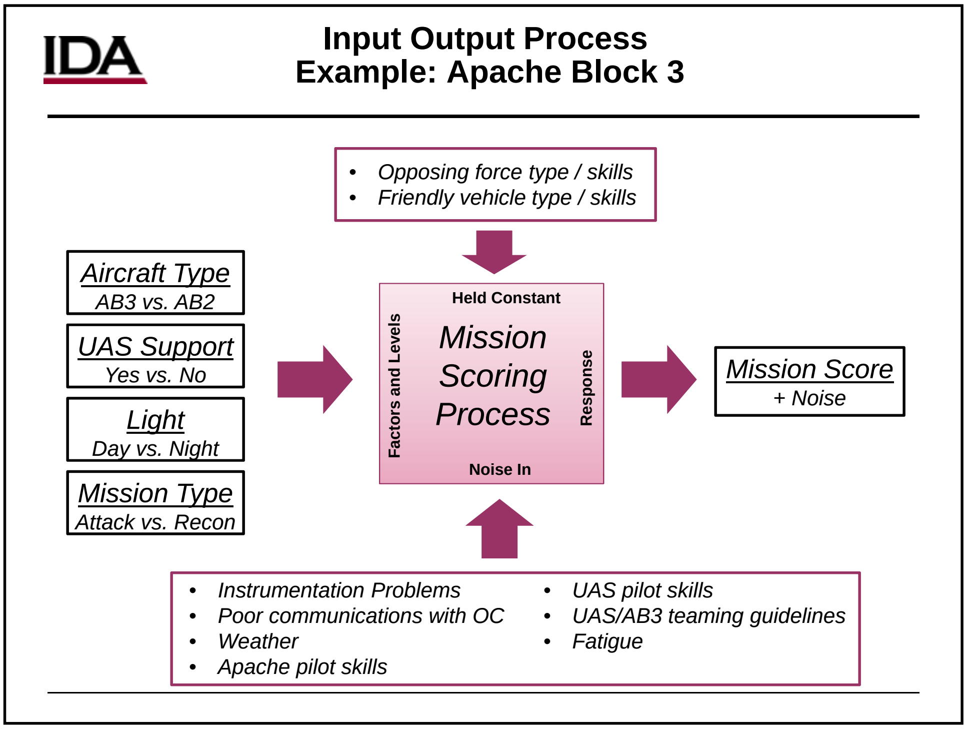
- Part of the AFOTEC Initial Test Design Process
- Recently added to COMOPTEVFOR's Operational Test Director Manual



TEMP and Test Plan Review: Integrated Testing

- **Action Officers should be able to answer the following questions when reviewing TEMPs/Test Plans:**
  – Is there a clear plan that identifies the test objectives, responses, and factors/levels for each phase of testing?

		Test Phase			
		DT	MS	IT	IOT
Critical Responses		Select MOE, MOP, MOS, KPP	Select MOE, MOP, MOS, KPP	Select MOE, MOP, MOS, KPP	Select MOE, MOP, MOS, KPP
Factors	Factor Levels				
Factor 1	Categorical 2 levels	Systematically Vary (SV)	SV	SV	Record (allow to vary with operational mission)
Factor 2	Continuous	Hold Constant (HC)	HC	SV	SV
Factor 3	Continuous	SV	SV	SV	SV
Factor 4	Categorical 6 levels	SV	SV	SV	SV



Input Output Process
Example: Apache Block 3

- *Opposing force type / skills*
- *Friendly vehicle type / skills*

Aircraft Type
AB3 vs. AB2

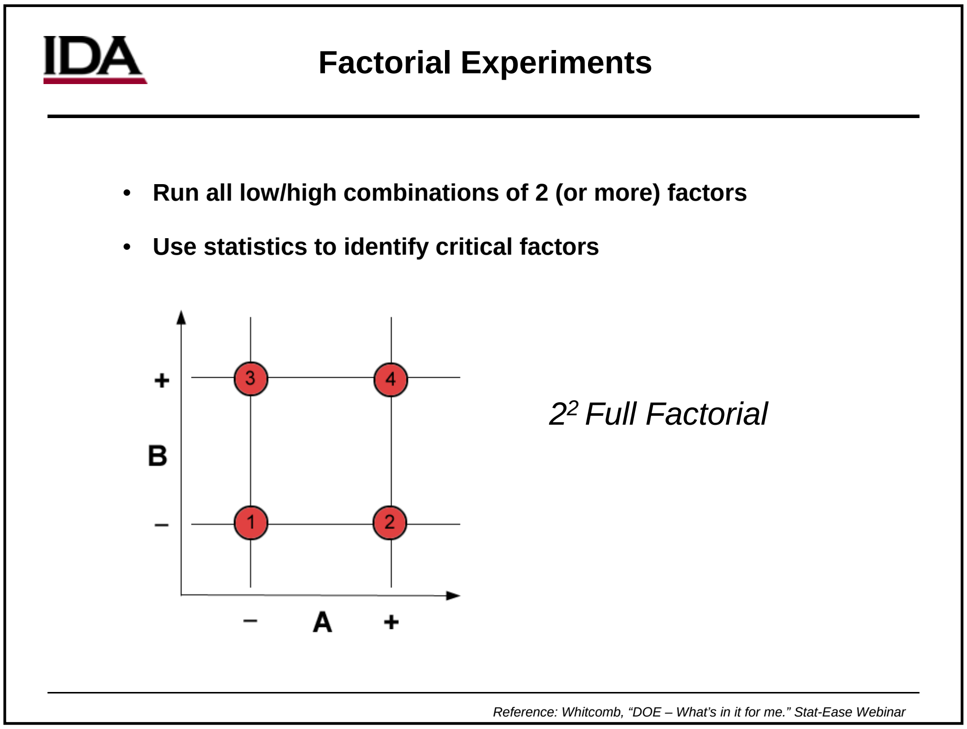
UAS Support
Yes vs. No

Light
Day vs. Night

Mission Type
Attack vs. Recon

Mission Score
+ Noise

- *Instrumentation Problems*
- *Poor communications with OC*
- *Weather*
- *Apache pilot skills*
- *UAS pilot skills*
- *UAS/AB3 teaming guidelines*
- *Fatigue*



Factorial Experiments

- **Run all low/high combinations of 2 (or more) factors**
- **Use statistics to identify critical factors**

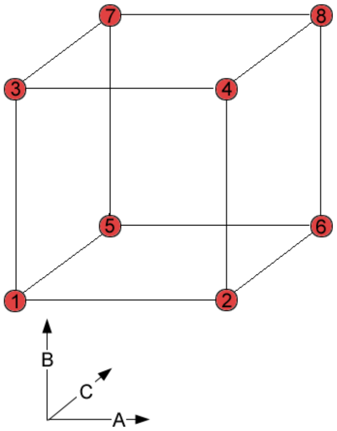
2^2 Full Factorial

Reference: Whitcomb, “DOE – What’s in it for me.” Stat-Ease Webinar

Factorial Experiments

Reference: Whitcomb, "DOE – What's in it for me." Stat-Ease Webinar

- **2^3 Full Factorial**



Run #	A	B	C	AB	AC	BC	ABC
1	-	-	-	+	+	+	-
2	+	-	-	-	-	+	+
3	-	+	-	-	+	-	+
4	+	+	-	+	-	-	-
5	-	-	+	+	-	-	+
6	+	-	+	-	+	-	-
7	-	+	+	-	-	+	-
8	+	+	+	+	+	+	+

- **With these 8 runs we can evaluate:**
  - three main effects
  - three two-factor interactions
  - one three factor interaction
  - and the overall average

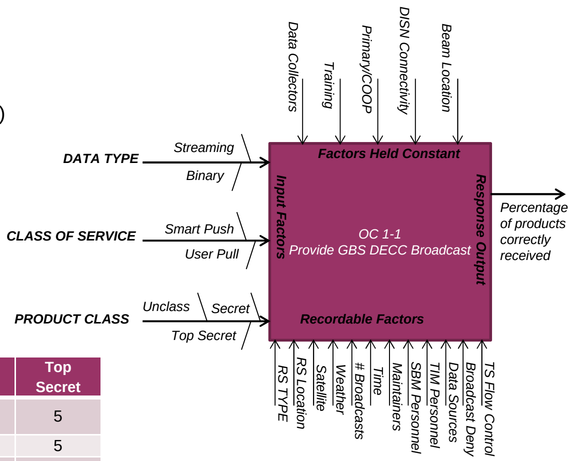
There are many other types of designs besides full factorials!

Global Broadcast Service (GBS)

- **GBS Background**
  - Provide a one-way, high-speed flow of near-real-time wideband information to forces garrisoned or deployed
  - High-capacity product dissemination (Imagery, UAS full-motion video, large data files)
  - Information transported through GBS supports a large variety of missions.
- **Experimental Objective**
  - Determine transmission completion within operational envelope
- **Response: Percentage of products received**
  - The products received can support a broad array of other missions
  - Lead operational test organization (17 TS) considered applicability of DOE to the system under test and arrived at a design for the "percentage of products correctly received" measure.

Global Broadcast Service (GBS)

- TEMP Content
  - *Factor selection process map (right)*
  - Factorial design matrix (below)
  - *Power calculations*
  - Proposed analysis method
- Binary Response Metric
- Possible Analyses:
  - Logistic regression
  - One and two-sample proportion significance tests



2x2x3 Factorial Design Matrix

		Unclass	Secret	Top Secret
Streaming	Smart Push	5	5	5
	User Pull	25	25	5
Binary	Smart Push	5	5	5
	User Pull	25	25	5

- *3 categorical factors*
- *Unbalanced design mimics operational expectations*

Consolidated Afloat Networks and Enterprise Services (CANES)

- **CANES Background**
  - consolidate and improve the networks on tactical platforms, largely through a common computing environment.
  - will modernize the IT infrastructure for ships, submarines, aircraft and selected shore sites
- **Test Objective**
  - Determine if CANES provides a timely and accurate display on the display terminal

Consolidated Afloat Networks and Enterprise Services (CANES)

- **Responses (all continuous):**
  - Chat latency (requirement: <=5 sec)
  - Time to display common operational picture
  - Time to download and display media on a CANES terminal (requirement: <=10 sec)
- **Factors (all categorical):**
  - Classification (Unclassified, SR, Secret, SCI)
  - Network Loading (Low, High)
  - Transmission Type (Internal, External)

4 x 2 x 2 General Factorial Design (32 runs)					
		Unclass	Secret	SR	TS-SCI
Internal	Low	2	2	2	2
	High	2	2	2	2
External	Low	2	2	2	2
	High	2	2	2	2

	Power		
Term	s2n=0.5	s2n=1	s2n=2
Classification (A)	0.313	0.597	0.975
Network Loading (B)	0.544	0.931	0.999
Transmission Type (C)	0.544	0.931	0.999
AB	0.313	0.597	0.975
AC	0.313	0.597	0.975
BC	0.544	0.931	0.999

Key Takeaways: Planning

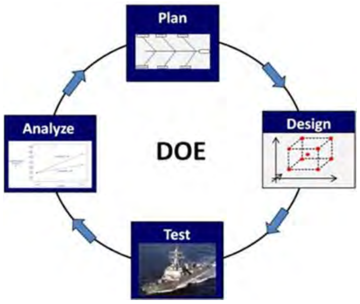
- **Identifying objectives, responses, and factors is an essential element of experimental design**
- **Objectives, responses, and factors should be clearly identified and linked**
- **Continuous responses (measures) are essential for cost efficient testing**
- **Don’t be overwhelmed by statistics, use operational experience to guide planning**

“By failing to prepare, you are preparing to fail.”
– ***Benjamin Franklin***

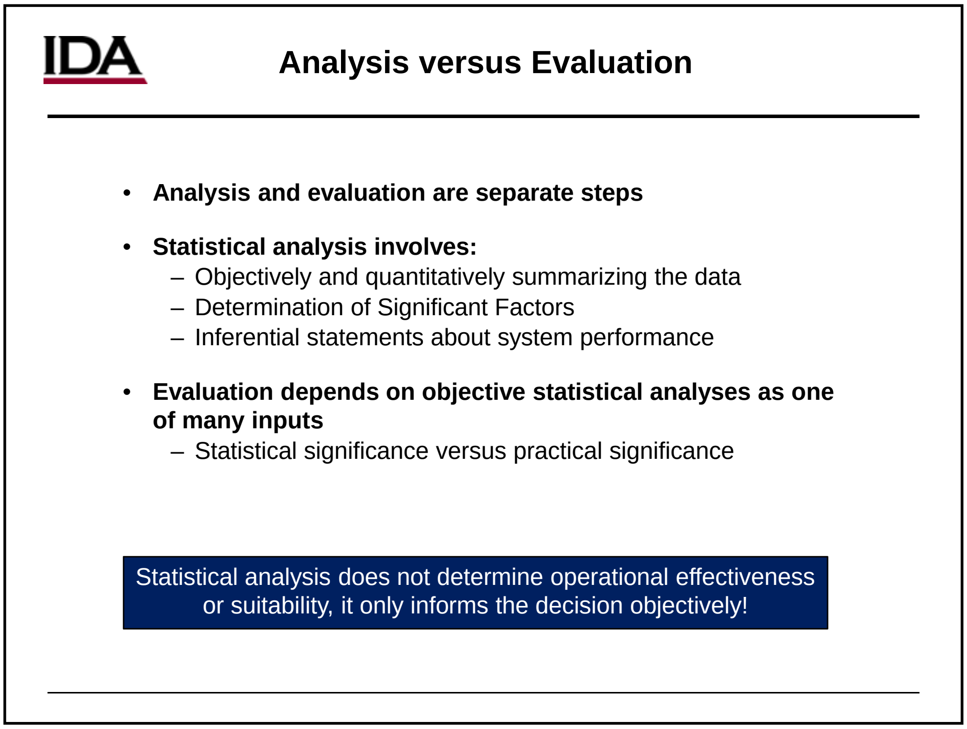
Design of Experiments:
Analysis

Steps in Testing and Evaluating a System (DOE Methodology)

1. Define the objective of the experiment
2. Select appropriate response variables
3. Choose factors, levels
4. Choose experimental design
5. Perform the test
6. **Statistically analyze the data**
7. **Draw conclusions (a.k.a. Evaluation)**



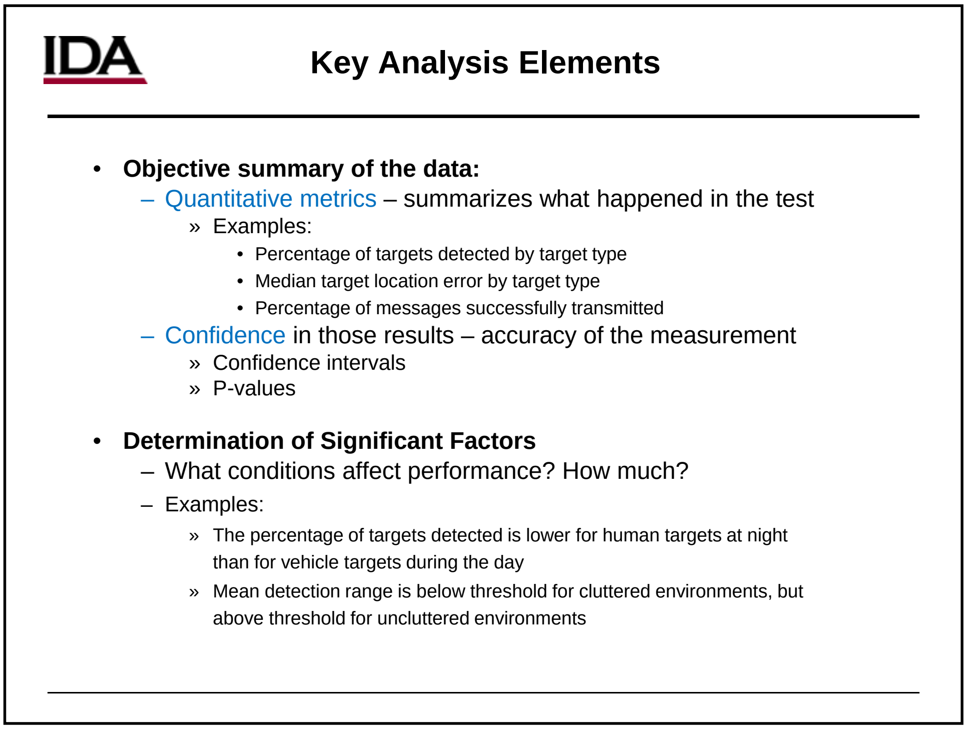
Steps are strategically linked into a defensible process!



Analysis versus Evaluation

- **Analysis and evaluation are separate steps**
- **Statistical analysis involves:**
  - Objectively and quantitatively summarizing the data
  - Determination of Significant Factors
  - Inferential statements about system performance
- **Evaluation depends on objective statistical analyses as one of many inputs**
  - Statistical significance versus practical significance

Statistical analysis does not determine operational effectiveness or suitability, it only informs the decision objectively!

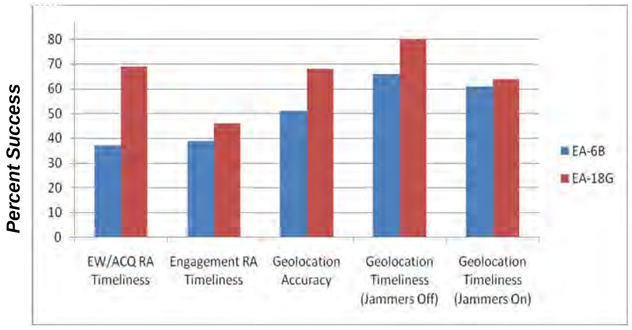


Key Analysis Elements

- **Objective summary of the data:**
  - Quantitative metrics – summarizes what happened in the test
    - » Examples:
      - Percentage of targets detected by target type
      - Median target location error by target type
      - Percentage of messages successfully transmitted
  - Confidence in those results – accuracy of the measurement
    - » Confidence intervals
    - » P-values
- **Determination of Significant Factors**
  - What conditions affect performance? How much?
  - Examples:
    - » The percentage of targets detected is lower for human targets at night than for vehicle targets during the day
    - » Mean detection range is below threshold for cluttered environments, but above threshold for uncluttered environments

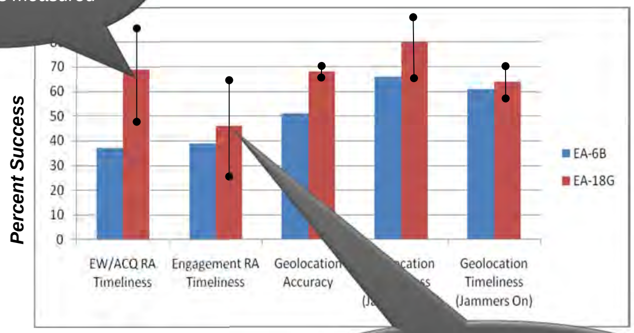
EA-18G/EA-6B Comparison Confidence Intervals

Figure from DOT&E EA-18G BLRIP



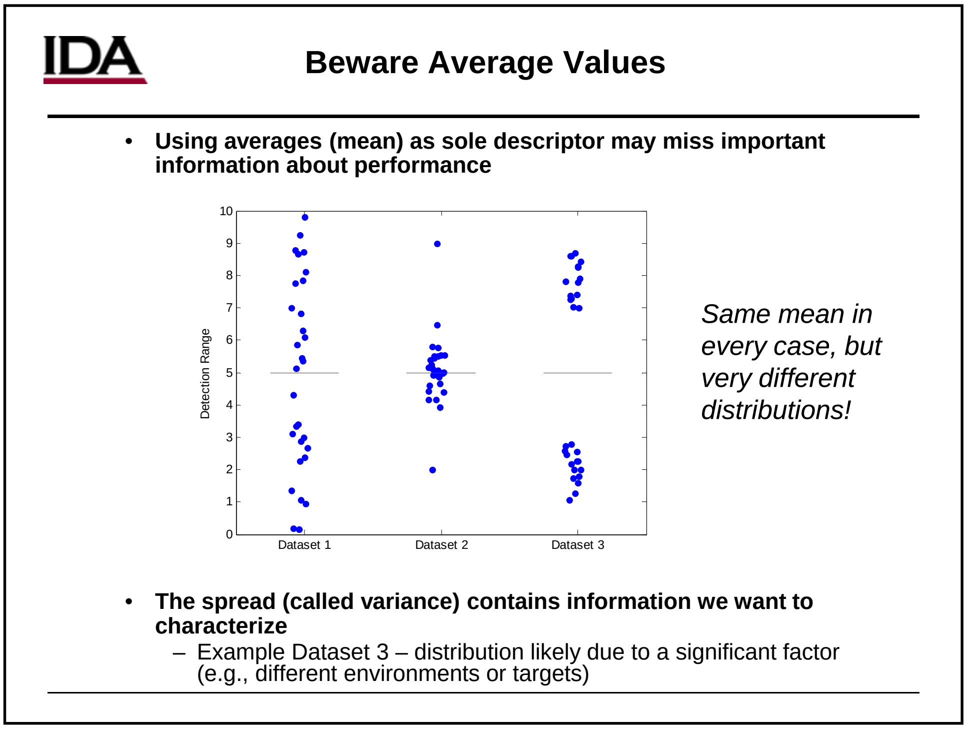
EA-18G/EA-6B Comparison Confidence Intervals

from DOT&E EA-18G BLRIP



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

Confidence intervals make it clear that performance is comparable

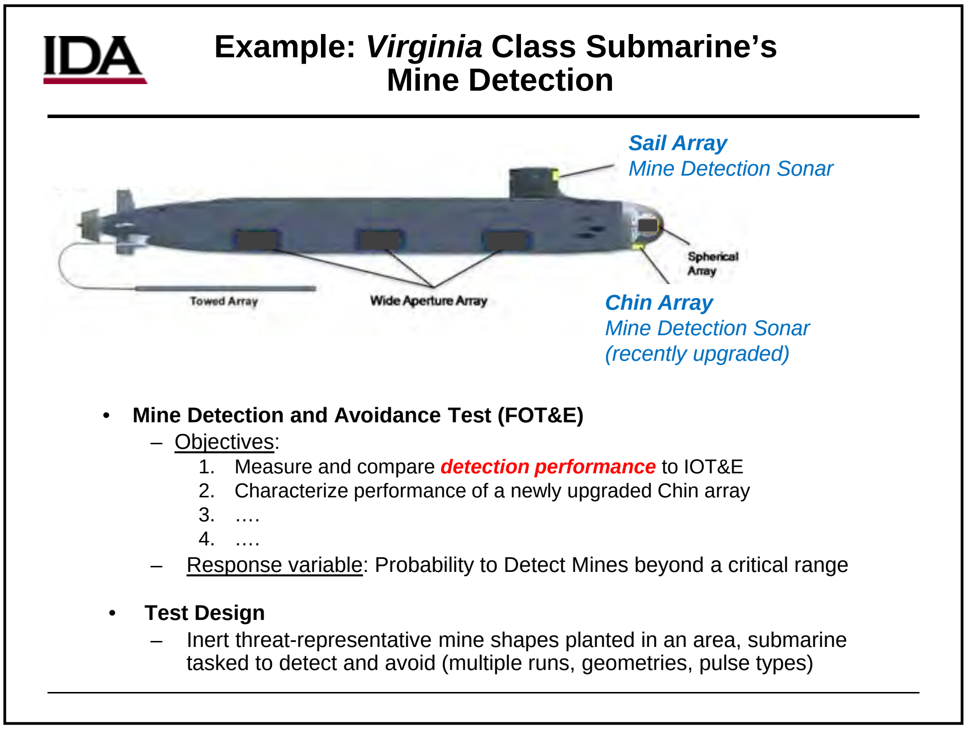


Beware Average Values

- **Using averages (mean) as sole descriptor may miss important information about performance**

Same mean in every case, but very different distributions!

- **The spread (called variance) contains information we want to characterize**
  - Example Dataset 3 – distribution likely due to a significant factor (e.g., different environments or targets)

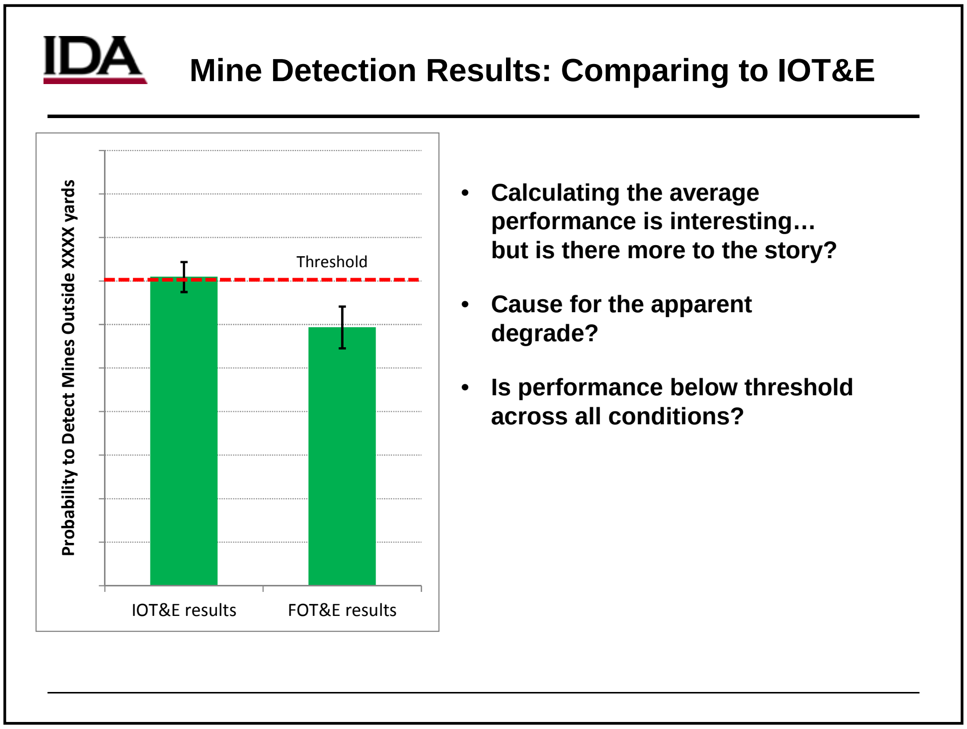


Example: *Virginia* Class Submarine's Mine Detection

Sail Array
Mine Detection Sonar

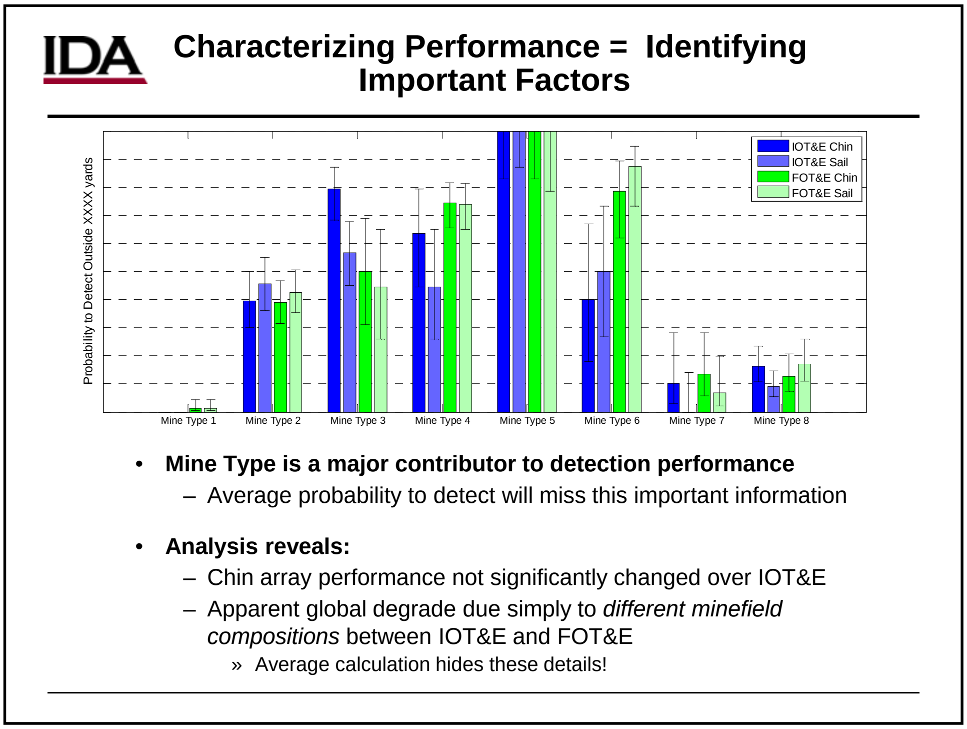
Chin Array
Mine Detection Sonar
(recently upgraded)

- **Mine Detection and Avoidance Test (FOT&E)**
  - Objectives:
    1. Measure and compare ***detection performance*** to IOT&E
    2. Characterize performance of a newly upgraded Chin array
    3. ….
    4. ….
  - Response variable: Probability to Detect Mines beyond a critical range
- **Test Design**
  - Inert threat-representative mine shapes planted in an area, submarine tasked to detect and avoid (multiple runs, geometries, pulse types)



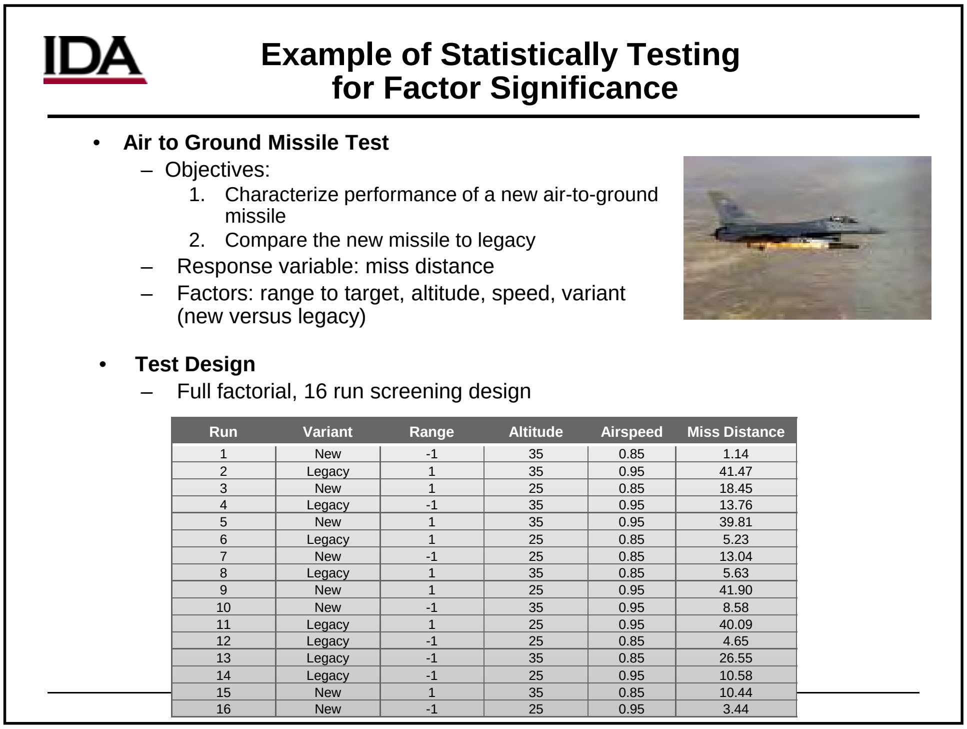
Mine Detection Results: Comparing to IOT&E

- Calculating the average performance is interesting… but is there more to the story?
- Cause for the apparent degrade?
- Is performance below threshold across all conditions?



Characterizing Performance = Identifying Important Factors

- **Mine Type is a major contributor to detection performance**
  - Average probability to detect will miss this important information
- **Analysis reveals:**
  - Chin array performance not significantly changed over IOT&E
  - Apparent global degrade due simply to *different minefield compositions* between IOT&E and FOT&E
    - » Average calculation hides these details!

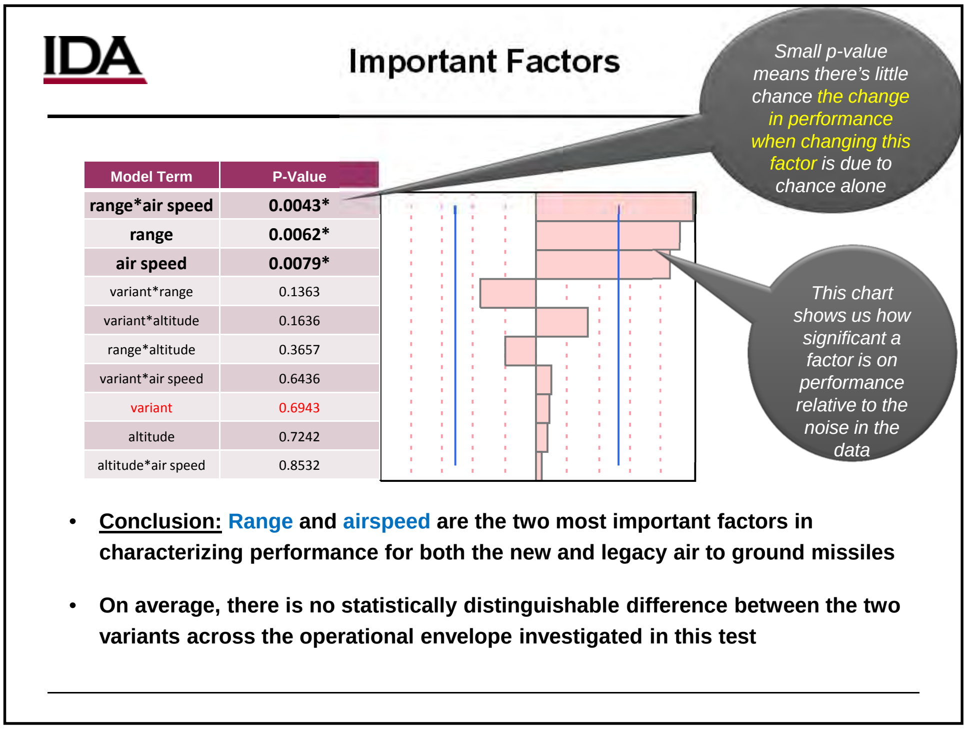


IDA

Example of Statistically Testing for Factor Significance

- **Air to Ground Missile Test**
  - Objectives:
    1. Characterize performance of a new air-to-ground missile
    2. Compare the new missile to legacy
  - Response variable: miss distance
  - Factors: range to target, altitude, speed, variant (new versus legacy)
- **Test Design**
  - Full factorial, 16 run screening design

Run	Variant	Range	Altitude	Airspeed	Miss Distance
1	New	-1	35	0.85	1.14
2	Legacy	1	35	0.95	41.47
3	New	1	25	0.85	18.45
4	Legacy	-1	35	0.95	13.76
5	New	1	35	0.95	39.81
6	Legacy	1	25	0.85	5.23
7	New	-1	25	0.85	13.04
8	Legacy	1	35	0.85	5.63
9	New	1	25	0.95	41.90
10	New	-1	35	0.95	8.58
11	Legacy	1	25	0.95	40.09
12	Legacy	-1	25	0.85	4.65
13	Legacy	-1	35	0.85	26.55
14	Legacy	-1	25	0.95	10.58
15	New	1	35	0.85	10.44
16	New	-1	25	0.95	3.44



IDA

Important Factors

Model Term	P-Value
range*air speed	**0.0043***
range	**0.0062***
air speed	**0.0079***
variant*range	0.1363
variant*altitude	0.1636
range*altitude	0.3657
variant*air speed	0.6436
variant	0.6943
altitude	0.7242
altitude*air speed	0.8532

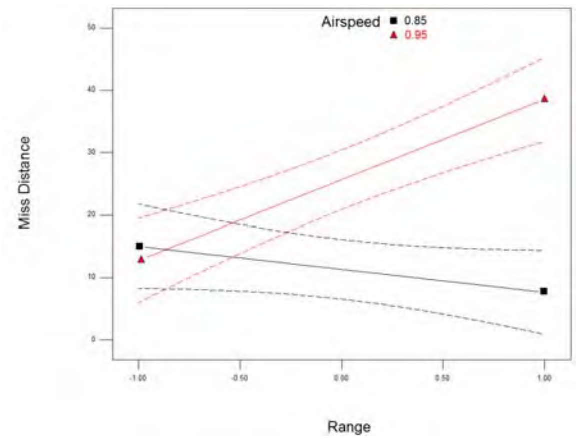
Small p-value means there's little chance the change in performance when changing this factor is due to chance alone

This chart shows us how significant a factor is on performance relative to the noise in the data

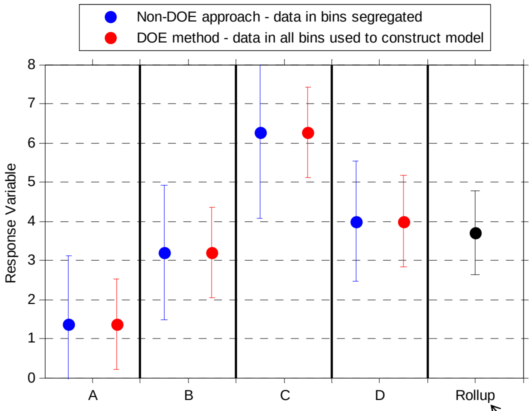
- **Conclusion: Range and airspeed are the two most important factors in characterizing performance for both the new and legacy air to ground missiles**
- **On average, there is no statistically distinguishable difference between the two variants across the operational envelope investigated in this test**

Graphical Presentation of Results

- **Interaction plots provide meaningful insights**
  - Miss distance increases with range at the higher airspeed



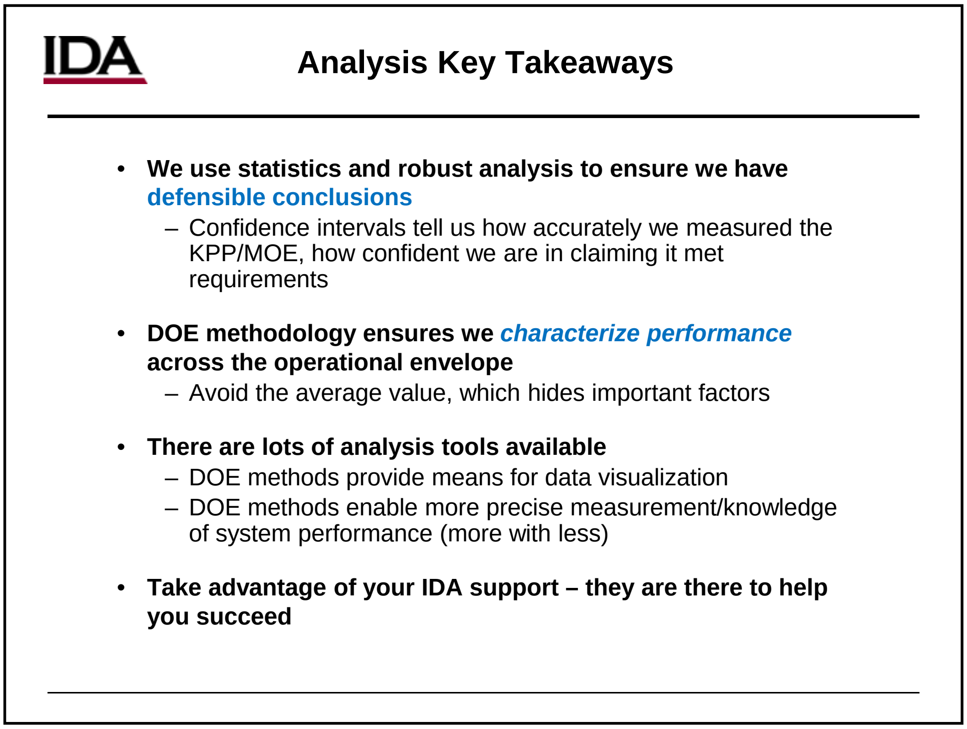
DOE versus Non-DOE Analysis



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

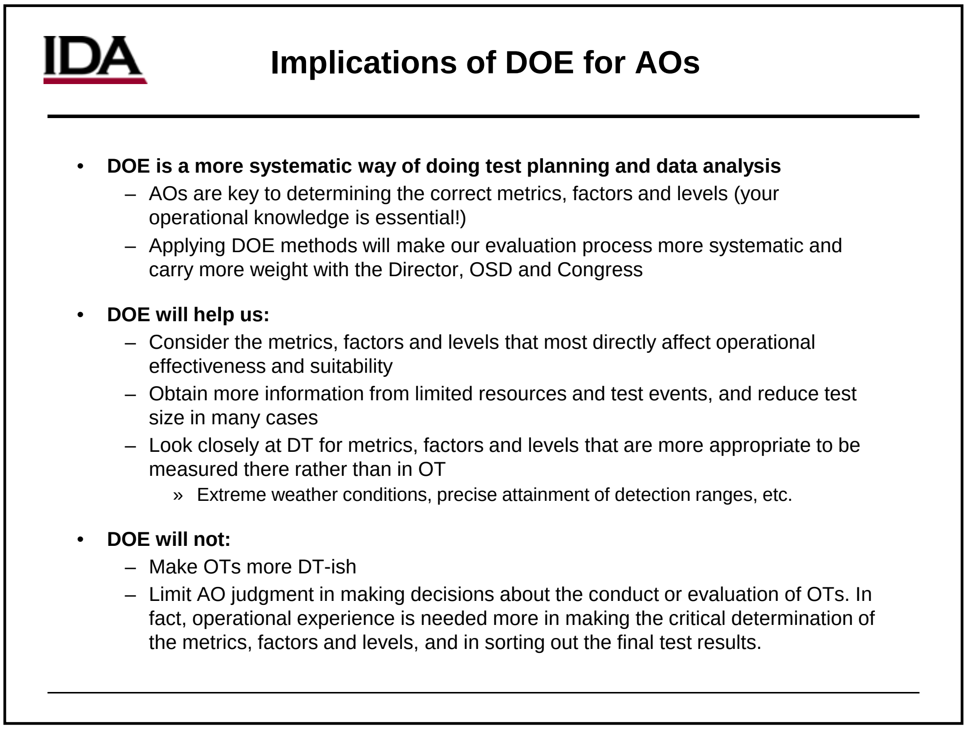
- **Non-DOE approach: calculate confidence intervals using only data collected under each condition**
- **DOE approach: construct a model (pool the data), use the model to estimate mean values in each condition**
  - Note the reduction in confidence interval size!
    - » In this case, intervals reduced 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***



Analysis Key Takeaways

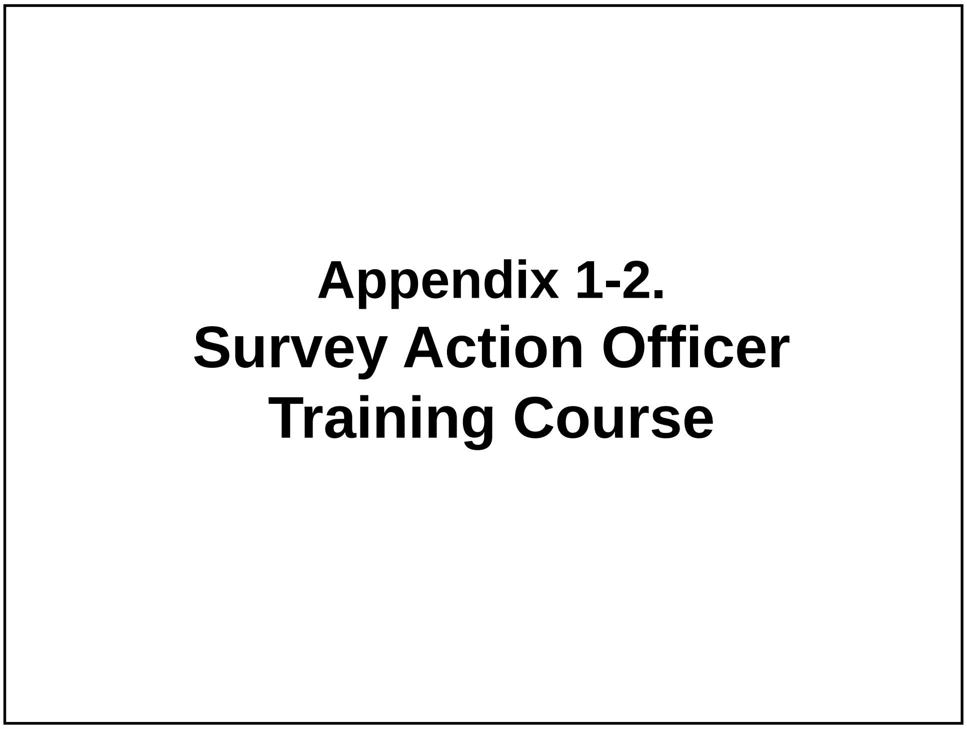
- **We use statistics and robust analysis to ensure we have defensible conclusions**
  - Confidence intervals tell us how accurately we measured the KPP/MOE, how confident we are in claiming it met requirements
- **DOE methodology ensures we *characterize performance* across the operational envelope**
  - Avoid the average value, which hides important factors
- **There are lots of analysis tools available**
  - DOE methods provide means for data visualization
  - DOE methods enable more precise measurement/knowledge of system performance (more with less)
- **Take advantage of your IDA support – they are there to help you succeed**



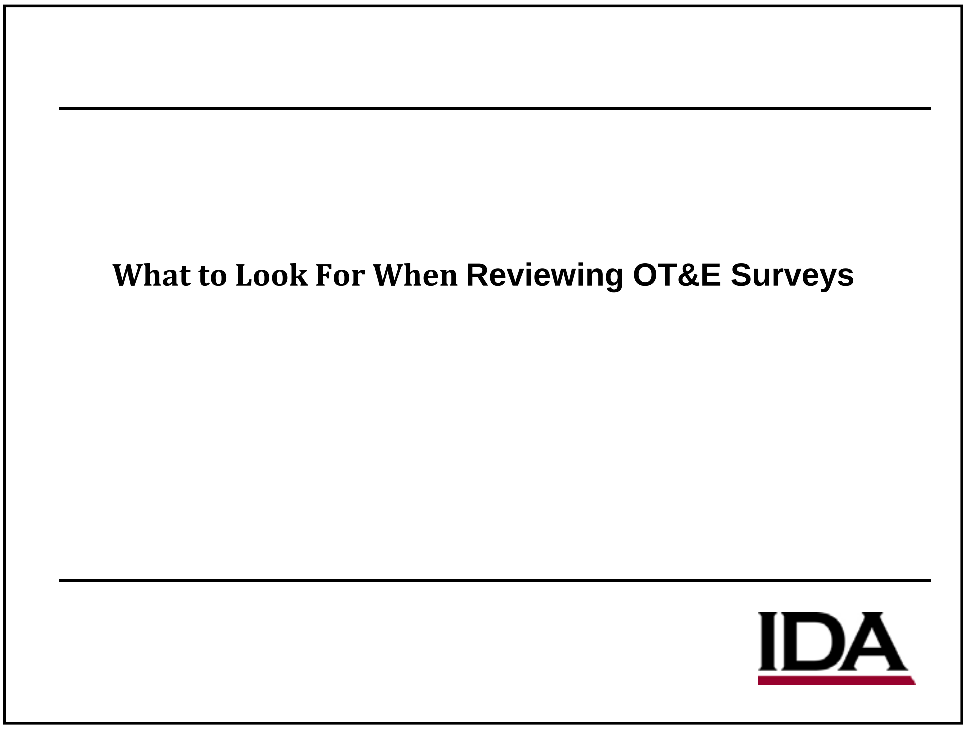
Implications of DOE for AOs

- **DOE is a more systematic way of doing test planning and data analysis**
  - AOs are key to determining the correct metrics, factors and levels (your operational knowledge is essential!)
  - Applying DOE methods will make our evaluation process more systematic and carry more weight with the Director, OSD and Congress
- **DOE will help us:**
  - Consider the metrics, factors and levels that most directly affect operational effectiveness and suitability
  - Obtain more information from limited resources and test events, and reduce test size in many cases
  - Look closely at DT for metrics, factors and levels that are more appropriate to be measured there rather than in OT
    - » Extreme weather conditions, precise attainment of detection ranges, etc.
- **DOE will not:**
  - Make OTs more DT-ish
  - Limit AO judgment in making decisions about the conduct or evaluation of OTs. In fact, operational experience is needed more in making the critical determination of the metrics, factors and levels, and in sorting out the final test results.

This page intentionally left blank.



Appendix 1-2.
Survey Action Officer Training Course

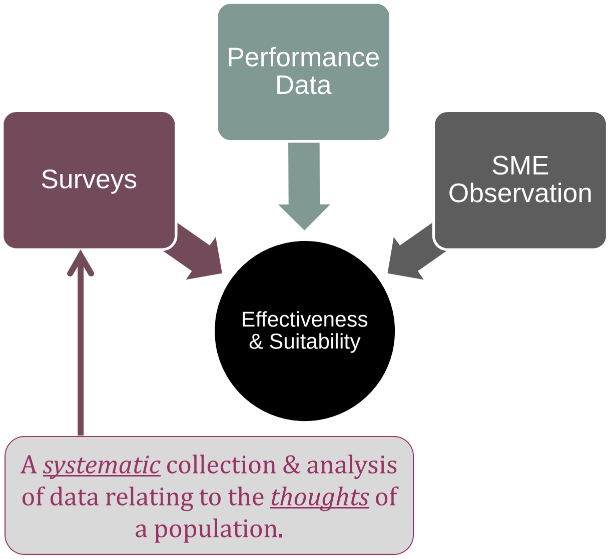


What to Look For When Reviewing OT&E Surveys

Goals

- When Measurement by Survey is Appropriate
- 5 Golden Rules for Writing Survey Items
- Appropriate & Effective Response Types
- Survey Formatting Best Practices
- System Usability Scale

Surveys: Important Part of OT&E



A *systematic* collection & analysis of data relating to the *thoughts* of a population.

Surveys DO NOT Measure:

- Physical Requirements
- Time
- Accuracy
- Situation Awareness

Surveys DO Measure:

- Usability*
- Workload**
- Demographics
- Opinions
- Observations
- Experience
- Judgments
- Knowledge

* System Usability Scale
** NASA TLX, Cooper-Harper

IDA

Surveys Are Not Always An Appropriate Measure

Joint High Speed Vessel

- **"JHSV has protective clothing for every crew member."**
  - Count the protective clothing & compare to number of crew.
- **"Engine exhaust levels in the mission bay do not exceed safety limits ..."**
  - Measure exhaust levels with Portable Emissions Measurement System & compare to safety limits
- **"Visual alarms … do not interfere with night vision."**
  - MIL-STD-1472 has requirements for brightness (measured by photometer), colors, & location of displays for use at night.
- **"Temperatures in primary work spaces were adequate."**
  - MIL-STD-1472 has requirements for temperatures (measured by thermometer).

IDA

5 Golden Rules of Writing Items

Singularity:	Only 1 Idea Per Question
User Friendly:	Items Do Not Require a Lot of Thought or Interpretation (e.g., short, clear, specific)
Neutrality:	Items Do Not Imply Value Judgments Items Are Not Emotionally Charged
Knowledge Liability:	Respondents Have Enough Information to Answer the Question
Independence:	Responses Will Not Affect Responses to Other Questions

IDA

Golden Rule Violations (1 of 2)

- Singularity: **“JHSV engineering drawings, commercial technical manuals, & technical support data are adequate for vessel operations & maintenance actions.”**
  - » The drawings could support vessel operations but not maintenance actions
  - » The manuals could support maintenance actions but not vessel operations.
- – 6 questions; e.g., Engineering drawings were helpful in conducting vessel operations.

- User Friendly: (JSF) **“Based on your experience, does any aspect of the aircraft, equipment, documentation, or procedures have the potential to compromise safety?”**
  - » You want the respondent to think about the answer, not what you are asking.
  - – Were there any potential safety issues?

- Neutrality: **“…The aviation enhancements of LHA 6 will more than offset the lack of surface connectors …”**
  - » Respondent knows there is a right answer to this question.
  - – The number of surface conductors was adequate.

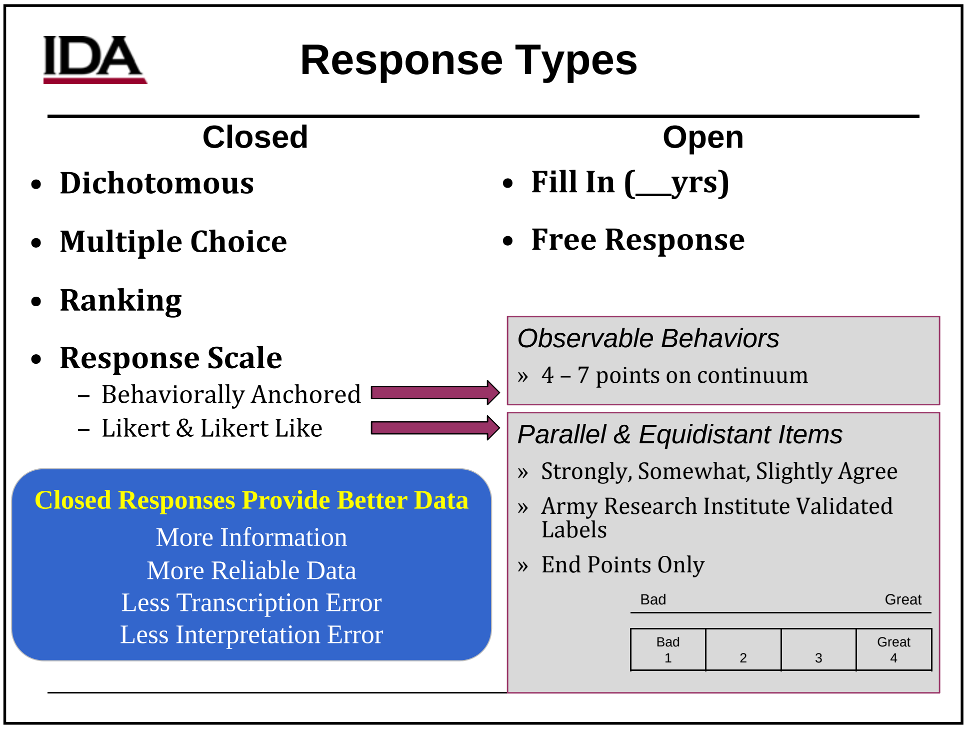
IDA

Golden Rule Violations (2 of 2)

- "Knowledge Liability: (ALR-69A) **“Were you able to identify each threat…”**
  - » If the respondent didn’t see the threat, how would s/he know it was there?
  - – Do not ask. Rather use SME observations or performance data.

- Independence (1): **“Based on your responses above, rate the acceptability of the ALR-69A”**
  - » If it is based on other questions, then it is redundant.
  - » What if the above questions were positive, but other things were wrong with the interface?
  - – The ALR-69A is easy to use.

- Independence (2): (JSF) **“(If not totally adequate) Rate the degree this deficiency impedes or degrades F-35A Block 1A.1 training effectiveness.”**
  - » Biases respondent to positive response on previous question.
  - – Utilize follow-up interviews to obtain information
  - – What changes would most improve training effectiveness?



IDA

Response Types

Closed

- **Dichotomous**
- **Multiple Choice**
- **Ranking**
- **Response Scale**
  - – Behaviorally Anchored
  - – Likert & Likert Like

Closed Responses Provide Better Data
More Information
More Reliable Data
Less Transcription Error
Less Interpretation Error

Open

- **Fill In (__yrs)**
- **Free Response**

Observable Behaviors

» 4 – 7 points on continuum

Parallel & Equidistant Items

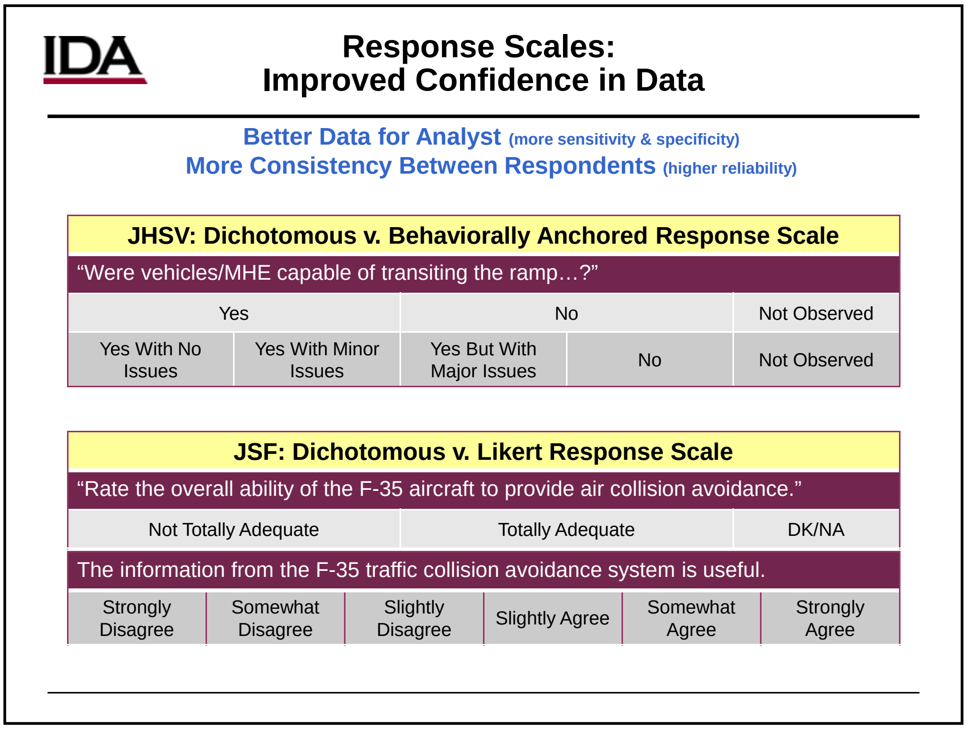
» Strongly, Somewhat, Slightly Agree

» Army Research Institute Validated Labels

» End Points Only

Bad Great

Bad 1	2	3	Great 4



IDA

Response Scales: Improved Confidence in Data

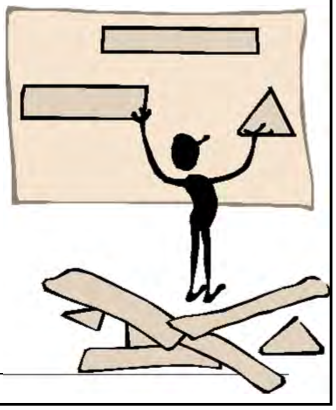
Better Data for Analyst (more sensitivity & specificity)
More Consistency Between Respondents (higher reliability)

JHSV: Dichotomous v. Behaviorally Anchored Response Scale				
"Were vehicles/MHE capable of transiting the ramp…?"				
Yes		No		Not Observed
Yes With No Issues	Yes With Minor Issues	Yes But With Major Issues	No	Not Observed

JSF: Dichotomous v. Likert Response Scale					
"Rate the overall ability of the F-35 aircraft to provide air collision avoidance."					
Not Totally Adequate			Totally Adequate		DK/NA
The information from the F-35 traffic collision avoidance system is useful.					
Strongly Disagree	Somewhat Disagree	Slightly Disagree	Slightly Agree	Somewhat Agree	Strongly Agree

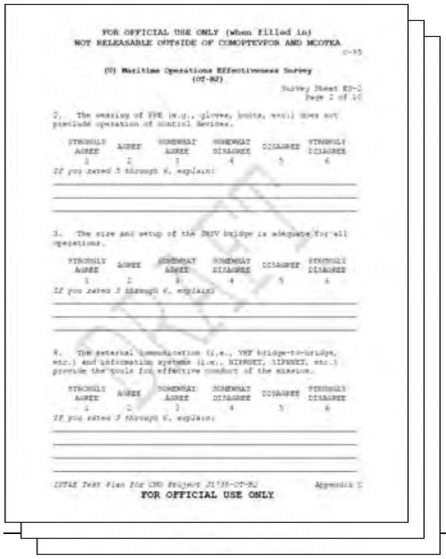
Formatting Surveys

- **Provide Brief Clear Introduction**
- **Logically Ordered Questions**
  - Grouped into Sections
  - Begin with interesting items that are clearly connected to the goals of the survey
  - Follow order of events
  - Within section: start generally and get more specific
- **Change Response Types** (but not too frequently)
  - Open ended always last
- **Alternate Response Scale**
  - If survey is very long
  - If respondent motivation is low
- **Consider Data Transfer and Analysis**
  - e.g., quick look questions first



Format Affects Respondent Motivation

More Than 3 Pages



FOR OFFICIAL USE ONLY (when filled in)
NOT RELEASABLE OUTSIDE OF COMOPTEVFOR AND MCOTEA

C-35

(U) Maritime Operations Effectiveness Survey
(OT-B2)

Survey Sheet ES-2
Page 2 of 10

2. The wearing of PPE (e.g., gloves, boots, etc.) does not preclude operation of control devices.

STRONGLY AGREE	AGREE	SOMEWHAT AGREE	SOMEWHAT DISAGREE	DISAGREE	STRONGLY DISAGREE
1	2	3	4	5	6

If you rated 3 through 6, explain:

3. The size and setup of the JHSV bridge is adequate for all operations.

STRONGLY AGREE	AGREE	SOMEWHAT AGREE	SOMEWHAT DISAGREE	DISAGREE	STRONGLY DISAGREE
1	2	3	4	5	6

If you rated 3 through 6, explain:

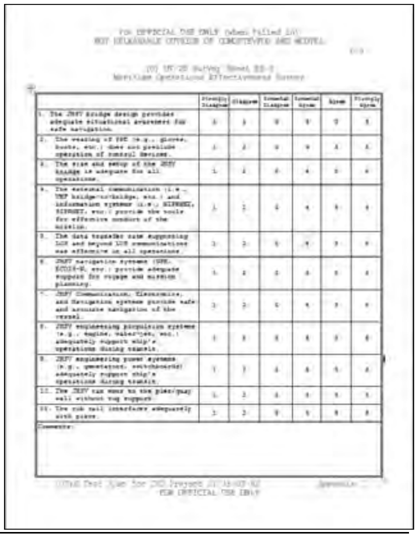
4. The external communication (i.e., VHF bridge-to-bridge, etc.) and information systems (i.e., NIPRNET, SIPRNET, etc.) provide the tools for effective conduct of the mission.

STRONGLY AGREE	AGREE	SOMEWHAT AGREE	SOMEWHAT DISAGREE	DISAGREE	STRONGLY DISAGREE
1	2	3	4	5	6

If you rated 3 through 6, explain:

JSTAS Test Plan for CNO Project 1735-OT-B2 Appendix C
FOR OFFICIAL USE ONLY

1 Page



FOR OFFICIAL USE ONLY (when filled in)
NOT RELEASABLE OUTSIDE OF COMOPTEVFOR AND MCOTEA

(U) DT/OT Survey Sheet ES-1
Maritime Operational Effectiveness Survey

	Strongly Disagree	Disagree	Somewhat Disagree	Somewhat Agree	Agree	Strongly Agree
1. The JHSV bridge design provides adequate situational awareness for safe navigation.	1	2	3	4	5	6
2. The wearing of PPE (e.g., gloves, boots, etc.) does not preclude operation of control devices.	1	2	3	4	5	6
3. The size and setup of the JHSV bridge is adequate for all operations.	1	2	3	4	5	6
4. The external communication (i.e., VHF bridge-to-bridge, etc.) and information systems (i.e., NIPRNET, SIPRNET, etc.) provide the tools for effective conduct of the mission.	1	2	3	4	5	6
5. The data transfer rate supporting LOS and beyond LOS communications was effective in all operations.	1	2	3	4	5	6
6. JHSV navigation systems (GPS, ECDIS-N, etc.) provide adequate support for voyage and mission planning.	1	2	3	4	5	6
7. JHSV Communication, Electronics, and Navigation systems provide safe and accurate navigation of the vessel.	1	2	3	4	5	6
8. JHSV engineering propulsion systems (e.g., engine, waterjet, etc.) adequately support ship's operations during transit.	1	2	3	4	5	6
9. JHSV engineering power systems (e.g., generators, switchboards) adequately support ship's operations during transit.	1	2	3	4	5	6
10. The JHSV can moor to the pier/quay wall without tug support.	1	2	3	4	5	6
11. The rub rail interfaces adequately with piers.	1	2	3	4	5	6
Comments:						

JSTAS Test Plan for CNO Project 1735-OT-B2 Appendix C
FOR OFFICIAL USE ONLY

IDA

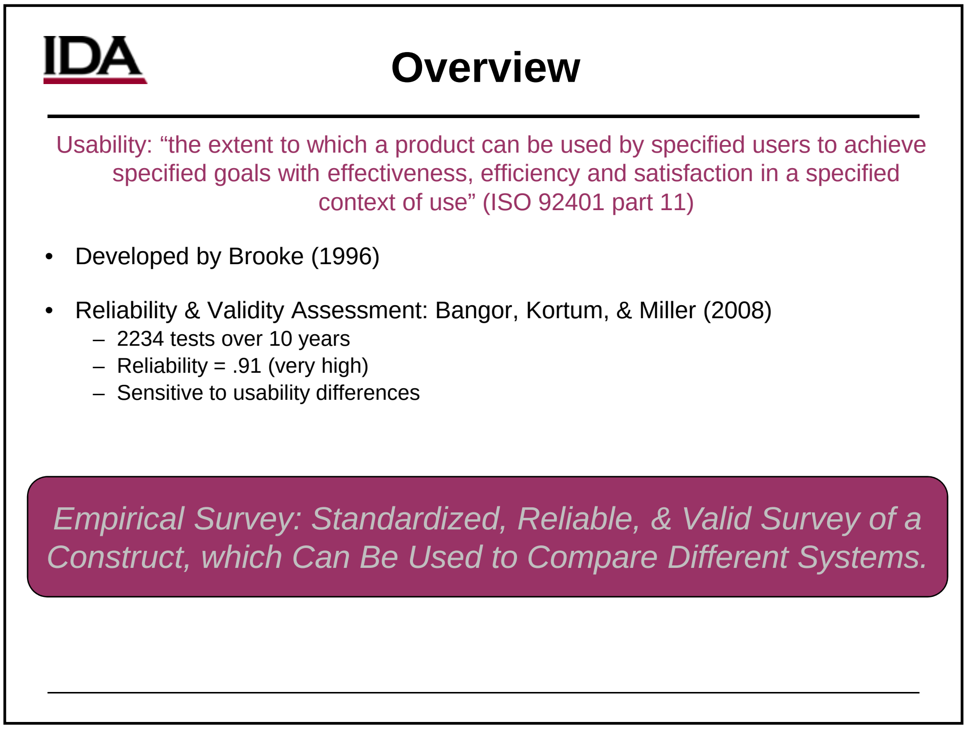
How to Review a Survey

- **Best Practices Met**
  - 5 Golden Rules
  - Appropriate Response Types
  - Effective Closed Response Options
  - Logical Question Order
- **Requirements Met**
  - Is Survey Best Method of Measuring?
  - Will Information Be Obtained?
  - Will the Collected Data Be Analyzable?
  - Is the Survey Length Appropriate for Conditions?
    (e.g., after every run or once during test)
  - Is it written at a 6th Grade Reading Level?
    (in Word: Flesch-Kincaid Grade Level)

Minimize Burden on Respondent; Maximize Data Accuracy

IDA

SYSTEM USABILITY SCALE (SUS)

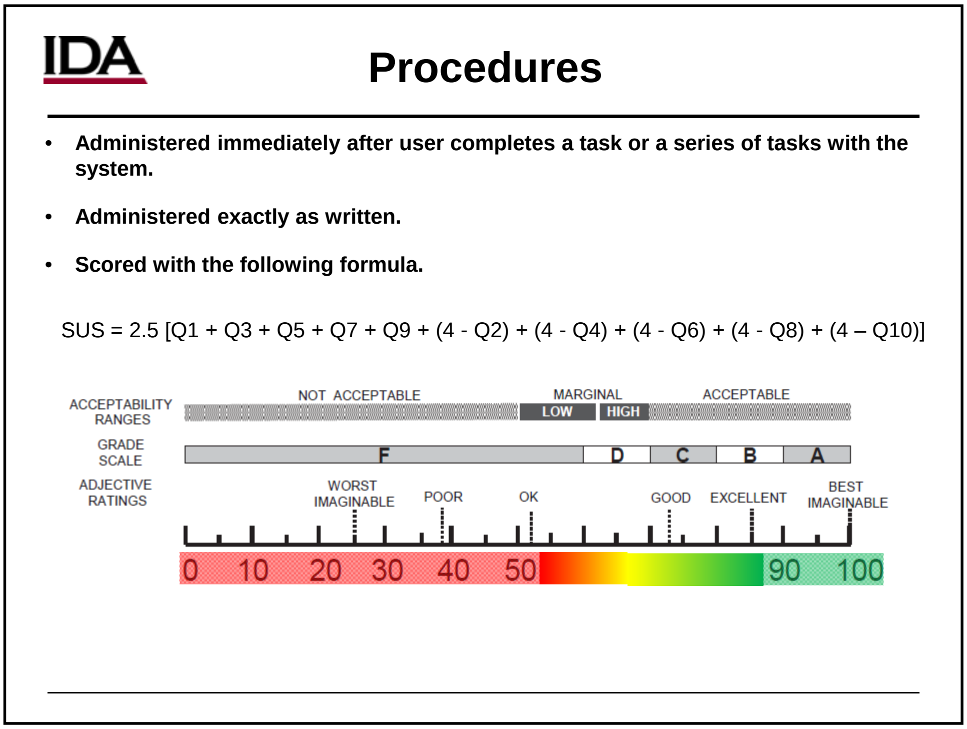


Overview

Usability: "the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use" (ISO 92401 part 11)

- Developed by Brooke (1996)
- Reliability & Validity Assessment: Bangor, Kortum, & Miller (2008)
  - 2234 tests over 10 years
  - Reliability = .91 (very high)
  - Sensitive to usability differences

Empirical Survey: Standardized, Reliable, & Valid Survey of a Construct, which Can Be Used to Compare Different Systems.

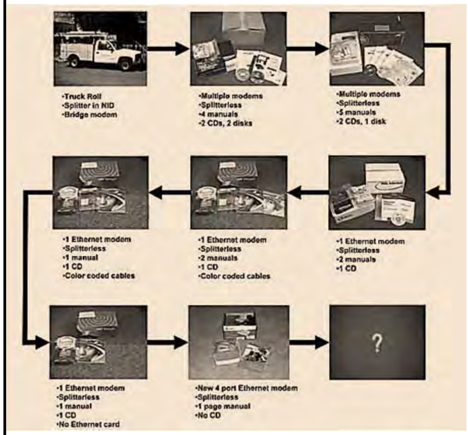


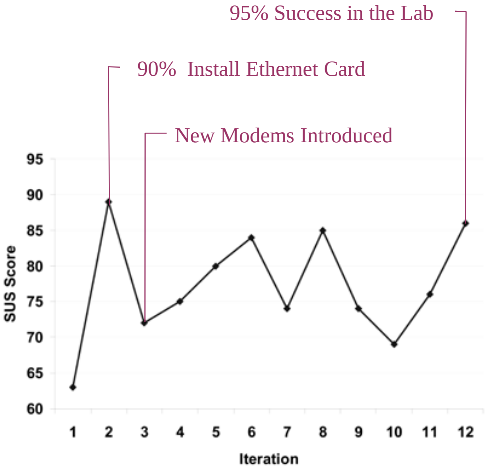
Procedures

- **Administered immediately after user completes a task or a series of tasks with the system.**
- **Administered exactly as written.**
- **Scored with the following formula.**

SUS = 2.5 [Q1 + Q3 + Q5 + Q7 + Q9 + (4 - Q2) + (4 - Q4) + (4 - Q6) + (4 - Q8) + (4 – Q10)]

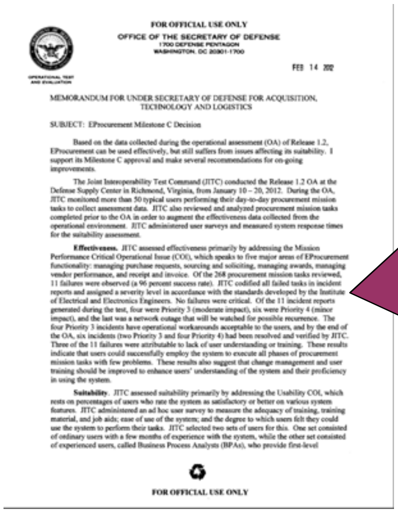
Using SUS to Compare Versions: DSL Self Install





Kortum, P., Grier, R. & Sullivan, M. (2009). DSL Self-installation: From Impossibility to Ubiquity. *Interfaces, 80*, 12-14.

SUS in OT&E



FOR OFFICIAL USE ONLY

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

OPERATIONAL TEST AND EVALUATION

FEB 14 2012

MEMORANDUM FOR UNDER SECRETARY OF DEFENSE FOR ACQUISITION, TECHNOLOGY AND LOGISTICS

SUBJECT: EProcurement Milestone C Decision

Based on the data collected during the operational assessment (OA) of Release 1.2, EProcurement can be used effectively, but still suffers from issues affecting its suitability. I support its Milestone C approval and make several recommendations for on-going improvements.

The Joint Interoperability Test Command (JITC) conducted the Release 1.2 OA at the Defense Supply Center in Richmond, Virginia, from January 10 – 20, 2012. During the OA, JITC monitored more than 50 typical users performing their day-to-day procurement mission tasks to collect assessment data. JITC also reviewed and analyzed procurement mission tasks completed prior to the OA in order to augment the effectiveness data collected from the operational environment. JITC administered user surveys and measured system response times for the suitability assessment.

Effectiveness. JITC assessed effectiveness primarily by addressing the Mission Performance Critical Operational Issue (COI), which speaks to five major areas of EProcurement functionality: managing purchase requests, sourcing and soliciting, managing awards, managing vendor performance, and receipt and invoice. Of the 268 procurement mission tasks reviewed, 11 failures were observed (a 96 percent success rate). JITC codified all failed tasks in incident reports and assigned a severity level in accordance with the standards developed by the Institute of Electrical and Electronics Engineers. No failures were critical. Of the 11 incident reports generated during the test, four were Priority 3 (moderate impact), six were Priority 4 (minor impact), and the last was a network outage that will be watched for possible recurrence. The four Priority 3 incidents have operational workarounds acceptable to the users, and by the end of the OA, six incidents (two Priority 3 and four Priority 4) had been resolved and verified by JITC. Three of the 11 failures were attributable to lack of user understanding or training. These results indicate that users could successfully employ the system to execute all phases of procurement mission tasks with few problems. These results also suggest that change management and user training should be improved to enhance users' understanding of the system and their proficiency in using the system.

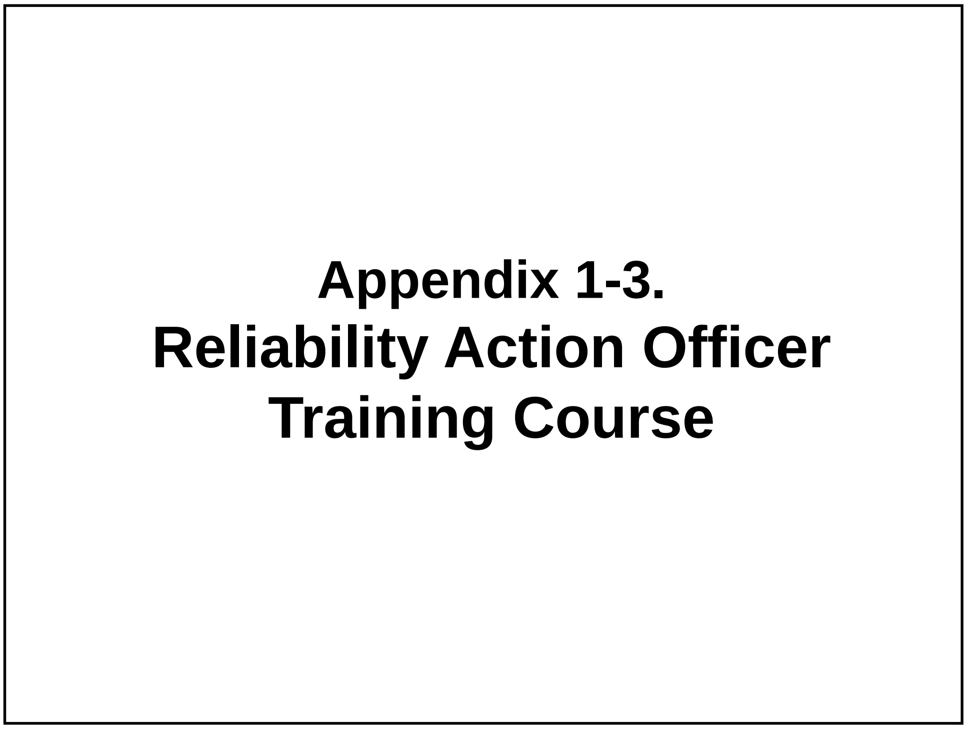
Suitability. JITC assessed suitability primarily by addressing the Usability COI, which rests on percentages of users who rate the system as satisfactory or better on various system features. JITC administered an ad hoc user survey to measure the adequacy of training, training material, and job aids; ease of use of the system; and the degree to which users felt they could use the system to perform their tasks. JITC selected two sets of users for this. One set consisted of ordinary users with a few months of experience with the system, while the other set consisted of experienced users, called Business Process Analysts (BPAs), who provide first-level

FOR OFFICIAL USE ONLY

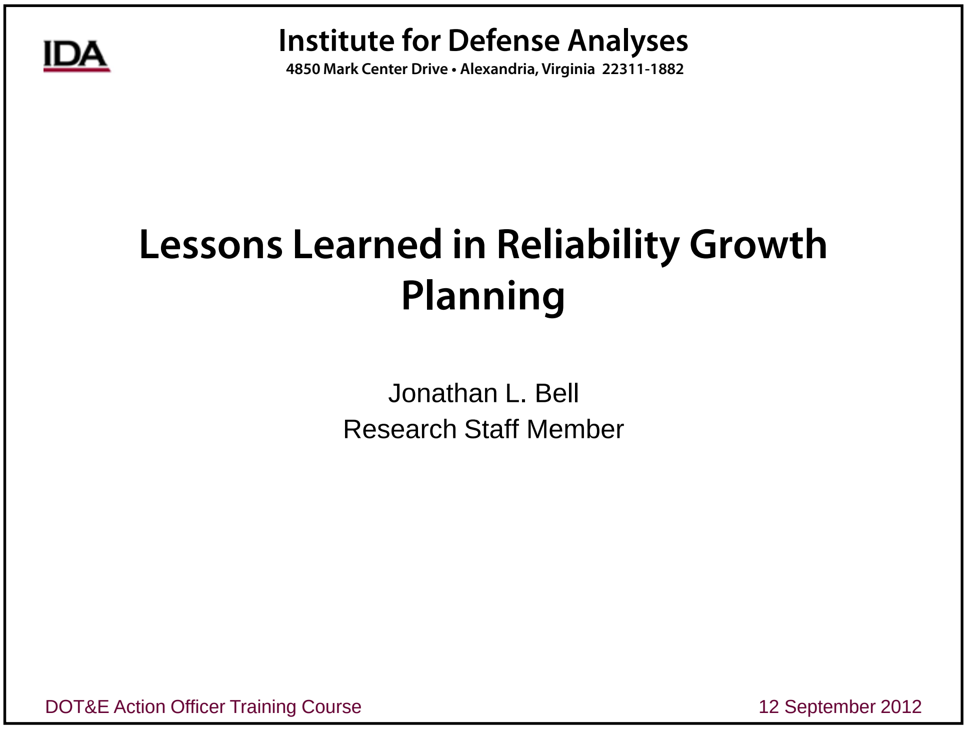
Memo from Dr. Gilmore to USD AT&L

JITC applied a professionally designed and academically studied survey, the System Usability Scale (SUS), to measure ease of use and the degree to which users felt they could use the system to perform their tasks (see attached references).

This page intentionally left blank.



Appendix 1-3.
Reliability Action Officer Training Course

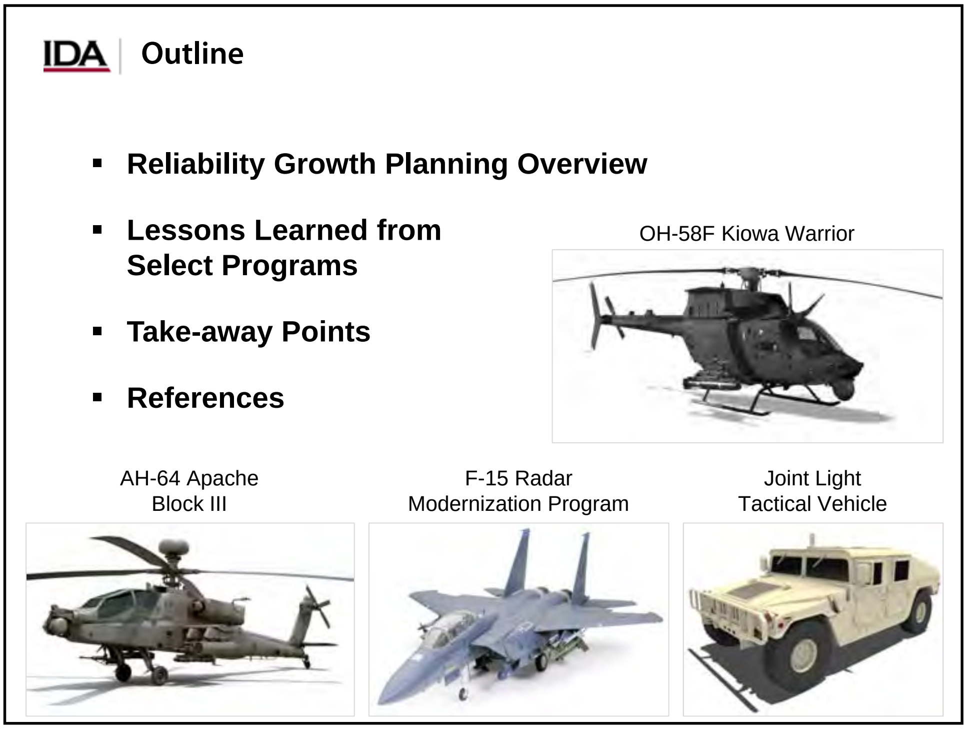


Institute for Defense Analyses
4850 Mark Center Drive • Alexandria, Virginia 22311-1882

Lessons Learned in Reliability Growth Planning

Jonathan L. Bell
Research Staff Member

DOT&E Action Officer Training Course 12 September 2012



IDA | Outline

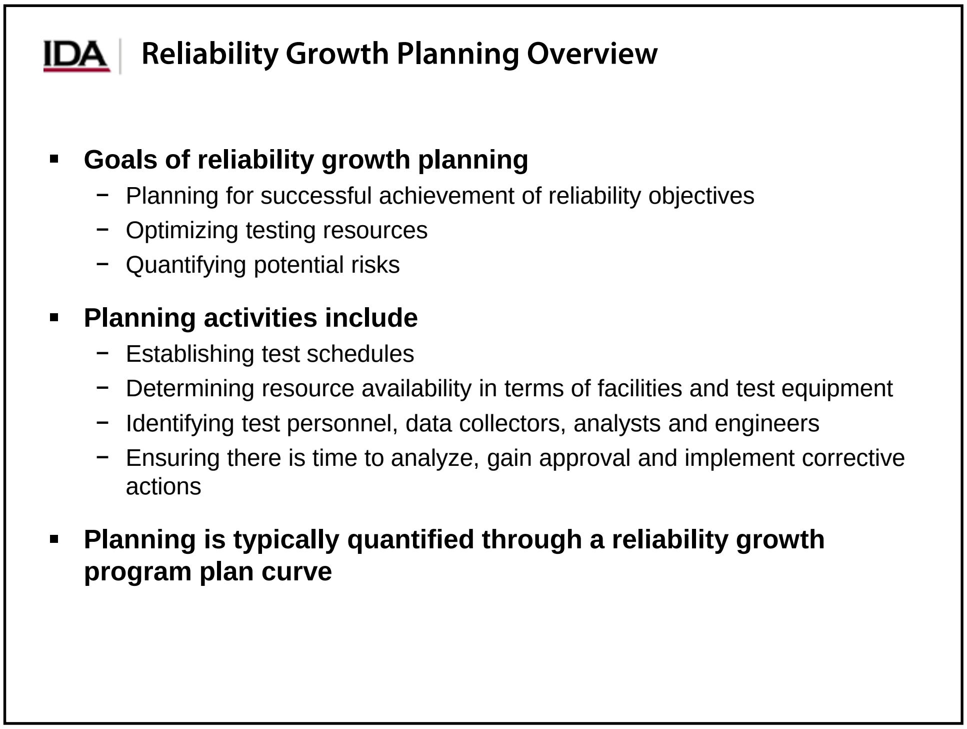
- **Reliability Growth Planning Overview**
- **Lessons Learned from Select Programs**
- **Take-away Points**
- **References**

OH-58F Kiowa Warrior

AH-64 Apache
Block III

F-15 Radar
Modernization Program

Joint Light
Tactical Vehicle

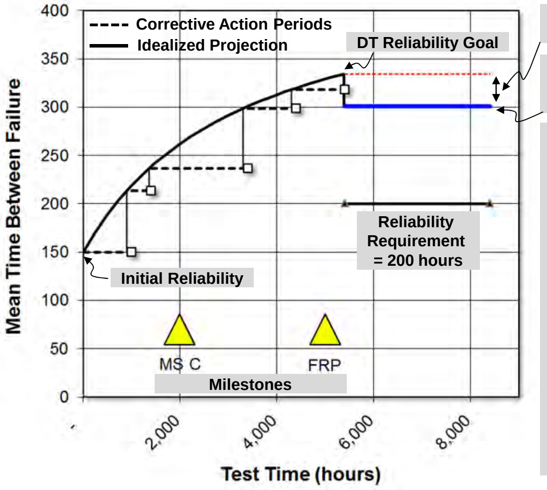


IDA | Reliability Growth Planning Overview

- **Goals of reliability growth planning**
  - Planning for successful achievement of reliability objectives
  - Optimizing testing resources
  - Quantifying potential risks
- **Planning activities include**
  - Establishing test schedules
  - Determining resource availability in terms of facilities and test equipment
  - Identifying test personnel, data collectors, analysts and engineers
  - Ensuring there is time to analyze, gain approval and implement corrective actions
- **Planning is typically quantified through a reliability growth program plan curve**

Reliability Growth Planning Overview

- **Typical Projection Methodology (PM2) Reliability Planning Curve**



Planned 10% reduction in DT MTBF due to OT environment

IOT&E planned reliability of 300 hours MTBF for Demonstrating 200 hours MTBF with 80% Confidence

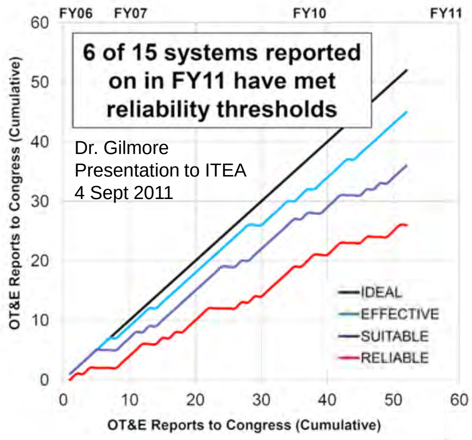
Other Model Parameters

- **Management Strategy - fraction of the initial system failure intensity due to failure modes that would receive corrective action**
- **Average Fix Effectiveness Factor - the reduction in the failure rate due to implementation of a corrective actions**
- **Growth Potential - theoretical upper limit on reliability which corresponds to the reliability that would result if all B-modes were surfaced and fixed with the realized failure mode FEF values**
- **A and B modes - failure modes that will (B modes) or will not (A modes) be addressed via corrective action**

"Department of Defense Handbook Reliability Growth Management," MIL-HDBK-189C, 24 June 2011.

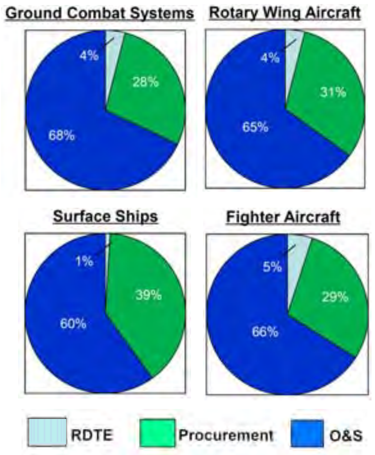
Reliability Growth Planning Overview

- **Why do it?**
  - Improve system reliability
  - Reduce O&S cost

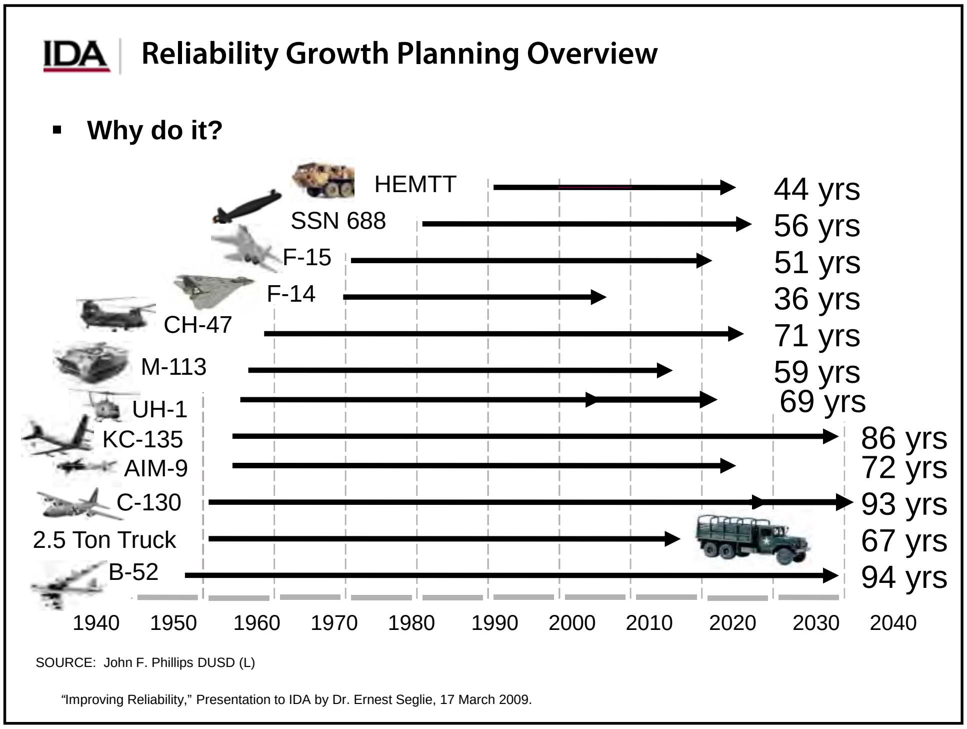


6 of 15 systems reported on in FY11 have met reliability thresholds

Dr. Gilmore
Presentation to ITEA
4 Sept 2011



"AEC/AMSAA Reliability Short Course Notes," 21 August 2011.

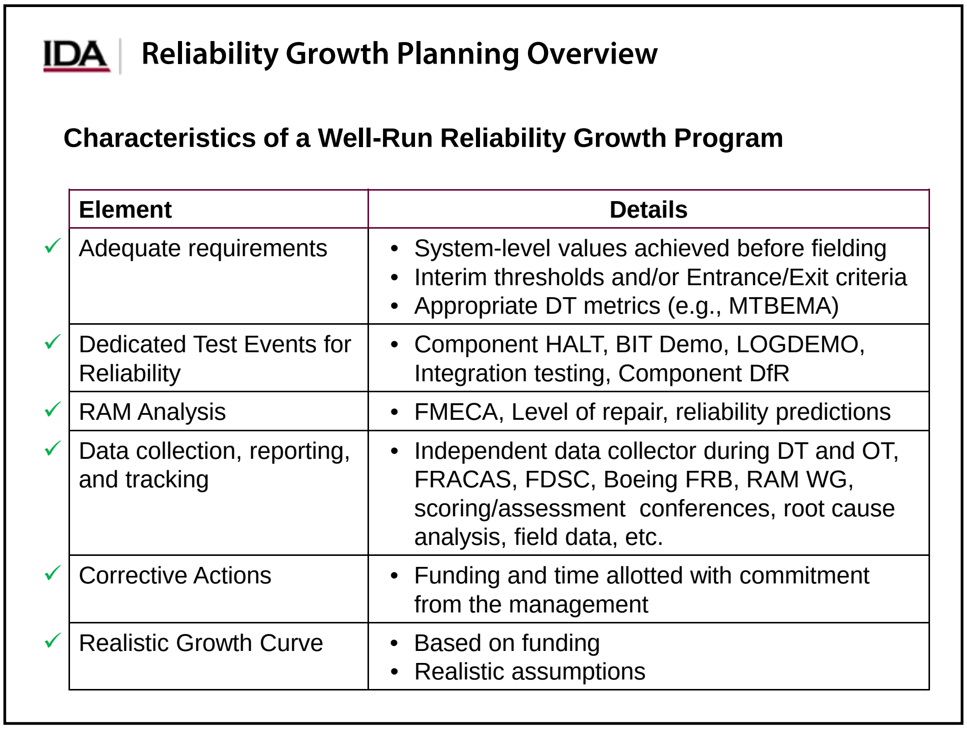


Reliability Growth Planning Overview

- **Why do it?**

SOURCE: John F. Phillips DUSD (L)

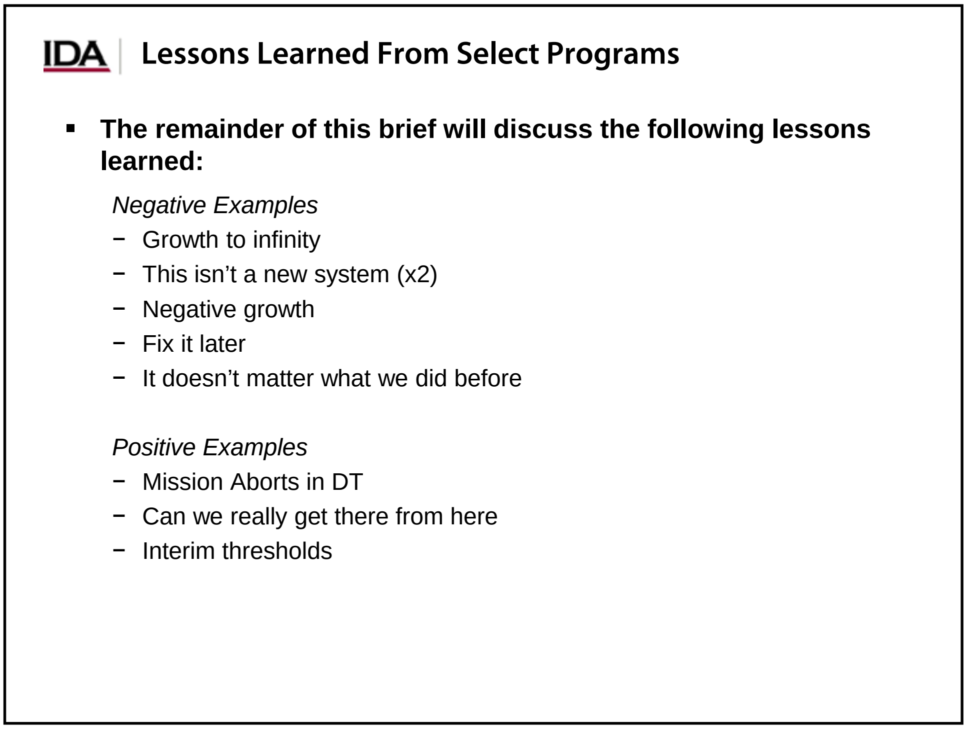
"Improving Reliability," Presentation to IDA by Dr. Ernest Seglie, 17 March 2009.



Reliability Growth Planning Overview

Characteristics of a Well-Run Reliability Growth Program

Element	Details
✓ Adequate requirements	• System-level values achieved before fielding • Interim thresholds and/or Entrance/Exit criteria • Appropriate DT metrics (e.g., MTBEMA)
✓ Dedicated Test Events for Reliability	• Component HALT, BIT Demo, LOGDEMO, Integration testing, Component DfR
✓ RAM Analysis	• FMECA, Level of repair, reliability predictions
✓ Data collection, reporting, and tracking	• Independent data collector during DT and OT, FRACAS, FDSC, Boeing FRB, RAM WG, scoring/assessment conferences, root cause analysis, field data, etc.
✓ Corrective Actions	• Funding and time allotted with commitment from the management
✓ Realistic Growth Curve	• Based on funding • Realistic assumptions

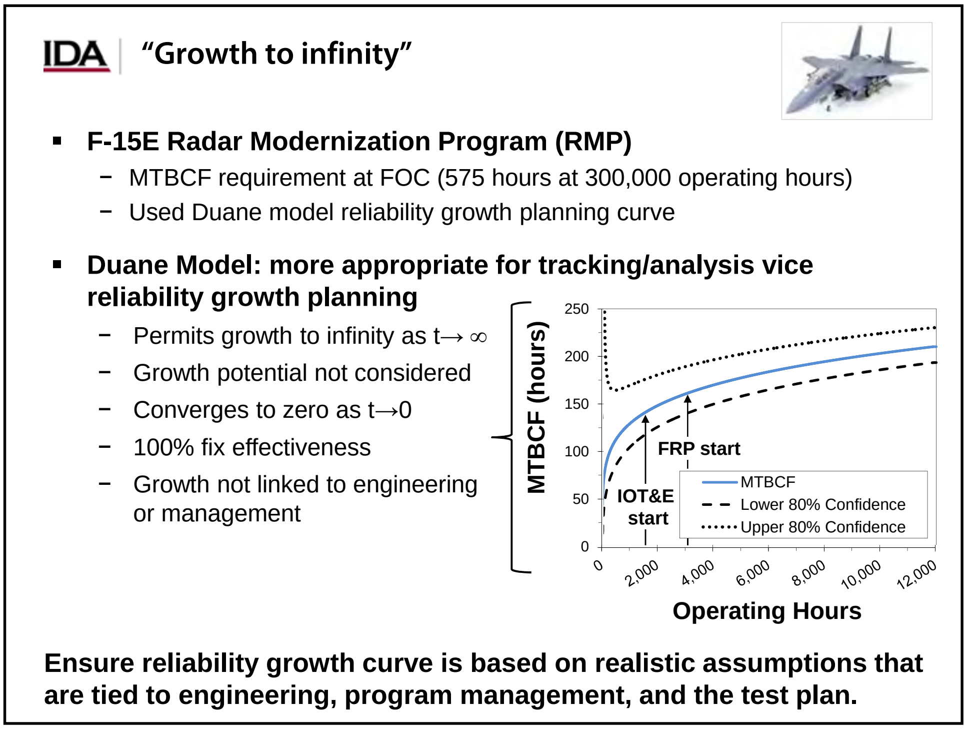


Lessons Learned From Select Programs

- **The remainder of this brief will discuss the following lessons learned:**

  Negative Examples
  - Growth to infinity
  - This isn't a new system (x2)
  - Negative growth
  - Fix it later
  - It doesn't matter what we did before

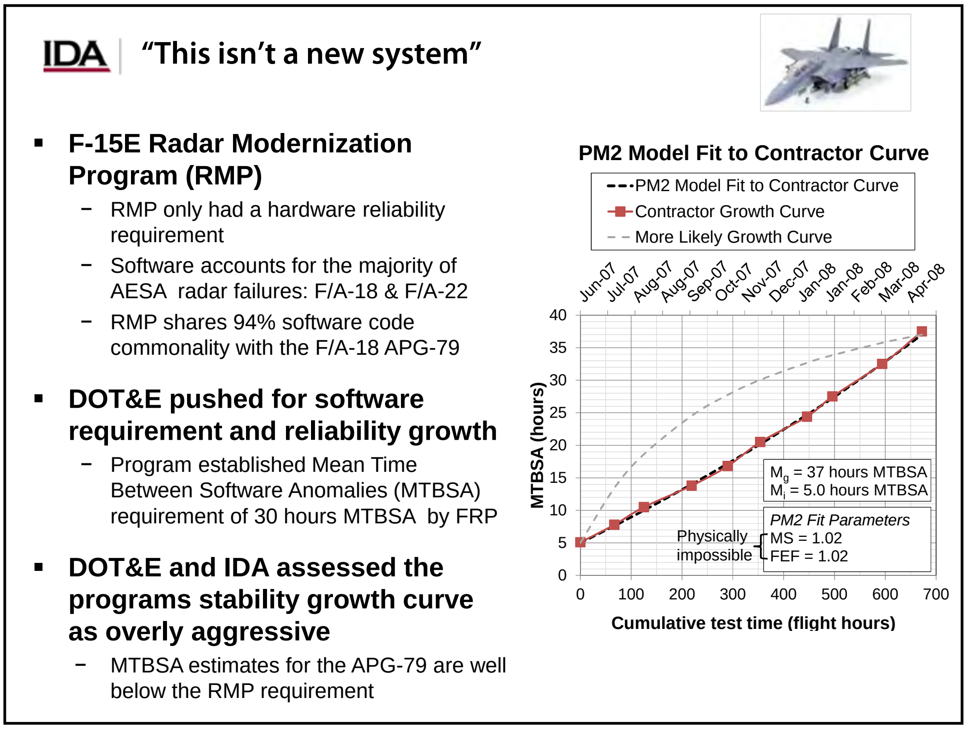
  Positive Examples
  - Mission Aborts in DT
  - Can we really get there from here
  - Interim thresholds



"Growth to infinity"

- **F-15E Radar Modernization Program (RMP)**
  - MTBCF requirement at FOC (575 hours at 300,000 operating hours)
  - Used Duane model reliability growth planning curve
- **Duane Model: more appropriate for tracking/analysis vice reliability growth planning**
  - Permits growth to infinity as t→ ∞
  - Growth potential not considered
  - Converges to zero as t→0
  - 100% fix effectiveness
  - Growth not linked to engineering or management

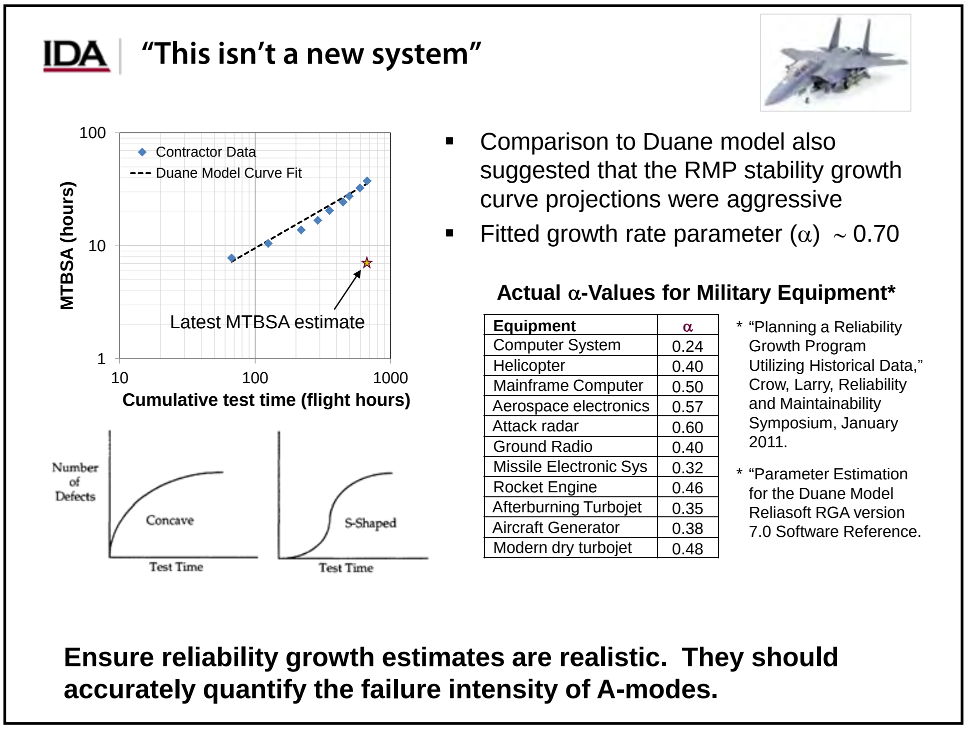
Ensure reliability growth curve is based on realistic assumptions that are tied to engineering, program management, and the test plan.



IDA | “This isn’t a new system”

- **F-15E Radar Modernization Program (RMP)**
  - RMP only had a hardware reliability requirement
  - Software accounts for the majority of AESA radar failures: F/A-18 & F/A-22
  - RMP shares 94% software code commonality with the F/A-18 APG-79
- **DOT&E pushed for software requirement and reliability growth**
  - Program established Mean Time Between Software Anomalies (MTBSA) requirement of 30 hours MTBSA by FRP
- **DOT&E and IDA assessed the programs stability growth curve as overly aggressive**
  - MTBSA estimates for the APG-79 are well below the RMP requirement

PM2 Model Fit to Contractor Curve



IDA | “This isn’t a new system”

- Comparison to Duane model also suggested that the RMP stability growth curve projections were aggressive
- Fitted growth rate parameter (α) ~ 0.70

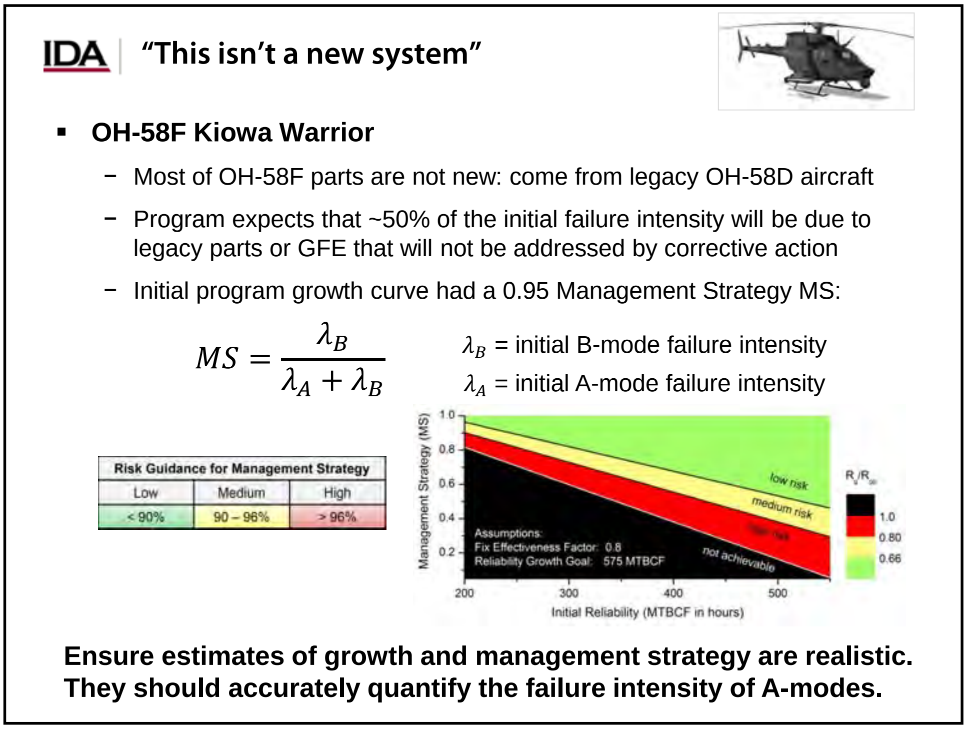
Actual α-Values for Military Equipment*

Equipment	α
Computer System	0.24
Helicopter	0.40
Mainframe Computer	0.50
Aerospace electronics	0.57
Attack radar	0.60
Ground Radio	0.40
Missile Electronic Sys	0.32
Rocket Engine	0.46
Afterburning Turbojet	0.35
Aircraft Generator	0.38
Modern dry turbojet	0.48

* “Planning a Reliability Growth Program Utilizing Historical Data,” Crow, Larry, Reliability and Maintainability Symposium, January 2011.

* “Parameter Estimation for the Duane Model Reliasoft RGA version 7.0 Software Reference.

Ensure reliability growth estimates are realistic. They should accurately quantify the failure intensity of A-modes.



IDA | “This isn’t a new system”

- **OH-58F Kiowa Warrior**
  - Most of OH-58F parts are not new: come from legacy OH-58D aircraft
  - Program expects that ~50% of the initial failure intensity will be due to legacy parts or GFE that will not be addressed by corrective action
  - Initial program growth curve had a 0.95 Management Strategy MS:

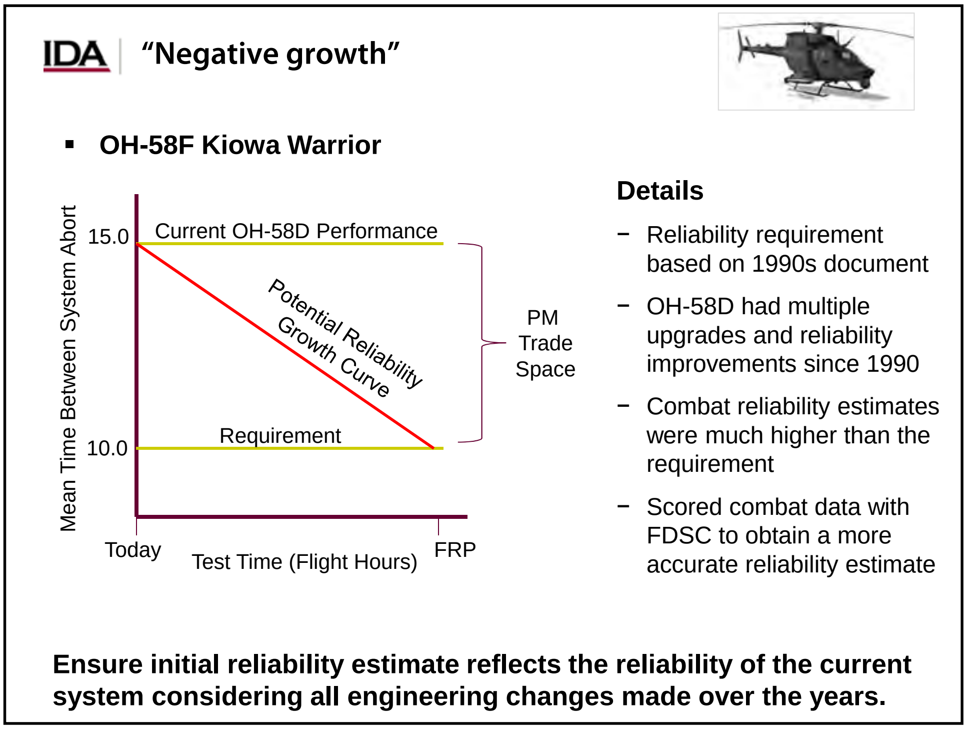
$$MS = \frac{\lambda_B}{\lambda_A + \lambda_B}$$

λ_B = initial B-mode failure intensity

λ_A = initial A-mode failure intensity

Risk Guidance for Management Strategy		
Low	Medium	High
< 90%	90 – 96%	> 96%

Ensure estimates of growth and management strategy are realistic. They should accurately quantify the failure intensity of A-modes.



IDA | “Negative growth”

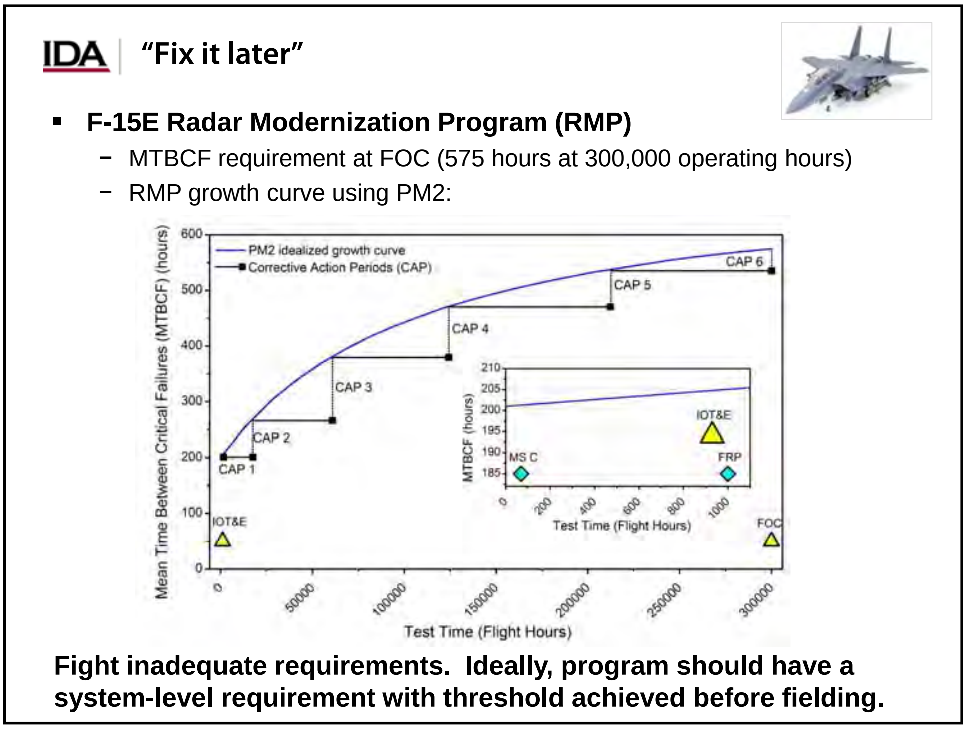
- **OH-58F Kiowa Warrior**

Mean Time Between System Abort
15.0
10.0
Current OH-58D Performance
Potential Reliability Growth Curve
Requirement
PM Trade Space
Today
Test Time (Flight Hours)
FRP

Details

- Reliability requirement based on 1990s document
- OH-58D had multiple upgrades and reliability improvements since 1990
- Combat reliability estimates were much higher than the requirement
- Scored combat data with FDSC to obtain a more accurate reliability estimate

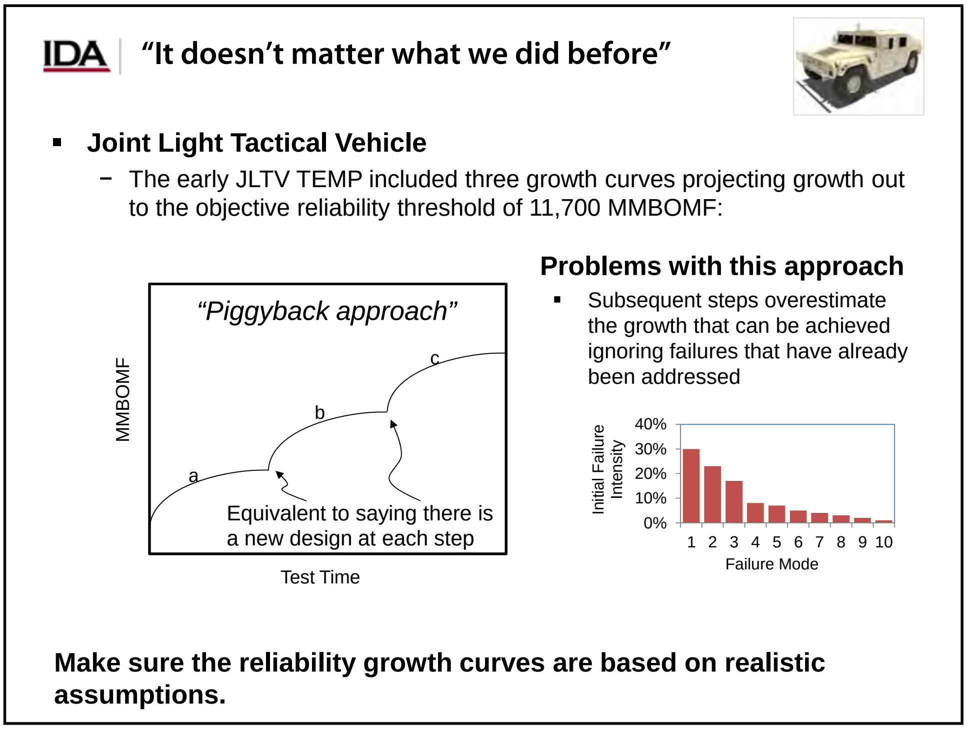
Ensure initial reliability estimate reflects the reliability of the current system considering all engineering changes made over the years.



“Fix it later”

- **F-15E Radar Modernization Program (RMP)**
  - MTBCF requirement at FOC (575 hours at 300,000 operating hours)
  - RMP growth curve using PM2:

Fight inadequate requirements. Ideally, program should have a system-level requirement with threshold achieved before fielding.



“It doesn’t matter what we did before”

- **Joint Light Tactical Vehicle**
  - The early JLTV TEMP included three growth curves projecting growth out to the objective reliability threshold of 11,700 MMBOMF:

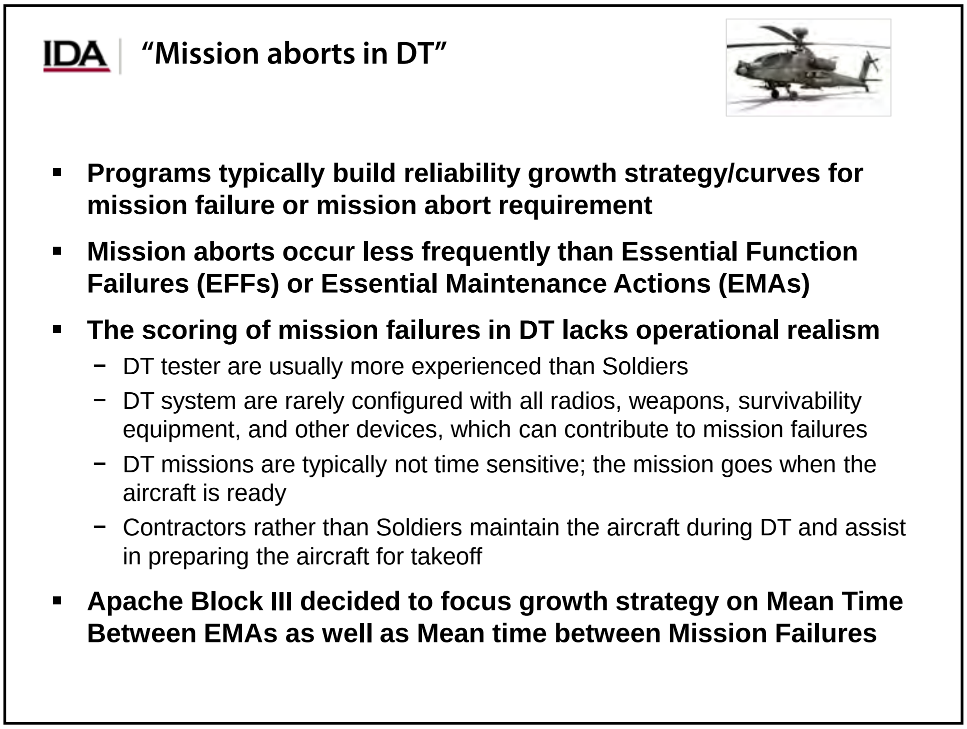
“Piggyback approach”
MMBOMF
a
b
c
Equivalent to saying there is a new design at each step
Test Time

Problems with this approach

- Subsequent steps overestimate the growth that can be achieved ignoring failures that have already been addressed

Initial Failure Intensity
0%
10%
20%
30%
40%
1
2
3
4
5
6
7
8
9
10
Failure Mode

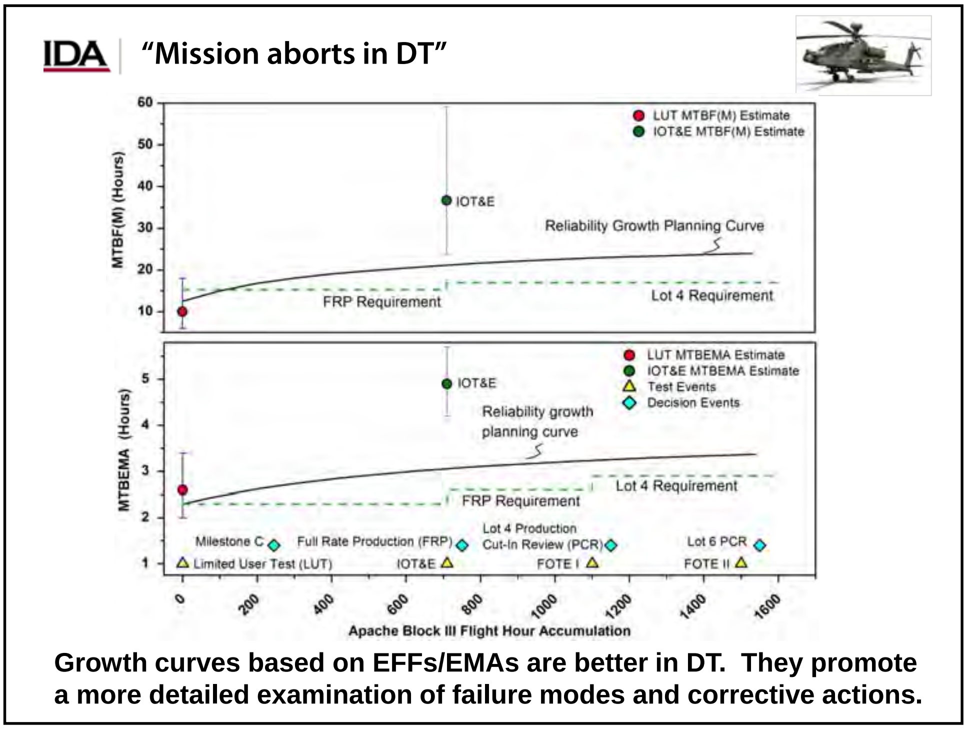
Make sure the reliability growth curves are based on realistic assumptions.



IDA

“Mission aborts in DT”

- **Programs typically build reliability growth strategy/curves for mission failure or mission abort requirement**
- **Mission aborts occur less frequently than Essential Function Failures (EFFs) or Essential Maintenance Actions (EMAs)**
- **The scoring of mission failures in DT lacks operational realism**
  - DT tester are usually more experienced than Soldiers
  - DT system are rarely configured with all radios, weapons, survivability equipment, and other devices, which can contribute to mission failures
  - DT missions are typically not time sensitive; the mission goes when the aircraft is ready
  - Contractors rather than Soldiers maintain the aircraft during DT and assist in preparing the aircraft for takeoff
- **Apache Block III decided to focus growth strategy on Mean Time Between EMAs as well as Mean time between Mission Failures**



IDA

“Mission aborts in DT”

Growth curves based on EFFs/EMAs are better in DT. They promote a more detailed examination of failure modes and corrective actions.

“Can we really get there from here”

- **Apache Block III program updated their initial reliability estimate and growth curves once data was available**

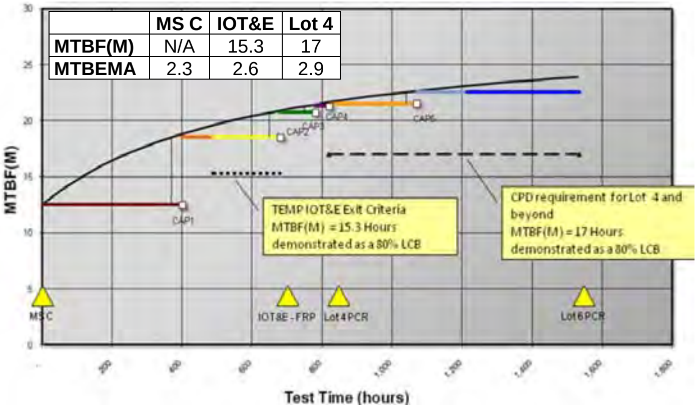
Model Input	Initial Estimated Value	Value from Early DT
Initial Reliability (M_I)	1.6 hours MTBMEA	2.3 hours MTBMEA
	9.0 hours MTBF(M)	12.5 hours MTBF(M)

- **Provided a more realistic value and assessment of program risk**
  - Increase in initial reliability actually lowered program risk in this case

Update growth curve assumptions once data is available, particularly for the initial reliability.

“Interim thresholds”

- **Apache Block III program developed interim reliability thresholds that were tied to the growth planning curve:**

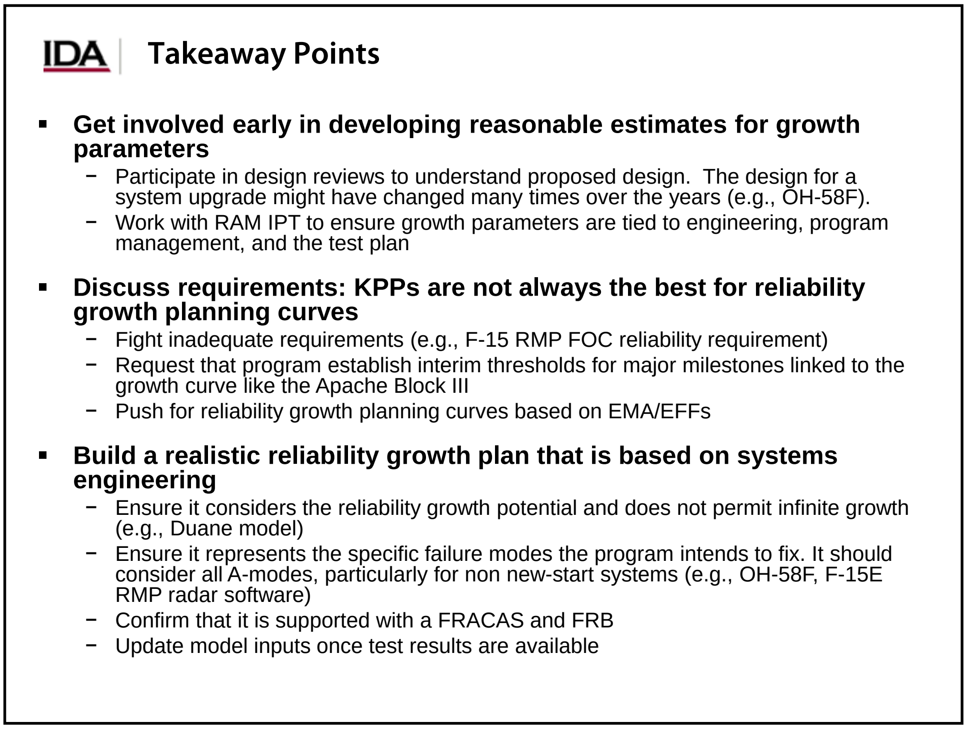


	MS C	IOT&E	Lot 4
MTBF(M)	N/A	15.3	17
MTBEMA	2.3	2.6	2.9

TEMP IOT&E Exit Criteria MTBF(M) = 15.3 Hours demonstrated as a 80% LCB

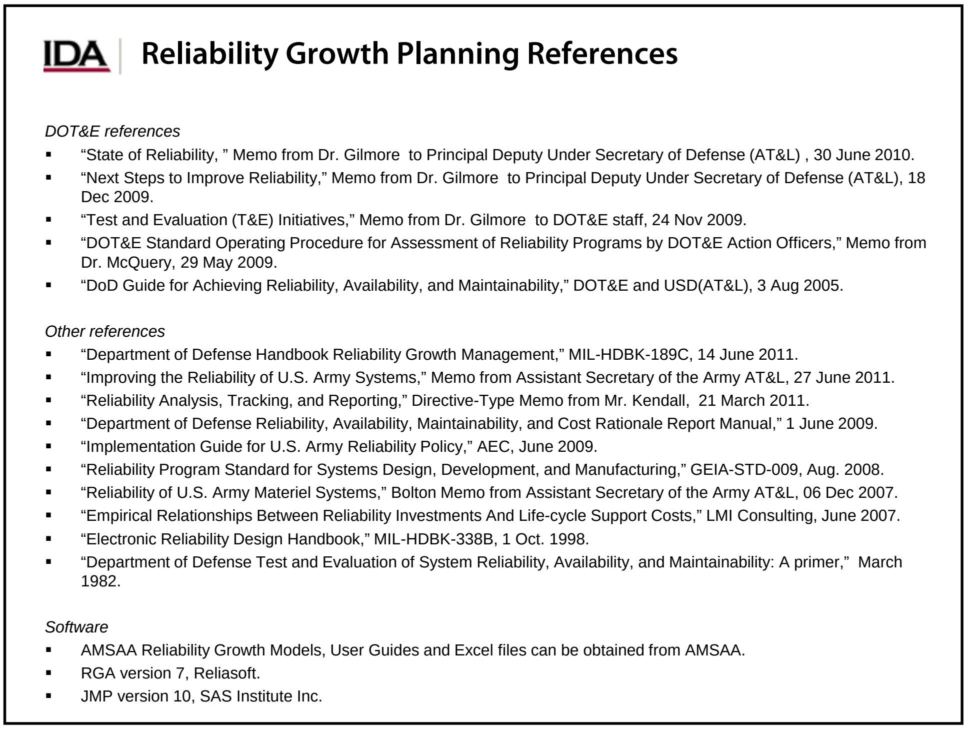
CPD requirement for Lot 4 and beyond MTBF(M) = 17 Hours demonstrated as a 80% LCB

Reliability growth plan should take into account major milestones and interim thresholds.



Takeaway Points

- **Get involved early in developing reasonable estimates for growth parameters**
  - Participate in design reviews to understand proposed design. The design for a system upgrade might have changed many times over the years (e.g., OH-58F).
  - Work with RAM IPT to ensure growth parameters are tied to engineering, program management, and the test plan
- **Discuss requirements: KPPs are not always the best for reliability growth planning curves**
  - Fight inadequate requirements (e.g., F-15 RMP FOC reliability requirement)
  - Request that program establish interim thresholds for major milestones linked to the growth curve like the Apache Block III
  - Push for reliability growth planning curves based on EMA/EFFs
- **Build a realistic reliability growth plan that is based on systems engineering**
  - Ensure it considers the reliability growth potential and does not permit infinite growth (e.g., Duane model)
  - Ensure it represents the specific failure modes the program intends to fix. It should consider all A-modes, particularly for non new-start systems (e.g., OH-58F, F-15E RMP radar software)
  - Confirm that it is supported with a FRACAS and FRB
  - Update model inputs once test results are available



Reliability Growth Planning References

DOT&E references

- "State of Reliability, " Memo from Dr. Gilmore to Principal Deputy Under Secretary of Defense (AT&L) , 30 June 2010.
- "Next Steps to Improve Reliability," Memo from Dr. Gilmore to Principal Deputy Under Secretary of Defense (AT&L), 18 Dec 2009.
- "Test and Evaluation (T&E) Initiatives," Memo from Dr. Gilmore to DOT&E staff, 24 Nov 2009.
- "DOT&E Standard Operating Procedure for Assessment of Reliability Programs by DOT&E Action Officers," Memo from Dr. McQuery, 29 May 2009.
- "DoD Guide for Achieving Reliability, Availability, and Maintainability," DOT&E and USD(AT&L), 3 Aug 2005.

Other references

- "Department of Defense Handbook Reliability Growth Management," MIL-HDBK-189C, 14 June 2011.
- "Improving the Reliability of U.S. Army Systems," Memo from Assistant Secretary of the Army AT&L, 27 June 2011.
- "Reliability Analysis, Tracking, and Reporting," Directive-Type Memo from Mr. Kendall, 21 March 2011.
- "Department of Defense Reliability, Availability, Maintainability, and Cost Rationale Report Manual," 1 June 2009.
- "Implementation Guide for U.S. Army Reliability Policy," AEC, June 2009.
- "Reliability Program Standard for Systems Design, Development, and Manufacturing," GEIA-STD-009, Aug. 2008.
- "Reliability of U.S. Army Materiel Systems," Bolton Memo from Assistant Secretary of the Army AT&L, 06 Dec 2007.
- "Empirical Relationships Between Reliability Investments And Life-cycle Support Costs," LMI Consulting, June 2007.
- "Electronic Reliability Design Handbook," MIL-HDBK-338B, 1 Oct. 1998.
- "Department of Defense Test and Evaluation of System Reliability, Availability, and Maintainability: A primer," March 1982.

Software

- AMSAA Reliability Growth Models, User Guides and Excel files can be obtained from AMSAA.
- RGA version 7, Reliasoft.
- JMP version 10, SAS Institute Inc.

This page intentionally left blank.

Appendix 1-4.
DOT&E Warfare
Brownbag Examples

DOT&E Design of Experiments Brownbag: Warfare Specific Examples

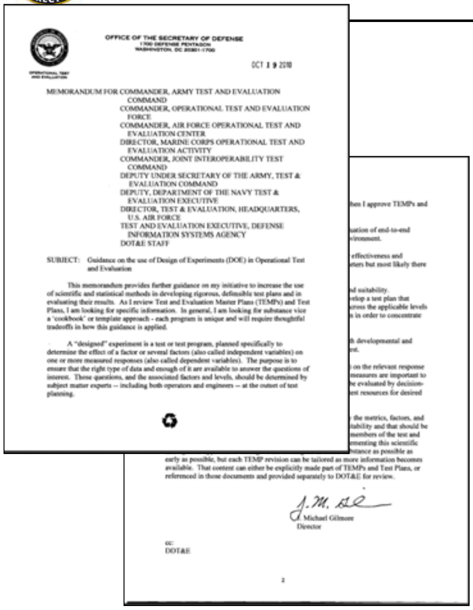
Dr. Catherine Warner
Science Advisor
Director, Operational Test & Evaluation
10 February 2012

Outline

- Design of Experiments Overview
- Warfare Area Specific Examples
  - Land Warfare
  - Net-Centric and Space System
  - Air Warfare
  - Naval Warfare
  - Live Fire

Guidance

Dr. Gilmore's October 19, 2010 Memo to OTAs



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

OPERATIONAL TEST AND EVALUATION

OCT 1 9 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.

hen I approve TEMPs and

uation of end-to-end
vironment.

effectiveness and
eters but most likely there

nd suitability.
velop a test plan that
cross the applicable levels
n in order to concentrate

th developmental and
est.

) on the relevant response
measures are important to
be evaluated by decision-
est resources for desired

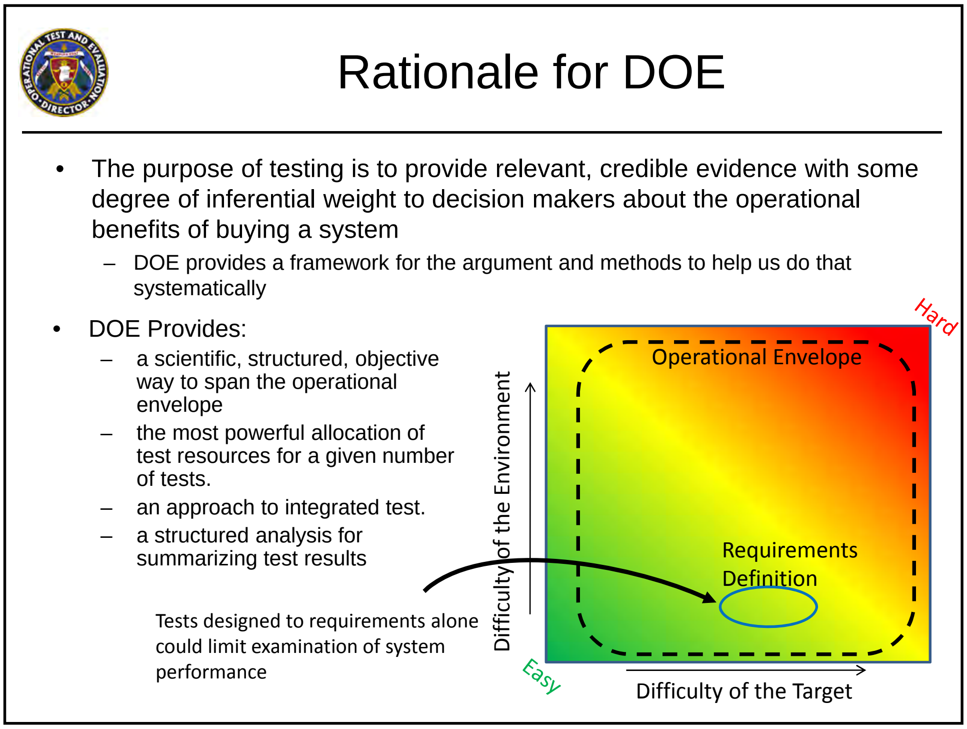
the metrics, factors, and
itability and that should be
members of the test and
ementing this scientific
stance as possible as
early as possible, but each TEMP revision can be tailored as more information becomes available. That content can either be explicitly made part of TEMPs and Test Plans, or referenced in those documents and provided separately to DOT&E for review.

J. M. Gil
J. Michael Gilmore
Director

cc:
DOT&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 understand "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.



Rationale for DOE

- 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
- DOE Provides:
  - a scientific, structured, objective way to span the operational envelope
  - the most powerful allocation of test resources for a given number of tests.
  - an approach to integrated test.
  - a structured analysis for summarizing test results

Tests designed to requirements alone could limit examination of system performance

Hard

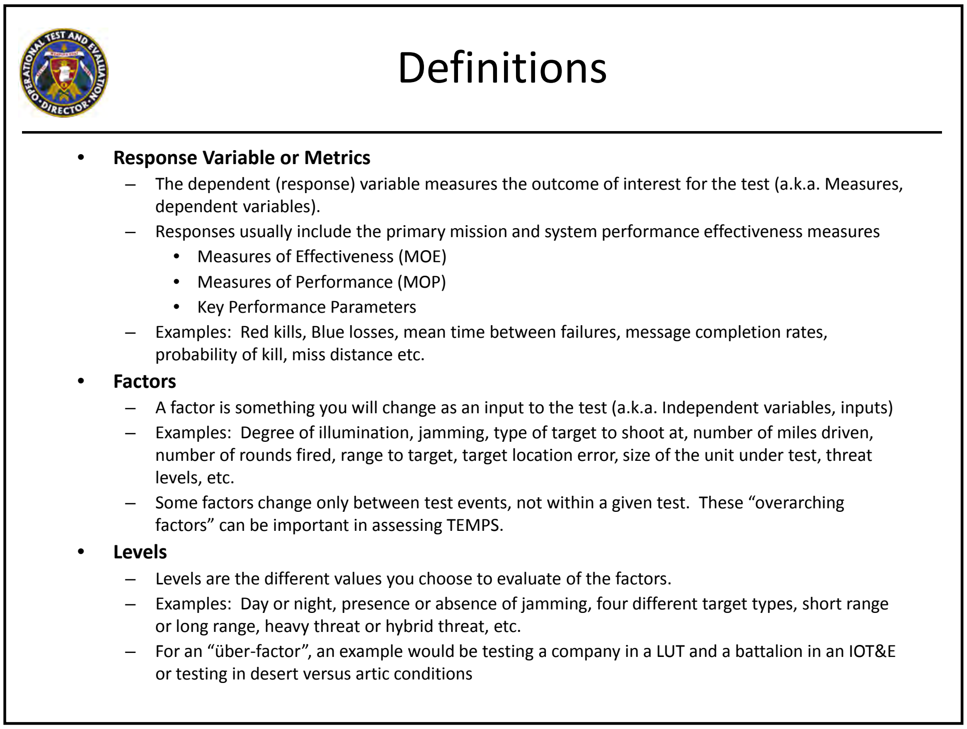
Operational Envelope

Requirements Definition

Difficulty of the Environment

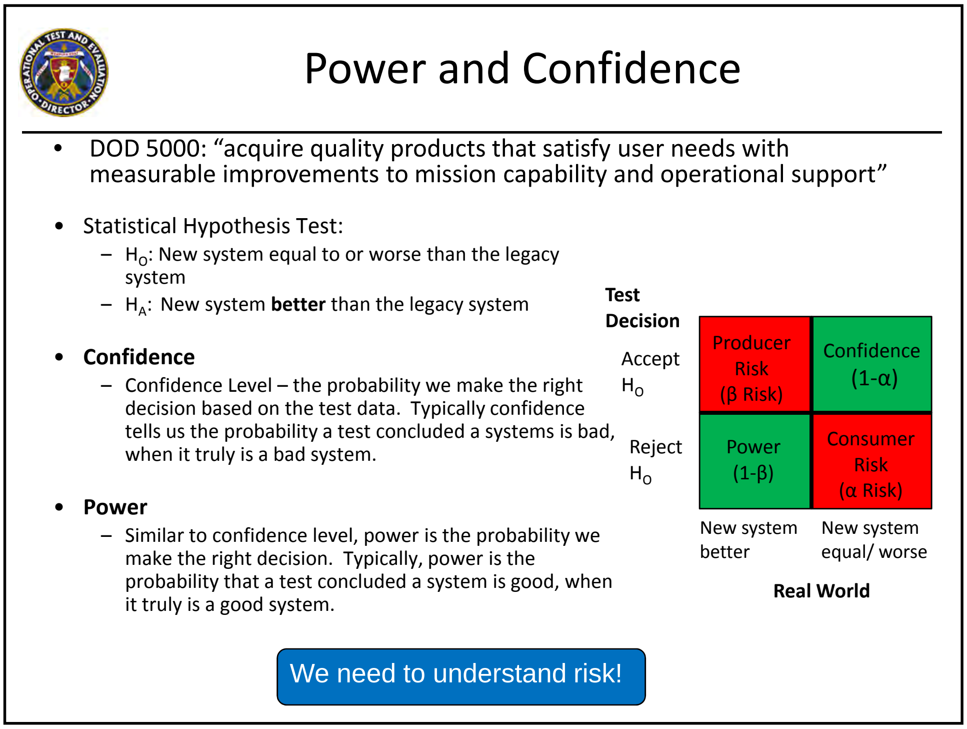
Easy

Difficulty of the Target



Definitions

- **Response Variable or Metrics**
  - The dependent (response) variable measures the outcome of interest for the test (a.k.a. Measures, dependent variables).
  - Responses usually include the primary mission and system performance effectiveness measures
    - Measures of Effectiveness (MOE)
    - Measures of Performance (MOP)
    - Key Performance Parameters
  - Examples: Red kills, Blue losses, mean time between failures, message completion rates, probability of kill, miss distance etc.
- **Factors**
  - A factor is something you will change as an input to the test (a.k.a. Independent variables, inputs)
  - Examples: Degree of illumination, jamming, type of target to shoot at, number of miles driven, number of rounds fired, range to target, target location error, size of the unit under test, threat levels, etc.
  - Some factors change only between test events, not within a given test. These “overarching factors” can be important in assessing TEMPS.
- **Levels**
  - Levels are the different values you choose to evaluate of the factors.
  - Examples: Day or night, presence or absence of jamming, four different target types, short range or long range, heavy threat or hybrid threat, etc.
  - For an “über-factor”, an example would be testing a company in a LUT and a battalion in an IOT&E or testing in desert versus artic conditions



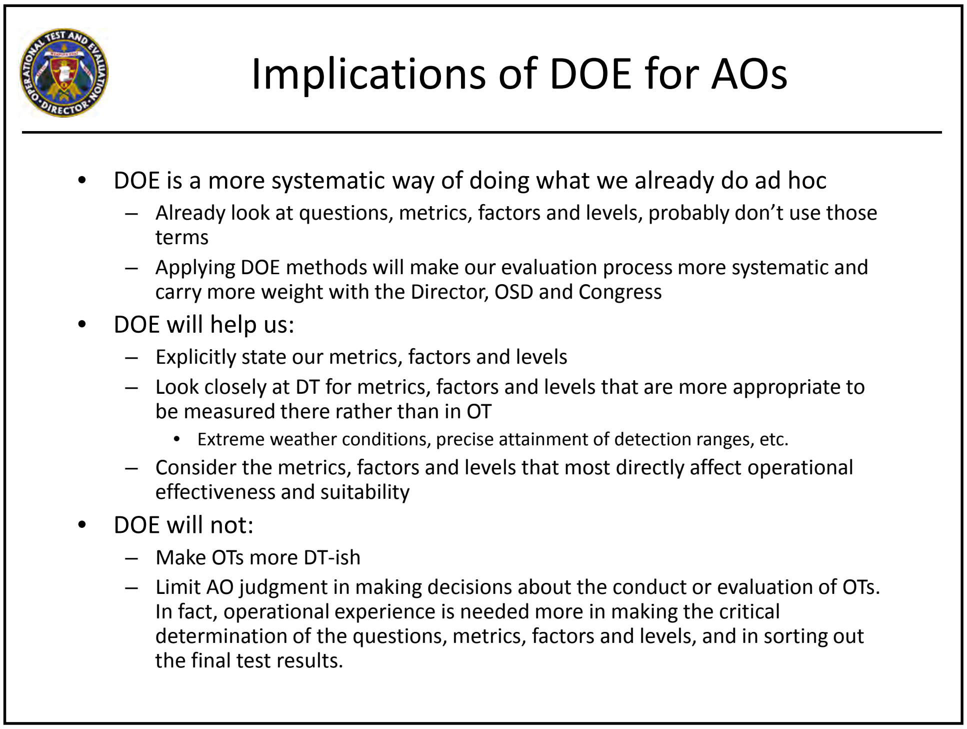
Power and Confidence

- DOD 5000: "acquire quality products that satisfy user needs with measurable improvements to mission capability and operational support"
- Statistical Hypothesis Test:
  - H_O: New system equal to or worse than the legacy system
  - H_A: New system **better** than the legacy system
- **Confidence**
  - Confidence Level – the probability we make the right decision based on the test data. Typically confidence tells us the probability a test concluded a systems is bad, when it truly is a bad system.
- **Power**
  - Similar to confidence level, power is the probability we make the right decision. Typically, power is the probability that a test concluded a system is good, when it truly is a good system.

Test Decision	New system better	New system equal/ worse
Accept H_O	🟥 Producer Risk (β Risk)	🟩 Confidence ($1-\alpha$)
Reject H_O	🟩 Power ($1-\beta$)	🟥 Consumer Risk (α Risk)

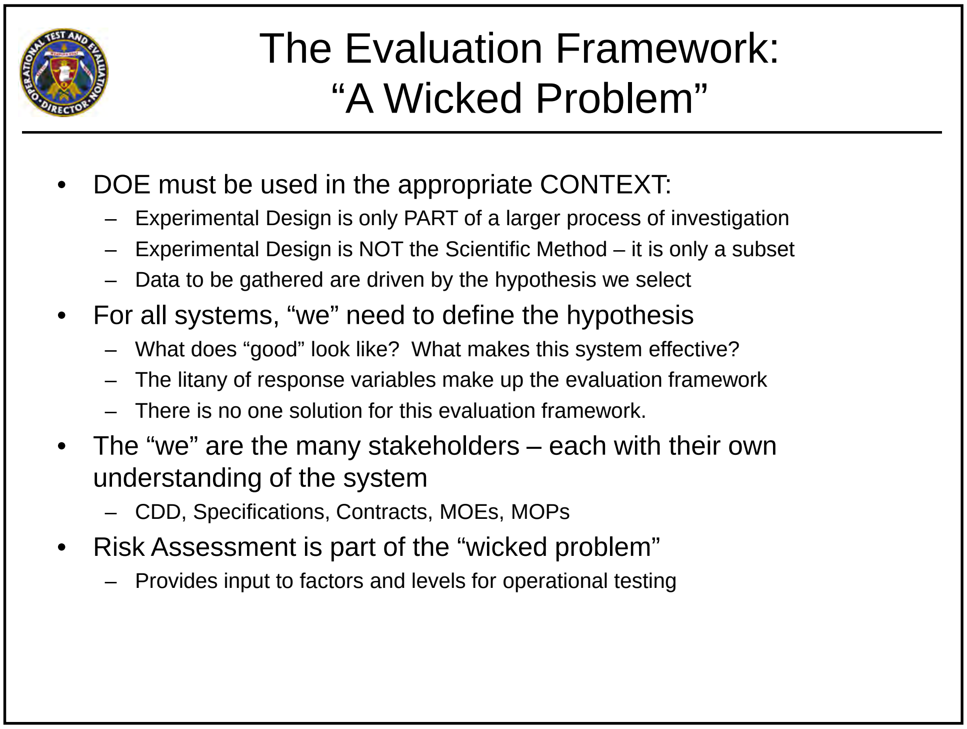
Real World

We need to understand risk!



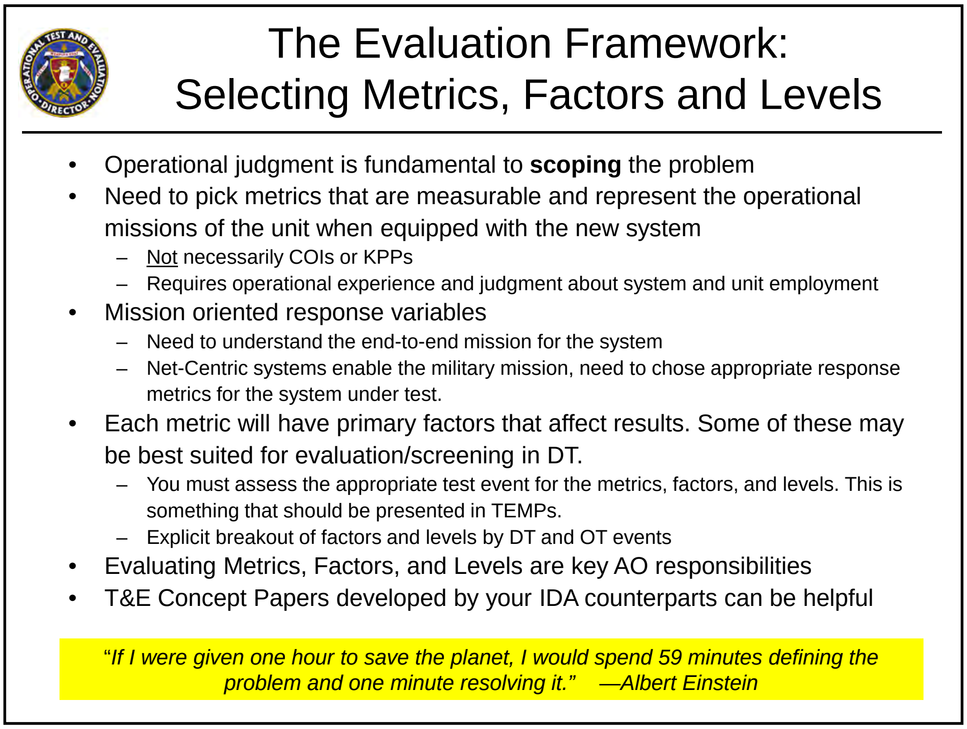
Implications of DOE for AOs

- DOE is a more systematic way of doing what we already do ad hoc
  - Already look at questions, metrics, factors and levels, probably don't use those terms
  - Applying DOE methods will make our evaluation process more systematic and carry more weight with the Director, OSD and Congress
- DOE will help us:
  - Explicitly state our metrics, factors and levels
  - Look closely at DT for metrics, factors and levels that are more appropriate to be measured there rather than in OT
    - Extreme weather conditions, precise attainment of detection ranges, etc.
  - Consider the metrics, factors and levels that most directly affect operational effectiveness and suitability
- DOE will not:
  - Make OTs more DT-ish
  - Limit AO judgment in making decisions about the conduct or evaluation of OTs. In fact, operational experience is needed more in making the critical determination of the questions, metrics, factors and levels, and in sorting out the final test results.



The Evaluation Framework: "A Wicked Problem"

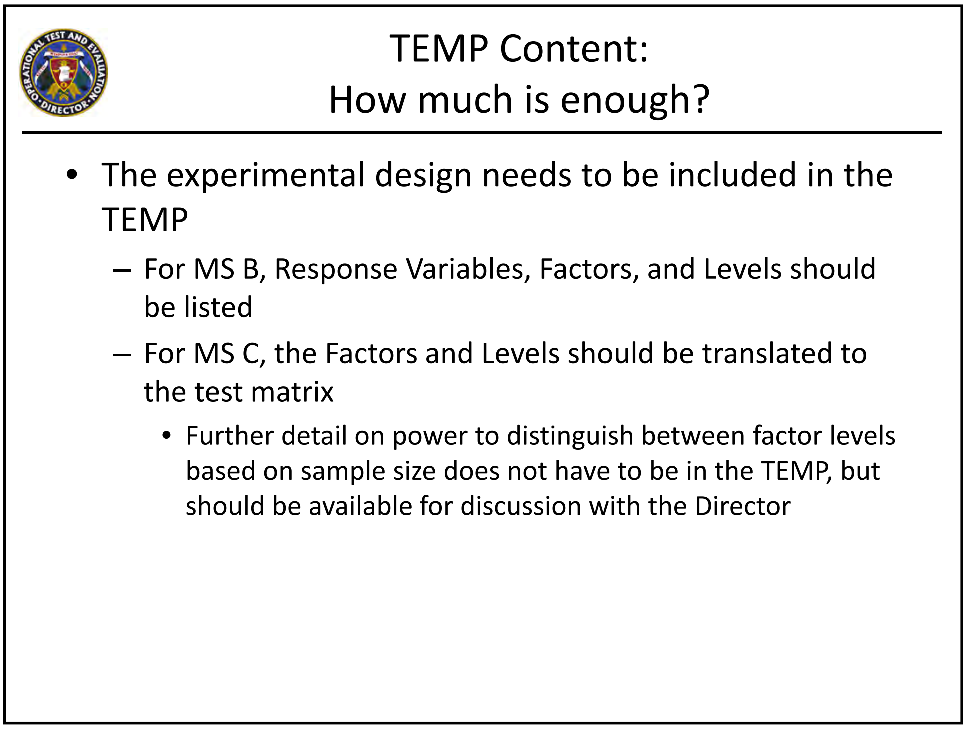
- DOE must be used in the appropriate CONTEXT:
  - Experimental Design is only PART of a larger process of investigation
  - Experimental Design is NOT the Scientific Method – it is only a subset
  - Data to be gathered are driven by the hypothesis we select
- For all systems, "we" need to define the hypothesis
  - What does "good" look like? What makes this system effective?
  - The litany of response variables make up the evaluation framework
  - There is no one solution for this evaluation framework.
- The "we" are the many stakeholders – each with their own understanding of the system
  - CDD, Specifications, Contracts, MOEs, MOPs
- Risk Assessment is part of the "wicked problem"
  - Provides input to factors and levels for operational testing



The Evaluation Framework: Selecting Metrics, Factors and Levels

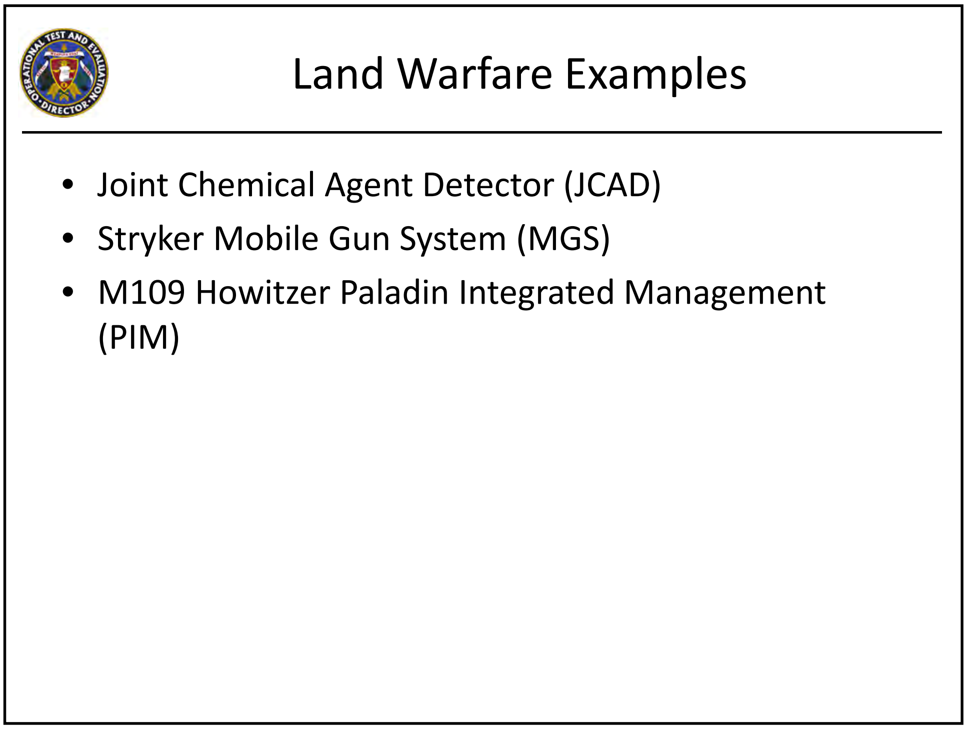
- Operational judgment is fundamental to **scoping** the problem
- Need to pick metrics that are measurable and represent the operational missions of the unit when equipped with the new system
  - Not necessarily COIs or KPPs
  - Requires operational experience and judgment about system and unit employment
- Mission oriented response variables
  - Need to understand the end-to-end mission for the system
  - Net-Centric systems enable the military mission, need to chose appropriate response metrics for the system under test.
- Each metric will have primary factors that affect results. Some of these may be best suited for evaluation/screening in DT.
  - You must assess the appropriate test event for the metrics, factors, and levels. This is something that should be presented in TEMPs.
  - Explicit breakout of factors and levels by DT and OT events
- Evaluating Metrics, Factors, and Levels are key AO responsibilities
- T&E Concept Papers developed by your IDA counterparts can be helpful

"If I were given one hour to save the planet, I would spend 59 minutes defining the problem and one minute resolving it." —Albert Einstein



TEMP Content: How much is enough?

- The experimental design needs to be included in the TEMP
  - For MS B, Response Variables, Factors, and Levels should be listed
  - For MS C, the Factors and Levels should be translated to the test matrix
    - Further detail on power to distinguish between factor levels based on sample size does not have to be in the TEMP, but should be available for discussion with the Director



Land Warfare Examples

- Joint Chemical Agent Detector (JCAD)
- Stryker Mobile Gun System (MGS)
- M109 Howitzer Paladin Integrated Management (PIM)

Joint Chemical Agent Detector

- JCAD has used DOE to characterize the detection performance envelope in DT events.
  - Currently on 4th iteration of DOE due to detector configuration changes (2006-2011).
  - Each test event has provided insight into ways the test design and evaluation can be improved.
  - DOE allows for a large amount of data to be analyzed in a short amount of time.

Generating DOE matrices

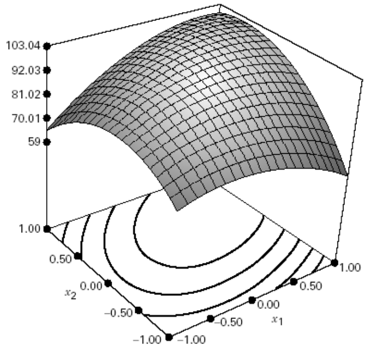
- Vendor (Smiths Detection) was initially a useful source of info on what factors would be important to consider.
  - Agent
  - Agent concentration
  - Temperature
  - Humidity
- Users provided initial “levels” of factors in CDD/CPD.
  - Required agents.
  - Minimum agent concentration for detection
  - Expected operating environment (generally -32°C - 49°C; 5-100% relative humidity), depending on agent.
- DPG test chamber constraints further refined levels of factors for matrix.
  - Chamber can’t go below 5°C or above 80% relative humidity.
  - T&E IPT agreed that chamber constraints would be test limitation.
- DOE matrix was generated by DPG (DTC) statistician using DOE design software (JMP, SAS, Design Expert). IDA support can also provide this.
  - Presented to T&E IPT (including power calculations).
  - They were refined to meet needs of all evaluators.
  - DOE design *and* evaluation plan were put into TEMP and DT/OT test plans.

Evaluating data

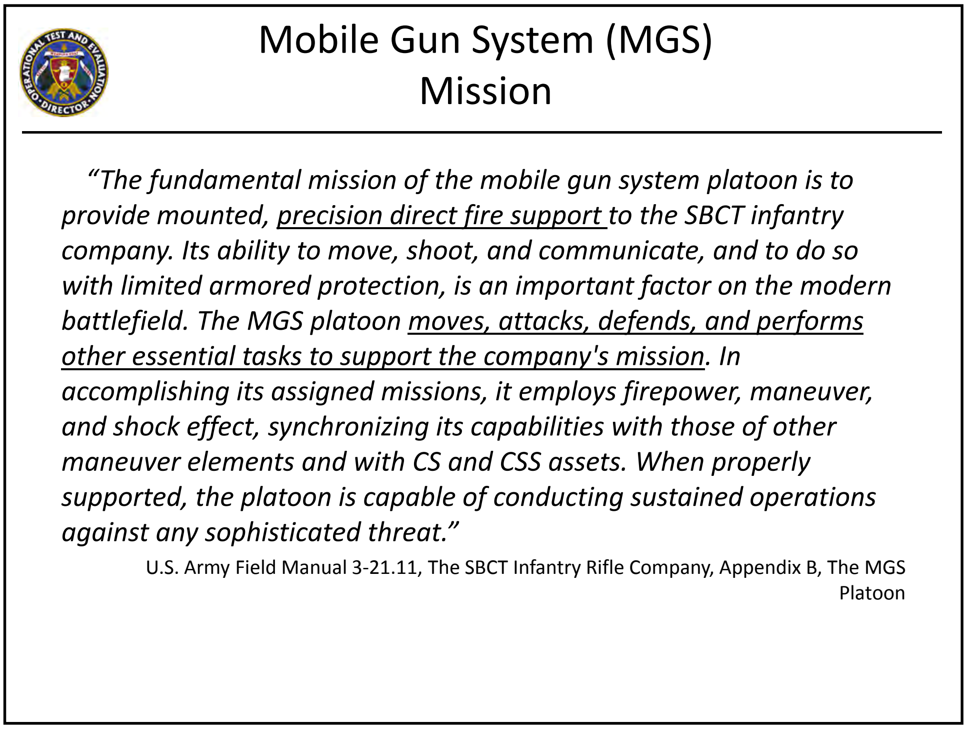
- 1st DOE iteration (2006-2007)
  - No modeling; simple P(d) and average time to alarm.
  - For JCAD, not modeling data made the evaluation harder.
    - Apples and oranges data points between agents (did not have same temperature/humidity combinations).
  - Lesson learned: Next time Evaluators will model
- 2nd DOE iteration (2009-2010)
  - Evaluators weren't fully comfortable with model going into test, so fall back plan was to calculate simple P(d). This led to many replicates (16 for each point).
  - 10,000 data points total.
  - Model was very statistically significant; was able to facilitate bivariate analysis(Time for 90% P(d)).
  - Lesson Learned: Test design was way too big. Model does not need to be that statistically significant to generate accurate results. Smaller test (fewer replicates) next time.
- 3rd DOE iteration (2011)
  - Fewer replicates per point: 6; <1000 total data points.
  - Model still statistically significant, still able to facilitate bivariate analysis.
  - Lesson Learned: Smaller tests can lead to similar results as larger tests.
    - Caveat: this may not always be possible for programs that don't have a good idea of system performance going into test. Evaluators had a good handle on the signal to noise ratio (for power calculation), which was learned in previous iterations.
- 4th DOE iteration (late 2011)
  - TBD; expect similar experience with data for modeling.

JCAD DOE lessons learned

- DOE includes not just the design but the end evaluation.
  - Evaluators need to state up front what the end evaluation will be to ensure an appropriate DOE design matrix is created.
    - TEMP or Test Plan should state matrices, power, and how the data will be evaluated.
- DOE Models can greatly speed up the end evaluation.
  - Rapid analysis
    - e.g. few hours for 10,000 data points
  - Give evaluators flexibility in what data to display it.
- A poor DOE design or a poor evaluation using a good DOE design will make life difficult.
  - Apples and oranges data points.



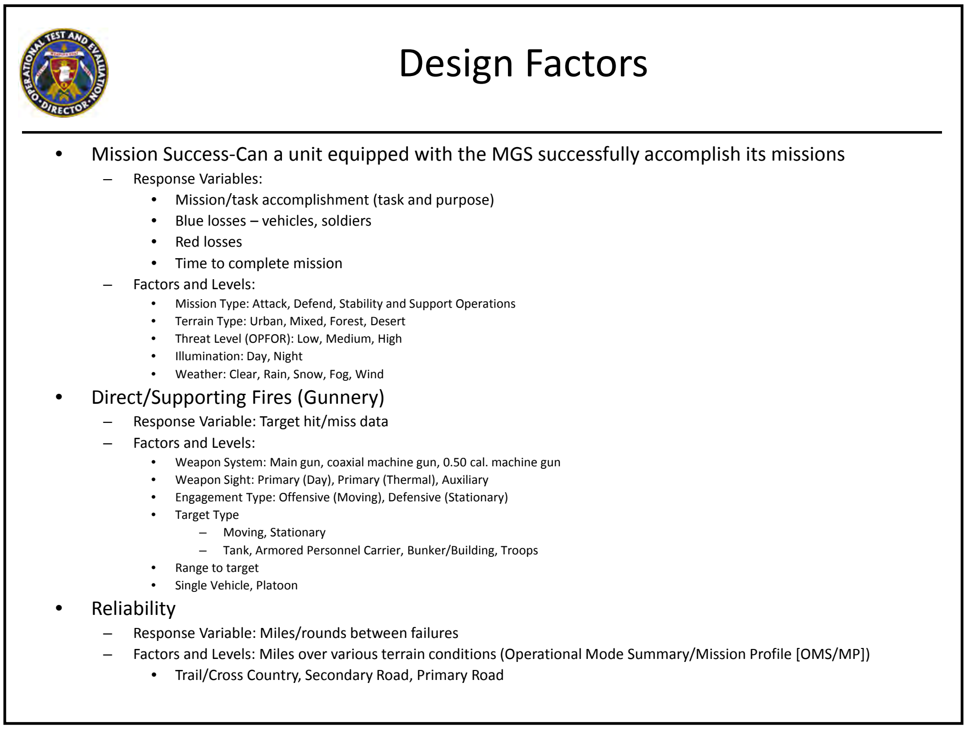
(*a*) Response surface



Mobile Gun System (MGS) Mission

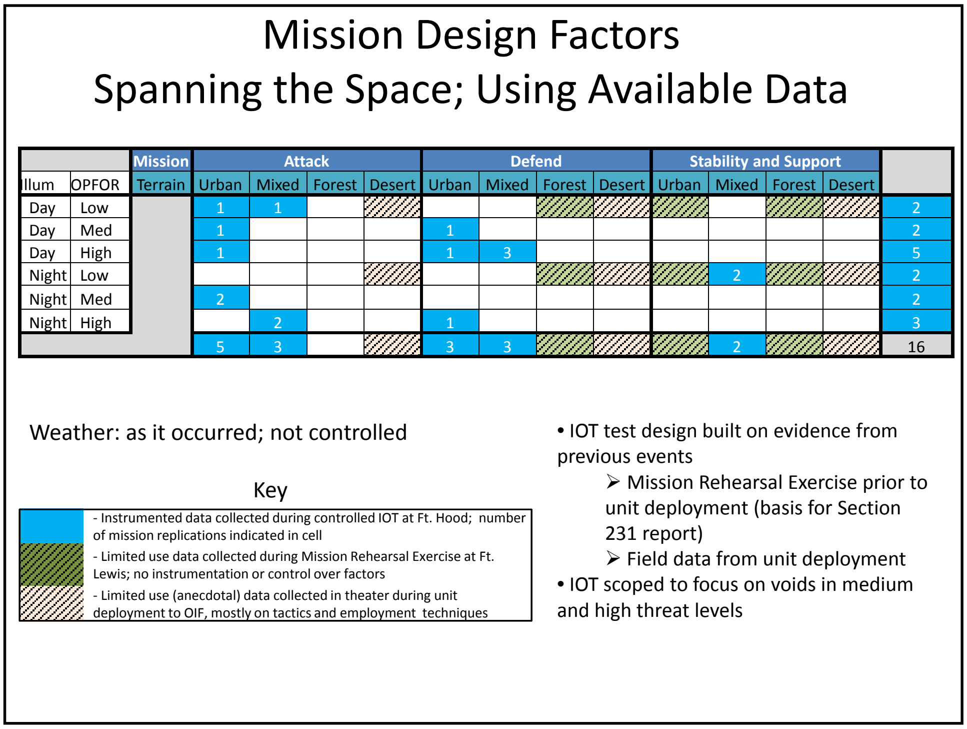
"The fundamental mission of the mobile gun system platoon is to provide mounted, precision direct fire support to the SBCT infantry company. Its ability to move, shoot, and communicate, and to do so with limited armored protection, is an important factor on the modern battlefield. The MGS platoon moves, attacks, defends, and performs other essential tasks to support the company's mission. In accomplishing its assigned missions, it employs firepower, maneuver, and shock effect, synchronizing its capabilities with those of other maneuver elements and with CS and CSS assets. When properly supported, the platoon is capable of conducting sustained operations against any sophisticated threat."

U.S. Army Field Manual 3-21.11, The SBCT Infantry Rifle Company, Appendix B, The MGS Platoon



Design Factors

- Mission Success-Can a unit equipped with the MGS successfully accomplish its missions
  - Response Variables:
    - Mission/task accomplishment (task and purpose)
    - Blue losses – vehicles, soldiers
    - Red losses
    - Time to complete mission
  - Factors and Levels:
    - Mission Type: Attack, Defend, Stability and Support Operations
    - Terrain Type: Urban, Mixed, Forest, Desert
    - Threat Level (OPFOR): Low, Medium, High
    - Illumination: Day, Night
    - Weather: Clear, Rain, Snow, Fog, Wind
- Direct/Supporting Fires (Gunnery)
  - Response Variable: Target hit/miss data
  - Factors and Levels:
    - Weapon System: Main gun, coaxial machine gun, 0.50 cal. machine gun
    - Weapon Sight: Primary (Day), Primary (Thermal), Auxiliary
    - Engagement Type: Offensive (Moving), Defensive (Stationary)
    - Target Type
      - Moving, Stationary
      - Tank, Armored Personnel Carrier, Bunker/Building, Troops
    - Range to target
    - Single Vehicle, Platoon
- Reliability
  - Response Variable: Miles/rounds between failures
  - Factors and Levels: Miles over various terrain conditions (Operational Mode Summary/Mission Profile [OMS/MP])
    - Trail/Cross Country, Secondary Road, Primary Road



Mission Design Factors
Spanning the Space; Using Available Data

		Mission	Attack				Defend				Stability and Support				
Illum	OPFOR	Terrain	Urban	Mixed	Forest	Desert	Urban	Mixed	Forest	Desert	Urban	Mixed	Forest	Desert	
Day	Low		1	1		▨			▩	▨	▩		▩	▨	2
Day	Med		1				1								2
Day	High		1				1	3							5
Night	Low					▨			▩	▨	▩	2	▩	▨	2
Night	Med		2												2
Night	High			2			1								3
			5	3		▨	3	3	▩	▨	▩	2	▩	▨	16

Weather: as it occurred; not controlled

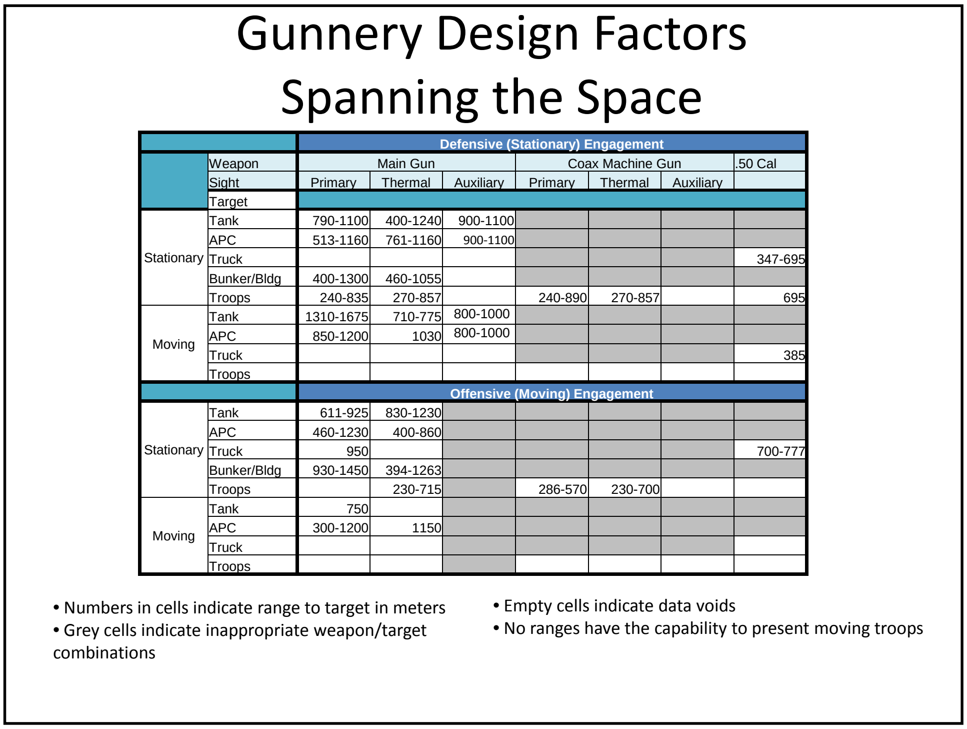
Key

🟦 - Instrumented data collected during controlled IOT at Ft. Hood; number of mission replications indicated in cell

▩ - Limited use data collected during Mission Rehearsal Exercise at Ft. Lewis; no instrumentation or control over factors

▨ - Limited use (anecdotal) data collected in theater during unit deployment to OIF, mostly on tactics and employment techniques

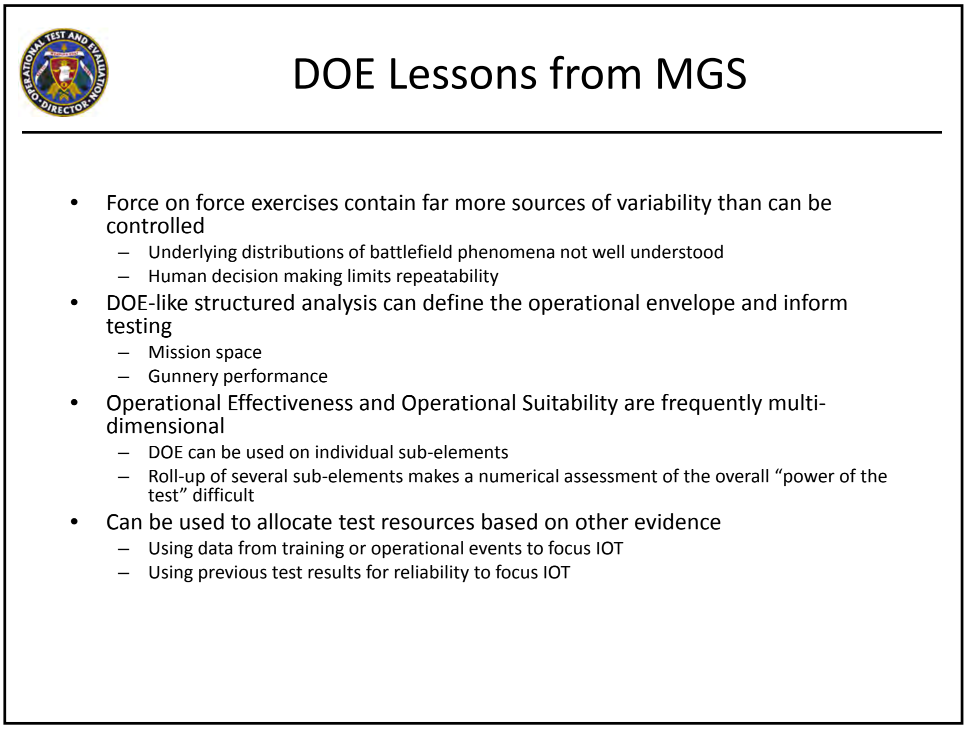
- IOT test design built on evidence from previous events
  - Mission Rehearsal Exercise prior to unit deployment (basis for Section 231 report)
  - Field data from unit deployment
- IOT scoped to focus on voids in medium and high threat levels



Gunnery Design Factors
Spanning the Space

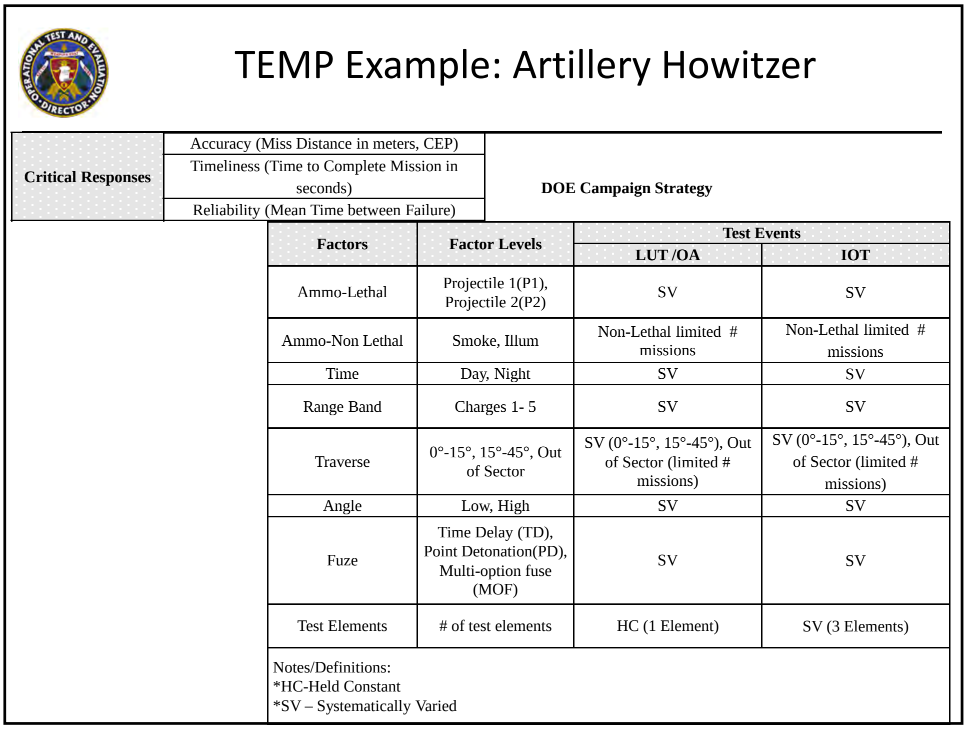
		Defensive (Stationary) Engagement						
	Weapon	Main Gun			Coax Machine Gun			.50 Cal
	Sight	Primary	Thermal	Auxiliary	Primary	Thermal	Auxiliary	
	Target							
Stationary	Tank	790-1100	400-1240	900-1100	▒	▒	▒	▒
	APC	513-1160	761-1160	900-1100	▒	▒	▒	▒
	Truck				▒	▒	▒	347-695
	Bunker/Bldg	400-1300	460-1055		▒	▒	▒	▒
	Troops	240-835	270-857		240-890	270-857		695
Moving	Tank	1310-1675	710-775	800-1000	▒	▒	▒	▒
	APC	850-1200	1030	800-1000	▒	▒	▒	▒
	Truck				▒	▒	▒	385
	Troops							
		Offensive (Moving) Engagement						
Stationary	Tank	611-925	830-1230	▒	▒	▒	▒	▒
	APC	460-1230	400-860	▒	▒	▒	▒	▒
	Truck	950		▒	▒	▒	▒	700-777
	Bunker/Bldg	930-1450	394-1263	▒	▒	▒	▒	▒
	Troops		230-715	▒	286-570	230-700		
Moving	Tank	750		▒	▒	▒	▒	▒
	APC	300-1200	1150	▒	▒	▒	▒	▒
	Truck			▒	▒	▒	▒	
	Troops			▒				

- Numbers in cells indicate range to target in meters
- Grey cells indicate inappropriate weapon/target combinations
- Empty cells indicate data voids
- No ranges have the capability to present moving troops



DOE Lessons from MGS

- Force on force exercises contain far more sources of variability than can be controlled
  - Underlying distributions of battlefield phenomena not well understood
  - Human decision making limits repeatability
- DOE-like structured analysis can define the operational envelope and inform testing
  - Mission space
  - Gunnery performance
- Operational Effectiveness and Operational Suitability are frequently multi-dimensional
  - DOE can be used on individual sub-elements
  - Roll-up of several sub-elements makes a numerical assessment of the overall “power of the test” difficult
- Can be used to allocate test resources based on other evidence
  - Using data from training or operational events to focus IOT
  - Using previous test results for reliability to focus IOT



TEMP Example: Artillery Howitzer

Critical Responses	Accuracy (Miss Distance in meters, CEP)
	Timeliness (Time to Complete Mission in seconds)
	Reliability (Mean Time between Failure)

DOE Campaign Strategy

Factors	**Factor Levels**	**Test Events**	
		LUT /OA	**IOT**
Ammo-Lethal	Projectile 1(P1), Projectile 2(P2)	SV	SV
Ammo-Non Lethal	Smoke, Illum	Non-Lethal limited # missions	Non-Lethal limited # missions
Time	Day, Night	SV	SV
Range Band	Charges 1- 5	SV	SV
Traverse	0°-15°, 15°-45°, Out of Sector	SV (0°-15°, 15°-45°), Out of Sector (limited # missions)	SV (0°-15°, 15°-45°), Out of Sector (limited # missions)
Angle	Low, High	SV	SV
Fuze	Time Delay (TD), Point Detonation(PD), Multi-option fuse (MOF)	SV	SV
Test Elements	# of test elements	HC (1 Element)	SV (3 Elements)

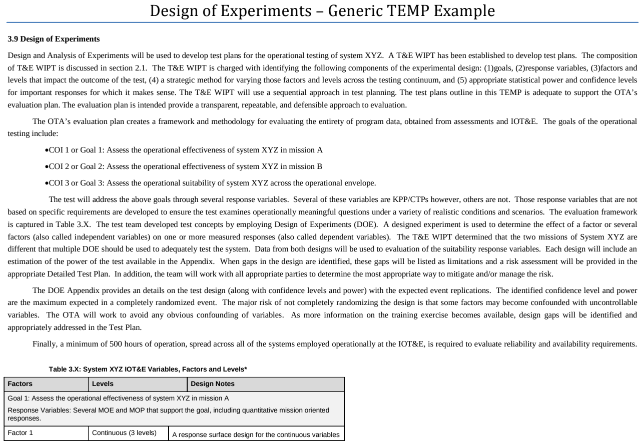
Notes/Definitions:
*HC-Held Constant
*SV – Systematically Varied

Example TEMP Language: How much is enough?

The DOE needs to be included in the TEMP

- Only an overview must be included in the TEMP but the detail should be available for discussion with the Director

Design of Experiments – Generic TEMP Example



3.9 Design of Experiments

Design and Analysis of Experiments will be used to develop test plans for the operational testing of system XYZ. A T&E WIPT has been established to develop test plans. The composition of T&E WIPT is discussed in section 2.1. The T&E WIPT is charged with identifying the following components of the experimental design: (1)goals, (2)response variables, (3)factors and levels that impact the outcome of the test, (4) a strategic method for varying those factors and levels across the testing continuum, and (5) appropriate statistical power and confidence levels for important responses for which it makes sense. The T&E WIPT will use a sequential approach in test planning. The test plans outline in this TEMP is adequate to support the OTA's evaluation plan. The evaluation plan is intended provide a transparent, repeatable, and defensible approach to evaluation.

The OTA's evaluation plan creates a framework and methodology for evaluating the entirety of program data, obtained from assessments and IOT&E. The goals of the operational testing include:

- COI 1 or Goal 1: Assess the operational effectiveness of system XYZ in mission A
- COI 2 or Goal 2: Assess the operational effectiveness of system XYZ in mission B
- COI 3 or Goal 3: Assess the operational suitability of system XYZ across the operational envelope.

The test will address the above goals through several response variables. Several of these variables are KPP/CTPs however, others are not. Those response variables that are not based on specific requirements are developed to ensure the test examines operationally meaningful questions under a variety of realistic conditions and scenarios. The evaluation framework is captured in Table 3.X. The test team developed test concepts by employing Design of Experiments (DOE). A designed experiment is used 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 T&E WIPT determined that the two missions of System XYZ are different that multiple DOE should be used to adequately test the system. Data from both designs will be used to evaluation of the suitability response variables. Each design will include an estimation of the power of the test available in the Appendix. When gaps in the design are identified, these gaps will be listed as limitations and a risk assessment will be provided in the appropriate Detailed Test Plan. In addition, the team will work with all appropriate parties to determine the most appropriate way to mitigate and/or manage the risk.

The DOE Appendix provides an details on the test design (along with confidence levels and power) with the expected event replications. The identified confidence level and power are the maximum expected in a completely randomized event. The major risk of not completely randomizing the design is that some factors may become confounded with uncontrollable variables. The OTA will work to avoid any obvious confounding of variables. As more information on the training exercise becomes available, design gaps will be identified and appropriately addressed in the Test Plan.

Finally, a minimum of 500 hours of operation, spread across all of the systems employed operationally at the IOT&E, is required to evaluate reliability and availability requirements.

Table 3.X: System XYZ IOT&E Variables, Factors and Levels*

Factors	Levels	Design Notes
Goal 1: Assess the operational effectiveness of system XYZ in mission A Response Variables: Several MOE and MOP that support the goal, including quantitative mission oriented responses.		
Factor 1	Continuous (3 levels)	A response surface design for the continuous variables

Net Centric and Space Examples

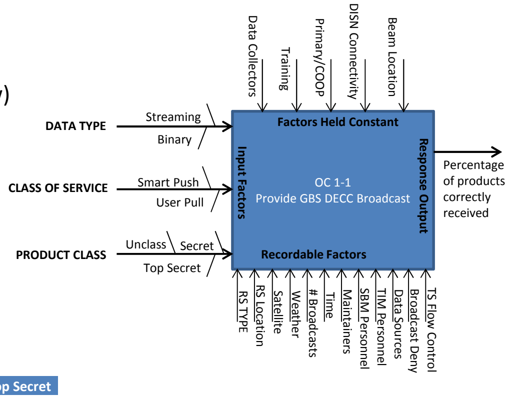
- Global Broadcast Service (GBS)
- Key Management Infrastructure (KMI)
- Consolidated Afloat Networks and Enterprise Services (CANES)
- A notional example of an Information Assurance (IA) system

Global Broadcast Service (GBS)

- GBS Background
  - Provide a one-way, high-speed flow of near-real-time wideband information to forces garrisoned or deployed
  - High-capacity product dissemination (Imagery, UAS full-motion video, large data files)
  - Information transported through GBS supports a large variety of missions.
- Experimental Objective
  - Determine transmission completion within operational envelope
- Response: Percentage of products received
  - The products received can support a broad array of other missions
  - Lead operational test organization (17 TS) considered applicability of DOE to the system under test and arrived at a design for the “percentage of products correctly received” measure.

Global Broadcast Service (GBS)

- TEMP Content
  - Factor selection process map (right)
  - Factorial design matrix (below)
  - Power calculations
  - Proposed analysis method
- Binary Response Metric
- Possible Analyses:
  - Logistic regression
  - One and two-sample proportion significance tests

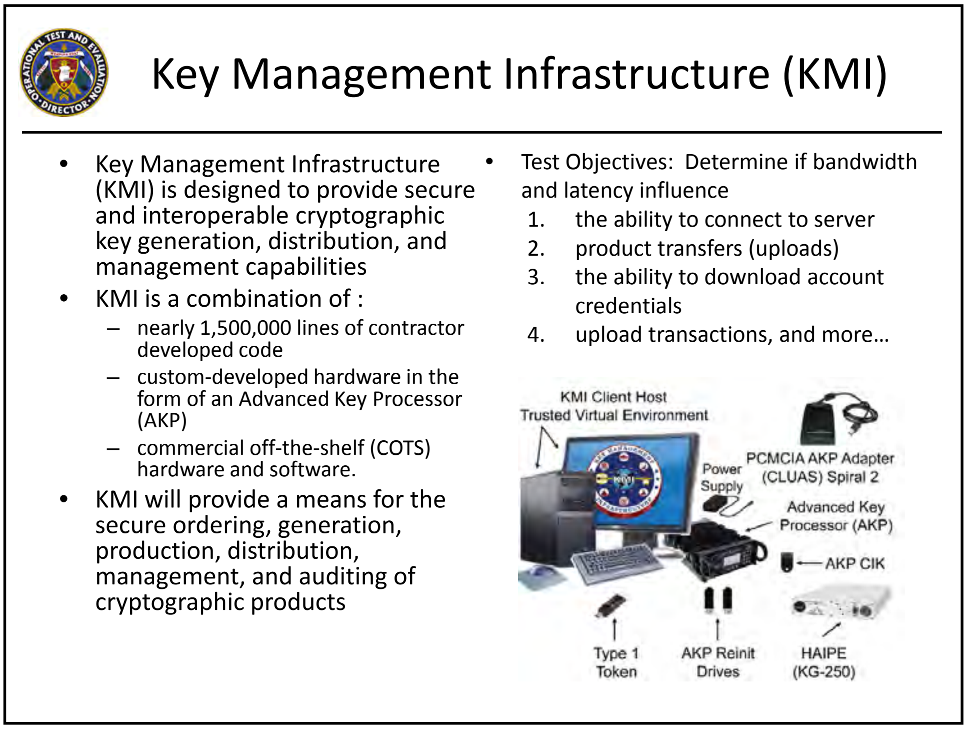


2x2x3 Factorial Design Matrix

		Unclass	Secret	Top Secret
Streaming	Smart Push	5	5	5
	User Pull	25	25	5
Binary	Smart Push	5	5	5
	User Pull	25	25	5

- 3 categorical factors
- Unbalanced design mimics operational expectations

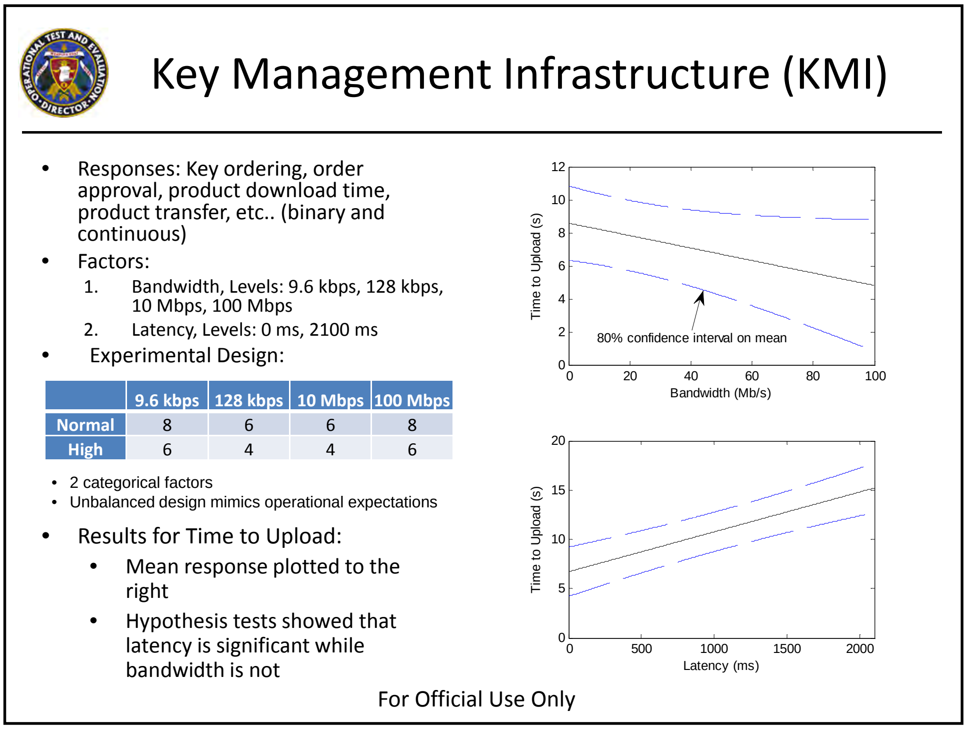
For Official Use Only



Key Management Infrastructure (KMI)

- Key Management Infrastructure (KMI) is designed to provide secure and interoperable cryptographic key generation, distribution, and management capabilities
- KMI is a combination of :
  - nearly 1,500,000 lines of contractor developed code
  - custom-developed hardware in the form of an Advanced Key Processor (AKP)
  - commercial off-the-shelf (COTS) hardware and software.
- KMI will provide a means for the secure ordering, generation, production, distribution, management, and auditing of cryptographic products

- Test Objectives: Determine if bandwidth and latency influence
  1. the ability to connect to server
  2. product transfers (uploads)
  3. the ability to download account credentials
  4. upload transactions, and more…



Key Management Infrastructure (KMI)

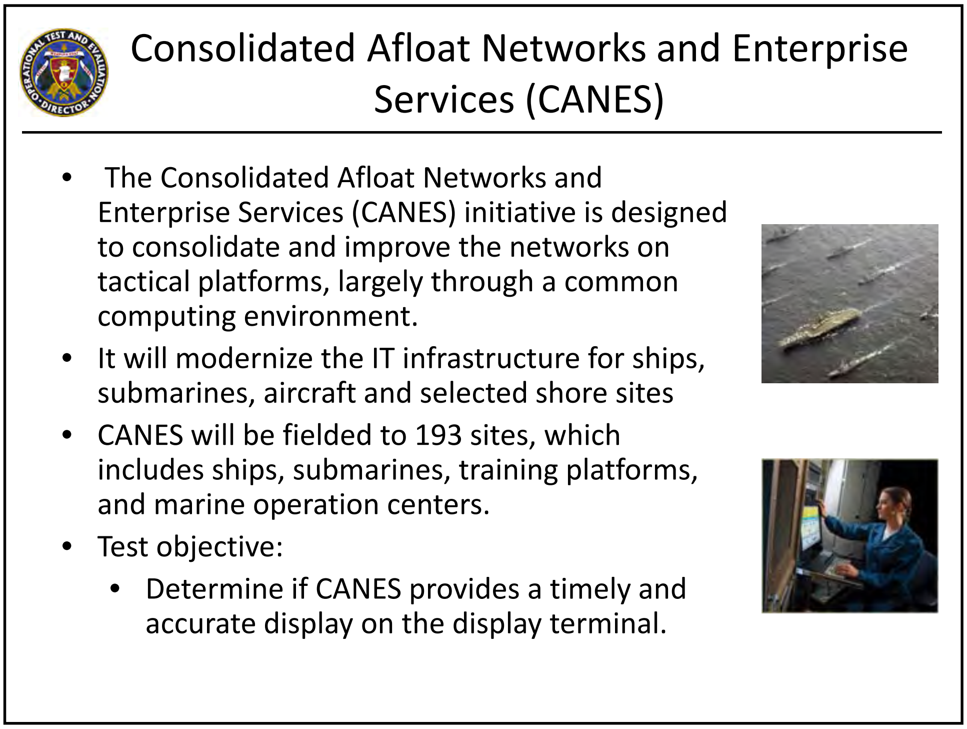
- Responses: Key ordering, order approval, product download time, product transfer, etc.. (binary and continuous)
- Factors:
  1. Bandwidth, Levels: 9.6 kbps, 128 kbps, 10 Mbps, 100 Mbps
  2. Latency, Levels: 0 ms, 2100 ms
- Experimental Design:

	9.6 kbps	128 kbps	10 Mbps	100 Mbps
Normal	8	6	6	8
High	6	4	4	6

- 2 categorical factors
- Unbalanced design mimics operational expectations

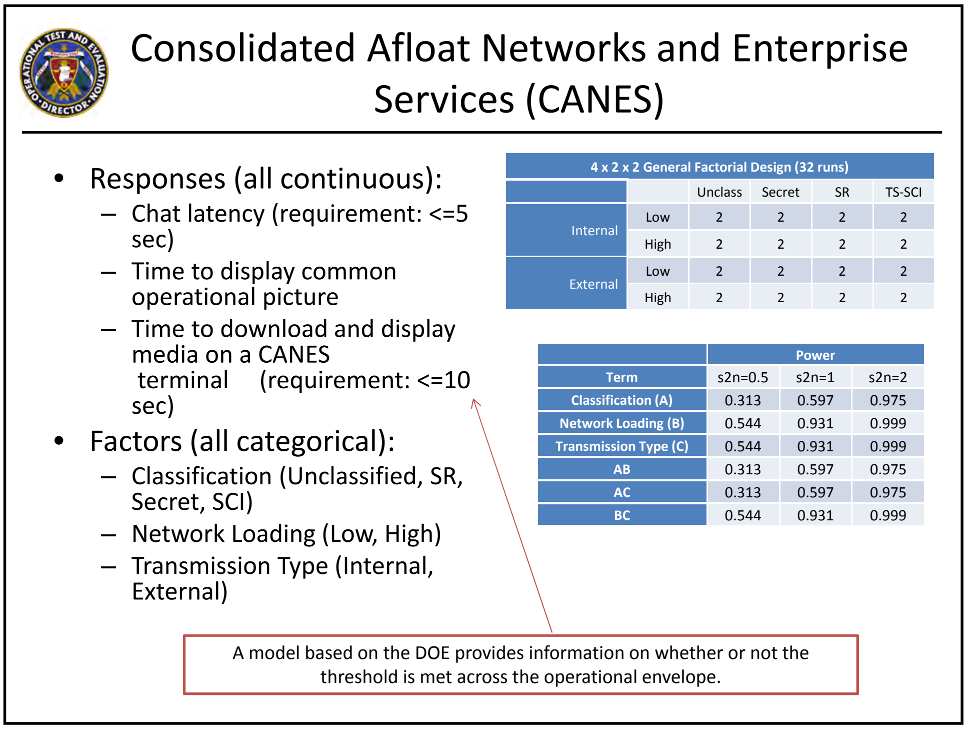
- Results for Time to Upload:
  - Mean response plotted to the right
  - Hypothesis tests showed that latency is significant while bandwidth is not

For Official Use Only



Consolidated Afloat Networks and Enterprise Services (CANES)

- The Consolidated Afloat Networks and Enterprise Services (CANES) initiative is designed to consolidate and improve the networks on tactical platforms, largely through a common computing environment.
- It will modernize the IT infrastructure for ships, submarines, aircraft and selected shore sites
- CANES will be fielded to 193 sites, which includes ships, submarines, training platforms, and marine operation centers.
- Test objective:
  - Determine if CANES provides a timely and accurate display on the display terminal.



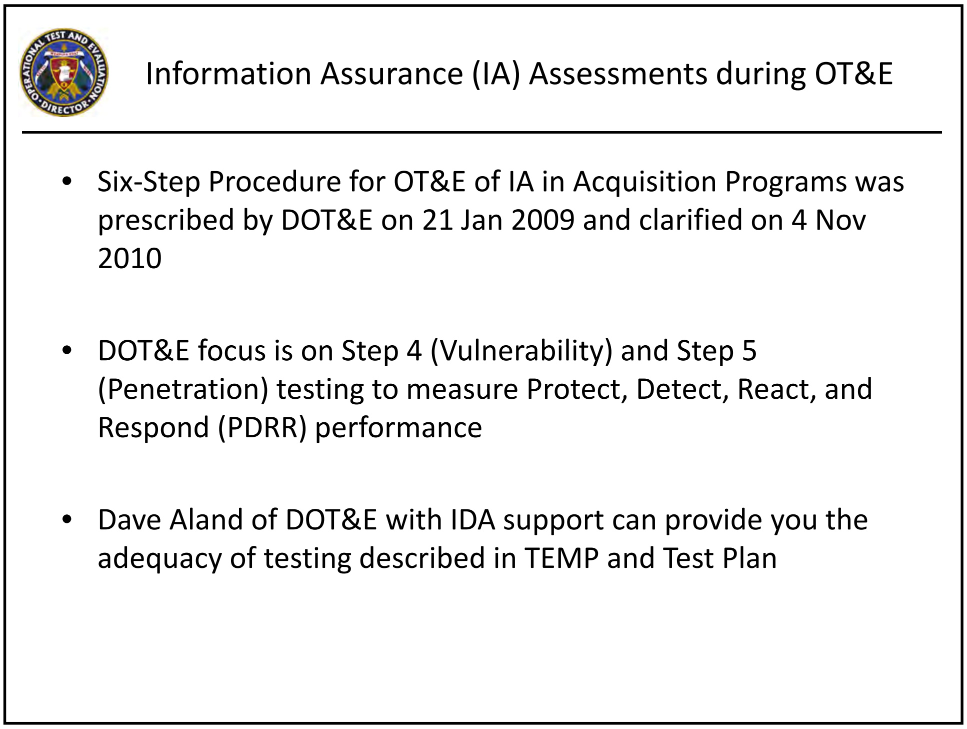
Consolidated Afloat Networks and Enterprise Services (CANES)

- Responses (all continuous):
  - Chat latency (requirement: <=5 sec)
  - Time to display common operational picture
  - Time to download and display media on a CANES terminal (requirement: <=10 sec)
- Factors (all categorical):
  - Classification (Unclassified, SR, Secret, SCI)
  - Network Loading (Low, High)
  - Transmission Type (Internal, External)

4 x 2 x 2 General Factorial Design (32 runs)					
		Unclass	Secret	SR	TS-SCI
Internal	Low	2	2	2	2
	High	2	2	2	2
External	Low	2	2	2	2
	High	2	2	2	2

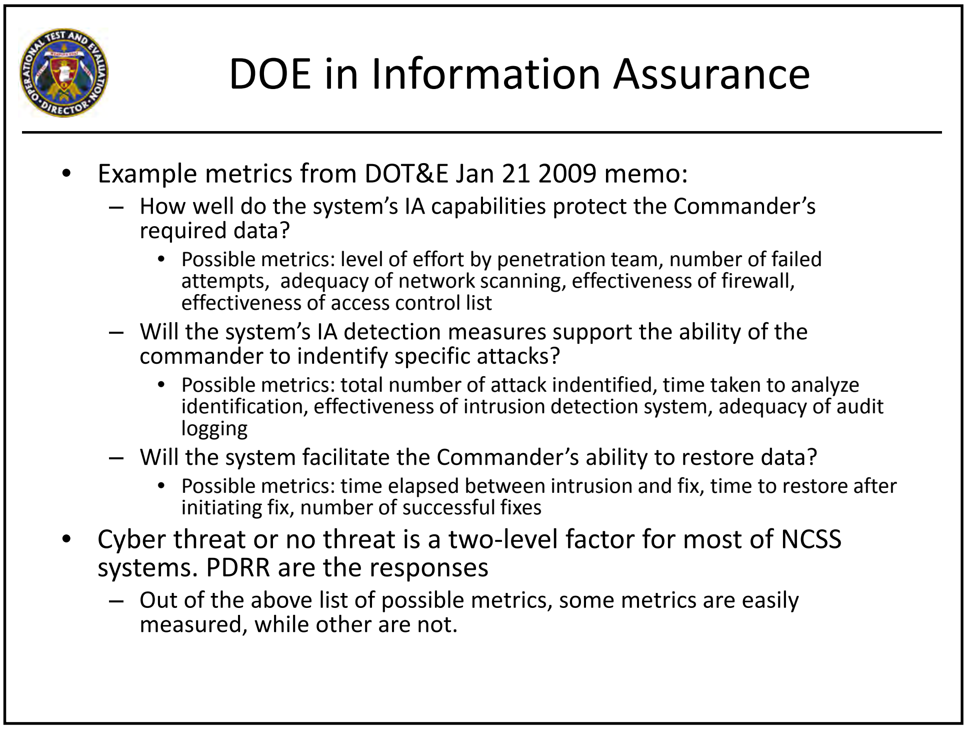
	Power		
Term	s2n=0.5	s2n=1	s2n=2
Classification (A)	0.313	0.597	0.975
Network Loading (B)	0.544	0.931	0.999
Transmission Type (C)	0.544	0.931	0.999
AB	0.313	0.597	0.975
AC	0.313	0.597	0.975
BC	0.544	0.931	0.999

A model based on the DOE provides information on whether or not the threshold is met across the operational envelope.



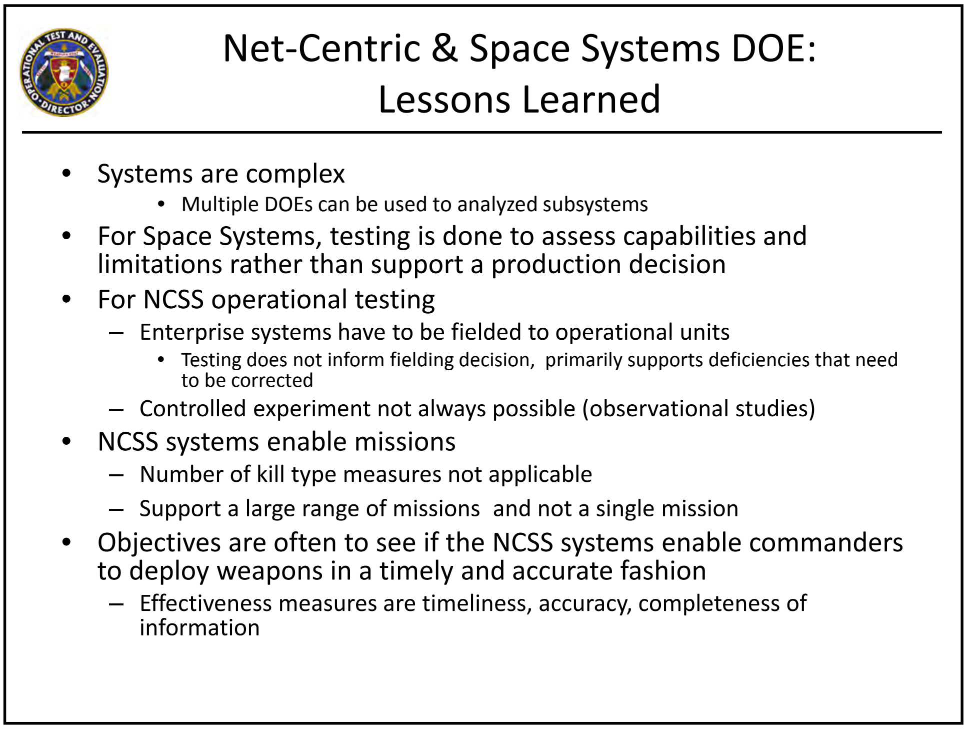
Information Assurance (IA) Assessments during OT&E

- Six-Step Procedure for OT&E of IA in Acquisition Programs was prescribed by DOT&E on 21 Jan 2009 and clarified on 4 Nov 2010
- DOT&E focus is on Step 4 (Vulnerability) and Step 5 (Penetration) testing to measure Protect, Detect, React, and Respond (PDRR) performance
- Dave Aland of DOT&E with IDA support can provide you the adequacy of testing described in TEMP and Test Plan



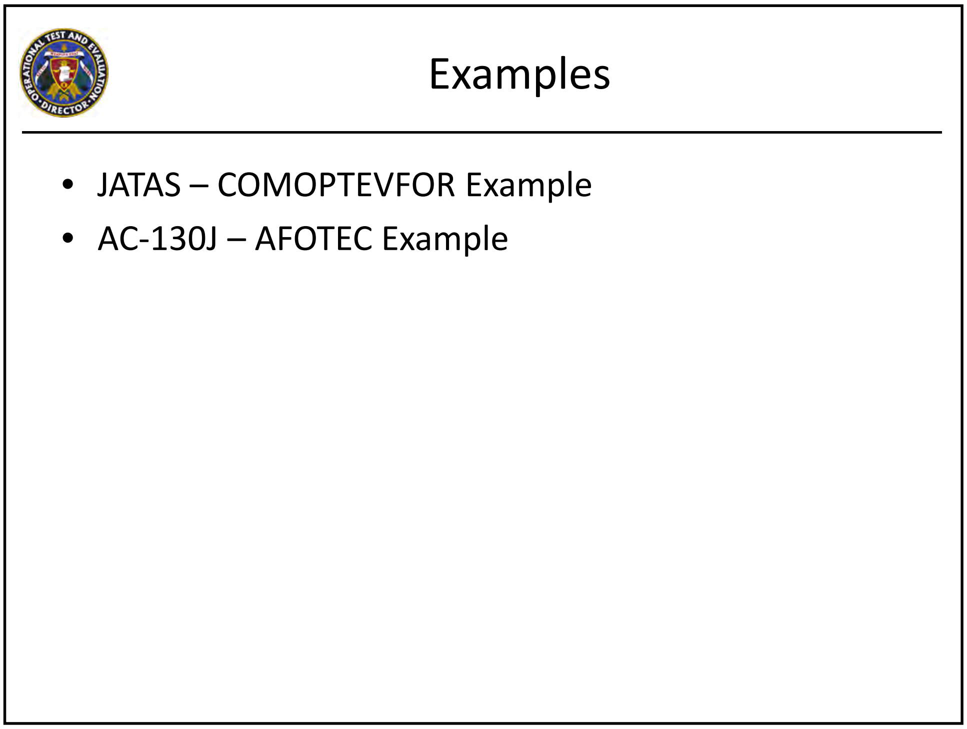
DOE in Information Assurance

- Example metrics from DOT&E Jan 21 2009 memo:
  - How well do the system's IA capabilities protect the Commander's required data?
    - Possible metrics: level of effort by penetration team, number of failed attempts, adequacy of network scanning, effectiveness of firewall, effectiveness of access control list
  - Will the system's IA detection measures support the ability of the commander to indentify specific attacks?
    - Possible metrics: total number of attack indentified, time taken to analyze identification, effectiveness of intrusion detection system, adequacy of audit logging
  - Will the system facilitate the Commander's ability to restore data?
    - Possible metrics: time elapsed between intrusion and fix, time to restore after initiating fix, number of successful fixes
- Cyber threat or no threat is a two-level factor for most of NCSS systems. PDRR are the responses
  - Out of the above list of possible metrics, some metrics are easily measured, while other are not.



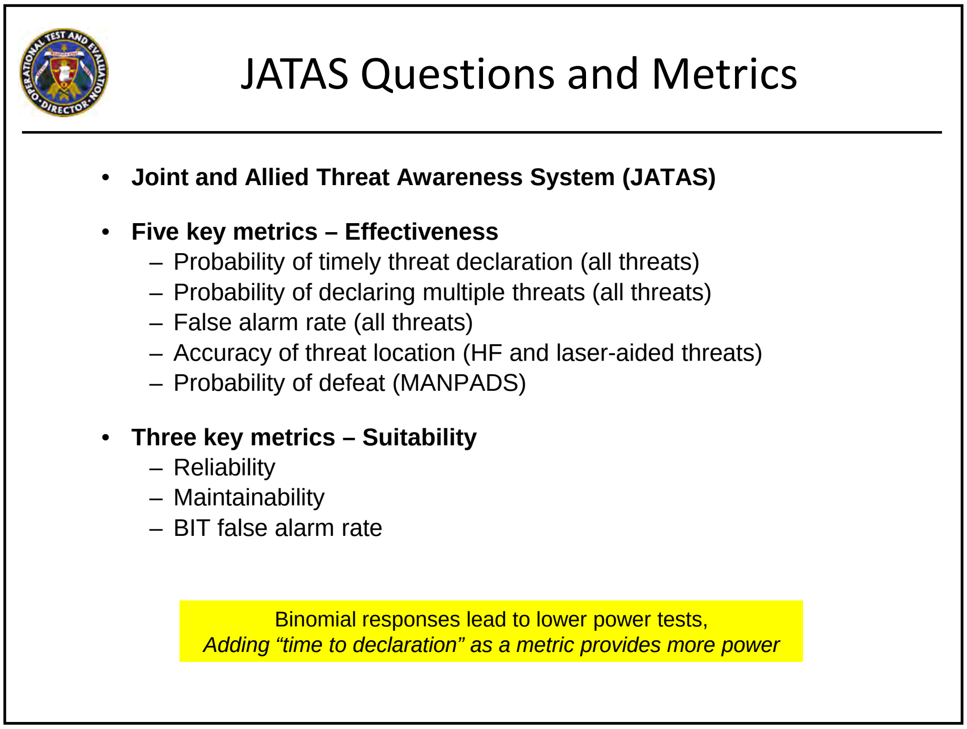
Net-Centric & Space Systems DOE: Lessons Learned

- Systems are complex
    - Multiple DOEs can be used to analyzed subsystems
- For Space Systems, testing is done to assess capabilities and limitations rather than support a production decision
- For NCSS operational testing
    - Enterprise systems have to be fielded to operational units
        - Testing does not inform fielding decision, primarily supports deficiencies that need to be corrected
    - Controlled experiment not always possible (observational studies)
- NCSS systems enable missions
    - Number of kill type measures not applicable
    - Support a large range of missions and not a single mission
- Objectives are often to see if the NCSS systems enable commanders to deploy weapons in a timely and accurate fashion
    - Effectiveness measures are timeliness, accuracy, completeness of information



Examples

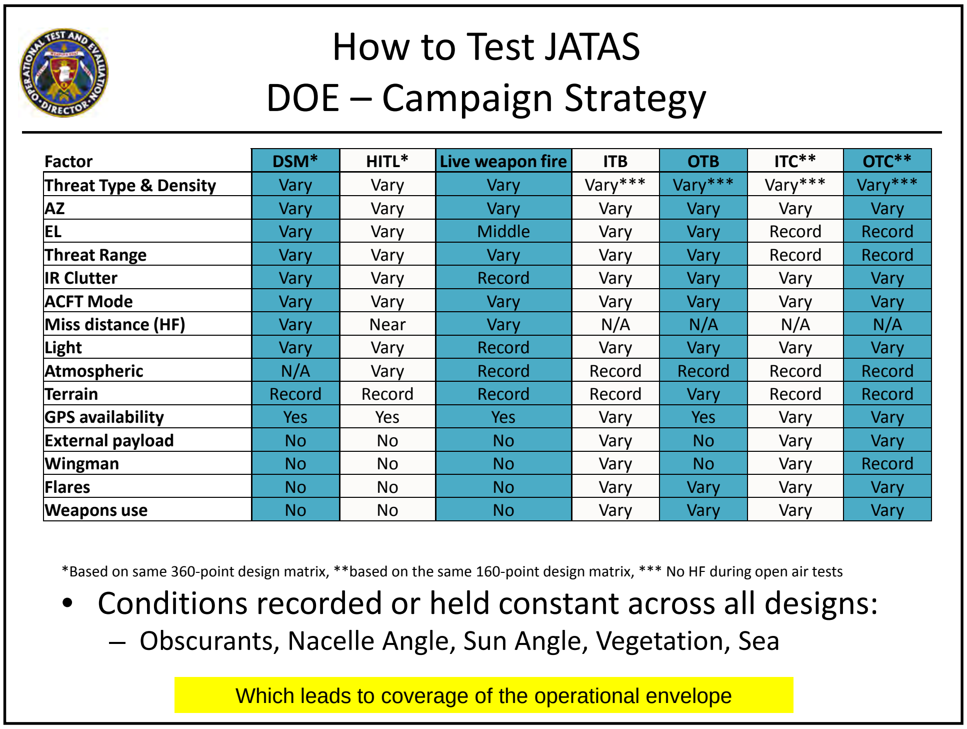
- JATAS – COMOPTEVFOR Example
- AC-130J – AFOTEC Example



JATAS Questions and Metrics

- **Joint and Allied Threat Awareness System (JATAS)**
- **Five key metrics – Effectiveness**
  - Probability of timely threat declaration (all threats)
  - Probability of declaring multiple threats (all threats)
  - False alarm rate (all threats)
  - Accuracy of threat location (HF and laser-aided threats)
  - Probability of defeat (MANPADS)
- **Three key metrics – Suitability**
  - Reliability
  - Maintainability
  - BIT false alarm rate

Binomial responses lead to lower power tests,
Adding "time to declaration" as a metric provides more power



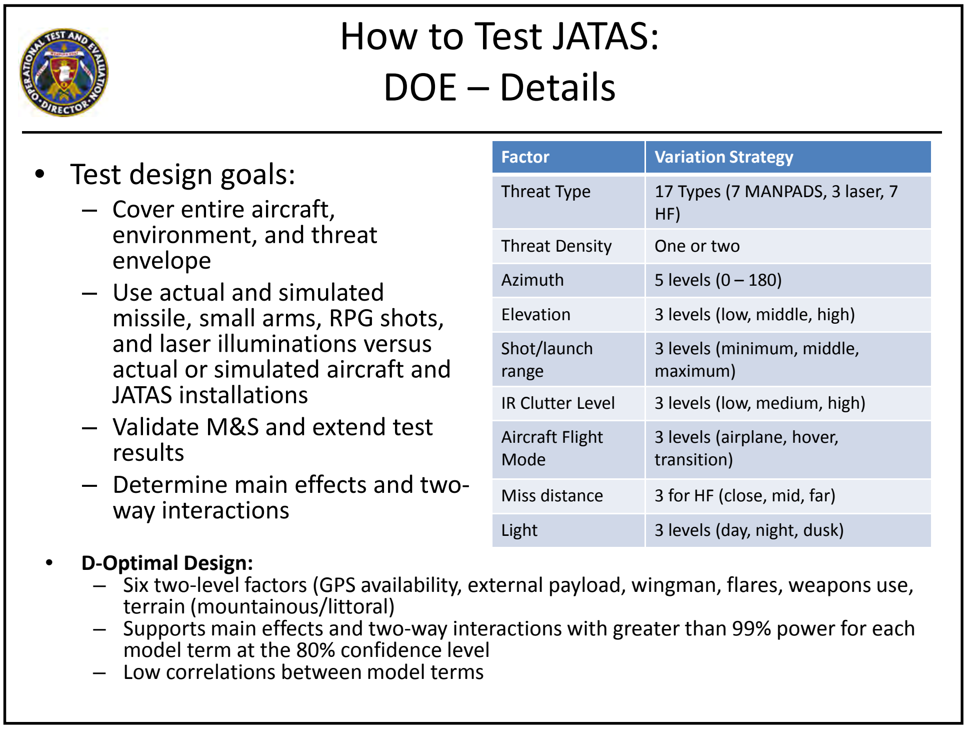
How to Test JATAS
DOE – Campaign Strategy

Factor	DSM*	HITL*	Live weapon fire	ITB	OTB	ITC**	OTC**
Threat Type & Density	Vary	Vary	Vary	Vary***	Vary***	Vary***	Vary***
AZ	Vary	Vary	Vary	Vary	Vary	Vary	Vary
EL	Vary	Vary	Middle	Vary	Vary	Record	Record
Threat Range	Vary	Vary	Vary	Vary	Vary	Record	Record
IR Clutter	Vary	Vary	Record	Vary	Vary	Vary	Vary
ACFT Mode	Vary	Vary	Vary	Vary	Vary	Vary	Vary
Miss distance (HF)	Vary	Near	Vary	N/A	N/A	N/A	N/A
Light	Vary	Vary	Record	Vary	Vary	Vary	Vary
Atmospheric	N/A	Vary	Record	Record	Record	Record	Record
Terrain	Record	Record	Record	Record	Vary	Record	Record
GPS availability	Yes	Yes	Yes	Vary	Yes	Vary	Vary
External payload	No	No	No	Vary	No	Vary	Vary
Wingman	No	No	No	Vary	No	Vary	Record
Flares	No	No	No	Vary	Vary	Vary	Vary
Weapons use	No	No	No	Vary	Vary	Vary	Vary

*Based on same 360-point design matrix, **based on the same 160-point design matrix, *** No HF during open air tests

- Conditions recorded or held constant across all designs:
  - Obscurants, Nacelle Angle, Sun Angle, Vegetation, Sea

Which leads to coverage of the operational envelope

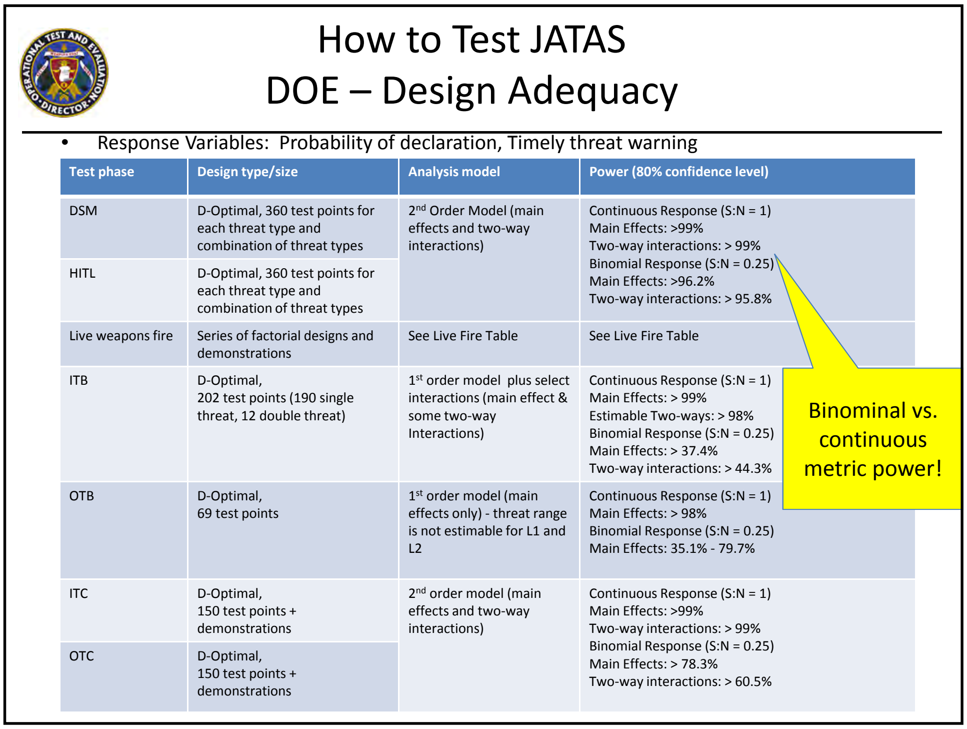


How to Test JATAS: DOE – Details

- Test design goals:
  - Cover entire aircraft, environment, and threat envelope
  - Use actual and simulated missile, small arms, RPG shots, and laser illuminations versus actual or simulated aircraft and JATAS installations
  - Validate M&S and extend test results
  - Determine main effects and two-way interactions

Factor	Variation Strategy
Threat Type	17 Types (7 MANPADS, 3 laser, 7 HF)
Threat Density	One or two
Azimuth	5 levels (0 – 180)
Elevation	3 levels (low, middle, high)
Shot/launch range	3 levels (minimum, middle, maximum)
IR Clutter Level	3 levels (low, medium, high)
Aircraft Flight Mode	3 levels (airplane, hover, transition)
Miss distance	3 for HF (close, mid, far)
Light	3 levels (day, night, dusk)

- **D-Optimal Design:**
  - Six two-level factors (GPS availability, external payload, wingman, flares, weapons use, terrain (mountainous/littoral)
  - Supports main effects and two-way interactions with greater than 99% power for each model term at the 80% confidence level
  - Low correlations between model terms

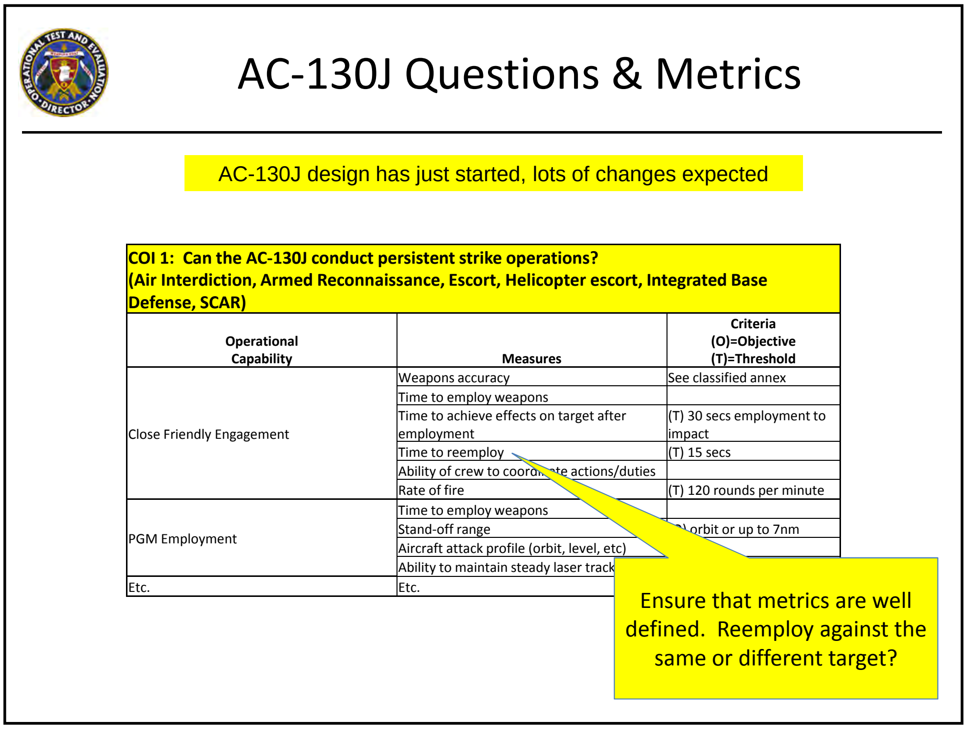


How to Test JATAS DOE – Design Adequacy

- Response Variables: Probability of declaration, Timely threat warning

Test phase	Design type/size	Analysis model	Power (80% confidence level)
DSM	D-Optimal, 360 test points for each threat type and combination of threat types	2nd Order Model (main effects and two-way interactions)	Continuous Response (S:N = 1) Main Effects: >99% Two-way interactions: > 99% Binomial Response (S:N = 0.25) Main Effects: >96.2% Two-way interactions: > 95.8%
HITL	D-Optimal, 360 test points for each threat type and combination of threat types		
Live weapons fire	Series of factorial designs and demonstrations	See Live Fire Table	See Live Fire Table
ITB	D-Optimal, 202 test points (190 single threat, 12 double threat)	1st order model plus select interactions (main effect & some two-way Interactions)	Continuous Response (S:N = 1) Main Effects: > 99% Estimable Two-ways: > 98% Binomial Response (S:N = 0.25) Main Effects: > 37.4% Two-way interactions: > 44.3%
OTB	D-Optimal, 69 test points	1st order model (main effects only) - threat range is not estimable for L1 and L2	Continuous Response (S:N = 1) Main Effects: > 98% Binomial Response (S:N = 0.25) Main Effects: 35.1% - 79.7%
ITC	D-Optimal, 150 test points + demonstrations	2nd order model (main effects and two-way interactions)	Continuous Response (S:N = 1) Main Effects: >99% Two-way interactions: > 99% Binomial Response (S:N = 0.25) Main Effects: > 78.3% Two-way interactions: > 60.5%
OTC	D-Optimal, 150 test points + demonstrations		

Binominal vs. continuous metric power!

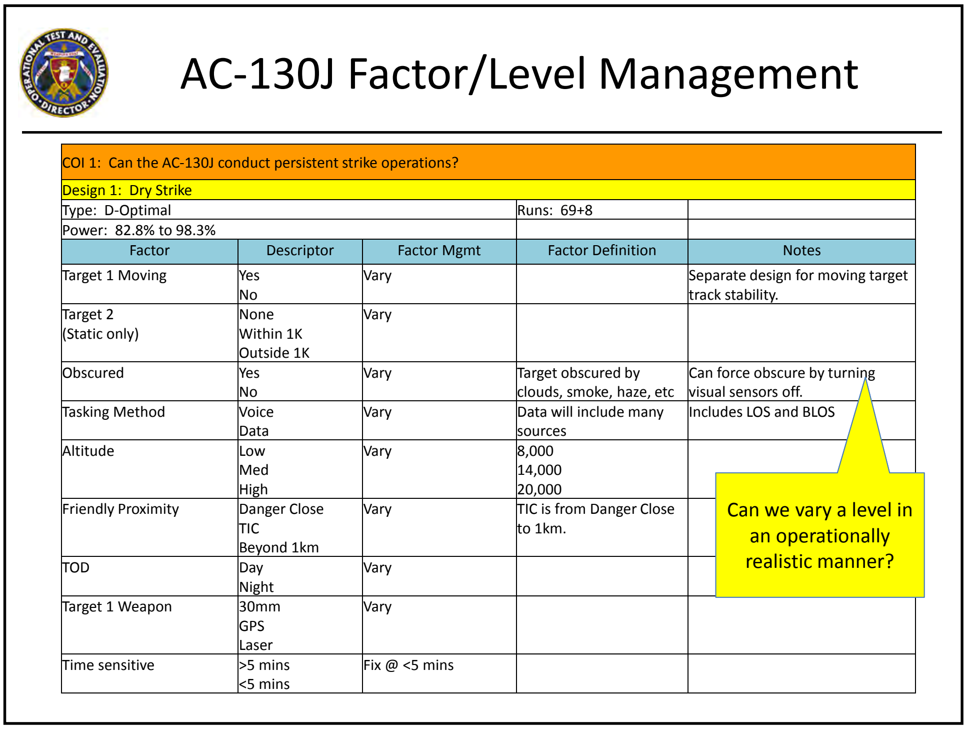


AC-130J Questions & Metrics

AC-130J design has just started, lots of changes expected

COI 1: Can the AC-130J conduct persistent strike operations? (Air Interdiction, Armed Reconnaissance, Escort, Helicopter escort, Integrated Base Defense, SCAR)		
Operational Capability	**Measures**	**Criteria (O)=Objective (T)=Threshold**
Close Friendly Engagement	Weapons accuracy	See classified annex
	Time to employ weapons	
	Time to achieve effects on target after employment	(T) 30 secs employment to impact
	Time to reemploy	(T) 15 secs
	Ability of crew to coordinate actions/duties	
	Rate of fire	(T) 120 rounds per minute
PGM Employment	Time to employ weapons	
	Stand-off range	(O) orbit or up to 7nm
	Aircraft attack profile (orbit, level, etc)	
	Ability to maintain steady laser track	
Etc.	Etc.	

Ensure that metrics are well defined. Reemploy against the same or different target?



AC-130J Factor/Level Management

COI 1: Can the AC-130J conduct persistent strike operations?				
Design 1: Dry Strike				
Type: D-Optimal			Runs: 69+8	
Power: 82.8% to 98.3%				
Factor	Descriptor	Factor Mgmt	Factor Definition	Notes
Target 1 Moving	Yes No	Vary		Separate design for moving target track stability.
Target 2 (Static only)	None Within 1K Outside 1K	Vary		
Obscured	Yes No	Vary	Target obscured by clouds, smoke, haze, etc	Can force obscure by turning visual sensors off.
Tasking Method	Voice Data	Vary	Data will include many sources	Includes LOS and BLOS
Altitude	Low Med High	Vary	8,000 14,000 20,000	
Friendly Proximity	Danger Close TIC Beyond 1km	Vary	TIC is from Danger Close to 1km.	
TOD	Day Night	Vary		
Target 1 Weapon	30mm GPS Laser	Vary		
Time sensitive	>5 mins <5 mins	Fix @ <5 mins		

Can we vary a level in an operationally realistic manner?

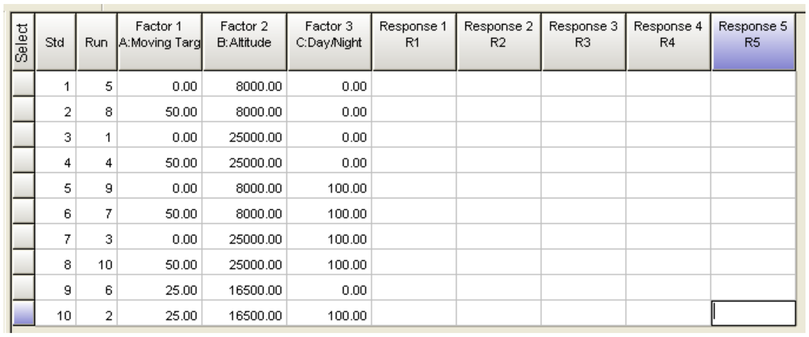
AFOTEC Factor Prioritization

		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

How do we prioritize the factors/levels?

AC-130J COI 1 – SOPGM Demo

DEMO Standoff Precision Guided Munitions (SOPGM)
Factorial, 2^3 w/2 center points
Signal/Noise = 2 for all responses
Power less than 80% for demo (65.7%)



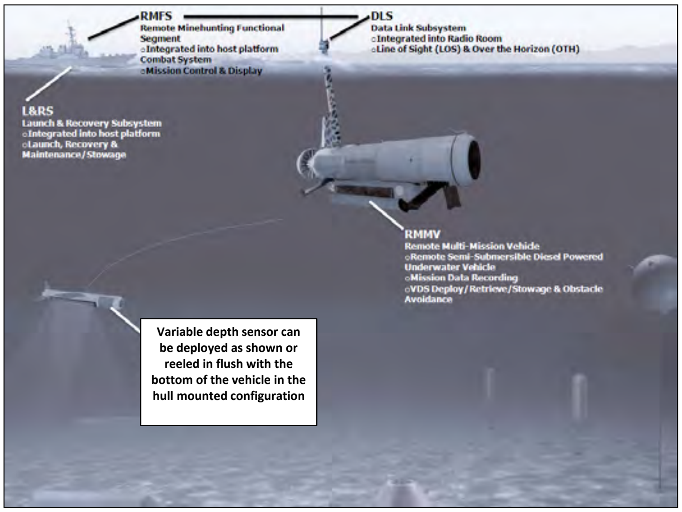
Select	Std	Run	Factor 1 A:Moving Targ	Factor 2 B:Altitude	Factor 3 C:Day/Night	Response 1 R1	Response 2 R2	Response 3 R3	Response 4 R4	Response 5 R5
	1	5	0.00	8000.00	0.00					
	2	8	50.00	8000.00	0.00					
	3	1	0.00	25000.00	0.00					
	4	4	50.00	25000.00	0.00					
	5	9	0.00	8000.00	100.00					
	6	7	50.00	8000.00	100.00					
	7	3	0.00	25000.00	100.00					
	8	10	50.00	25000.00	100.00					
	9	6	25.00	16500.00	0.00					
	10	2	25.00	16500.00	100.00					

Power maybe low in a demo, but we still want to know what it is!

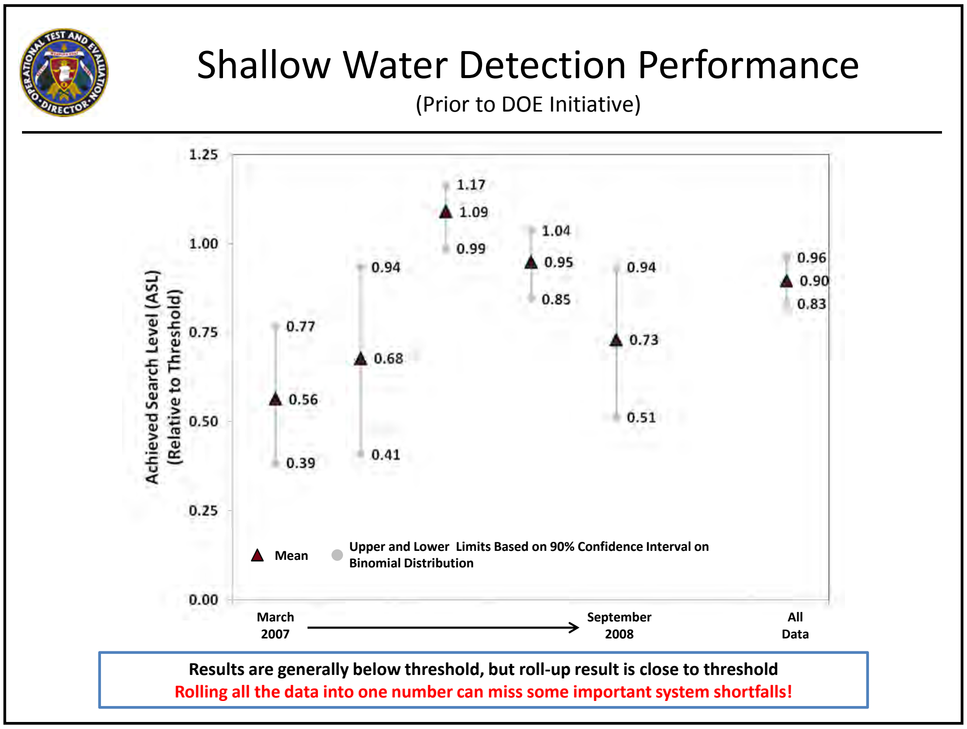
Examples

- Remote Mine-Hunting System (RMS)
- Cargo ship testing using Advanced Mine Simulation System (AMISS)

Remote Mine-hunting System


RMFS
Remote Minehunting Functional Segment
○Integrated into host platform Combat System
○Mission Control & Display
DLS
Data Link Subsystem
○Integrated into Radio Room
○Line of Sight (LOS) & Over the Horizon (OTH)
L&RS
Launch & Recovery Subsystem
○Integrated into host platform
○Launch, Recovery & Maintenance/Stowage
RMMV
Remote Multi-Mission Vehicle
○Remote Semi-Submersible Diesel Powered Underwater Vehicle
○Mission Data Recording
○VDS Deploy/Retrieve/Stowage & Obstacle Avoidance
Variable depth sensor can be deployed as shown or reeled in flush with the bottom of the vehicle in the hull mounted configuration

Does system detection performance support minehunting objectives?

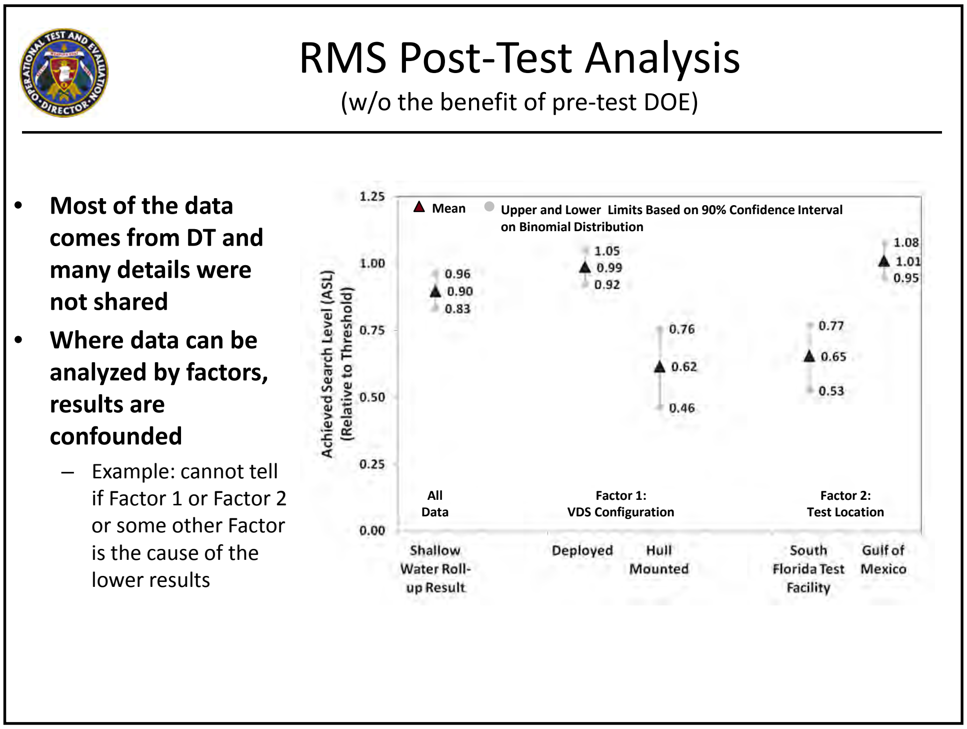


Shallow Water Detection Performance

(Prior to DOE Initiative)

Results are generally below threshold, but roll-up result is close to threshold

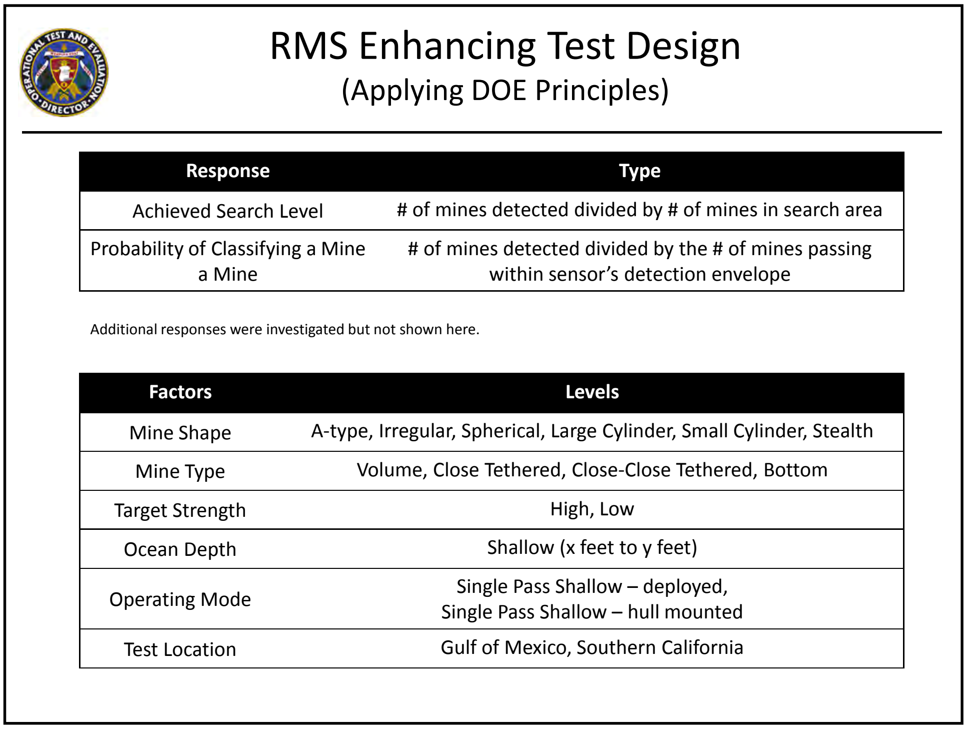
Rolling all the data into one number can miss some important system shortfalls!



RMS Post-Test Analysis

(w/o the benefit of pre-test DOE)

- **Most of the data comes from DT and many details were not shared**
- **Where data can be analyzed by factors, results are confounded**
  - Example: cannot tell if Factor 1 or Factor 2 or some other Factor is the cause of the lower results



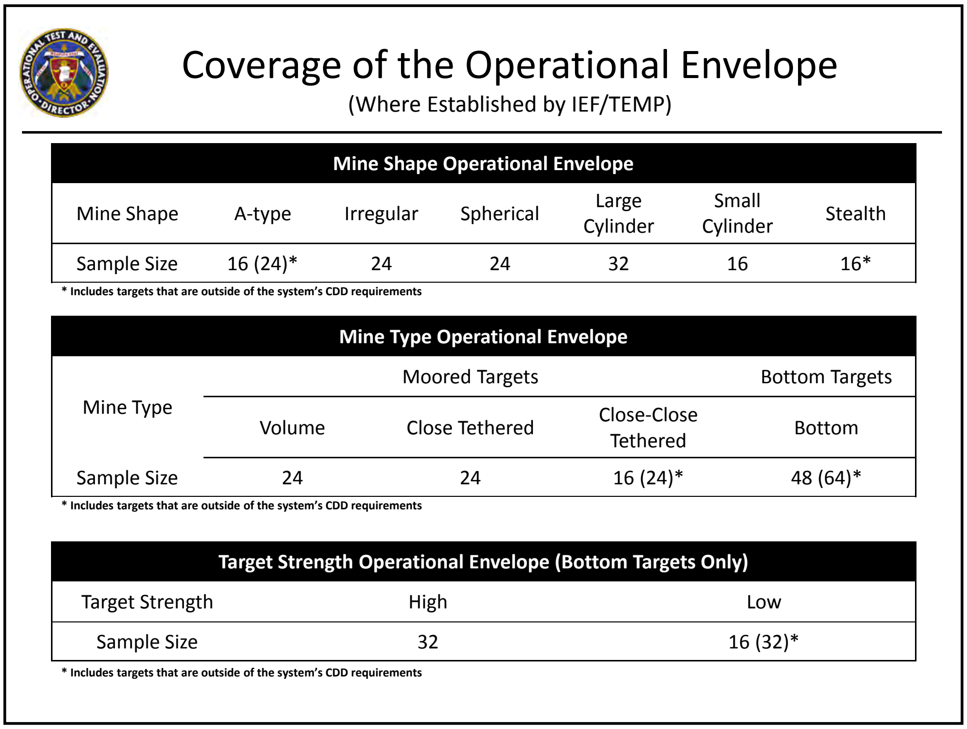
RMS Enhancing Test Design

(Applying DOE Principles)

Response	Type
Achieved Search Level	# of mines detected divided by # of mines in search area
Probability of Classifying a Mine a Mine	# of mines detected divided by the # of mines passing within sensor's detection envelope

Additional responses were investigated but not shown here.

Factors	Levels
Mine Shape	A-type, Irregular, Spherical, Large Cylinder, Small Cylinder, Stealth
Mine Type	Volume, Close Tethered, Close-Close Tethered, Bottom
Target Strength	High, Low
Ocean Depth	Shallow (x feet to y feet)
Operating Mode	Single Pass Shallow – deployed, Single Pass Shallow – hull mounted
Test Location	Gulf of Mexico, Southern California



Coverage of the Operational Envelope

(Where Established by IEF/TEMP)

Mine Shape Operational Envelope

Mine Shape	A-type	Irregular	Spherical	Large Cylinder	Small Cylinder	Stealth
Sample Size	16 (24)*	24	24	32	16	16*

* Includes targets that are outside of the system's CDD requirements

Mine Type Operational Envelope

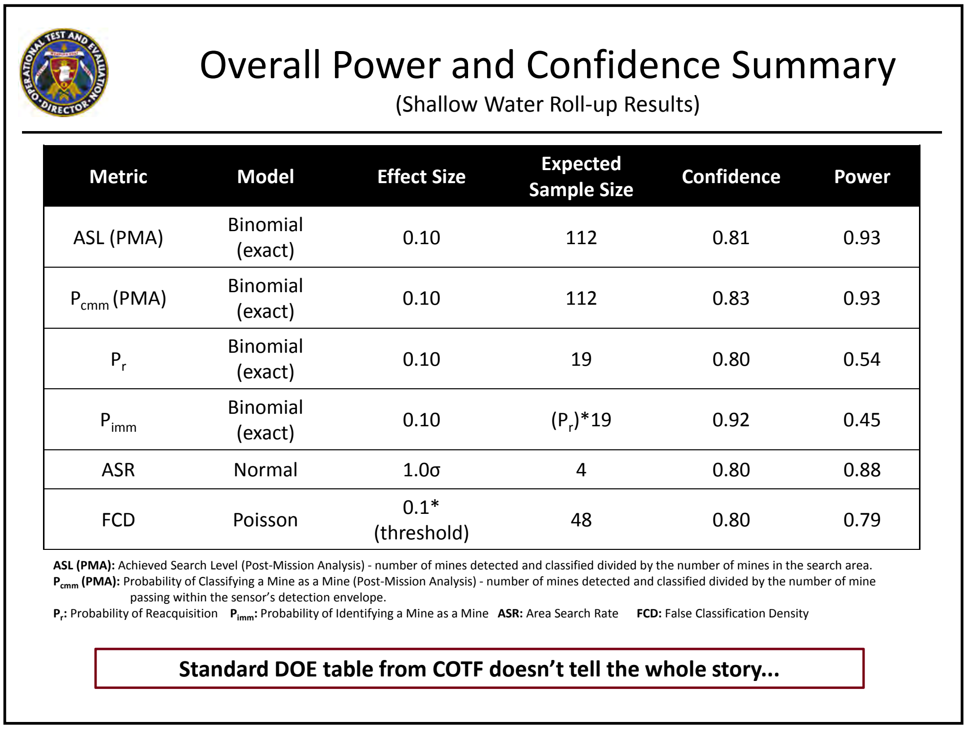
Mine Type	Moored Targets			Bottom Targets
	Volume	Close Tethered	Close-Close Tethered	Bottom
Sample Size	24	24	16 (24)*	48 (64)*

* Includes targets that are outside of the system's CDD requirements

Target Strength Operational Envelope (Bottom Targets Only)

Target Strength	High	Low
Sample Size	32	16 (32)*

* Includes targets that are outside of the system's CDD requirements



Overall Power and Confidence Summary

(Shallow Water Roll-up Results)

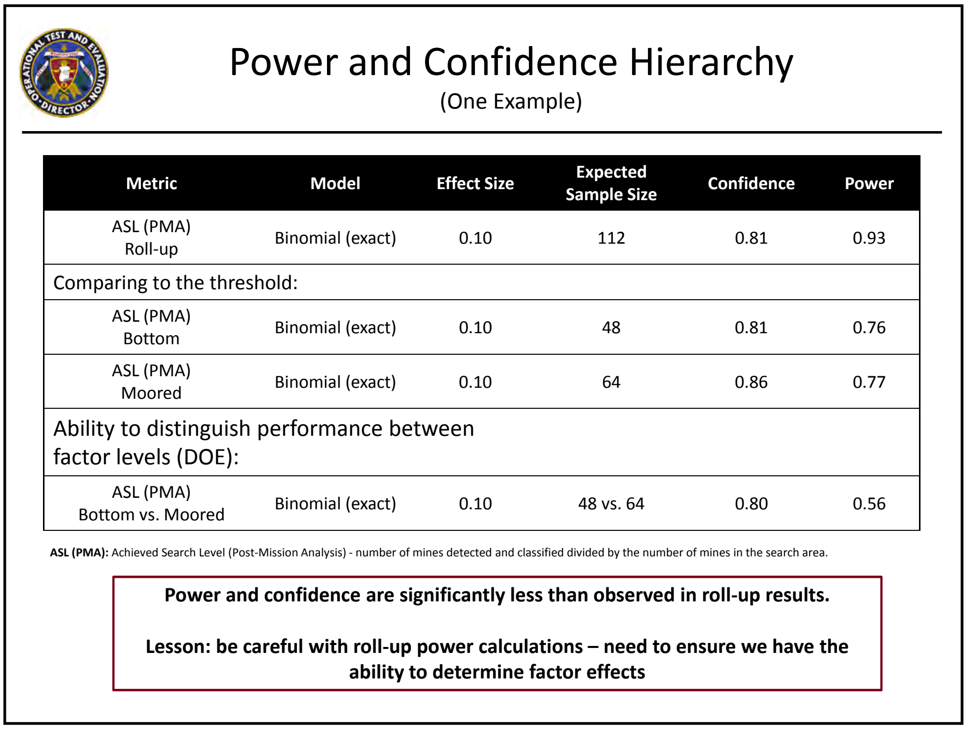
Metric	Model	Effect Size	Expected Sample Size	Confidence	Power
ASL (PMA)	Binomial (exact)	0.10	112	0.81	0.93
P_{cmm} (PMA)	Binomial (exact)	0.10	112	0.83	0.93
P_r	Binomial (exact)	0.10	19	0.80	0.54
P_{imm}	Binomial (exact)	0.10	(P_r)*19	0.92	0.45
ASR	Normal	1.0σ	4	0.80	0.88
FCD	Poisson	0.1* (threshold)	48	0.80	0.79

ASL (PMA): Achieved Search Level (Post-Mission Analysis) - number of mines detected and classified divided by the number of mines in the search area.

P_{cmm} (PMA): Probability of Classifying a Mine as a Mine (Post-Mission Analysis) - number of mines detected and classified divided by the number of mine passing within the sensor's detection envelope.

P_r: Probability of Reacquisition **P_{imm}:** Probability of Identifying a Mine as a Mine **ASR:** Area Search Rate **FCD:** False Classification Density

Standard DOE table from COTF doesn't tell the whole story...



Power and Confidence Hierarchy

(One Example)

Metric	Model	Effect Size	Expected Sample Size	Confidence	Power
ASL (PMA) Roll-up	Binomial (exact)	0.10	112	0.81	0.93
Comparing to the threshold:					
ASL (PMA) Bottom	Binomial (exact)	0.10	48	0.81	0.76
ASL (PMA) Moored	Binomial (exact)	0.10	64	0.86	0.77
Ability to distinguish performance between factor levels (DOE):					
ASL (PMA) Bottom vs. Moored	Binomial (exact)	0.10	48 vs. 64	0.80	0.56

ASL (PMA): Achieved Search Level (Post-Mission Analysis) - number of mines detected and classified divided by the number of mines in the search area.

Power and confidence are significantly less than observed in roll-up results.

Lesson: be careful with roll-up power calculations – need to ensure we have the ability to determine factor effects

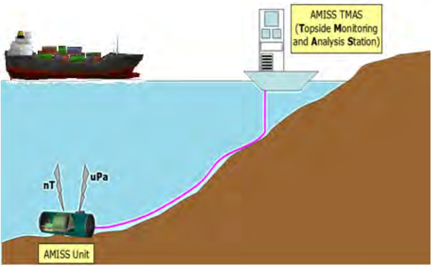
RMS Considerations for Test Plans

(DOE related)

- Ocean depth, shallow water operating mode, and test location are known performance factors not specifically addressed by the DOE in the system’s IEF
  - Test plans need to ensure appropriate coverage of the operational envelope
    - May need to spread target resources over multiple test fields in order to accommodate detection opportunities at various water/case depths
    - Directly related to decision to conduct operations in hull mounted or deployed configuration (i.e., operating mode)
    - Can be augmented by DT data provided it is operationally representative (including appropriate DT data will produce a more powerful assessment)
  - Test plans need to vary test location in order to assess system performance in different environments
    - TEMP resources call for one Operational Test in Gulf of Mexico and another Operational Test in Southern California Operating Area (not executing tests in different locations will produce a less powerful assessment)

Example: Adequate Test Plans for Mine Susceptibility

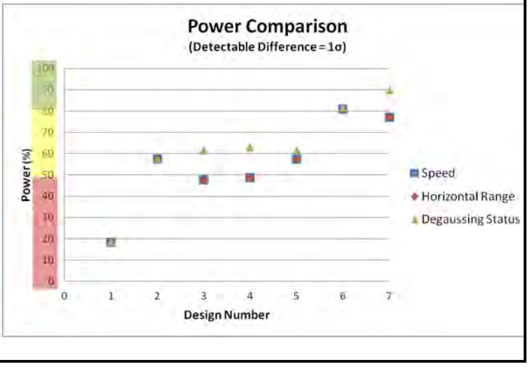
- Goal:
  - Develop an adequate test to assess the susceptibility of a cargo ship against a variety of mine types using the Advanced Mine Simulation System (AMISS).
- Responses:
  - Magnetic signature, acoustic signature, pressure
- Factors:
  - Speed, range, degaussing s



Design of Experiments Solution

- A reasonable test size was considered to be between 15 and 30 runs
- Compared several statistical designs and selected a replicated central composite design
  - Maximized power across all three factors
  - Provides 5 levels of range for maximum flexibility

	Design Type	Number of Runs	Model Terms
1	Full Factorial (2-level)	8	6
2	Full Factorial (2-level) replicated	16	7
3	General Factorial (3x3x2)	18	9
4	Central Composite Design (w/ 1 center point)	18	9
5	Central Composite Design (replicated center point)	20	9
6	Central composite Design with replicated factorial points (Large CCD)	28	9
7	Replicated General Factorial	36	9



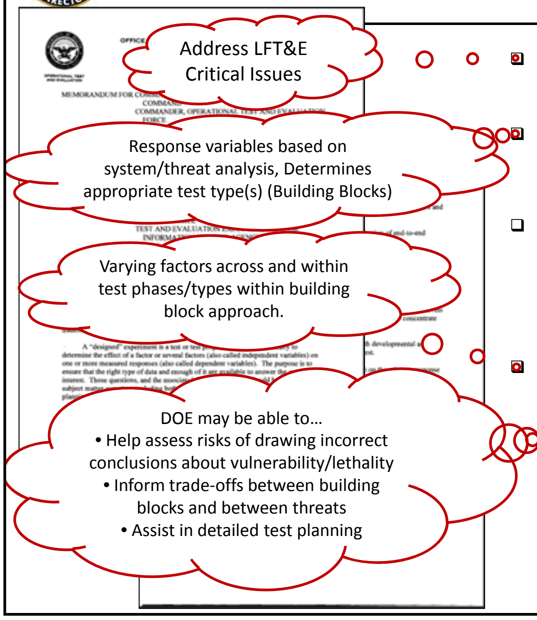
Live Fire Examples

- DOE for Armored Vehicles
- DOE for Fixed Wing Aircraft
- DOE for Ships

Examples are vulnerability focused, could also apply to other areas of LFT&E.

Guidance

Dr. Gilmore's October 19, 2010 Memo to OTAs



Address LFT&E Critical Issues

Response variables based on system/threat analysis, Determines appropriate test type(s) (Building Blocks)

Varying factors across and within test phases/types within building block approach.

DOE may be able to…
- Help assess risks of drawing incorrect conclusions about vulnerability/lethality
- Inform trade-offs between building blocks and between threats
- Assist in detailed test planning

☒ **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.)

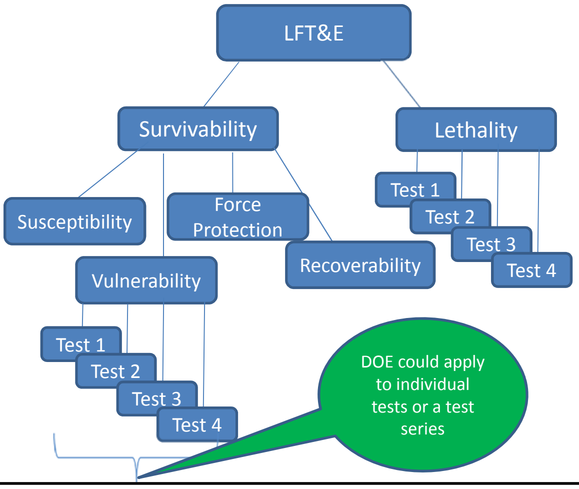
❑ **Factors** that affect those measures of effectiveness and suitability. Systematically, in a rigorous and structured way, develop a test plan that provides good breadth of coverage of those factors across the applicable levels of the factors, taking into account known information in order to concentrate on the factors of most interest.

☒ **A method for strategically varying factors** across both developmental and operational testing with respect to responses of interest.

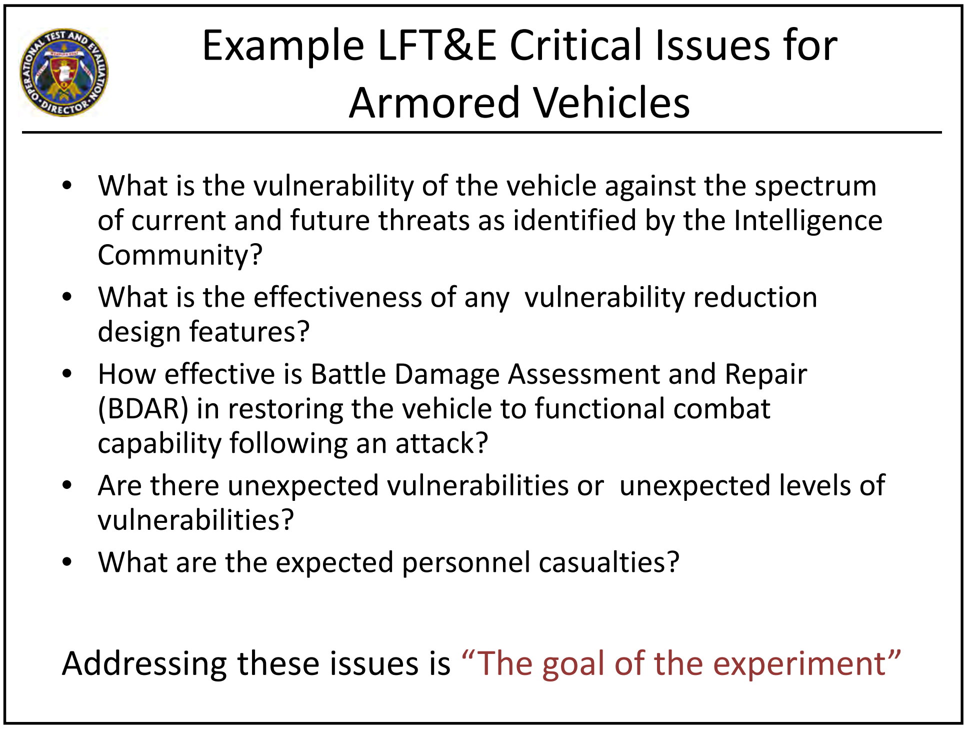
☒ **Statistical measures of merit (power and confidence)** on the relevant response variables for which it makes sense. These statistical measures are important to understand "how much testing is enough?" and can be evaluated by decision makers on a quantitative basis so they can trade off test resources for desired confidence in results.

DOE Applied to LFT&E

- DOE could apply within each test "building block" or across multiple building blocks



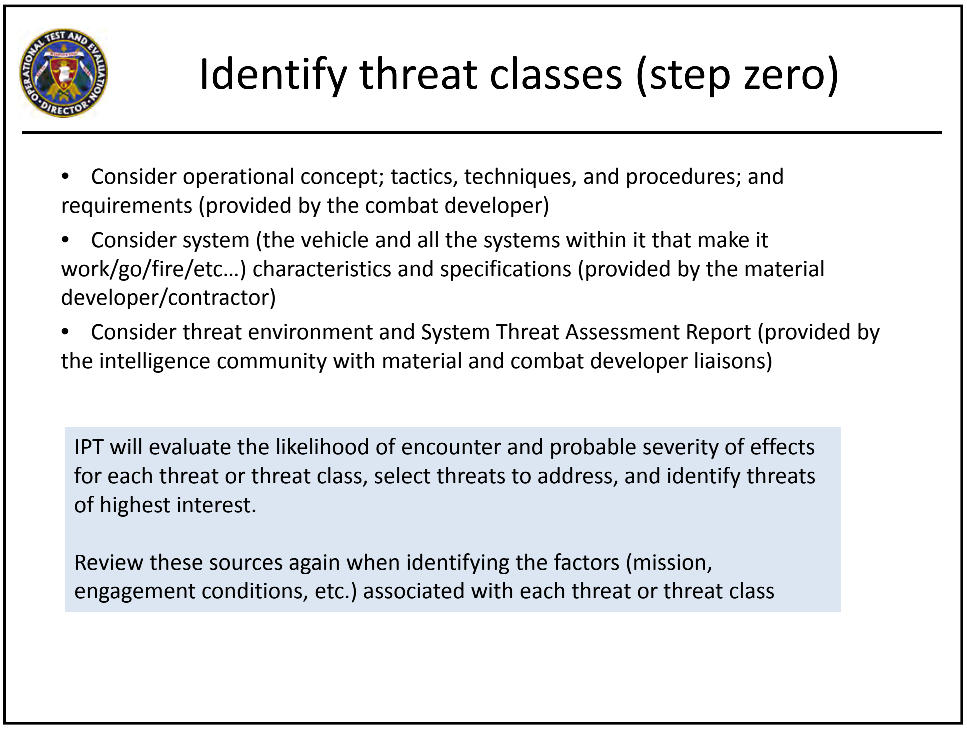
DOE could apply to individual tests or a test series



Example LFT&E Critical Issues for Armored Vehicles

- What is the vulnerability of the vehicle against the spectrum of current and future threats as identified by the Intelligence Community?
- What is the effectiveness of any vulnerability reduction design features?
- How effective is Battle Damage Assessment and Repair (BDAR) in restoring the vehicle to functional combat capability following an attack?
- Are there unexpected vulnerabilities or unexpected levels of vulnerabilities?
- What are the expected personnel casualties?

Addressing these issues is “The goal of the experiment”

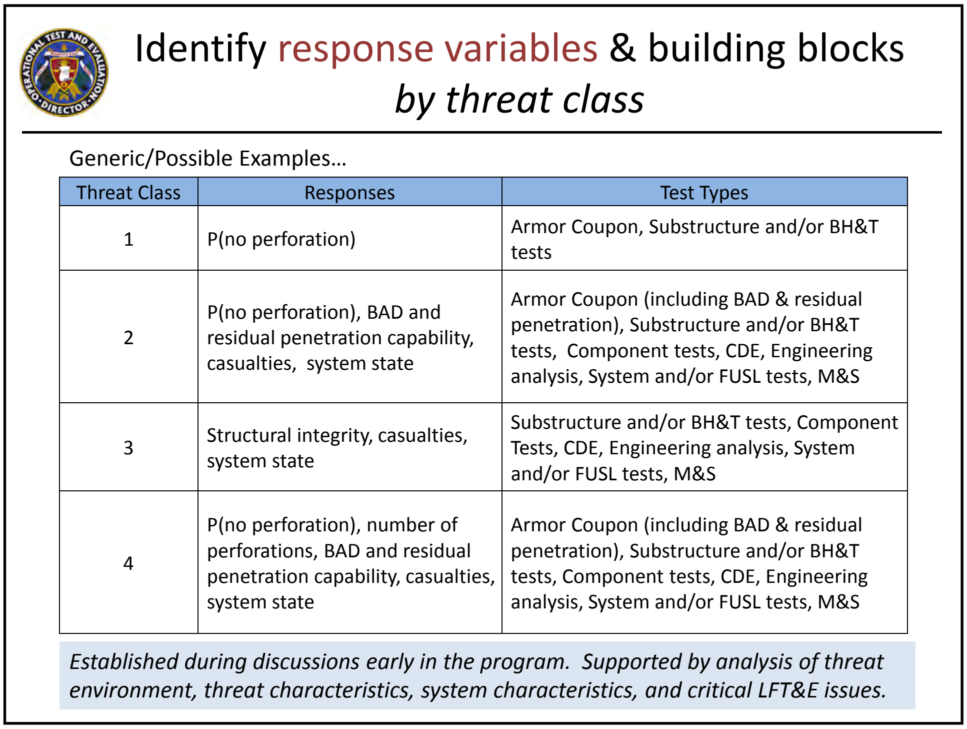


Identify threat classes (step zero)

- Consider operational concept; tactics, techniques, and procedures; and requirements (provided by the combat developer)
- Consider system (the vehicle and all the systems within it that make it work/go/fire/etc…) characteristics and specifications (provided by the material developer/contractor)
- Consider threat environment and System Threat Assessment Report (provided by the intelligence community with material and combat developer liaisons)

IPT will evaluate the likelihood of encounter and probable severity of effects for each threat or threat class, select threats to address, and identify threats of highest interest.

Review these sources again when identifying the factors (mission, engagement conditions, etc.) associated with each threat or threat class

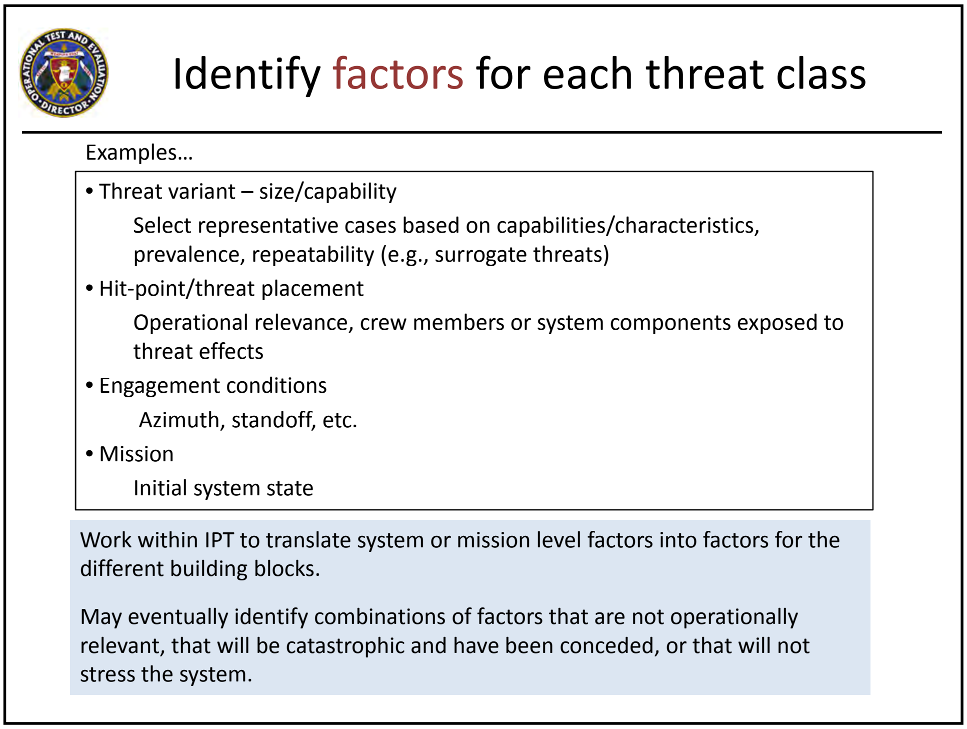


Identify response variables & building blocks *by threat class*

Generic/Possible Examples…

Threat Class	Responses	Test Types
1	P(no perforation)	Armor Coupon, Substructure and/or BH&T tests
2	P(no perforation), BAD and residual penetration capability, casualties, system state	Armor Coupon (including BAD & residual penetration), Substructure and/or BH&T tests, Component tests, CDE, Engineering analysis, System and/or FUSL tests, M&S
3	Structural integrity, casualties, system state	Substructure and/or BH&T tests, Component Tests, CDE, Engineering analysis, System and/or FUSL tests, M&S
4	P(no perforation), number of perforations, BAD and residual penetration capability, casualties, system state	Armor Coupon (including BAD & residual penetration), Substructure and/or BH&T tests, Component tests, CDE, Engineering analysis, System and/or FUSL tests, M&S

Established during discussions early in the program. Supported by analysis of threat environment, threat characteristics, system characteristics, and critical LFT&E issues.



Identify factors for each threat class

Examples…

- Threat variant – size/capability

  Select representative cases based on capabilities/characteristics, prevalence, repeatability (e.g., surrogate threats)
- Hit-point/threat placement

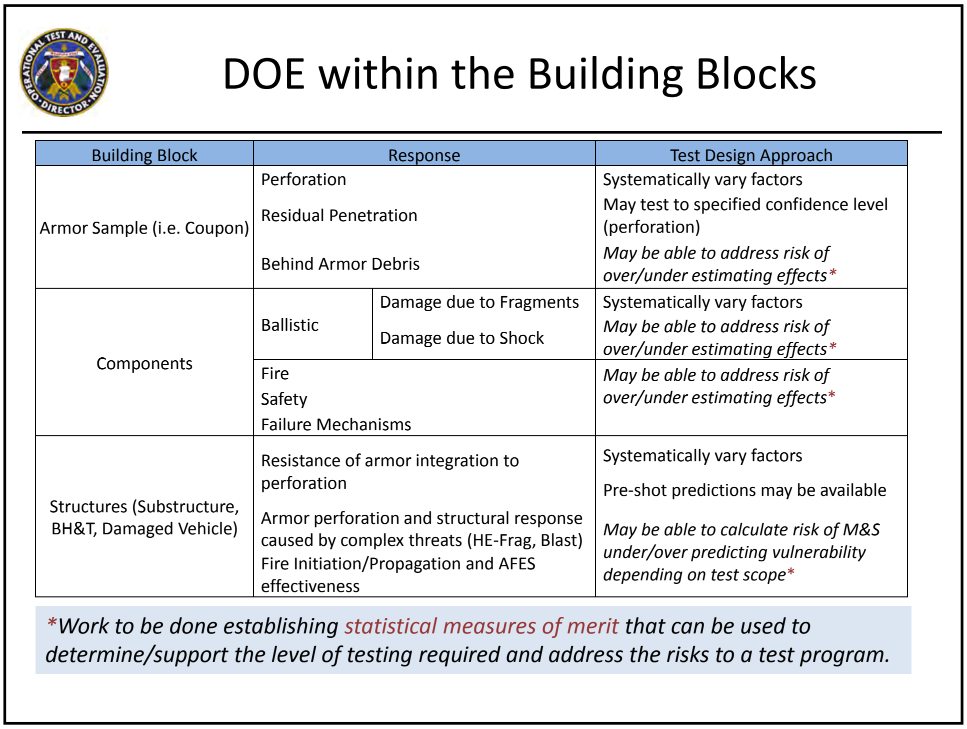
  Operational relevance, crew members or system components exposed to threat effects
- Engagement conditions

  Azimuth, standoff, etc.
- Mission

  Initial system state

Work within IPT to translate system or mission level factors into factors for the different building blocks.

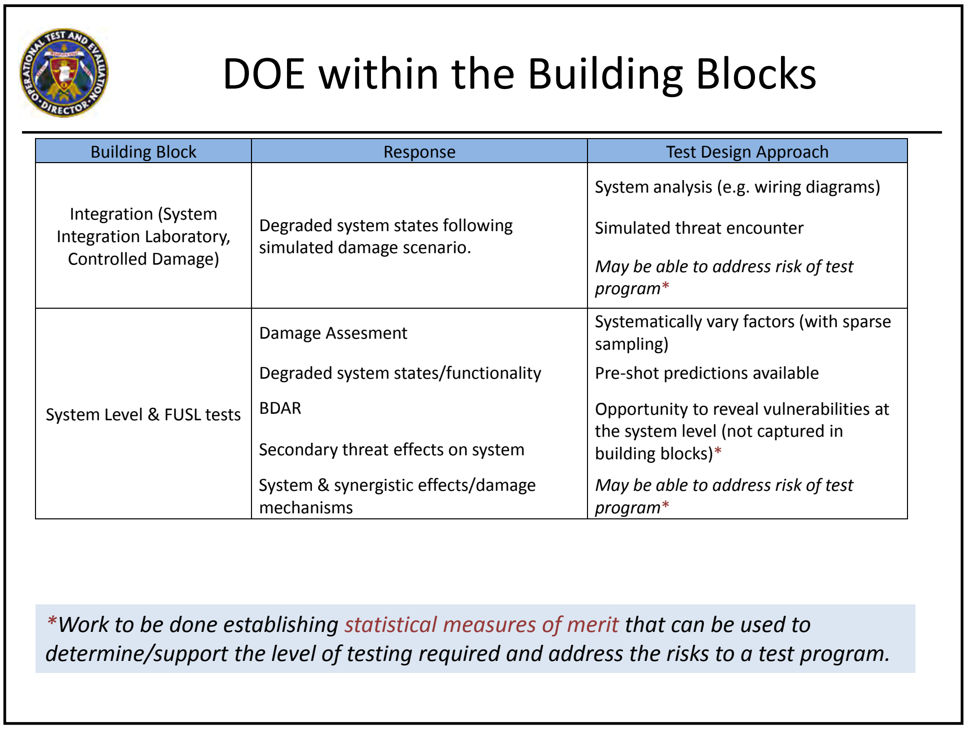
May eventually identify combinations of factors that are not operationally relevant, that will be catastrophic and have been conceded, or that will not stress the system.



DOE within the Building Blocks

Building Block	Response		Test Design Approach
Armor Sample (i.e. Coupon)	Perforation Residual Penetration Behind Armor Debris		Systematically vary factors May test to specified confidence level (perforation) _May be able to address risk of over/under estimating effects_*
Components	Ballistic	Damage due to Fragments Damage due to Shock	Systematically vary factors _May be able to address risk of over/under estimating effects_*
	Fire Safety Failure Mechanisms		_May be able to address risk of over/under estimating effects_*
Structures (Substructure, BH&T, Damaged Vehicle)	Resistance of armor integration to perforation Armor perforation and structural response caused by complex threats (HE-Frag, Blast) Fire Initiation/Propagation and AFES effectiveness		Systematically vary factors Pre-shot predictions may be available _May be able to calculate risk of M&S under/over predicting vulnerability depending on test scope_*

*Work to be done establishing statistical measures of merit that can be used to determine/support the level of testing required and address the risks to a test program.



DOE within the Building Blocks

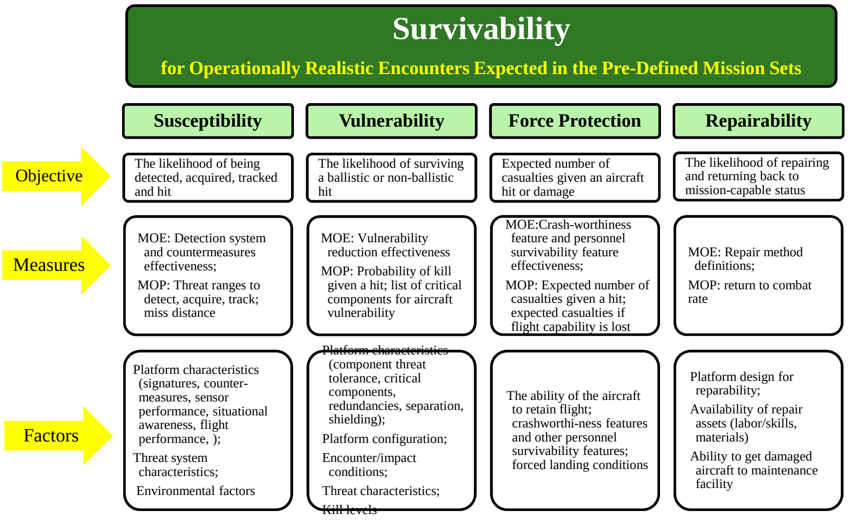
Building Block	Response	Test Design Approach
Integration (System Integration Laboratory, Controlled Damage)	Degraded system states following simulated damage scenario.	System analysis (e.g. wiring diagrams) Simulated threat encounter _May be able to address risk of test program_*
System Level & FUSL tests	Damage Assesment Degraded system states/functionality BDAR Secondary threat effects on system System & synergistic effects/damage mechanisms	Systematically vary factors (with sparse sampling) Pre-shot predictions available Opportunity to reveal vulnerabilities at the system level (not captured in building blocks)* _May be able to address risk of test program_*

*Work to be done establishing statistical measures of merit that can be used to determine/support the level of testing required and address the risks to a test program.

Examples

- DOE for Armored Vehicles
- DOE for Fixed Wing Aircraft
- DOE for Ships

Design of Experiments (DOE) for Aircraft Survivability



Survivability

for Operationally Realistic Encounters Expected in the Pre-Defined Mission Sets

	Susceptibility	Vulnerability	Force Protection	Repairability
Objective	The likelihood of being detected, acquired, tracked and hit	The likelihood of surviving a ballistic or non-ballistic hit	Expected number of casualties given an aircraft hit or damage	The likelihood of repairing and returning back to mission-capable status
Measures	MOE: Detection system and countermeasures effectiveness; MOP: Threat ranges to detect, acquire, track; miss distance	MOE: Vulnerability reduction effectiveness MOP: Probability of kill given a hit; list of critical components for aircraft vulnerability	MOE:Crash-worthiness feature and personnel survivability feature effectiveness; MOP: Expected number of casualties given a hit; expected casualties if flight capability is lost	MOE: Repair method definitions; MOP: return to combat rate

Factors	Platform characteristics (signatures, counter-measures, sensor performance, situational awareness, flight performance,);
	Threat system characteristics;
	Environmental factors
		Platform characteristics (component threat tolerance, critical components, redundancies, separation, shielding);
		Platform configuration;
		Encounter/impact conditions;
		Threat characteristics;
		Kill levels
			The ability of the aircraft to retain flight; crashworthi-ness features and other personnel survivability features; forced landing conditions	Platform design for reparability;
				Availability of repair assets (labor/skills, materials)
				Ability to get damaged aircraft to maintenance facility

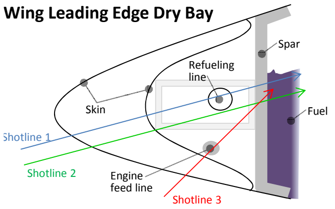
Example
Critical Issue: Vulnerability of the Aircraft to Threat-Induced Fires

Typical Assessment Approach

1. Empirical Data

I. Test design (60 tests) based on conditions specific to a particular platform. Fire potential is investigated through geometric analysis by particular components that are impacted – determine the number of shotlines.

Wing Leading Edge Dry Bay



For example, Shotline 1 is chosen to assess whether threat will function on either skin, refueling line or spar and whether that function is sufficient to ignite fuel from the line or from the tank.

Problem: Geometry variations are specific to this wing section and are rarely applicable to other programs. Each program is forced to re-assess this issue even though all variables (factors) affecting this issue are the same. Determining P_{fire} for the wide range of variations in geometry and shotlines requires many tests– only a subset can be tested and this approach cannot be used to optimize the test matrix.

II. Results primarily discussed as a function of test location. For example, "threat penetrated the lower wing leading edge skin and impacted the front spars. The threat did not function and no fire occurred".

Problem: Such data reduction only partially answers "what really happened and why?" – our understanding of the issue is limited to a very specific set of conditions

2. Physics-based Fire Prediction Model (FPM)

I. Should cover the design space that was not tested but correlations between test data/model predictions are poor.

II. Doesn't support a test process to improve confidence in the results.

III. The tool is not reliable and we are left with an incomplete assessment of this issue – recurring concern.

DOE-Based Assessment Approach

1. Empirical Data

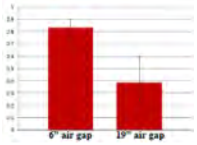
I. Test design (58 tests) is based on a clearly stated range of factors that are systematically changed to observe which factors or their combinations contributed most to the fire starting potential .

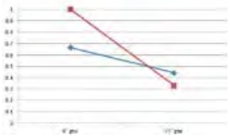
FACTORS	LEVELS	
Shotline obliquity	20 degrees	45 degrees
Skin thickness	0.07 inches	0.15 inches
Air gap distance	6 to 8 inches	18 to 24 inches
Airflow speed	50 cubic ft/min	160 cubic ft/min
Fragment velocity	4000 fps	7000 fps
Fragment mass	40 grain	75 grain
Fuel tank material	Aluminum	BMI

For example, shotlines are chosen to assess how the probability of fire changes with the air gap distance.

Advantage: Other programs can use these data and add their own tests (e.g., other levels) to improve confidence in their areas of interest.

II. Results are analyzed and discussed as a function of factors. For example, "probability of fire decreases with larger air gap distance. The magnitude of decrease is greater for higher threat velocity".





Advantage: Results are more meaningful; they reveal unexpected behavior the source of which can be more reliably identified. Used to build a mathematical model that predicts the response for specified factor settings.

2. Data can be used to Build a Response Model

I. Fits the data points and can be used to predict fire ignition at other points within the analysis space. For this example, 95% C.I. on the probability of fire ignition of 0.03 while FPM had no prediction capabilities.

Detailed Test Plan/Report Framework Concerns

	Typical Approach	DOE – BASED Approach
OBJECTIVE	Generate the necessary data to allow assessment of the system vulnerability to ballistic threat-induced fire. Confidence levels not considered.	Generate the necessary data to allow assessment of the system vulnerability to ballistic threat-induced fire with a specified level of confidence

RESPONSE	MOE - likelihood of sustained fire: threat functioning characteristics ; release of the flammable fluids ; fire sustainment; structural damage measurement. Not all responses are considered; some are not measurable.	Probability of fire; Fire duration;
		Time to First Fuel Spurt;
		Forward Face Flash; Back Face Flash.

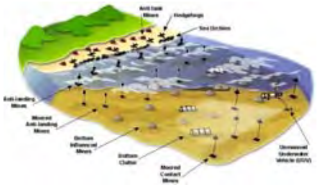
FACTORS	Only one mission scenario segment considered. Factors not always explicitly stated – the rationale behind using the ones tested are typically not explained.	Considers all possible variables: Threat (type, size, velocity, attitude); Impact conditions ; Fuel (type, temp., quant., pressure); Dry bay airflow (velocity, pressure, temp); Ambient conditions (temp, pressure)
LEVELS	Levels are not explicitly stated.	Two levels typically considered.
MATRIX	Rationale not provided - chosen with an effort to maximize the number of tests possible for the selected threats. Assumptions necessary to extrapolate the results to other conditions. Does not isolate well variables of importance.	Designed to test hypotheses about unique or combined effects. Designed to maximize the collection of valuable data in the minimum number of possible tests. Explores multiple conditions while retaining power and confidence to get the right answer.
ANALYSIS	Minimal - an assessment is made based on temp, pressure histories and a video review as to the type of fire which occurred (no fire, self-extinguishing fire, or sustained fire). Unexpected behavior difficult to address. Confidence intervals, power not discussed.	Explains the impact of factors on identified responses. Can be used to build a model to address the response at other test points. Provides confidence levels.

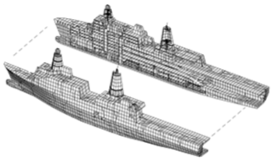
Examples

- DOE for Armored Vehicles
- DOE for Fixed Wing Aircraft
- DOE for Ships

Four Basic Elements of Ship LFT&E

- Component and surrogate testing to discover weaknesses
- Damage-scenario-based engineering and other analyses to assess the actual ship (includes use of validated M&S)
  - $P_{k/h}$ studies
- Full Ship Shock Trial
  - Not a full-up, system-level test
- Total Ship Survivability Trial to assess a ship's ability to control damage and recover mission capabilities





Reported in periodic Vulnerability Assessment Reports or Survivability Assessment Reports

What DOE Can Do to Help Ship LFT&E?

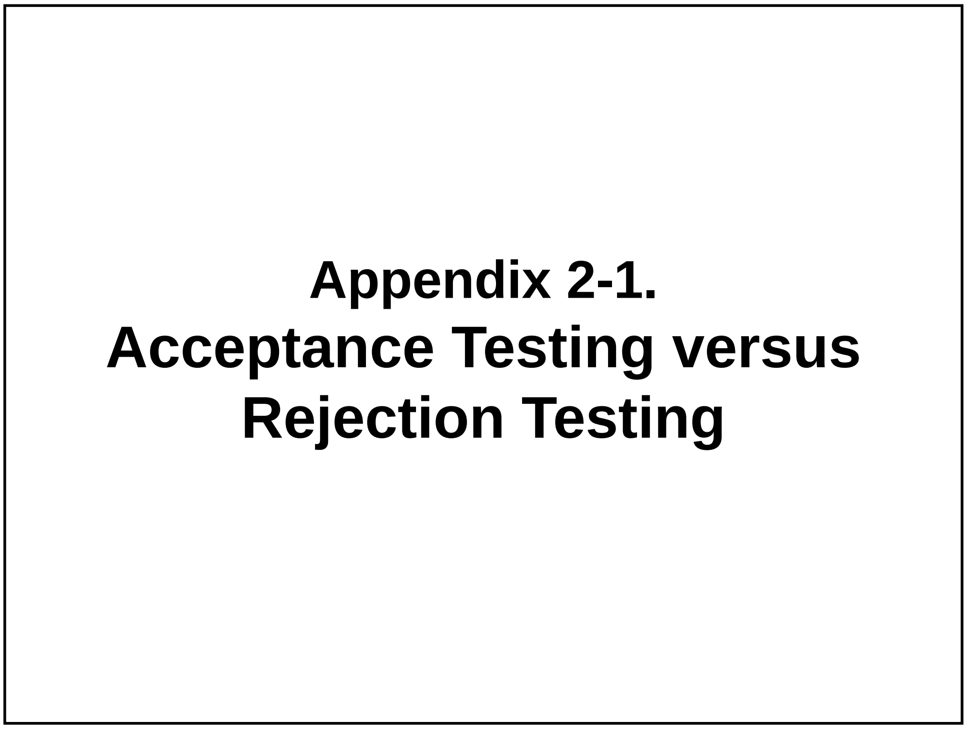
- Current LFT&E planning based on ad hoc scientific methods and intuition
- DOE could:
  - ➢ Influence decisions on number of tests or modeling iterations
  - ➢ Influence scope of test and test planning by providing objective data
  - ➢ Where limited data is available can help determine confidence in test results
  - ➢ Provide an input to establish confidence in M&S that uses test data
  - ➢ Influence where limited test resources can best be used
- Examples
  - ➢ LHA weapons effects testing
  - ➢ Ship to shore connector
    - 1/10th scale model test

LHA 2 Weapons Effects Test

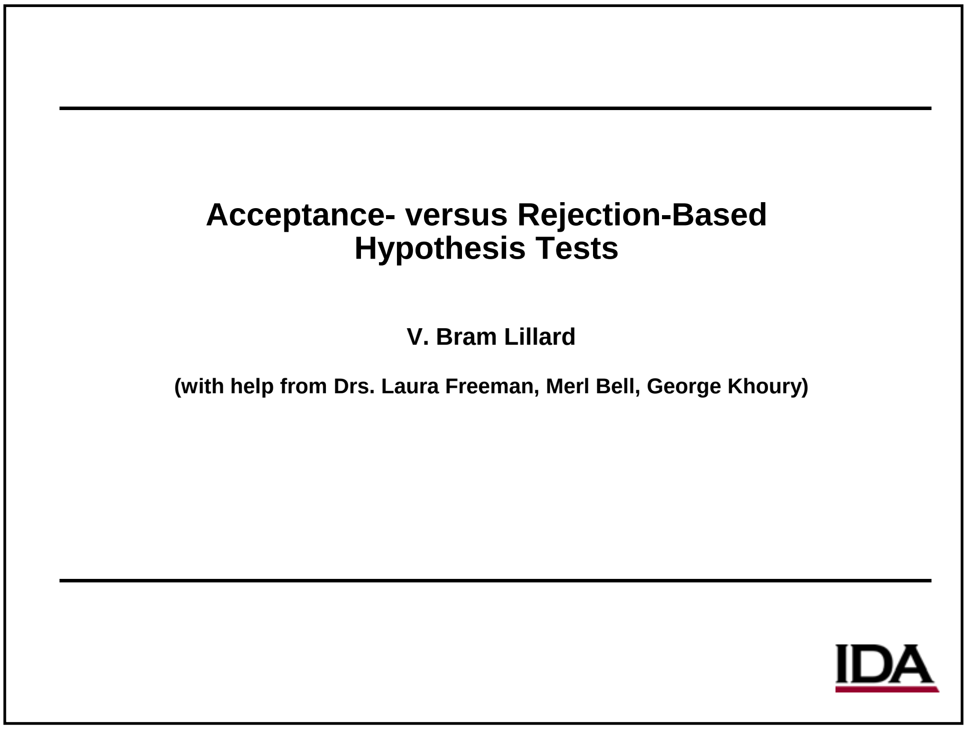
LHA 2 WET is Surrogate Test for LHA 6 LFT&E

Appendix 2
Tutorials

This page intentionally left blank.



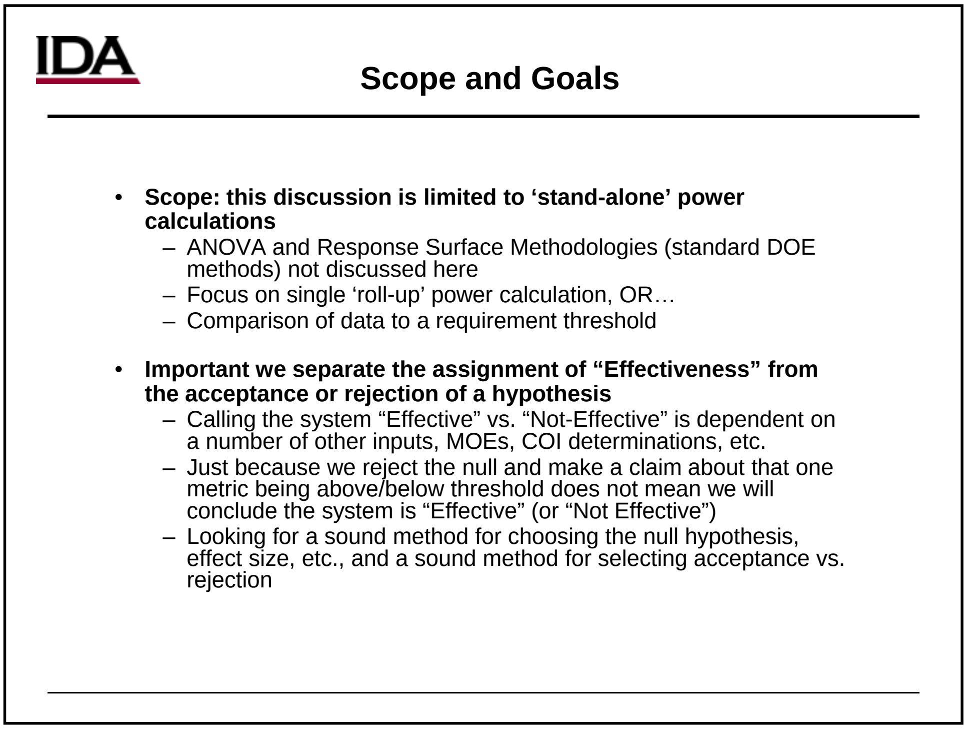
Appendix 2-1.
Acceptance Testing versus Rejection Testing



Acceptance- versus Rejection-Based Hypothesis Tests

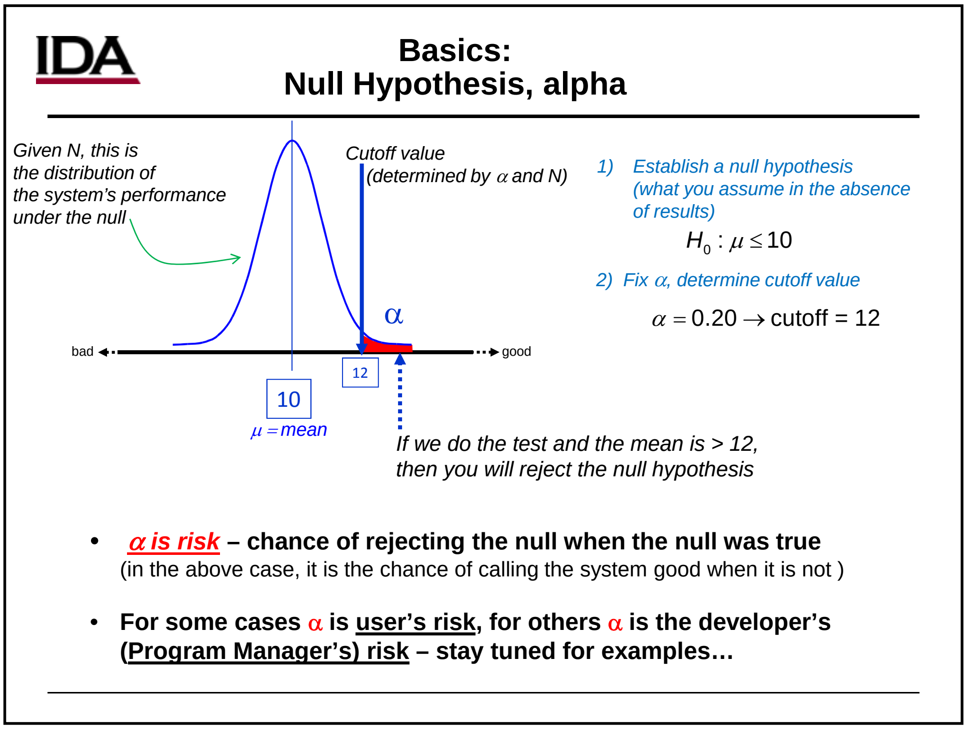
V. Bram Lillard

(with help from Drs. Laura Freeman, Merl Bell, George Khoury)



Scope and Goals

- **Scope: this discussion is limited to 'stand-alone' power calculations**
  - ANOVA and Response Surface Methodologies (standard DOE methods) not discussed here
  - Focus on single 'roll-up' power calculation, OR…
  - Comparison of data to a requirement threshold
- **Important we separate the assignment of "Effectiveness" from the acceptance or rejection of a hypothesis**
  - Calling the system "Effective" vs. "Not-Effective" is dependent on a number of other inputs, MOEs, COI determinations, etc.
  - Just because we reject the null and make a claim about that one metric being above/below threshold does not mean we will conclude the system is "Effective" (or "Not Effective")
  - Looking for a sound method for choosing the null hypothesis, effect size, etc., and a sound method for selecting acceptance vs. rejection



Basics:
Null Hypothesis, alpha

Given N, this is the distribution of the system's performance under the null

Cutoff value (determined by α and N)

If we do the test and the mean is > 12, then you will reject the null hypothesis

1) *Establish a null hypothesis (what you assume in the absence of results)*

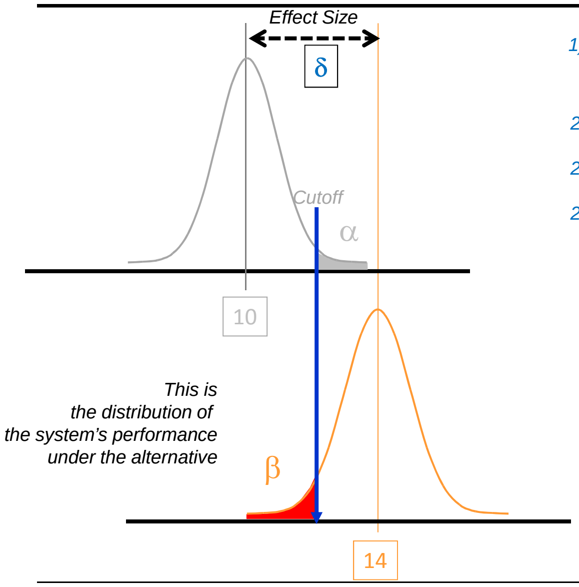
$$H_0 : \mu \le 10$$

2) *Fix α, determine cutoff value*

$$\alpha = 0.20 \rightarrow \text{cutoff} = 12$$

- ***α is risk*** **– chance of rejecting the null when the null was true**
  (in the above case, it is the chance of calling the system good when it is not)
- **For some cases α is user's risk, for others α is the developer's (Program Manager's) risk – stay tuned for examples…**

Basics:
Alternative Hypothesis, beta


Effect Size
δ
Cutoff
α
10
This is the distribution of the system's performance under the alternative
β
14

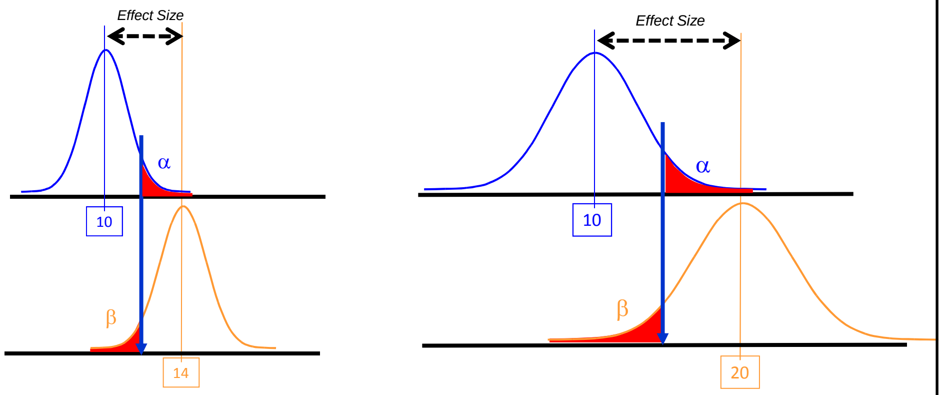
1) Establish an alternative hypothesis

$$H_a : \mu > 10 + \delta$$

2a) Fix β (or power) and δ, determine N
OR….
2b) Fix δ and N, determine β
OR….
2c) Fix β and N, determine δ

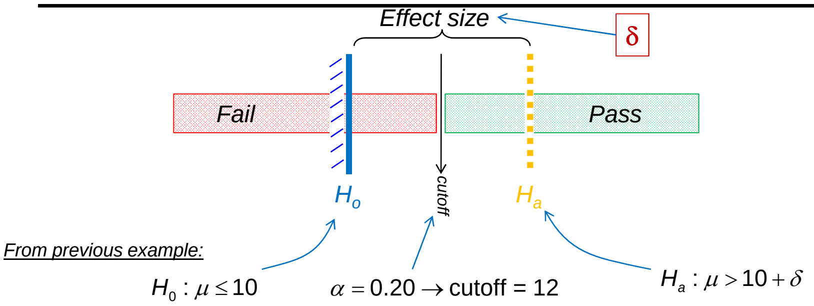
- ***β is risk*** **– chance of failing to accept the alternative when the alternative is true** (in this case, it is the chance of calling the system bad when it is actually good)
- Typically, trade off sample size (N) to affect power and/or effect size

Illustration:
Fixed α and β, different N



- **With large N, can achieve small effect sizes**
- **With small N, effect size is large in order to maintain power**

Simplified Nomenclature



From previous example:

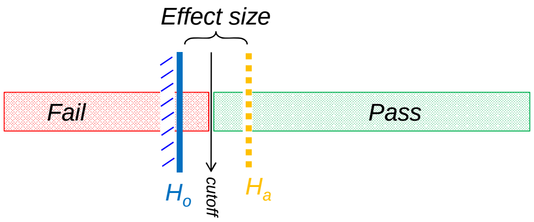
$H_0 : \mu \leq 10$

$\alpha = 0.20 \rightarrow$ cutoff = 12

$H_a : \mu > 10 + \delta$

- Now discuss different cases (choosing nulls/alternatives)
- Pros and cons of each case
- Determine methodology for selecting which method is appropriate

Effect of more samples (higher N)

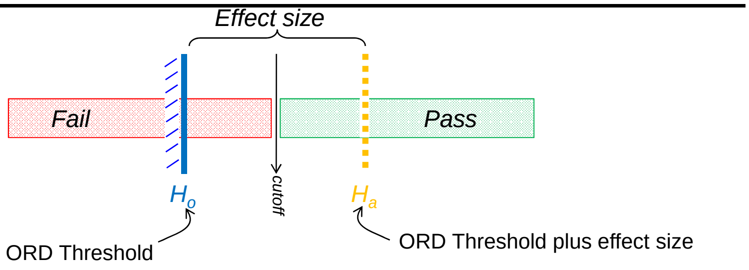


Options

(different applications of the standard textbook methods)

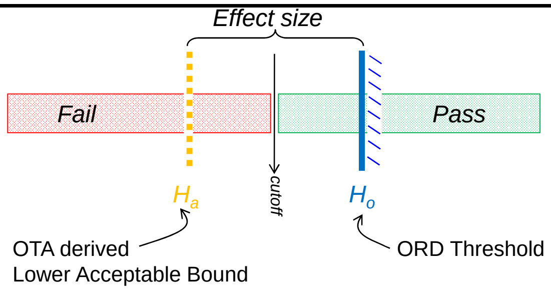
- **Strict Rejection Test:**
  - Assume system is at or below threshold
  - To reject null (pass the system), performance must be significantly above threshold
- **Strict Acceptance Test:**
  - Assume system is at or above threshold
  - To reject null (fail the system), performance must be significantly below threshold
- **Modified Rejection Test:**
  - Assume system's performance is significantly below threshold
  - Critical value is set at the threshold – reject the null (pass the system) if performance at or above threshold
- **Regression-Focused Acceptance Test:**
  - Assume system is at or above past performance
  - To reject null (fail the system), performance must be significantly below past performance
  - Alternate use: assume system is at or above ORD Objective

Strict Rejection Test



- **Pros:**
  - Requires testing to prove system is good (philosophically sound)
  - When we fix α, warfighter's risk is fixed – focus on negotiating PM's risk using effect size and power to determine N
  - PM is inclined to fund more testing (higher N) in order to make the test easier to pass (i.e., cutoff value moves closer to the ORD threshold)
  - Null and confidence values represent the testers' focus on the operator (avoid accepting bad systems)
- **Cons:**
  - If system is only designed to meet the ORD threshold, low-power tests are likely to fail the system
  - Can be above ORD threshold but fail to reject null

Strict Acceptance Test



Effect size

Fail

Pass

H_a

cutoff

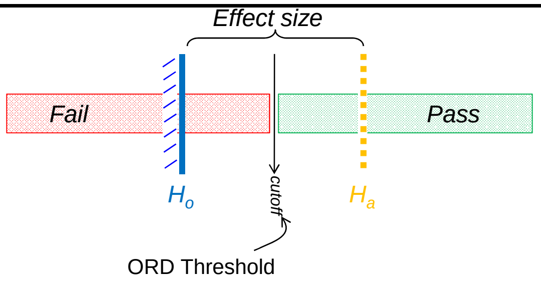
H_o

OTA derived
Lower Acceptable Bound

ORD Threshold

- **Pros:**
  - Reflects the reality that systems are often only designed to meet threshold requirements
- **Cons:**
  - No testing or little testing required to pass the system (easy test to pass) – effect size and power negotiations to determine N push risk onto the warfighter
  - No leverage in negotiation with program manager over resources. PM has no incentive to provide additional test time or targets, etc.; in fact, PM is motivated to de-fund the testing (decrease N) in order to make the test easier to pass.
  - By design, this test does not treat threshold as a true threshold – lower performance is acceptable (appears to be setting requirements)
  - Not consistent with post-test reporting – will only reject null (state system is below threshold) if confidence is high enough

Modified Rejection Test



Effect size

Fail

Pass

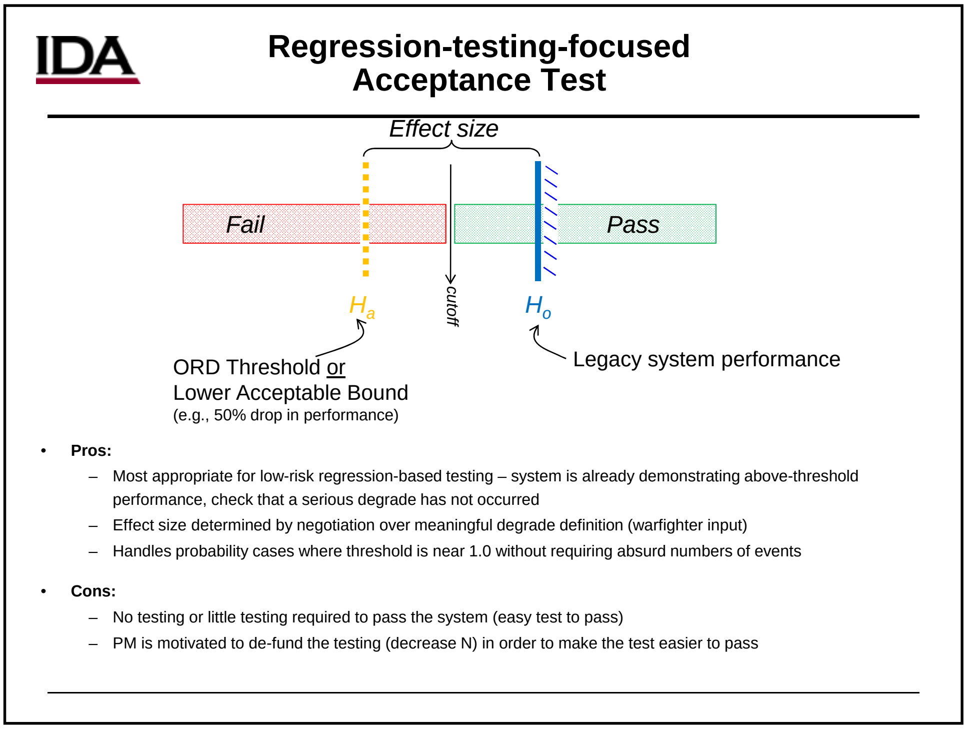
H_o

cutoff

H_a

ORD Threshold

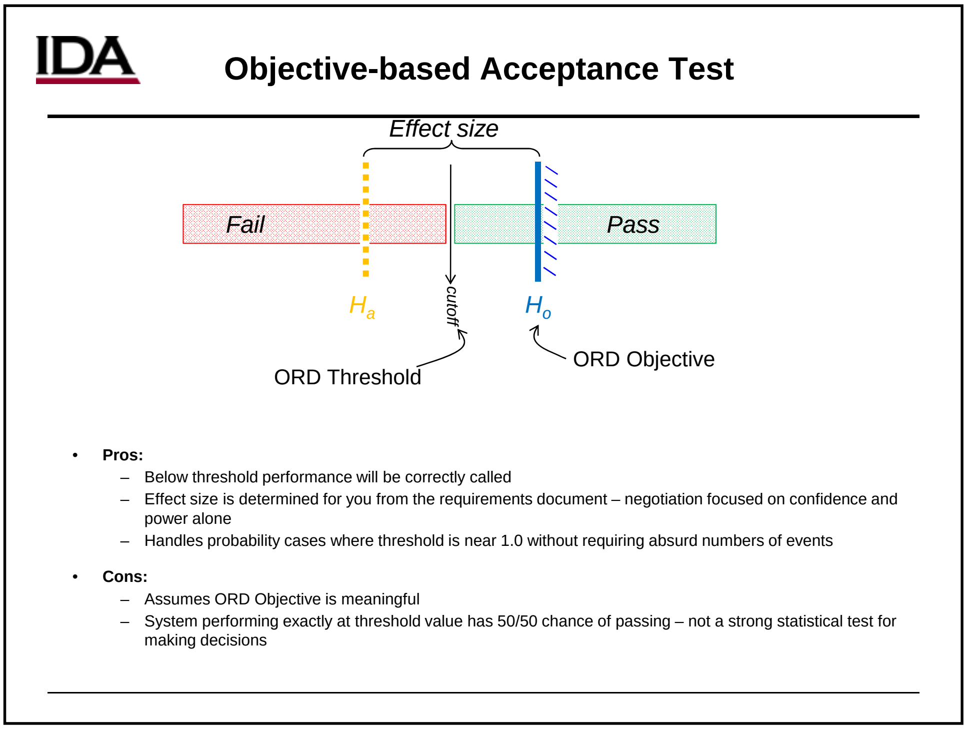
- **Pros:**
  - Reflects the reality that systems are often designed only to meet threshold requirements, but maintains many of the benefits of the strict rejection test method (slide 8)
  - Below threshold performance will be correctly called when close to the threshold – less of a chance for incorrect calls as was the case with strict rejection test (note, you give up statistical confidence in the call!)
  - PM is more inclined to fund more (higher N) testing so his system doesn't fail
  - Above threshold performance will cause rejection of null, acceptance of alternative (matches our statements about above/below threshold performance)
- **Cons:**
  - Stating H_o will seem like testers are making up requirements
  - Conclusions about system being above or below ORD thresholds ignores statistical confidence in those statements
  - α no longer represents the risk of passing a below-threshold system



Regression-testing-focused Acceptance Test

Effect size
Fail
Pass
H_a
cutoff
H_o
ORD Threshold or
Lower Acceptable Bound
(e.g., 50% drop in performance)
Legacy system performance

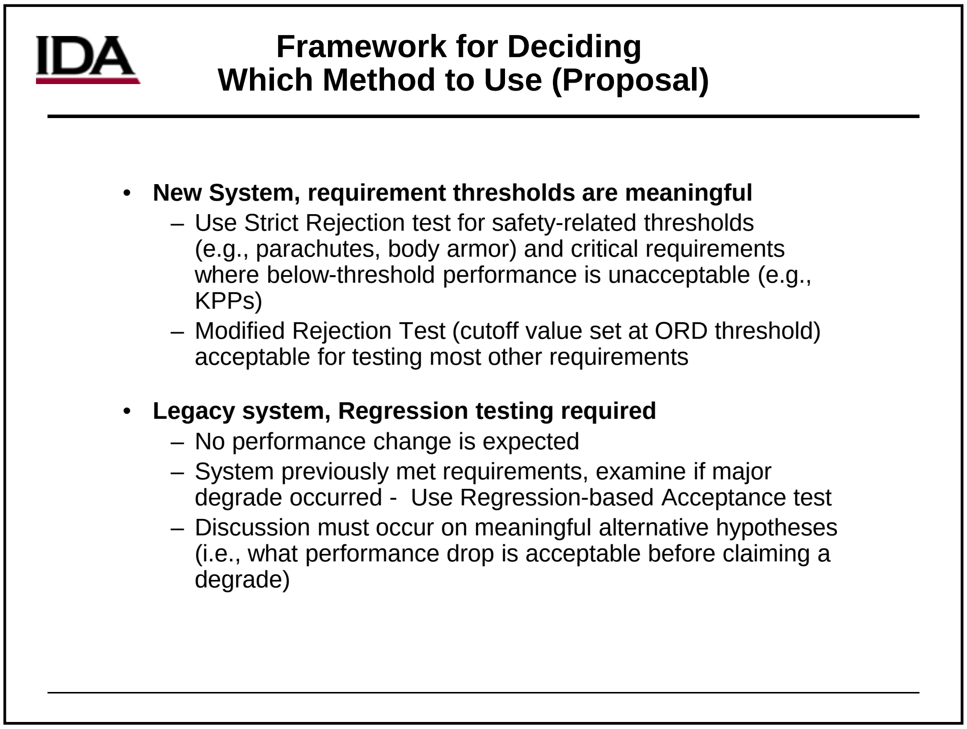
- **Pros:**
  - Most appropriate for low-risk regression-based testing – system is already demonstrating above-threshold performance, check that a serious degrade has not occurred
  - Effect size determined by negotiation over meaningful degrade definition (warfighter input)
  - Handles probability cases where threshold is near 1.0 without requiring absurd numbers of events
- **Cons:**
  - No testing or little testing required to pass the system (easy test to pass)
  - PM is motivated to de-fund the testing (decrease N) in order to make the test easier to pass



Objective-based Acceptance Test

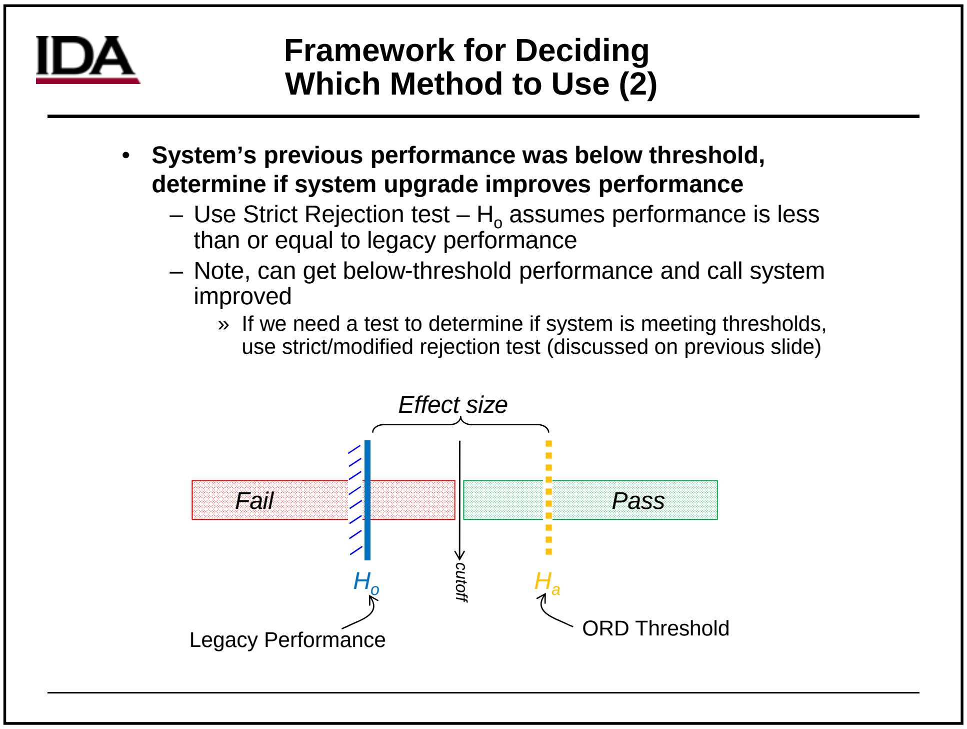
Effect size
Fail
Pass
H_a
cutoff
H_o
ORD Threshold
ORD Objective

- **Pros:**
  - Below threshold performance will be correctly called
  - Effect size is determined for you from the requirements document – negotiation focused on confidence and power alone
  - Handles probability cases where threshold is near 1.0 without requiring absurd numbers of events
- **Cons:**
  - Assumes ORD Objective is meaningful
  - System performing exactly at threshold value has 50/50 chance of passing – not a strong statistical test for making decisions



Framework for Deciding Which Method to Use (Proposal)

- **New System, requirement thresholds are meaningful**
  - Use Strict Rejection test for safety-related thresholds (e.g., parachutes, body armor) and critical requirements where below-threshold performance is unacceptable (e.g., KPPs)
  - Modified Rejection Test (cutoff value set at ORD threshold) acceptable for testing most other requirements
- **Legacy system, Regression testing required**
  - No performance change is expected
  - System previously met requirements, examine if major degrade occurred - Use Regression-based Acceptance test
  - Discussion must occur on meaningful alternative hypotheses (i.e., what performance drop is acceptable before claiming a degrade)



Framework for Deciding Which Method to Use (2)

- **System's previous performance was below threshold, determine if system upgrade improves performance**
  - Use Strict Rejection test – H_o assumes performance is less than or equal to legacy performance
  - Note, can get below-threshold performance and call system improved
    - » If we need a test to determine if system is meeting thresholds, use strict/modified rejection test (discussed on previous slide)

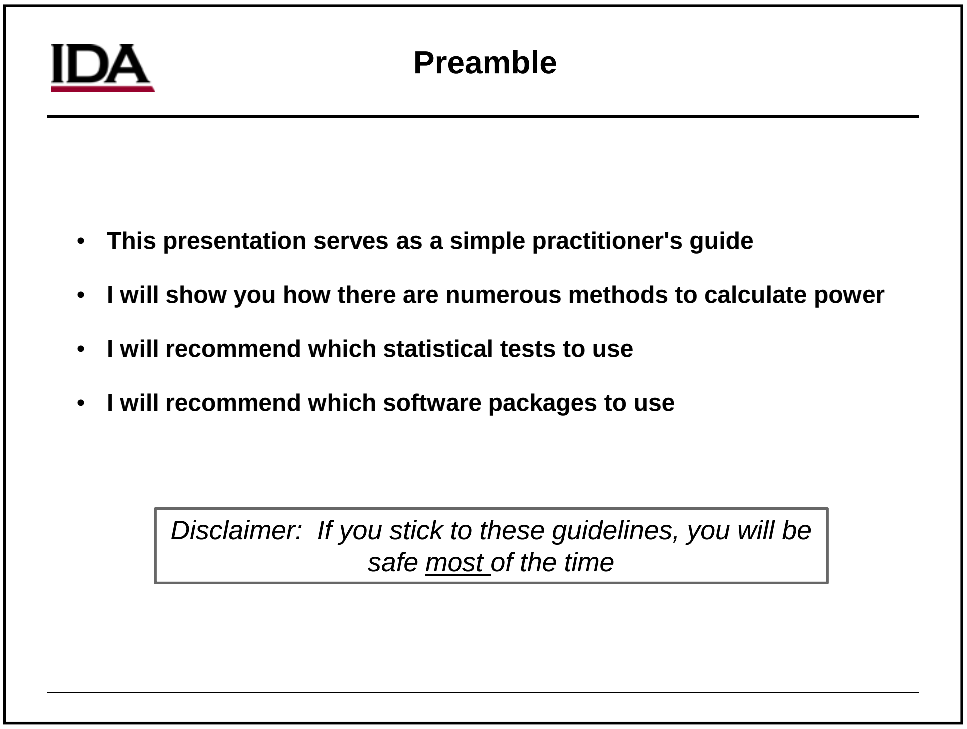
Appendix 2-2.
Power Calculations

Power Calculations:
Software Differences, Challenges and Recommendations

Thomas H. Johnson

Laura Freeman

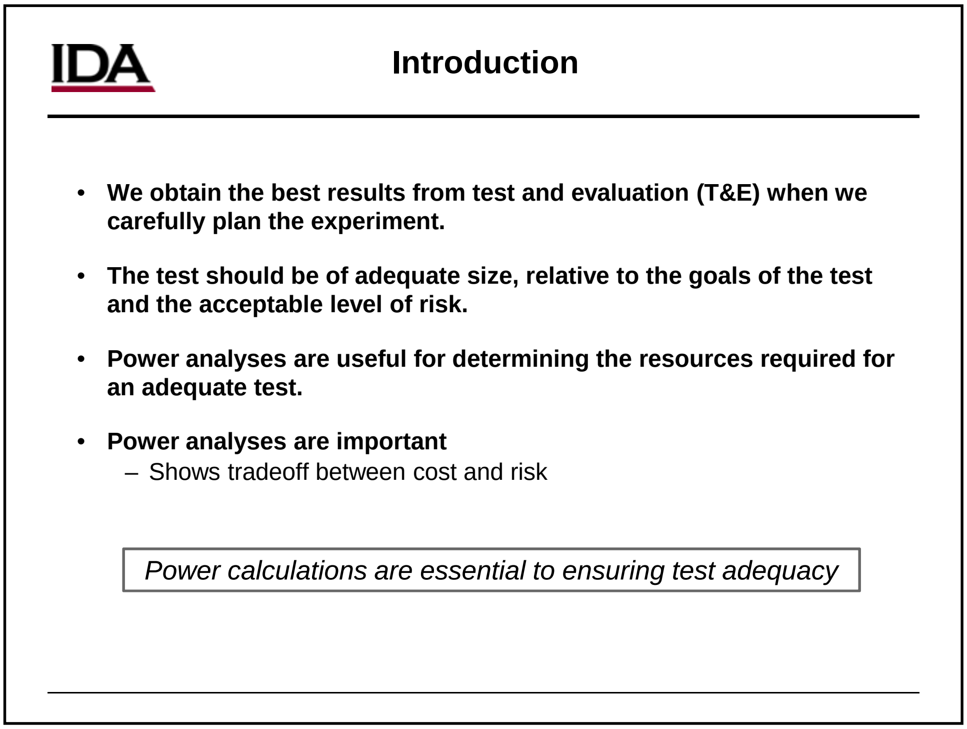
IDA



Preamble

- This presentation serves as a simple practitioner's guide
- I will show you how there are numerous methods to calculate power
- I will recommend which statistical tests to use
- I will recommend which software packages to use

Disclaimer: If you stick to these guidelines, you will be safe most of the time



Introduction

- We obtain the best results from test and evaluation (T&E) when we carefully plan the experiment.
- The test should be of adequate size, relative to the goals of the test and the acceptable level of risk.
- Power analyses are useful for determining the resources required for an adequate test.
- Power analyses are important
  - Shows tradeoff between cost and risk

Power calculations are essential to ensuring test adequacy

Introduction

- **Comparing the results of power calculations made by different test organizations can be challenging because:**
  - Different types of hypotheses tests require different methods of calculation
  - There are multiple correct methods to calculate power for each type of hypothesis test
  - There are numerous statistical software packages available, each of which uses different assumptions
- **Some software packages allow you to calculate results in cases where assumptions are invalid**
- **It is important to understand assumptions of power calculations to prevent mistakes in test design**

Different power results made from different software packages leads to debate!!!

Software Packages

- **Software packages used in this presentation**
  - JMP
  - Russ Lenth's Tool (online)
  - Design Expert
  - GPower
- **There are many other packages that could be used**

Outline

- **I will show which software packages to use and which options to select for different examples**
- **An example is provided for each of the following**
  - Test of One Proportion
  - Test of Two Proportions
  - One-Sample t-Test
  - Design of Experiments
    - Continuous factors and categorical factors
- **Summary chart of recommendations**

Test of One Proportion:
Missile Firing Example

- **A missile is required to have a probability of hit (P_H) of 80%**
- **How many missiles do we need to fire to have an 80% confidence level and power to detect a 10% difference in probability of hit?**
- **Use test of one proportion**
  - Null Hypothesis

    $$H_0 : P_H = 0.8$$
  - Alternative Hypothesis

    $$H_1 : P_H < 0.8$$

- **Test to see if the outcome of the experiment is significantly lower than a hypothesized value**

One Proportion Test

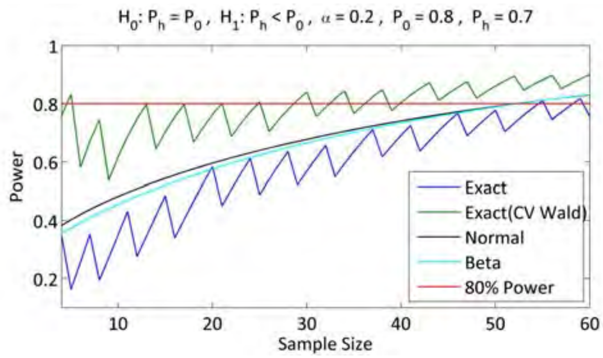
- **Common calculation methods:**

Design Expert cannot handle binomial responses

	JMP	Russ Lenth	Design Expert	GPower
Normal Approx		✓		
Beta Approx		✓		
Exact		✓		✓
Exact Wald	✓	✓		

One Proportion Test

- **Back to the missile firing example**
- **Power calculations differ depending on the method used**
- **What is the "right" sample size to achieve 80% Power?**


H_0: $P_h = P_0$, H_1: $P_h < P_0$, $\alpha = 0.2$, $P_0 = 0.8$, $P_h = 0.7$
Power
1
0.8
0.6
0.4
0.2
10
20
30
40
50
60
Sample Size
Exact
Exact(CV Wald)
Normal
Beta
80% Power

IDA

One Proportion Test

- **Sample size results**

	Type I Error (α)	Power	Sample Size
Normal Approx	20.0%	80.4%	53
Beta Approx	20.0%	79.8%	52
Exact	19.7%	81.0%	55
⚠ Exact Wald	67.2%	83.2%	5

- **Recommendation**

Use Russ Lenth's exact method for one proportion tests

IDA

Test of Two Proportions:
Fighter Upgrade Example

- **An existing fighter jet shot down enemy aircrafts with a 57% success rate in a previous test consisting of 56 runs**
- **An upgrade is applied to the existing fighter and it is required that the new system performs better than the old one**
- **How many runs do we need in the second test to have an 80% confidence level and 70% power to detect a 15% difference from the outcome of the first test?**
- **Use test of two proportions**

$$H_0 : P_{\text{success (new system)}} = P_{\text{success (old system)}}$$

$$H_1 : P_{\text{success (new system)}} > P_{\text{success (old system)}}$$

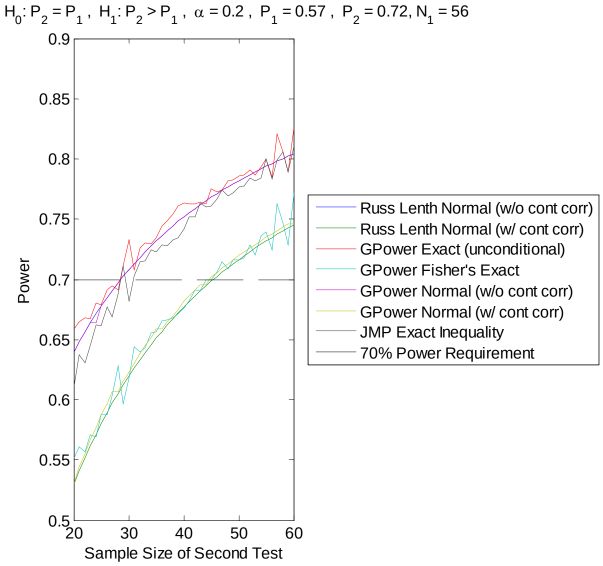
Two Proportions Test

- **Two Proportion Test Calculation Methods**

		JMP	Russ Lenth	Design Expert	GPower
Exact	Inequality (McNemar)				✓
	Inequality (Fisher's)				✓
	Inequality (unconditional)				✓
	Inequality with Offset	✓			✓
	Test Statistic Options				✓
Normal Approx.	Continuity Correction		✓		✓
	Arcsin Correction				✓

Two Proportions Test

- **Each calculation method leads to different results**
- **What is the "right" sample size to achieve 70% Power?**



Two Proportions Test

- **Sample size results**

Software	Method	Option	Type I Error (α)	Power	Sample Size (N2)
Gpower	Exact	unconditional	20.2%	69.4%	27
Gpower	Exact	Fisher's	14.2%	70.0%	45
Gpower	Normal Approx	w/o continuity	N/A	69.7%	28
Gpower	Normal Approx	w/ continuity	N/A	70.2%	45
JMP	Exact		N/A	70.3%	31
Russ Lenth	Normal Approx	w/ continuity	N/A	70.2%	46
Russ Lenth	Normal Approx	w/o continuity	N/A	70.3%	29

⚠ *GPower shows exact Type I error, while JMP does not*

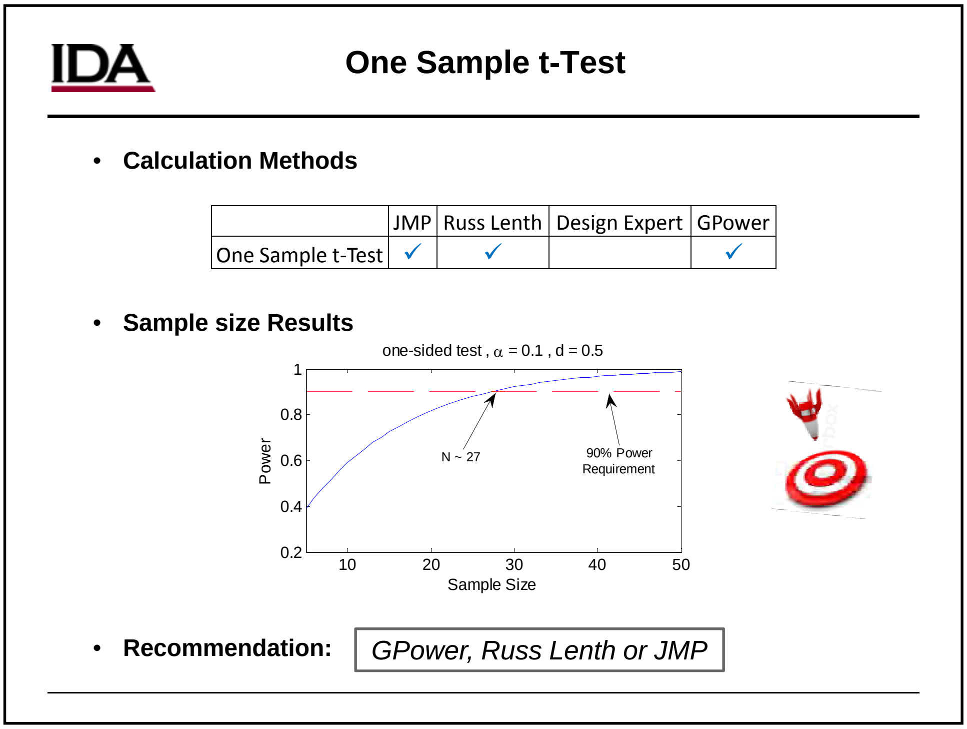
- **Recommendation**

Use GPower’s exact (unconditional) test for two proportions

One-Sample t-Test Example:
Bomb Drop

- **A new type of bomb is being drop tested and we are interested to see if the mean miss distance is greater than 3 meters**
- **From previous testing of a similar bomb, the standard deviation of miss distance was found to be 6 meters**
- **How many bombs do we need to drop to have an 90% confidence level and 90% power to detect a 3 meter miss distance from the target?**
- **Use a One-Sample t-Test**

Null Hypothesis	Alternative Hypothesis
$H_0: \mu = 3$	$H_1: \mu > 3$



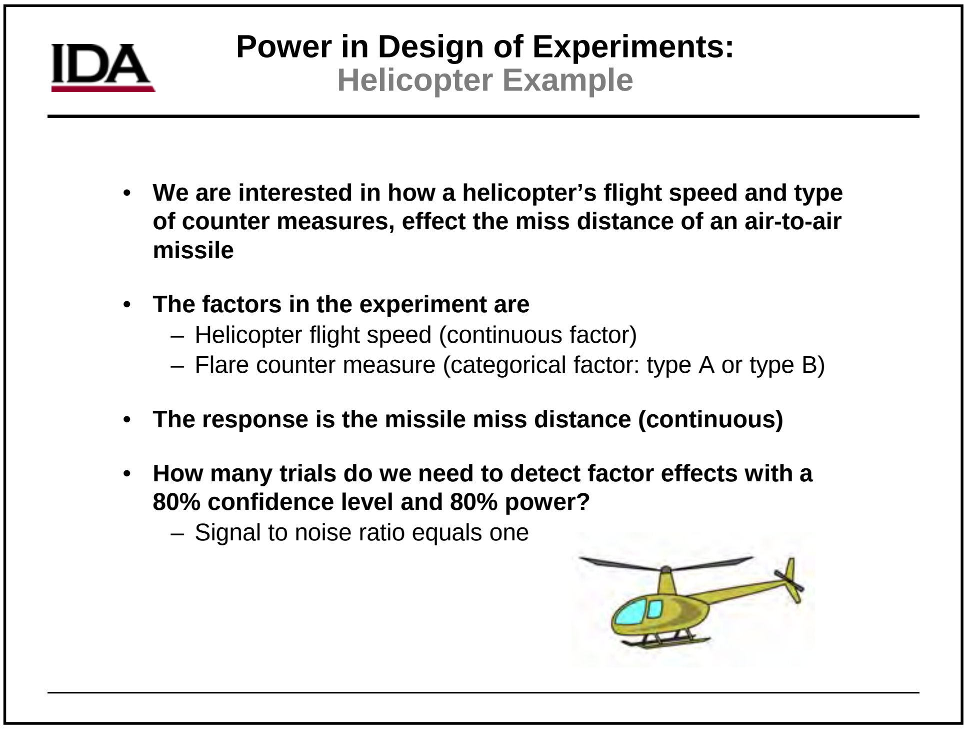
One Sample t-Test

- **Calculation Methods**

	JMP	Russ Lenth	Design Expert	GPower
One Sample t-Test	✓	✓		✓

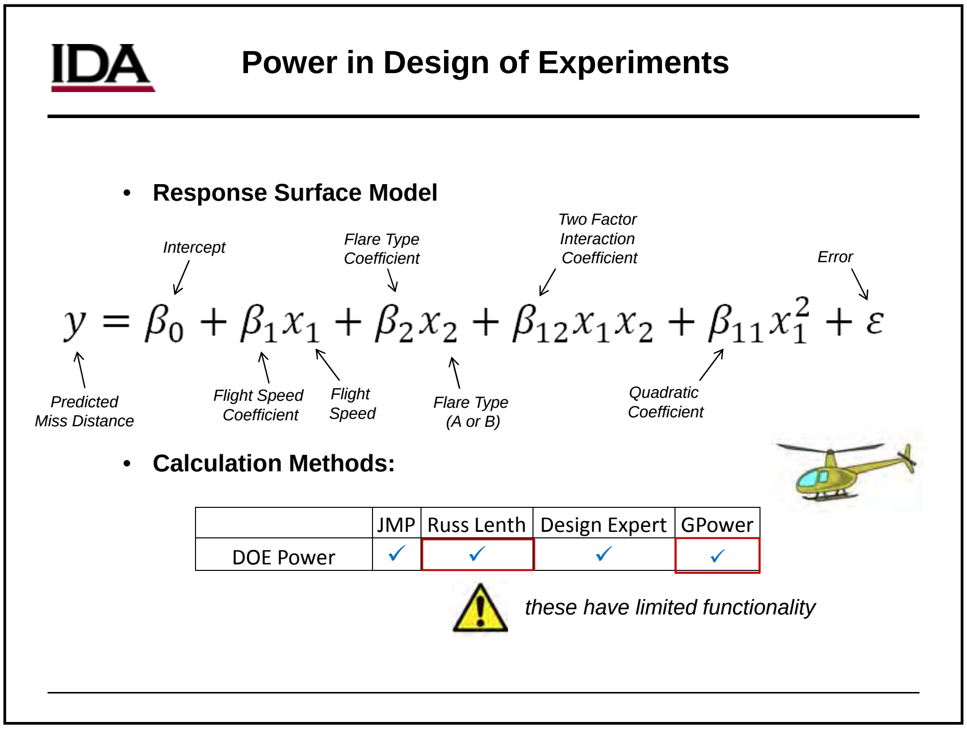
- **Sample size Results**

- **Recommendation:** *GPower, Russ Lenth or JMP*



Power in Design of Experiments: Helicopter Example

- **We are interested in how a helicopter's flight speed and type of counter measures, effect the miss distance of an air-to-air missile**
- **The factors in the experiment are**
  - Helicopter flight speed (continuous factor)
  - Flare counter measure (categorical factor: type A or type B)
- **The response is the missile miss distance (continuous)**
- **How many trials do we need to detect factor effects with a 80% confidence level and 80% power?**
  - Signal to noise ratio equals one



Power in Design of Experiments

- **Response Surface Model**

Intercept / Flare Type Coefficient / Two Factor Interaction Coefficient / Error

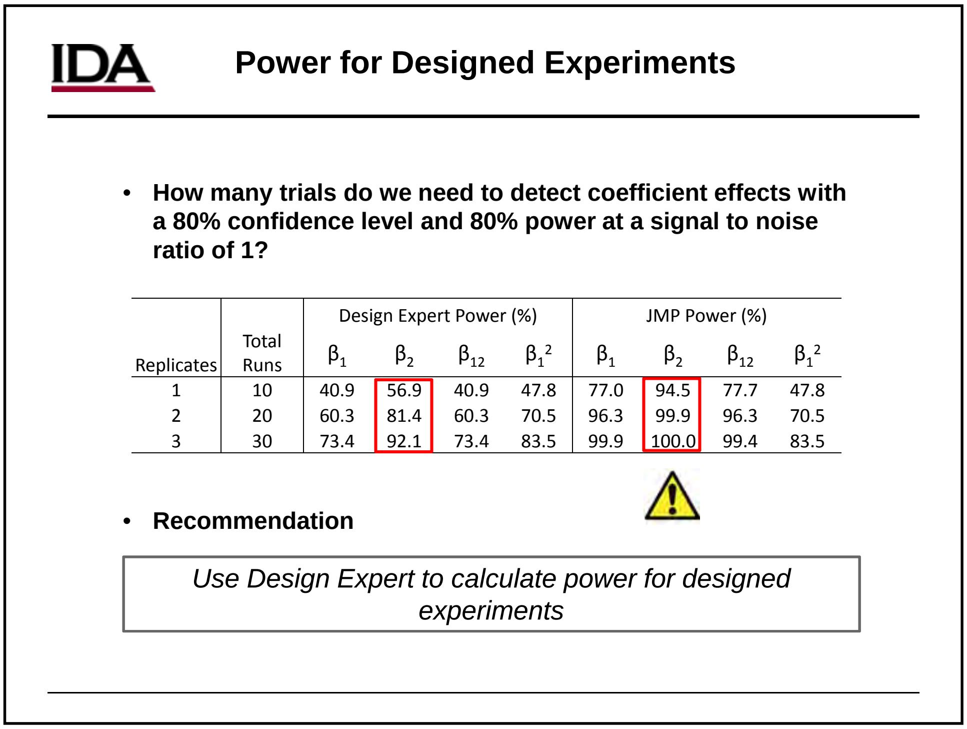
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{11} x_1^2 + \varepsilon$$

Predicted Miss Distance / Flight Speed Coefficient / Flight Speed / Flare Type (A or B) / Quadratic Coefficient

- **Calculation Methods:**

	JMP	Russ Lenth	Design Expert	GPower
DOE Power	✓	✓	✓	✓

⚠ *these have limited functionality*



Power for Designed Experiments

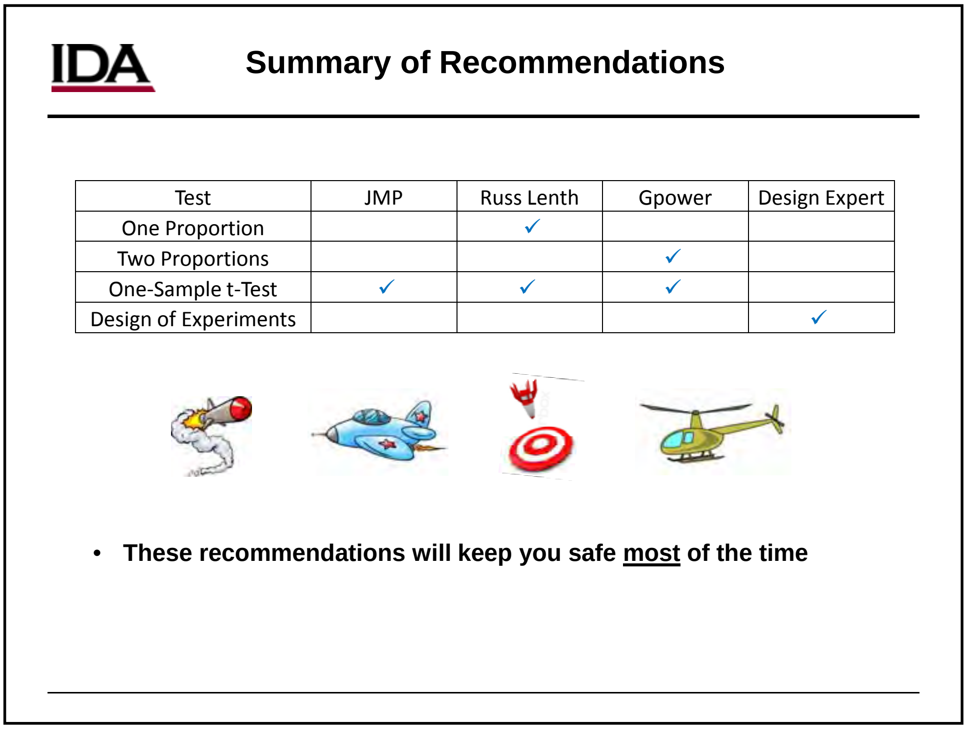
- **How many trials do we need to detect coefficient effects with a 80% confidence level and 80% power at a signal to noise ratio of 1?**

		Design Expert Power (%)				JMP Power (%)			
Replicates	Total Runs	β_1	β_2	β_{12}	β_1^2	β_1	β_2	β_{12}	β_1^2
1	10	40.9	56.9	40.9	47.8	77.0	94.5	77.7	47.8
2	20	60.3	81.4	60.3	70.5	96.3	99.9	96.3	70.5
3	30	73.4	92.1	73.4	83.5	99.9	100.0	99.4	83.5

⚠

- **Recommendation**

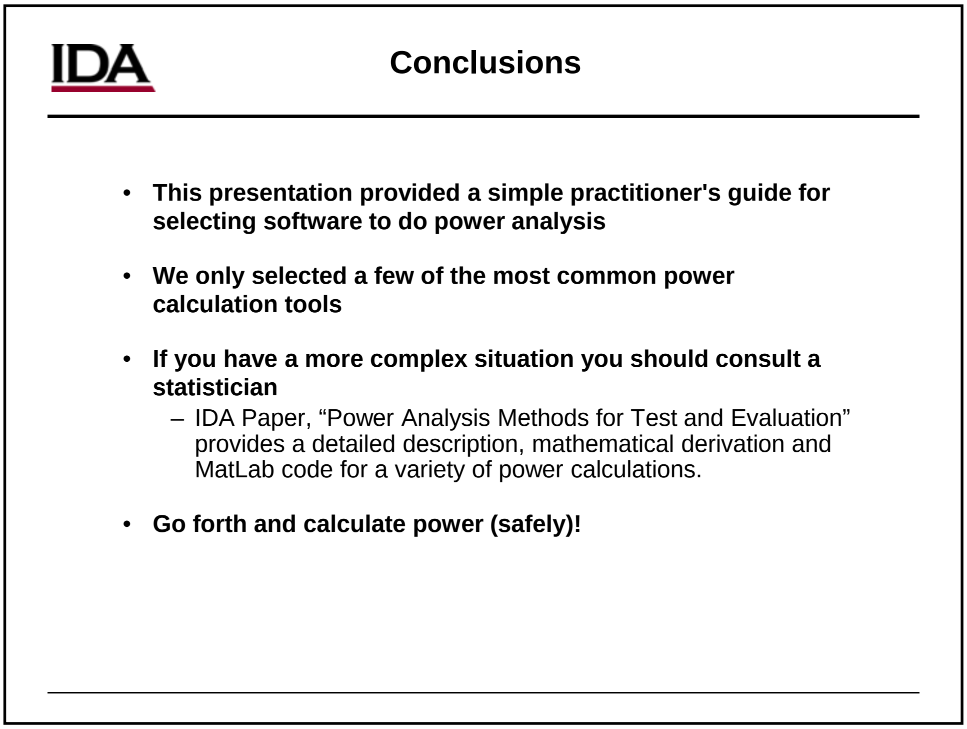
Use Design Expert to calculate power for designed experiments



Summary of Recommendations

Test	JMP	Russ Lenth	Gpower	Design Expert
One Proportion		✓		
Two Proportions			✓	
One-Sample t-Test	✓	✓	✓	
Design of Experiments				✓

- **These recommendations will keep you safe <u>most</u> of the time**



Conclusions

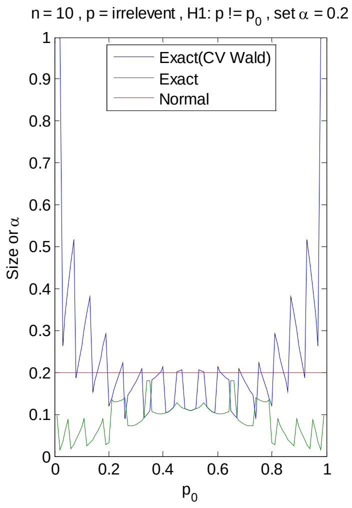
- **This presentation provided a simple practitioner's guide for selecting software to do power analysis**
- **We only selected a few of the most common power calculation tools**
- **If you have a more complex situation you should consult a statistician**
  - IDA Paper, “Power Analysis Methods for Test and Evaluation” provides a detailed description, mathematical derivation and MatLab code for a variety of power calculations.
- **Go forth and calculate power (safely)!**

Backup Slides

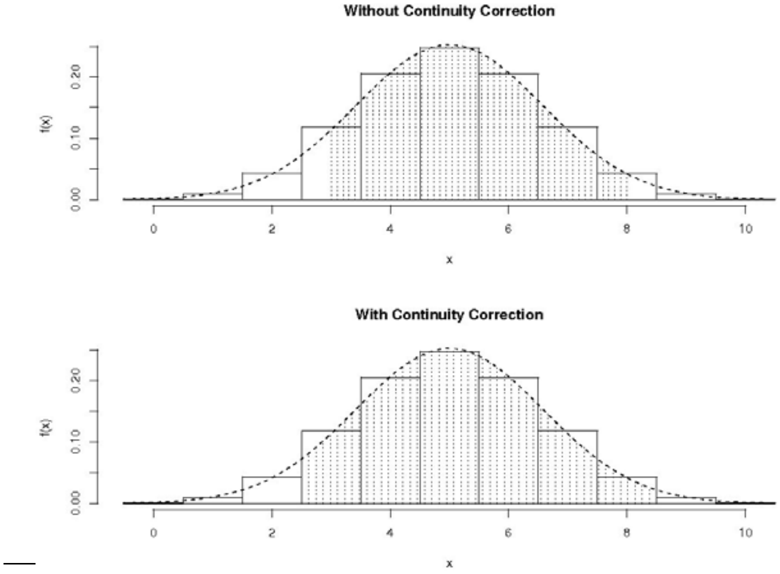
- **These recommendations will keep you safe most of the time**

One Proportion Test

- **Exact Methods tend to be more conservative**
- **However, the Type I error rate is not constant for exact methods.**
  - As a result the test design may be more or less risky than originally planned.
- **Be wary of the Exact Wald calculations**
  - Especially for low/high probabilities
- **Selection of confidence level can be misleading**
  - JMP does not provide the user with the actual size of the test.
  - User sets alpha, but the power is calculated at an alpha different than what was set
  - Figure shows how alpha varies with null hypothesis
  - exact is more stable than exact Wald near the extremes of the null hypothesis



Two Proportion Test


Without Continuity Correction

With Continuity Correction

ANOVA Comparisons

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2$$

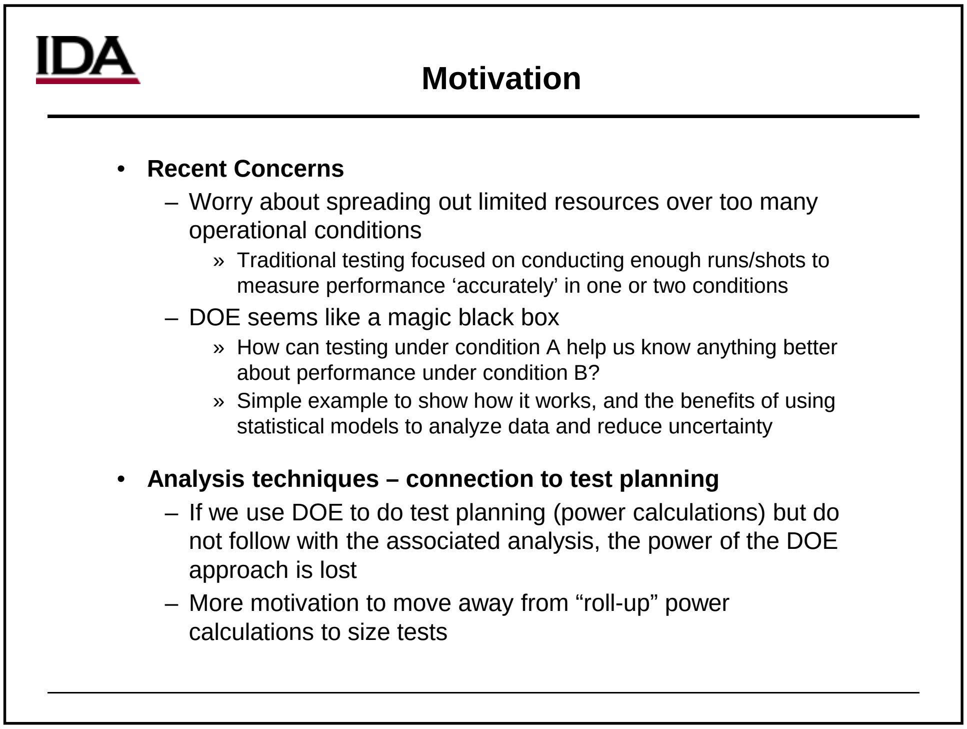
	β_1	β_2	β_{12}	β_{11}	β_{22}
JMP	1	1	1	1	1
Design Expert	1/2	1/2	1/2	1	1
Russ Lenth's Tool	$\sqrt{2}/2$	$\sqrt{2}/2$	1/2	N/A*	N/A*

This page intentionally left blank.

Appendix 2-3.
What Does DOE Buy Us?

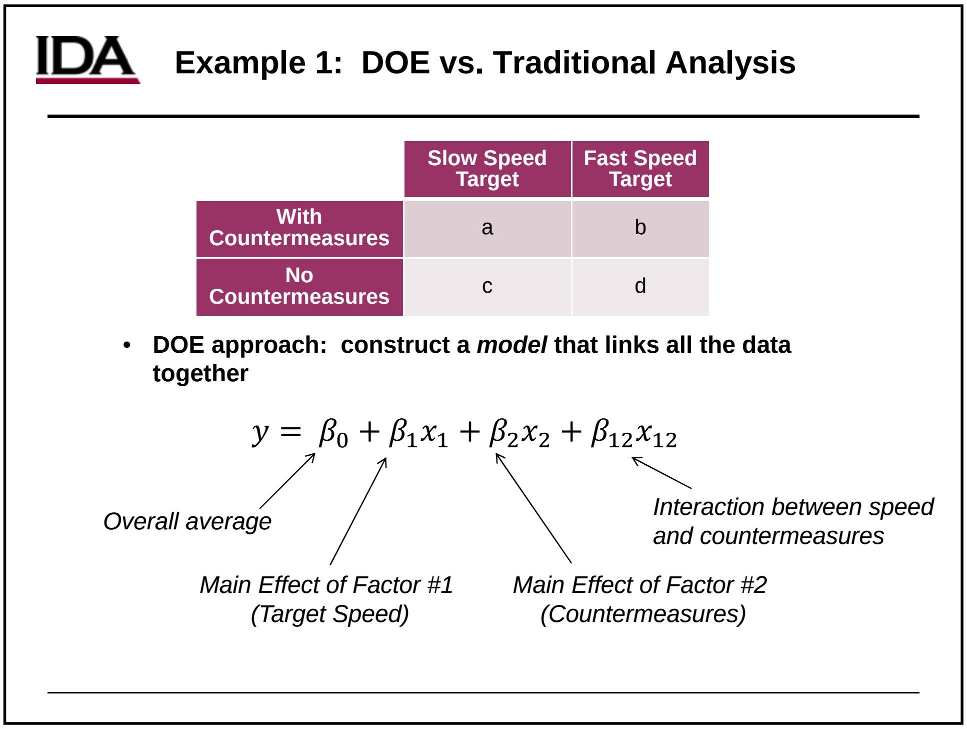
What Does DOE Buy Us?
(Examples to Illustrate the Value of Using DOE)

V. Bram Lillard
Laura Freeman



Motivation

- **Recent Concerns**
  - Worry about spreading out limited resources over too many operational conditions
    - Traditional testing focused on conducting enough runs/shots to measure performance ‘accurately’ in one or two conditions
  - DOE seems like a magic black box
    - How can testing under condition A help us know anything better about performance under condition B?
    - Simple example to show how it works, and the benefits of using statistical models to analyze data and reduce uncertainty
- **Analysis techniques – connection to test planning**
  - If we use DOE to do test planning (power calculations) but do not follow with the associated analysis, the power of the DOE approach is lost
  - More motivation to move away from “roll-up” power calculations to size tests



Example 1: DOE vs. Traditional Analysis

	Slow Speed Target	Fast Speed Target
With Countermeasures	a	b
No Countermeasures	c	d

- **DOE approach: construct a *model* that links all the data together**

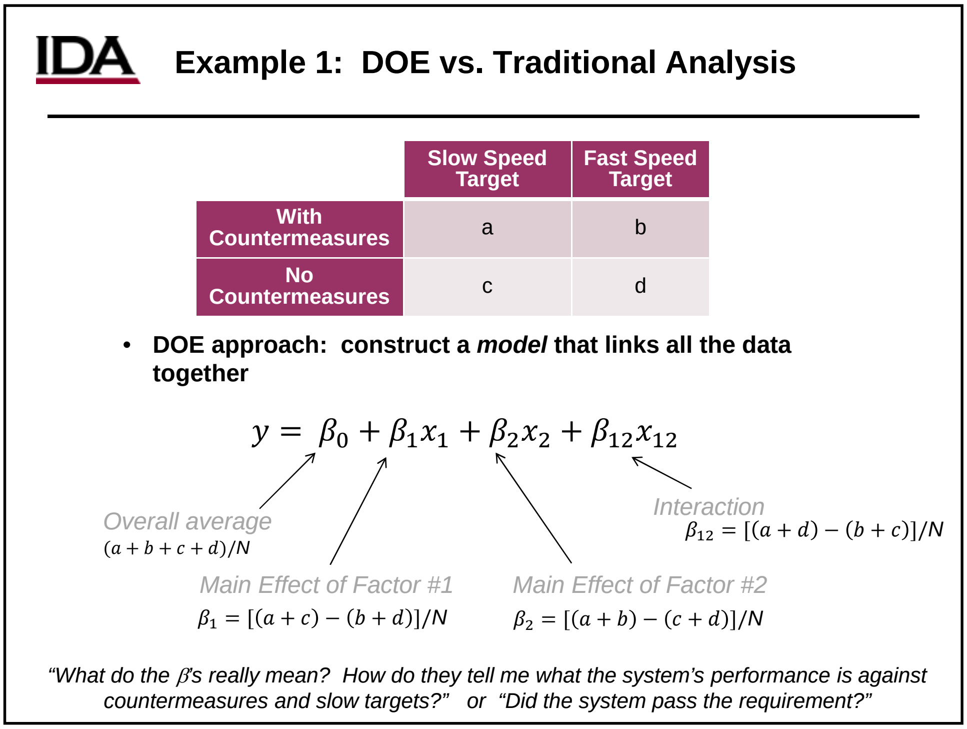
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_{12}$$

Overall average

Interaction between speed and countermeasures

Main Effect of Factor #1 (Target Speed)

Main Effect of Factor #2 (Countermeasures)



Example 1: DOE vs. Traditional Analysis

	Slow Speed Target	Fast Speed Target
With Countermeasures	a	b
No Countermeasures	c	d

- **DOE approach: construct a *model* that links all the data together**

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_{12}$$

Overall average

$(a+b+c+d)/N$

Interaction

$\beta_{12} = [(a+d)-(b+c)]/N$

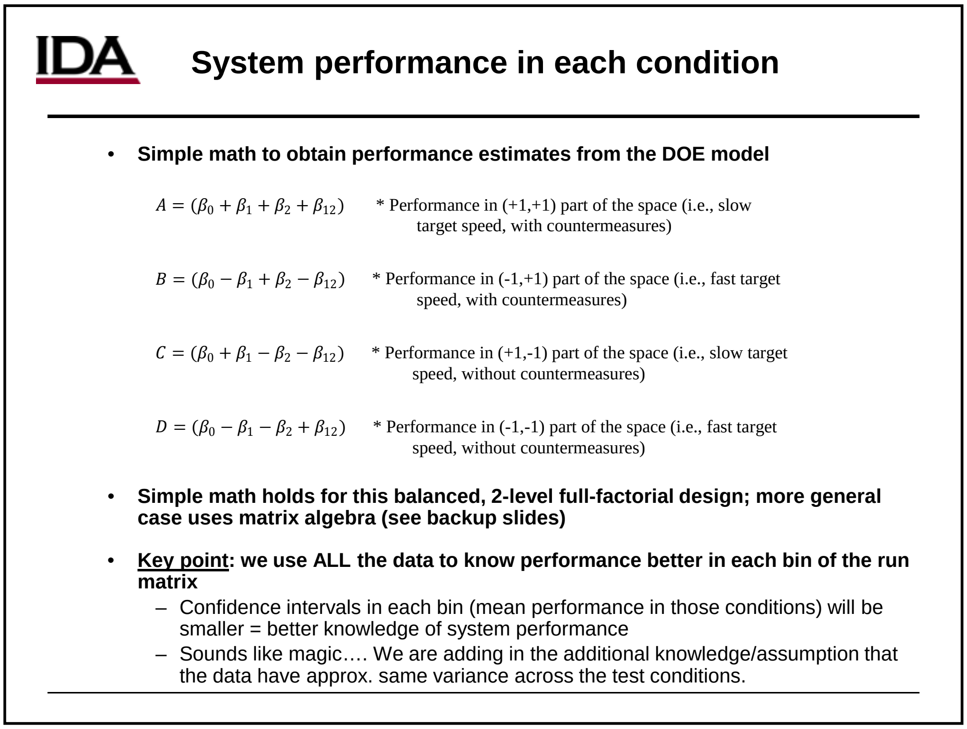
Main Effect of Factor #1

$\beta_1 = [(a+c)-(b+d)]/N$

Main Effect of Factor #2

$\beta_2 = [(a+b)-(c+d)]/N$

"What do the β's really mean? How do they tell me what the system's performance is against countermeasures and slow targets?" or "Did the system pass the requirement?"



System performance in each condition

- **Simple math to obtain performance estimates from the DOE model**

$A = (\beta_0 + \beta_1 + \beta_2 + \beta_{12})$ * Performance in (+1,+1) part of the space (i.e., slow target speed, with countermeasures)

$B = (\beta_0 - \beta_1 + \beta_2 - \beta_{12})$ * Performance in (-1,+1) part of the space (i.e., fast target speed, with countermeasures)

$C = (\beta_0 + \beta_1 - \beta_2 - \beta_{12})$ * Performance in (+1,-1) part of the space (i.e., slow target speed, without countermeasures)

$D = (\beta_0 - \beta_1 - \beta_2 + \beta_{12})$ * Performance in (-1,-1) part of the space (i.e., fast target speed, without countermeasures)

- **Simple math holds for this balanced, 2-level full-factorial design; more general case uses matrix algebra (see backup slides)**
- **Key point: we use ALL the data to know performance better in each bin of the run matrix**
  - Confidence intervals in each bin (mean performance in those conditions) will be smaller = better knowledge of system performance
  - Sounds like magic…. We are adding in the additional knowledge/assumption that the data have approx. same variance across the test conditions.

IDA

Example 1: with data

	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

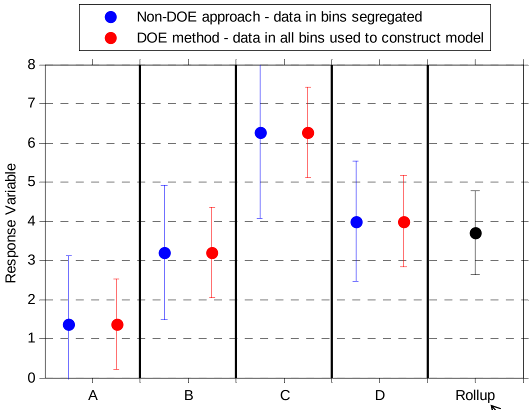
- **IOT&E of a system: 12 runs, 3 in each condition,**
  measured the detection range (response variable)
  - Average performance: 3.7
  - Wide spread in performance, as expected due to the different operational conditions: [0.2 to 7.5]
  - Large confidence intervals expected (traditional view) in each condition since we only have 3 runs in each case

Means and std. deviations (note stdev is approx. same in all bins)

	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)

IDA

DOE versus Non-DOE Analysis



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

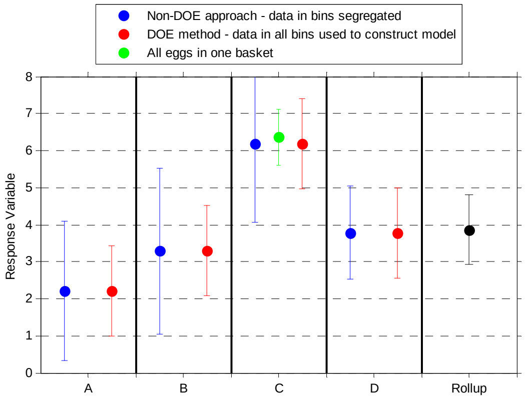
- **Non-DOE approach: calculate confidence intervals using only data collected under each condition**
- **DOE approach: construct a model (pool the data), use the model to estimate mean values in each condition**
  - Note the reduction in confidence interval size!
    - » In this case, intervals reduced 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***

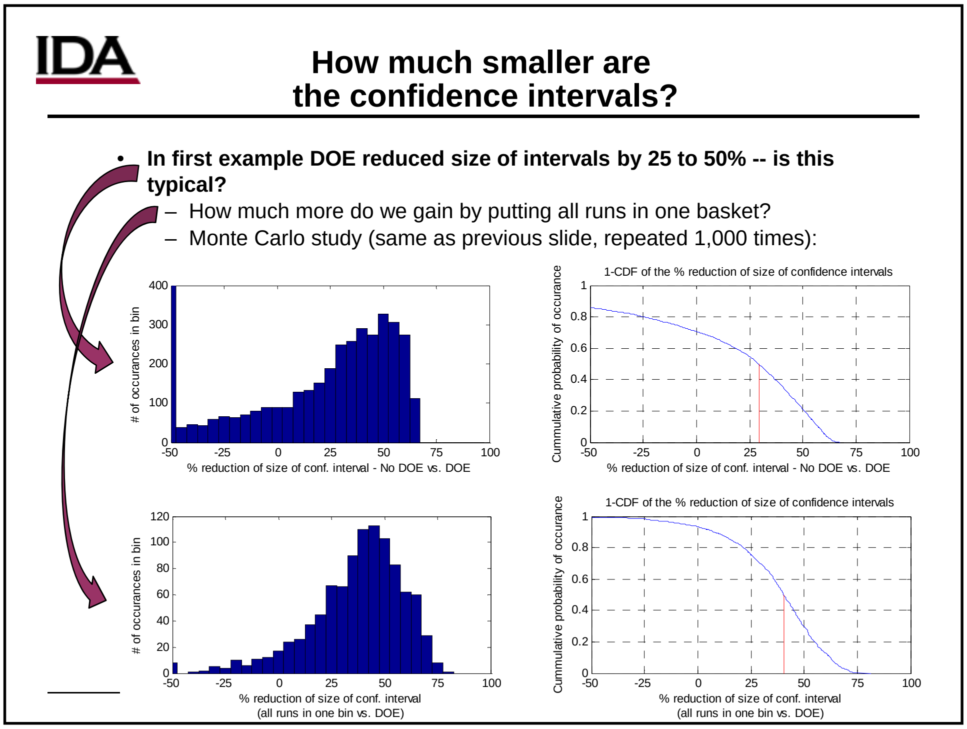
Example 1 Modified: Consolidate resources to one bin

- **Should we have allocated our resources and conducted all 12 runs in one set of conditions?**
  - Understood that we lose ability to know performance in other conditions, but at least we'd have an accurate measure in one case.
- **To do the comparison, must do a Monte Carlo study, sampling from a known distribution**
  - Compare 12 runs in one bin vs. 3 runs in 4 bins

Example 1 Modified: Consolidate resources to one bin

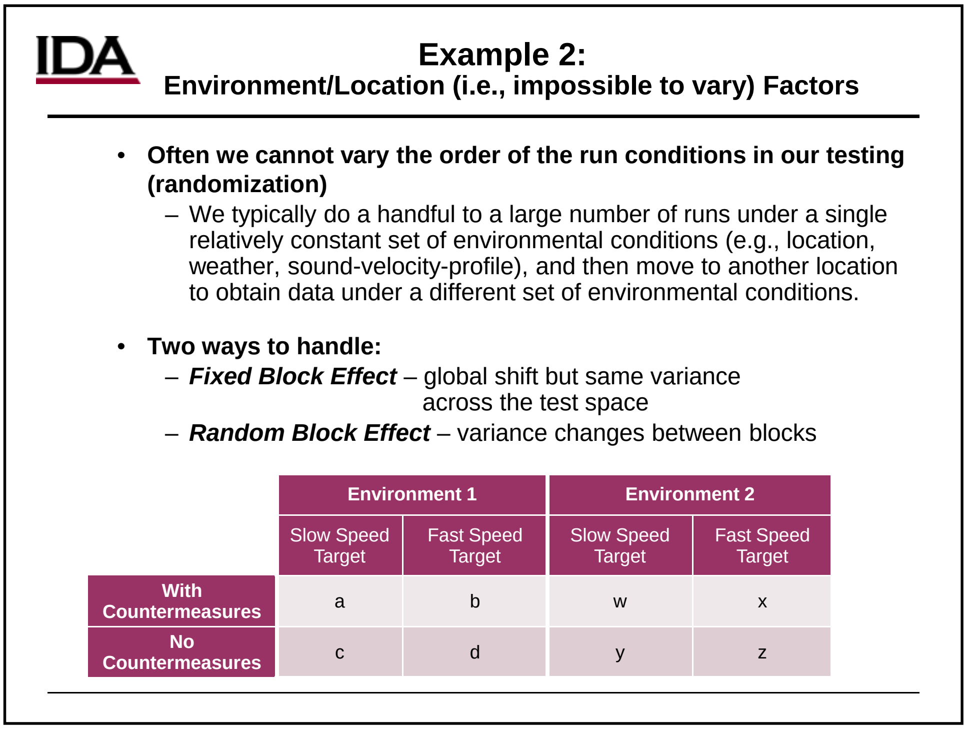


- **As expected, confidence interval is smaller for 12-in-one-bin case**
  - But is this a better test strategy?
    - » Worth the loss of information in other conditions?
    - » That precise a measurement necessary?



How much smaller are the confidence intervals?

- **In first example DOE reduced size of intervals by 25 to 50% -- is this typical?**
  - How much more do we gain by putting all runs in one basket?
  - Monte Carlo study (same as previous slide, repeated 1,000 times):

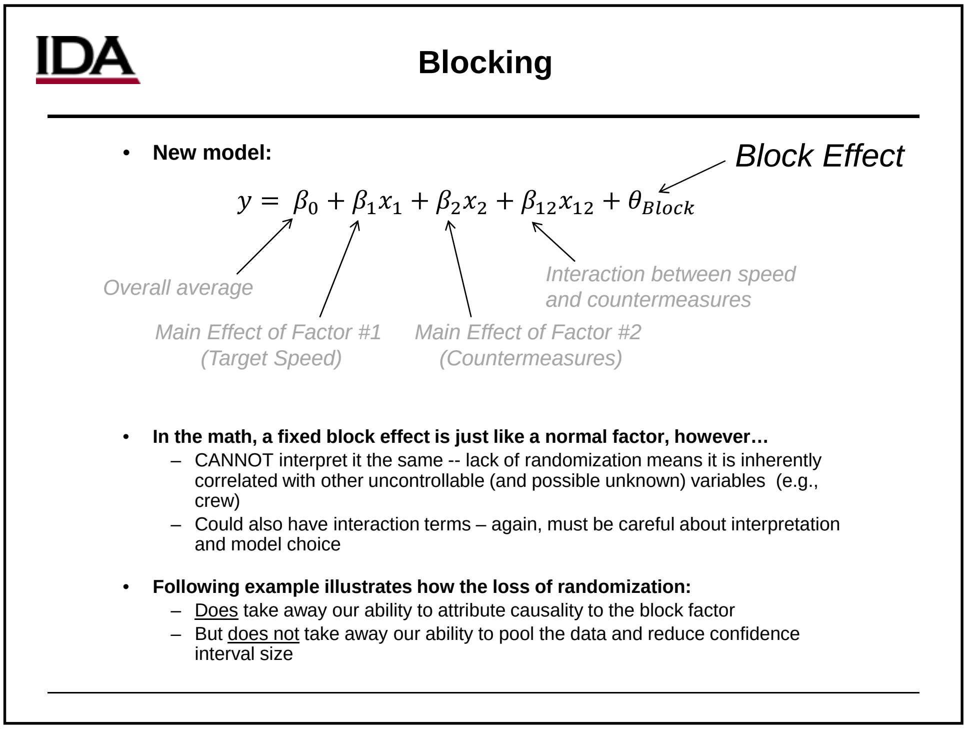


Example 2:
Environment/Location (i.e., impossible to vary) Factors

- **Often we cannot vary the order of the run conditions in our testing (randomization)**
  - We typically do a handful to a large number of runs under a single relatively constant set of environmental conditions (e.g., location, weather, sound-velocity-profile), and then move to another location to obtain data under a different set of environmental conditions.

- **Two ways to handle:**
  - ***Fixed Block Effect*** – global shift but same variance across the test space
  - ***Random Block Effect*** – variance changes between blocks

	Environment 1		**Environment 2**	
	Slow Speed Target	Fast Speed Target	Slow Speed Target	Fast Speed Target
With Countermeasures	a	b	w	x
No Countermeasures	c	d	y	z



Blocking

- **New model:**

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_{12} + \theta_{Block}$$

Block Effect

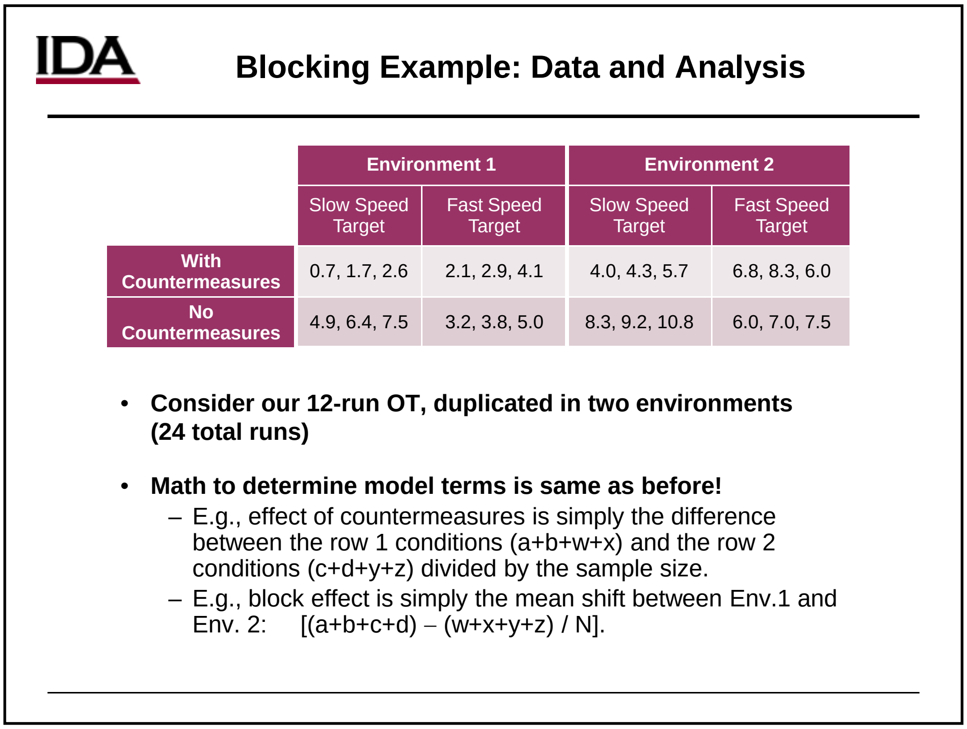
Overall average

Main Effect of Factor #1 (Target Speed)

Main Effect of Factor #2 (Countermeasures)

Interaction between speed and countermeasures

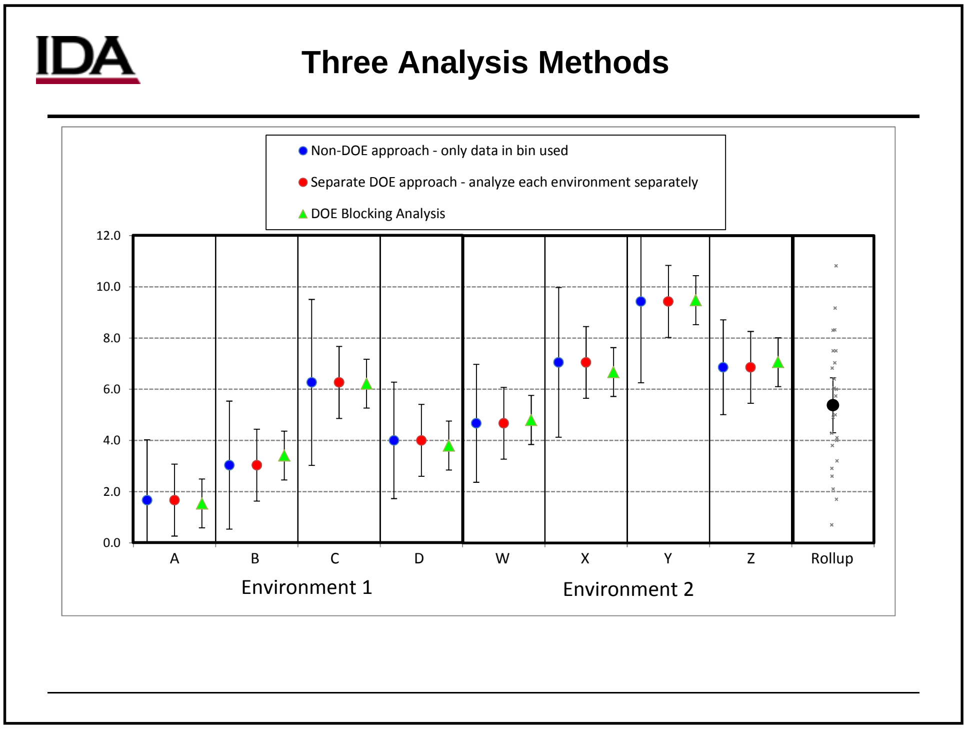
- **In the math, a fixed block effect is just like a normal factor, however…**
  - CANNOT interpret it the same -- lack of randomization means it is inherently correlated with other uncontrollable (and possible unknown) variables (e.g., crew)
  - Could also have interaction terms – again, must be careful about interpretation and model choice
- **Following example illustrates how the loss of randomization:**
  - Does take away our ability to attribute causality to the block factor
  - But does not take away our ability to pool the data and reduce confidence interval size



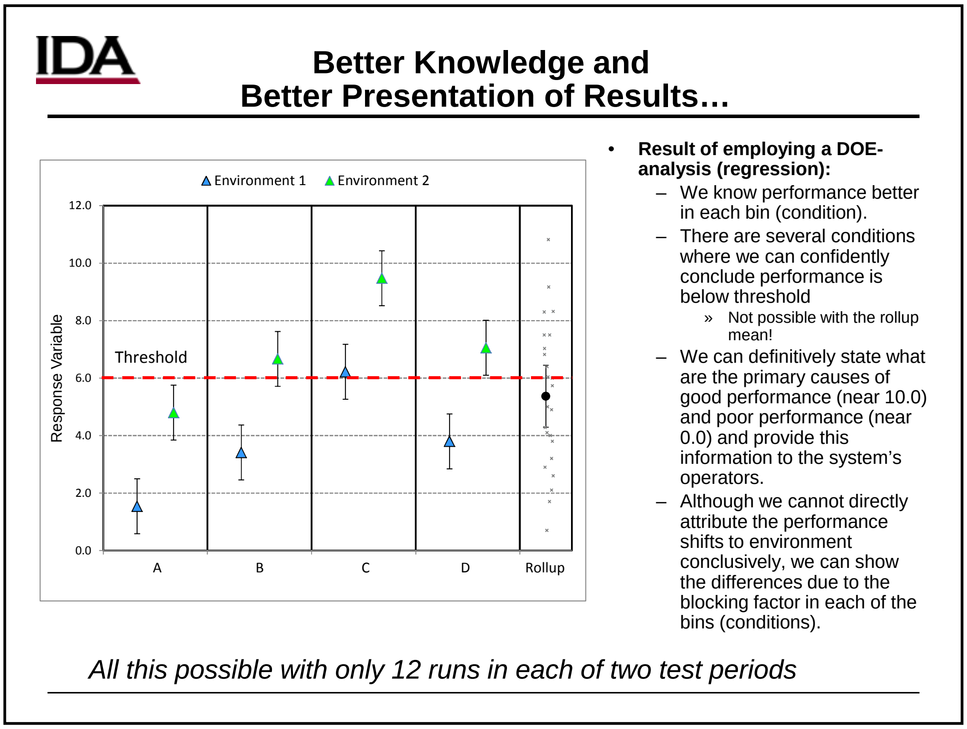
Blocking Example: Data and Analysis

	Environment 1		Environment 2	
	Slow Speed Target	Fast Speed Target	Slow Speed Target	Fast Speed Target
With Countermeasures	0.7, 1.7, 2.6	2.1, 2.9, 4.1	4.0, 4.3, 5.7	6.8, 8.3, 6.0
No Countermeasures	4.9, 6.4, 7.5	3.2, 3.8, 5.0	8.3, 9.2, 10.8	6.0, 7.0, 7.5

- **Consider our 12-run OT, duplicated in two environments (24 total runs)**
- **Math to determine model terms is same as before!**
  - E.g., effect of countermeasures is simply the difference between the row 1 conditions (a+b+w+x) and the row 2 conditions (c+d+y+z) divided by the sample size.
  - E.g., block effect is simply the mean shift between Env.1 and Env. 2: [(a+b+c+d) – (w+x+y+z) / N].



Three Analysis Methods



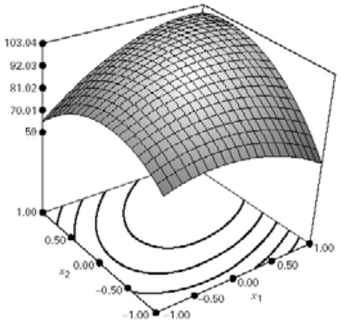
Better Knowledge and Better Presentation of Results…

- **Result of employing a DOE-analysis (regression):**
  - We know performance better in each bin (condition).
  - There are several conditions where we can confidently conclude performance is below threshold
    - » Not possible with the rollup mean!
  - We can definitively state what are the primary causes of good performance (near 10.0) and poor performance (near 0.0) and provide this information to the system's operators.
  - Although we cannot directly attribute the performance shifts to environment conclusively, we can show the differences due to the blocking factor in each of the bins (conditions).

All this possible with only 12 runs in each of two test periods

Other Benefits of DOE Analysis

- **Interpolation and Prediction**
  - We didn’t test in every possible condition: with continuous factors, and an understanding of the response, we can estimate performance in other sectors of the space
- **Data drives the analysis methodology**
  - Significance tests for factor effects can be used to determine what order model is significant, insignificant terms can be dropped from the model improving the modeling power
- **Structured Methodology for planning tests to characterize performance**



(*a*) Response surface

- **Model provides useful tool for data visualization**
  - Pareto charts
  - Contour plots
  - Interaction graphs

Caveats and Conclusions

- **These examples worked because the *assumptions* were valid and the test design supported the analysis**
  - » Normally distributed data
  - » Nearly constant variance
  - » Randomization
  - » Orthogonal design
  - » Powerful enough test for the effect size (~2σ)
- **Employing a regression analysis in concert with a carefully planned test with DOE concepts results in ability to say a lot with a little**
  - Much better reporting than global means with confidence intervals alone
  - Better reporting than bin-by-bin means as well
- **Planning a test using DOE concepts is a good start – need to follow with associated analysis**
- **Tomorrow’s session to detail additional analysis techniques…**

Backup

Math behind the curtain

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_{12} + \varepsilon$$

Write the model equation as a matrix (one row for each run)

$$\mathbf{y} = \boldsymbol{X} \cdot \boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

"Design Matrix"

Number of observations

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_{12} \end{bmatrix} \downarrow$$

Model terms →

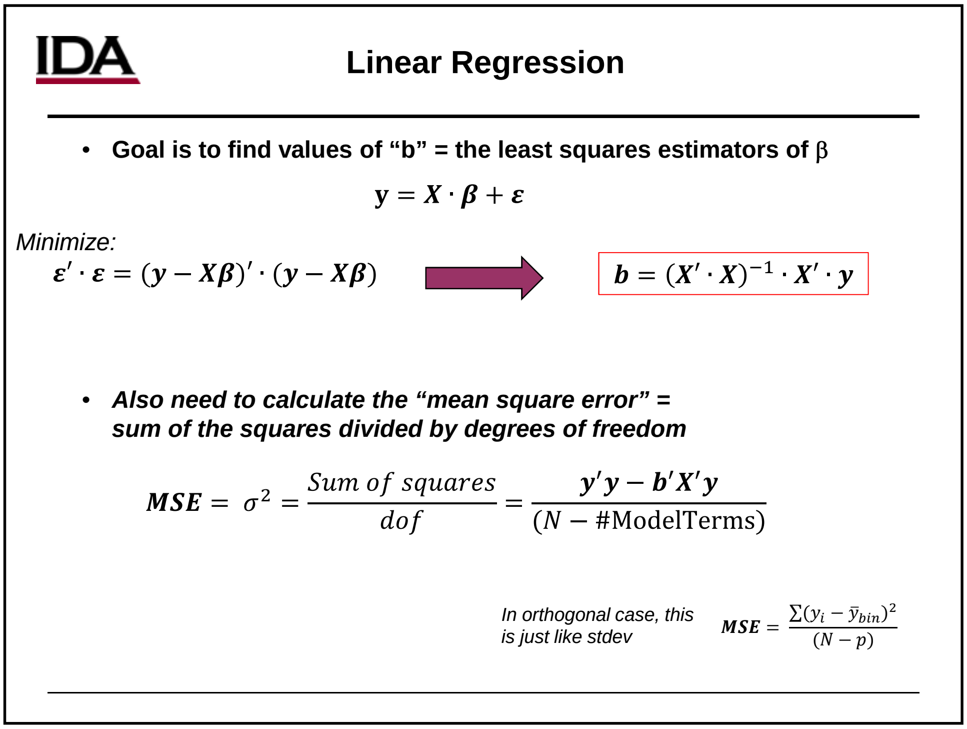
Number of runs ↓

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{1,1} & x_{2,1} & (x_1 x_2)_1 \\ 1 & x_{1,2} & x_{2,2} & (x_1 x_2)_2 \\ 1 & x_{1,3} & x_{2,3} & (x_1 x_2)_3 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1,12} & x_{2,12} & (x_1 x_2)_{12} \end{bmatrix}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_{12} \end{bmatrix}$$

- **Typically put $x_{i,j}$ in "coded" units: i.e., the point in the design space where you make a measurement**
  - Example, run number 1 was done at the (x_1 = +1, x_2 = +1) part of the DOE matrix

	Slow Speed Target	Fast Speed Target
With Countermeasures	(+1, +1)	(-1, +1)
No Countermeasures	(+1, -1)	(-1, -1)



IDA

Linear Regression

- **Goal is to find values of "b" = the least squares estimators of β**

$$\mathbf{y} = \boldsymbol{X}\cdot\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Minimize:

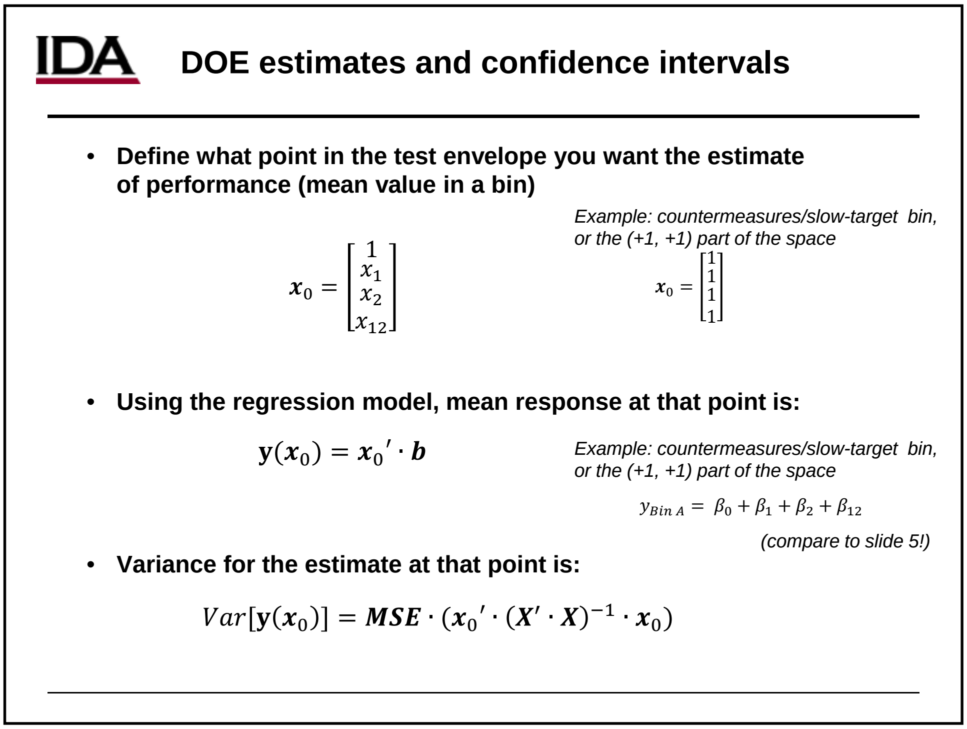
$$\boldsymbol{\varepsilon}'\cdot\boldsymbol{\varepsilon} = (\boldsymbol{y} - \boldsymbol{X\beta})'\cdot(\boldsymbol{y} - \boldsymbol{X\beta}) \quad \Rightarrow \quad \boxed{\boldsymbol{b} = (\boldsymbol{X}'\cdot\boldsymbol{X})^{-1}\cdot\boldsymbol{X}'\cdot\boldsymbol{y}}$$

- ***Also need to calculate the "mean square error" = sum of the squares divided by degrees of freedom***

$$\boldsymbol{MSE} = \sigma^2 = \frac{Sum\ of\ squares}{dof} = \frac{\boldsymbol{y}'\boldsymbol{y} - \boldsymbol{b}'\boldsymbol{X}'\boldsymbol{y}}{(N - \#\mathrm{ModelTerms})}$$

In orthogonal case, this is just like stdev

$$\boldsymbol{MSE} = \frac{\Sigma(y_i - \bar{y}_{bin})^2}{(N-p)}$$



IDA

DOE estimates and confidence intervals

- **Define what point in the test envelope you want the estimate of performance (mean value in a bin)**

$$\boldsymbol{x}_0 = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ x_{12} \end{bmatrix}$$

Example: countermeasures/slow-target bin, or the (+1, +1) part of the space

$$\boldsymbol{x}_0 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

- **Using the regression model, mean response at that point is:**

$$\mathbf{y}(\boldsymbol{x}_0) = \boldsymbol{x}_0{}'\cdot\boldsymbol{b}$$

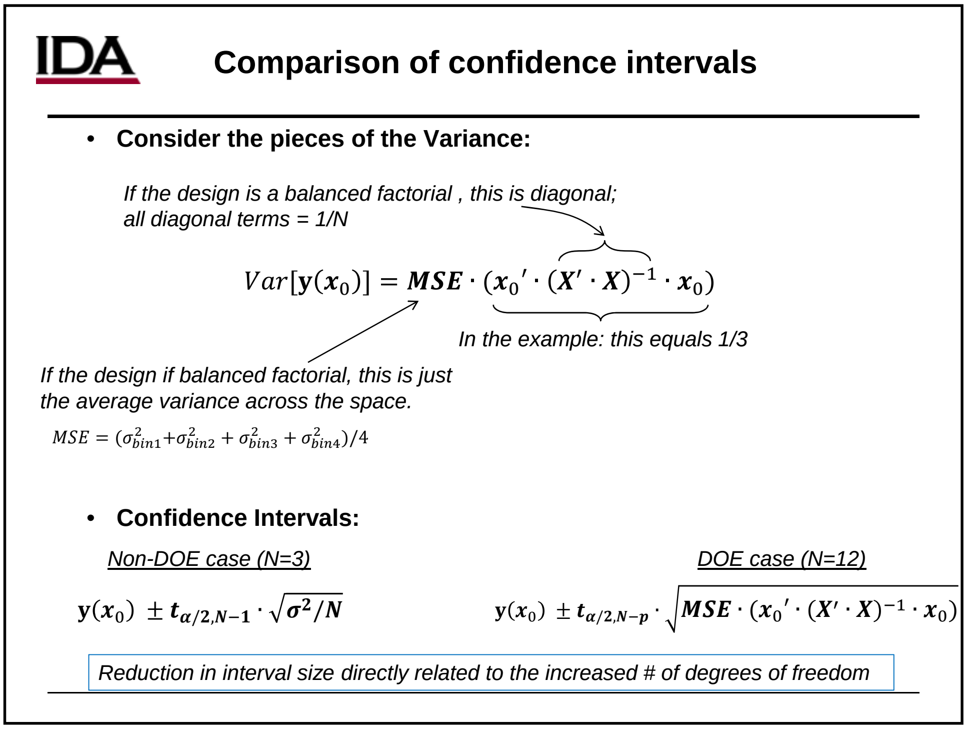
Example: countermeasures/slow-target bin, or the (+1, +1) part of the space

$$y_{Bin\ A} = \beta_0 + \beta_1 + \beta_2 + \beta_{12}$$

(compare to slide 5!)

- **Variance for the estimate at that point is:**

$$Var[\mathbf{y}(\boldsymbol{x}_0)] = \boldsymbol{MSE}\cdot(\boldsymbol{x}_0{}'\cdot(\boldsymbol{X}'\cdot\boldsymbol{X})^{-1}\cdot\boldsymbol{x}_0)$$



Comparison of confidence intervals

- **Consider the pieces of the Variance:**

If the design is a balanced factorial , this is diagonal;
all diagonal terms = 1/N

$$Var[\mathbf{y}(\boldsymbol{x}_0)] = \boldsymbol{MSE} \cdot (\underbrace{\boldsymbol{x}_0{}' \cdot \overbrace{(\boldsymbol{X}' \cdot \boldsymbol{X})^{-1}} \cdot \boldsymbol{x}_0})$$

In the example: this equals 1/3

If the design if balanced factorial, this is just
the average variance across the space.

$MSE = (\sigma_{bin1}^2 + \sigma_{bin2}^2 + \sigma_{bin3}^2 + \sigma_{bin4}^2)/4$

- **Confidence Intervals:**

Non-DOE case (N=3)

$$\mathbf{y}(\boldsymbol{x}_0) \pm \boldsymbol{t}_{\boldsymbol{\alpha/2,N-1}} \cdot \sqrt{\boldsymbol{\sigma^2/N}}$$

DOE case (N=12)

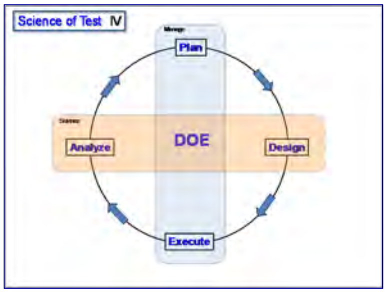
$$\mathbf{y}(\boldsymbol{x}_0) \pm \boldsymbol{t}_{\boldsymbol{\alpha/2,N-p}} \cdot \sqrt{\boldsymbol{MSE} \cdot (\boldsymbol{x}_0{}' \cdot (\boldsymbol{X}' \cdot \boldsymbol{X})^{-1} \cdot \boldsymbol{x}_0)}$$

Reduction in interval size directly related to the increased # of degrees of freedom

Appendix 3
Roadmap Case Studies

This page intentionally left blank.

Appendix 3-1.
Examples of DOE Applied in Air Warfare OT

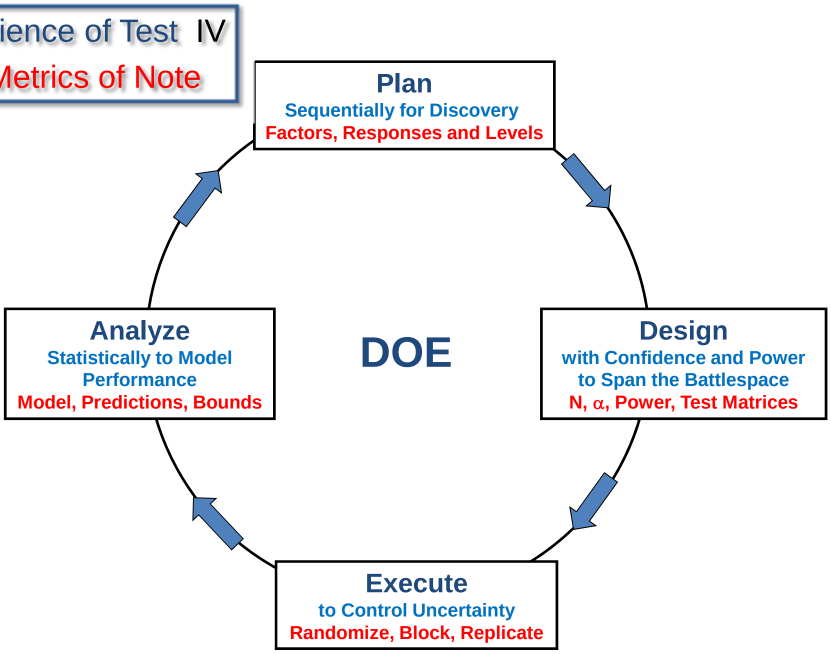


Examples of DOE Applied in OT

Matt Kowalski, 53d Wing
Greg Hutto 46 TW
Jim Simpson, 53d Wing

Science of Test IV

Metrics of Note



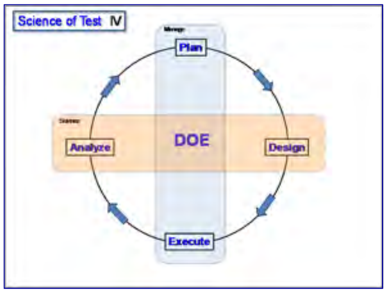
Plan
Sequentially for Discovery
Factors, Responses and Levels

Design
with Confidence and Power
to Span the Battlespace
N, α, Power, Test Matrices

DOE

Execute
to Control Uncertainty
Randomize, Block, Replicate

Analyze
Statistically to Model Performance
Model, Predictions, Bounds



Air-to-Ground Missile

Maverick H/K FDE

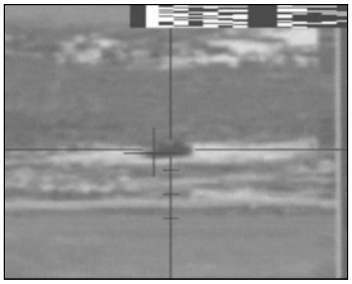
Purpose of Test

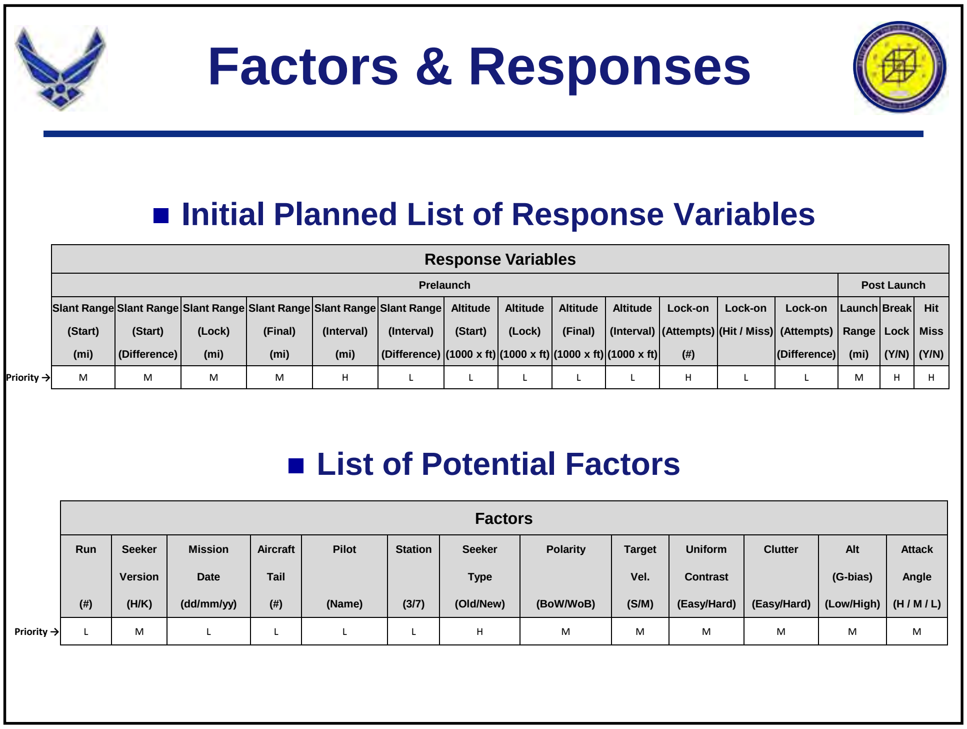
- WSEP found problems with certain target conditions
- Raytheon made enhancements to software
- Upon Fielding Recommendation Raytheon to retrofit existing Maverick AGM-65H/K inventory with new software

System Description

- Maverick Air to Ground Missile (AGM)-65H/K
  - Electro-Optical (E/O) guidance

    (Black & White television camera)
  - AGM-65H 125 lb warhead
    - Molten Aluminum Projectile
    - Used in armor penetration
  - AGM-65K 300 lb warhead
    - Blast and Frag Projectile
- Seeker software attempts to bound target by analyzing black and white contrast between target and background





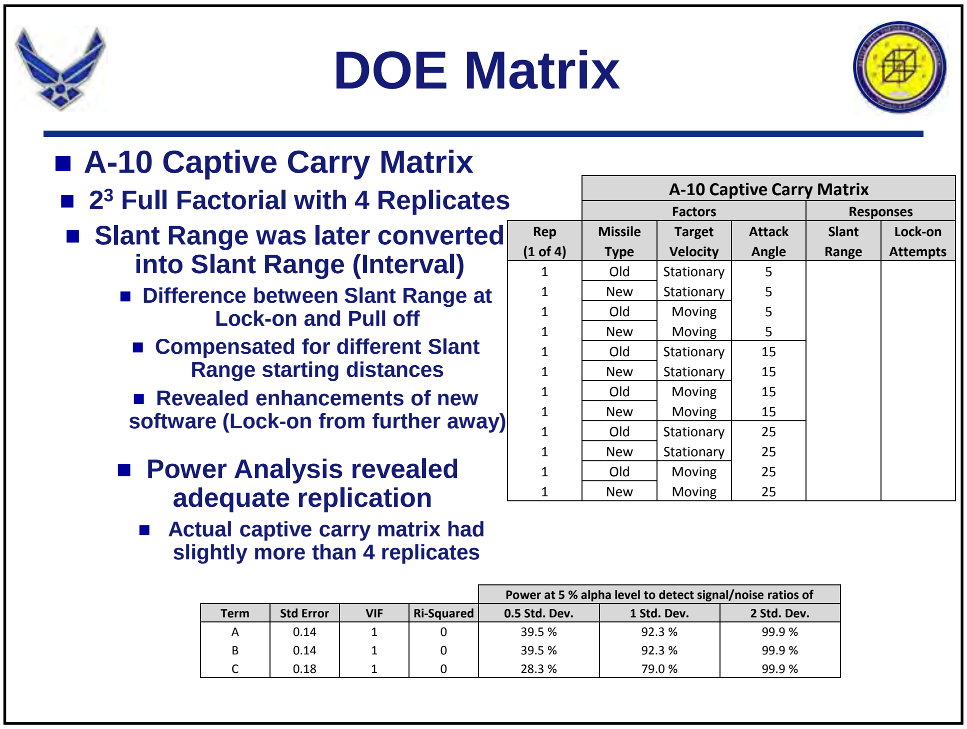
Factors & Responses

- Initial Planned List of Response Variables

	Response Variables															
	Prelaunch													Post Launch		
	Slant Range (Start) (mi)	Slant Range (Start) (Difference)	Slant Range (Lock) (mi)	Slant Range (Final) (mi)	Slant Range (Interval) (mi)	Slant Range (Interval) (Difference)	Altitude (Start) (1000 x ft)	Altitude (Lock) (1000 x ft)	Altitude (Final) (1000 x ft)	Altitude (Interval) (1000 x ft)	Lock-on (Attempts) (#)	Lock-on (Hit / Miss)	Lock-on (Attempts) (Difference)	Launch Range (mi)	Break Lock (Y/N)	Hit Miss (Y/N)
Priority →	M	M	M	M	H	L	L	L	L	L	H	L	L	M	H	H

- List of Potential Factors

	Factors												
	Run (#)	Seeker Version (H/K)	Mission Date (dd/mm/yy)	Aircraft Tail (#)	Pilot (Name)	Station (3/7)	Seeker Type (Old/New)	Polarity (BoW/WoB)	Target Vel. (S/M)	Uniform Contrast (Easy/Hard)	Clutter (Easy/Hard)	Alt (G-bias) (Low/High)	Attack Angle (H / M / L)
Priority →	L	M	L	L	L	L	H	M	M	M	M	M	M

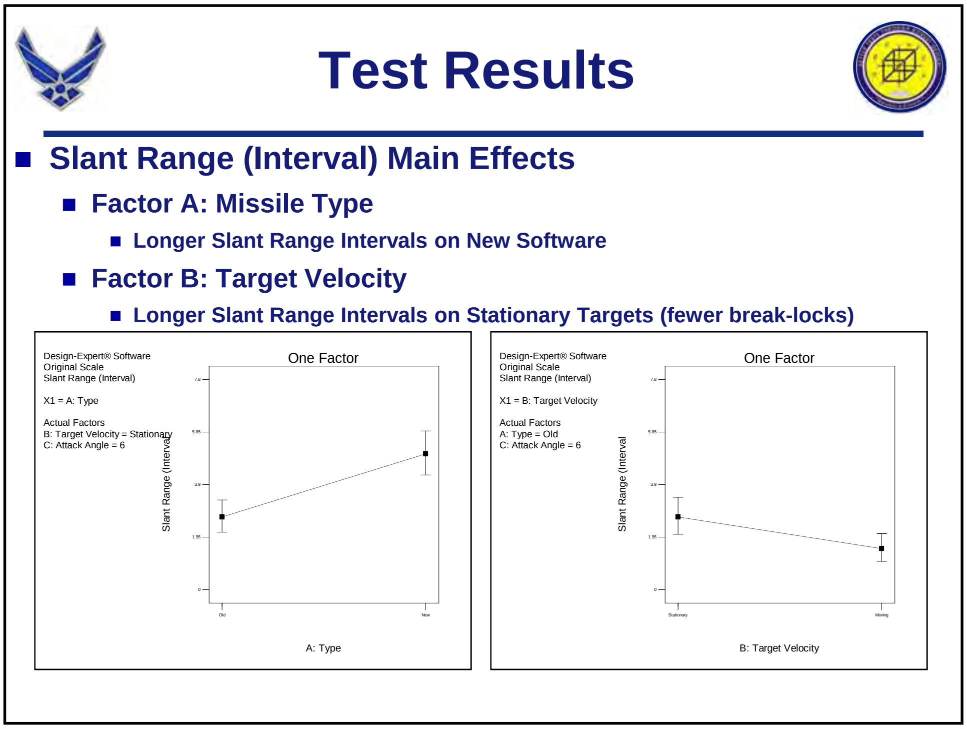


DOE Matrix

- A-10 Captive Carry Matrix
  - 2^3 Full Factorial with 4 Replicates
  - Slant Range was later converted into Slant Range (Interval)
    - Difference between Slant Range at Lock-on and Pull off
    - Compensated for different Slant Range starting distances
    - Revealed enhancements of new software (Lock-on from further away)
  - Power Analysis revealed adequate replication
    - Actual captive carry matrix had slightly more than 4 replicates

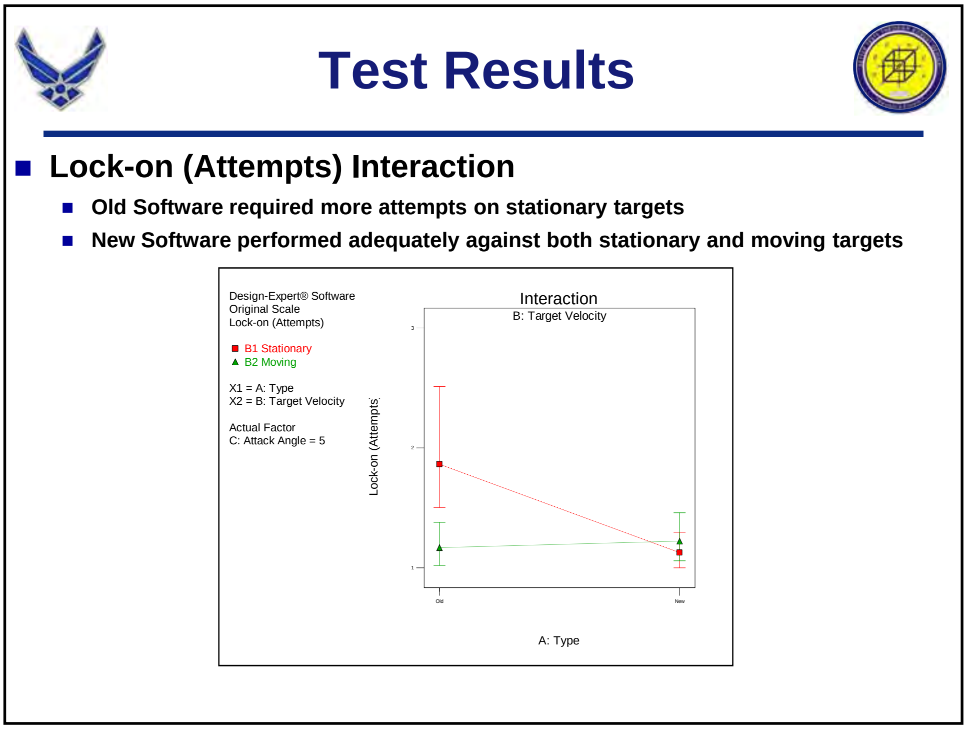
	A-10 Captive Carry Matrix				
	Factors			Responses	
Rep (1 of 4)	Missile Type	Target Velocity	Attack Angle	Slant Range	Lock-on Attempts
1	Old	Stationary	5		
1	New	Stationary	5		
1	Old	Moving	5		
1	New	Moving	5		
1	Old	Stationary	15		
1	New	Stationary	15		
1	Old	Moving	15		
1	New	Moving	15		
1	Old	Stationary	25		
1	New	Stationary	25		
1	Old	Moving	25		
1	New	Moving	25		

				Power at 5 % alpha level to detect signal/noise ratios of		
Term	Std Error	VIF	Ri-Squared	0.5 Std. Dev.	1 Std. Dev.	2 Std. Dev.
A	0.14	1	0	39.5 %	92.3 %	99.9 %
B	0.14	1	0	39.5 %	92.3 %	99.9 %
C	0.18	1	0	28.3 %	79.0 %	99.9 %



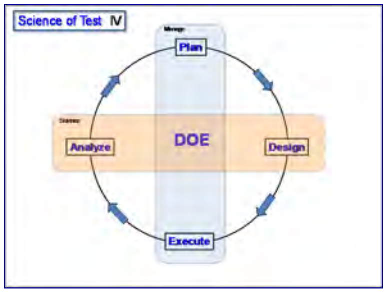
Test Results

- Slant Range (Interval) Main Effects
  - Factor A: Missile Type
    - Longer Slant Range Intervals on New Software
  - Factor B: Target Velocity
    - Longer Slant Range Intervals on Stationary Targets (fewer break-locks)



Test Results

- Lock-on (Attempts) Interaction
  - Old Software required more attempts on stationary targets
  - New Software performed adequately against both stationary and moving targets



Fighter OFP System

Integrated KT/DT/OT

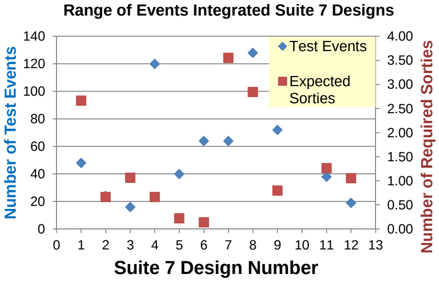
Suite 7 F-15C/E OFP Upgrade

Test Objective

- Diverse stakeholders – Boeing, 46 TW, 53d Wing with multiple systems
- Radar GMTT/TI, EA/EP, SAR, TEWS, Sniper, AMRAAM, etc.
- Constrained sorties for Suite 7 test program
- Solution: **custom design** for each objective

DOE Approach

- Partner with Boeing to answer both DT and OT questions
- Augment experience and engineering judgment w/ series of designed experiments
- Chart shows wide variety of events and sorties totaling about 15 sorties
- Represents about 10% savings, but goal was statistically defensible tests



Summary of 12 Designs

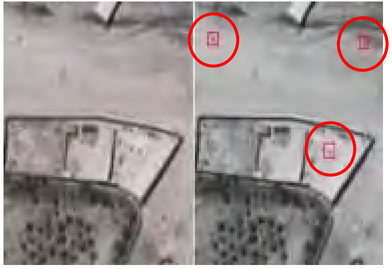
				Power				
Design	Topic	N Variables (Vars)	Var Levels	1 sigma	2 sigma	Model	Design Strategy	Test Events
1	Air to Air Jam Protection (EA)	6 x 2^4 fraction	Mixed	20-80	70-99	ME+2FI	D-optimal	48
2	Velocity Sweep Excursion	3x2 - 4 reps full	Mixed		92-99	ME+2FI	Gen Factorial	24
3	Other EA Mode Excursion	4x2^2 - 1 rep	Mixed		84-99	ME+2FI	Gen Factorial	16
4	WVR AutoAcquire categoric	4x2 30 reps	Mixed		90	ME+2FI	Gen Factorial	120
5	WVR AutoAcquire numeric	5x5 2 reps	Mixed	47	96	Quadratic	D-opt RSM	40
6	WVR AutoAcquire WEZ check	2 cat x 3 numeric	Mixed	80-97	99	Quadratic	CCD RSM	64
7	SAR Map (EHRM) Matrix	6 x 2 level vars	2^k	97	99	ME+2FI	Full Factorial	64
8	Air-to-Ground Mov'g Tgt Trk	2^7 Vars	2^k	99	99	ME+2FI	Full Factorial	128
9	IFF Mode 5 Design 2 reps	2^2 x 3^2 4 vars	Mixed	86-96	99	ME+2FI	Full Factorial	72
11	Sniper Targeting Pod Tgt Loc Er	2^7 level vars	2^k +cp	65-70	99	Rev ME+2FI	1/4 fraction	38
12	Sniper Air-Gnd Movers track	2^5 level vars	2^k +cp	30	80	ME+2FI	1/2 Fraction	19

- **In total, about 6-8 days work over four months with KT/DT/OT team**
- **Each design tuned to perceived risk, expense, complexity of battlespace**

Glossary:
Vars – separate test conditions (alt, range, EA tech)
ME -- Main Effects, vars acting alone
2FI – 2 factor interaction, 2 vars acting together
WVR - within visual range engagements e.g. <5 nmi

Radar GMTI

2-Level Fractional Factorial Design



Design Metrics

Metric Name	Metric value
2 σ Power @ 95% Confid	99.9
Pred SD Accuracy @75% FDS	.31
Variables Considered	7
All Combinations	128
Test Set Points	32 + 16+4
Fraction of All Combos	38%
Model Order Supported	7 FI
Aliasing -	None-Full Resolution

Run Set Objective

- Can Suite7 Radar Indicate Moving Ground Targets?
- ID factors that influence detect/display

DOE Approach

- Many factors to begin – 7-9 variables
- Screen these down to the most important factors

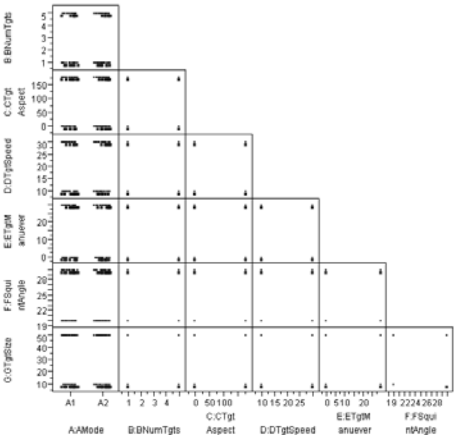
Pros and Cons of this set:

- Screening design with follow on additional runs
- Very robust to missing data – even 30-40%
- Efficient and learn as you test
- Excellent power and confidence
- Sequential experimentation – stop early

Visualizing the Input Space

GMTI Matrix

Factor	Name	Units	Type	Low	High
A	AMode	Type	Categoric	Mode 1	Mode 2
B	BNumTgts	Count	Numeric	1	5
C	CTgtAspect	Degrees	Numeric	0	180
D	DTgtSpeed	Knots	Numeric	10	30
E	TgtManueve	Degrees	Numeric	0	30
F	SquintAngl	Degrees	Numeric	20	30
G	GTgtSize	Feet	Numeric	10	50



- Primarily Screening Design
- Inexpensive Points – Place more targets on range
- Multiple test points per pass

Electronic Attack

Mixed Level Fraction

Design Metrics

Metric Name	Metric value
2 σ Power @ 95% Confid	95%+ but 55% EA
Pred SD Accuracy @75% FDS	1.45
Variables Considered	5 2^4 x 6 level
All Combinations	96
Test Set Points	48
Fraction of All Combos	50%
Model Order Supported	Main Eff + 2 FI (ex EA)
Aliasing -	Extensive, moderate

Run Set Objective

- Can Suite7 Radar Defend Electronic Attack?
- ID techniques that influence detect/display

DOE Approach

- Multiple EA techniques – dozens to begin
- Focus on discipline to examine "most important"
- Cannot achieve power for all levels

Pros and Cons of this set:

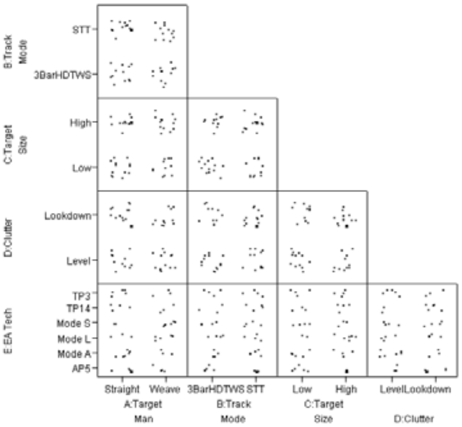
- Screening design with good resolution
- Will ID EA with strong impact (3 sigma)
- Low power EA (2 sigma) and EA interactions
- Robust to missing data – even 10-20%
- Good orthogonality (term isolation)
- A target for augmenting in strong EA techs
- Also – sequential experimentation – redesign

Input Space

EA Matrix

Factor	Name	Type	Low Actual	High Actual	
A	Target Man	Categoric	Straight	Weave	2 Levels:
B	Track Mode	Categoric	3BarHDTWS	STT	2 Levels:
C	Target Size	Categoric	Low	High	2 Levels:
D	Clutter	Categoric	Level	Lookdown	2 Levels:
E	EA Tech	Categoric	AP5	TP14	6 Levels:

- Desire broad look at EA Techniques – bottom row
- More Expensive Points – Each one a pass
- Single test point per pass



AMRAAM High Off-Boresight

Response Surface Design

Design Metrics

Metric Name	Metric value
2 σ Power @ 95% Confid	99%
Pred SD Accuracy @75% FDS	.47
Variables Considered	5 – 2^2 cat x 3^3 numeric
All Combinations	108
Test Set Points	64
Fraction of All Combos	59%
Model Order Supported	Quadratic
Aliasing -	Full resolution

Run Set Objective:

- Can Suite7 supply correct Weapon Engagement Zone for complex shots?
- ID conditions causing inaccurate displays

DOE Approach

- Multiple radar modes and AMRAAM types
- 3 var CCD crossed with 2 categoric vars in face-centered CCD
- Design can easily be expanded

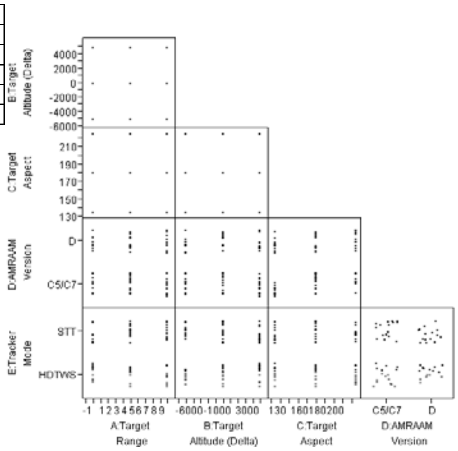
Pros and Cons of this set:

- Nicely handles geometric variables across three levels; if add more, go to fractional factorial
- Two categoric variables as well – expanding to 3 levels is possible
- Design could be trimmed if desired – points are cheap, however
- Good power and coverage of space

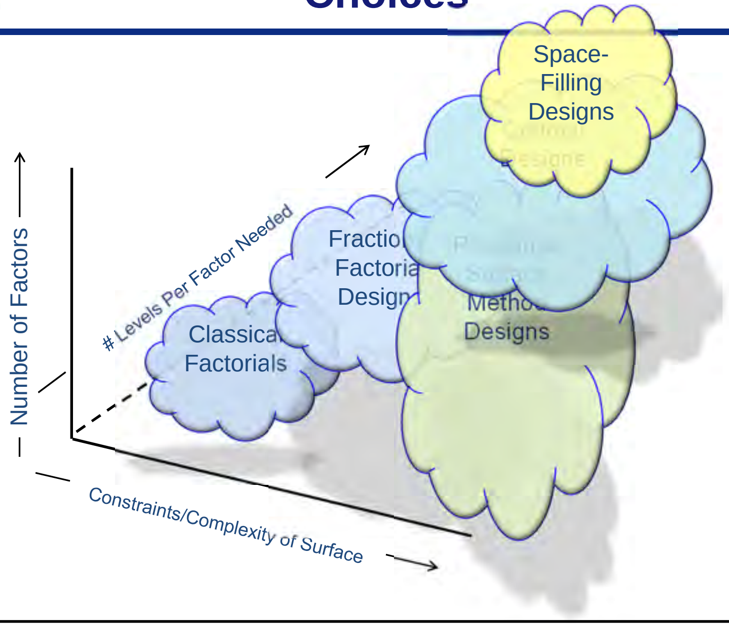
Input Space
HOBS WEZ

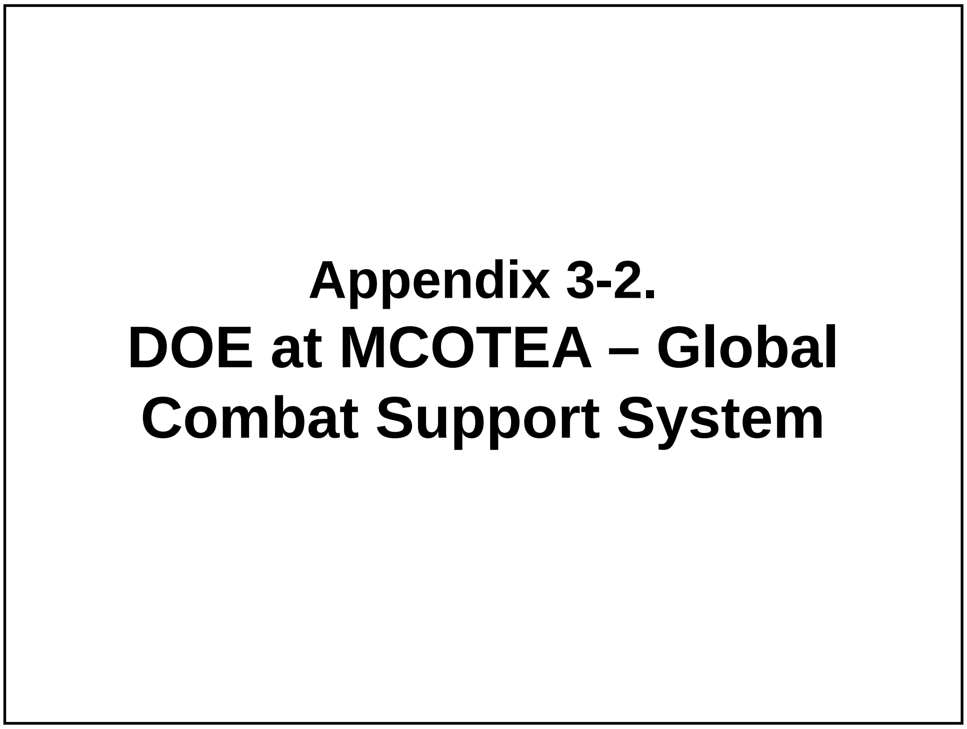
Factor	Name	Units	Type	Low Actual	High Actual	Levels:
A	Target Range	nm	Numeric	0	10	3
B	Target Altitude (Delta)	ft	Numeric	-5000	5000	3
C	Target Aspect	deg	Numeric	135	225	3
D	AMRAAM Version	cat	Categoric	C5/C7	D	2
E	Tracker Mode	cat	Categoric	STT	HDTWS	2

- Desire broad look at how WEZ behaves over geometry & version
- Good system-system interface: AMRAAM, radar, F-15 OFP, Link 16

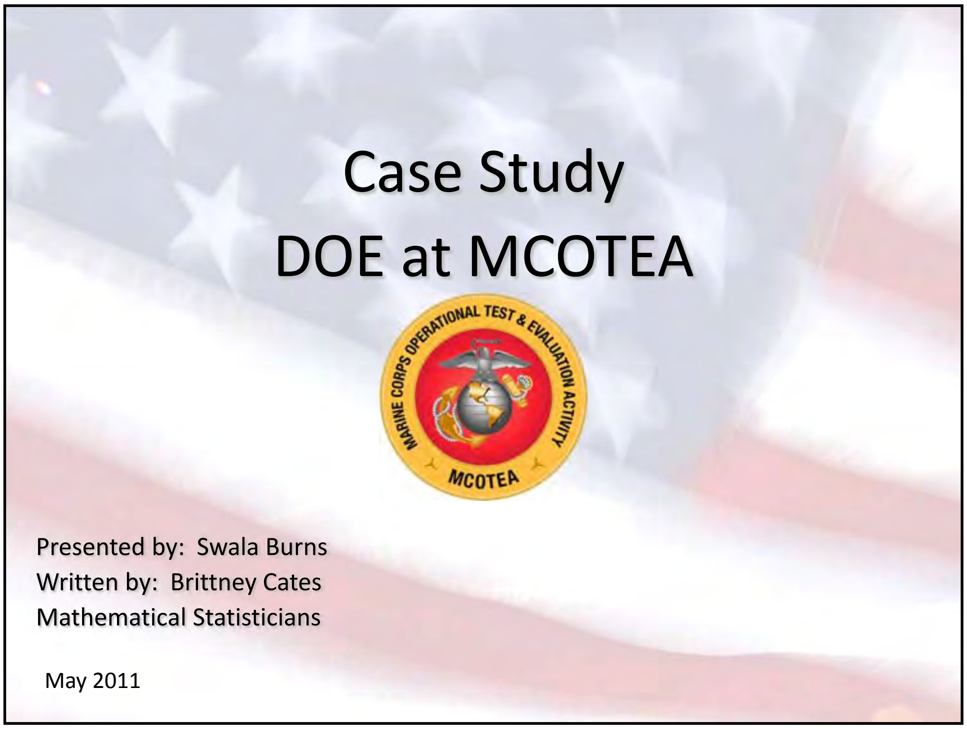


Problem Context Guides Design Choices


Space-Filling Designs
Optimal Designs
Response Surface Method Designs
Fractional Factorial Designs
Classical Factorials
Number of Factors
# Levels Per Factor Needed
Constraints/Complexity of Surface



Appendix 3-2.
DOE at MCOTEA – Global Combat Support System



Case Study
DOE at MCOTEA

MARINE CORPS OPERATIONAL TEST & EVALUATION ACTIVITY
MCOTEA

Presented by: Swala Burns
Written by: Brittney Cates
Mathematical Statisticians

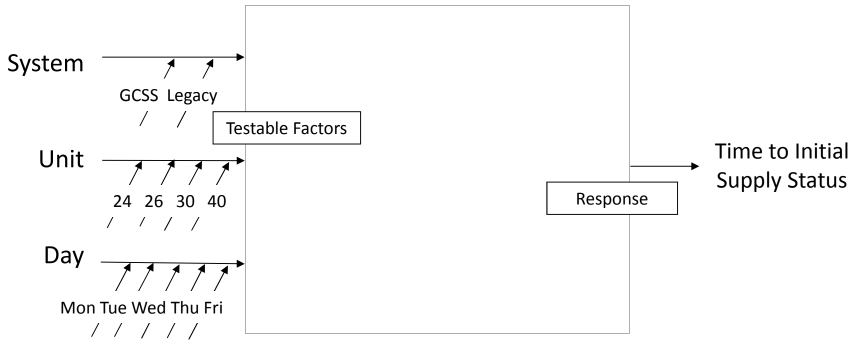
May 2011

Global Combat Support System – Marine Corps (GCSS-MC)

- Physical implementation of enterprise information technology architecture for Combat Service Support (CSS) functions
- Comparable to “Amazon.com”

Capabilities:

- Gain visibility of equipment readiness and position
- Track the location of inbound supplies
- Streamline the Warfighter’s procedures for requesting support



System

GCSS Legacy

Testable Factors

Unit

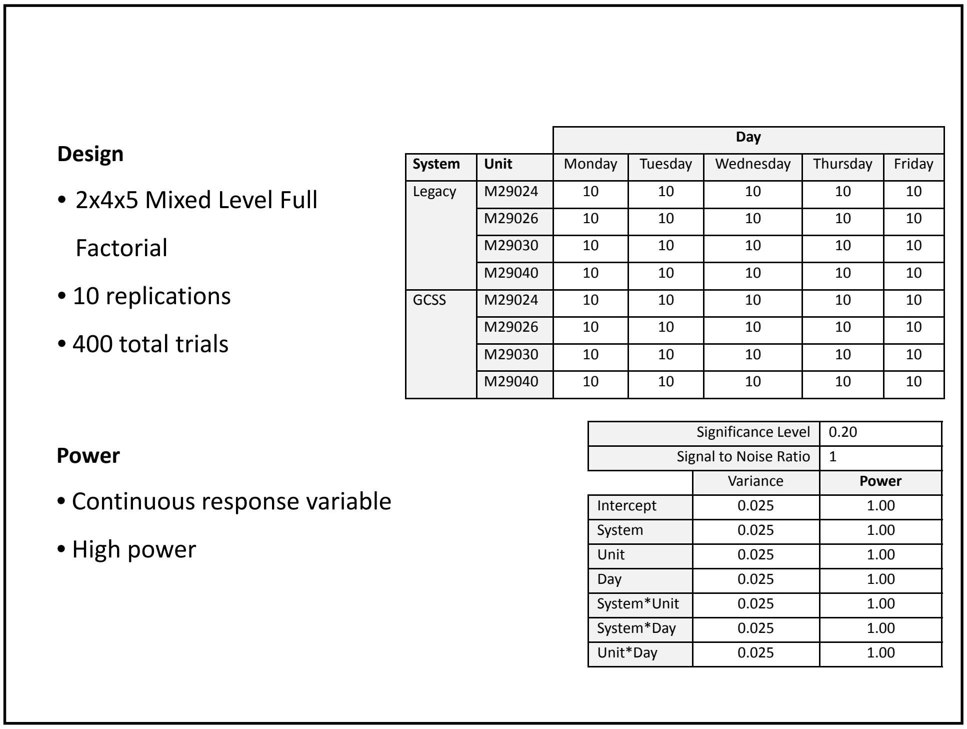
24 26 30 40

Day

Mon Tue Wed Thu Fri

Response

Time to Initial Supply Status



Design

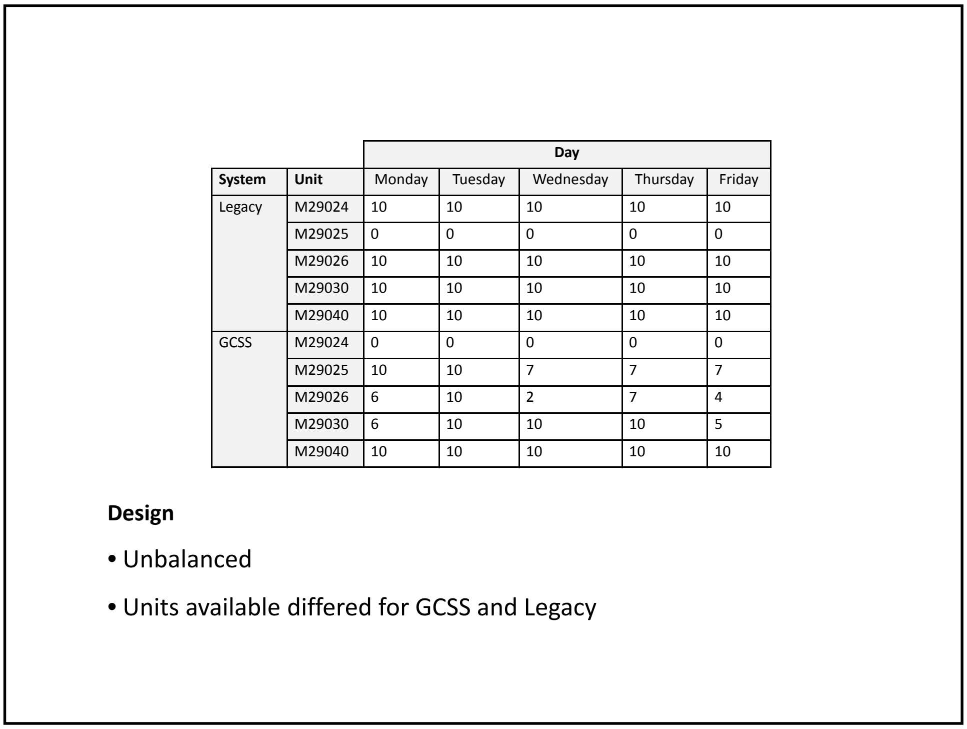
- 2x4x5 Mixed Level Full Factorial
- 10 replications
- 400 total trials

		Day				
System	**Unit**	Monday	Tuesday	Wednesday	Thursday	Friday
Legacy	M29024	10	10	10	10	10
	M29026	10	10	10	10	10
	M29030	10	10	10	10	10
	M29040	10	10	10	10	10
GCSS	M29024	10	10	10	10	10
	M29026	10	10	10	10	10
	M29030	10	10	10	10	10
	M29040	10	10	10	10	10

Power

- Continuous response variable
- High power

Significance Level	0.20	
Signal to Noise Ratio	1	
	Variance	**Power**
Intercept	0.025	1.00
System	0.025	1.00
Unit	0.025	1.00
Day	0.025	1.00
System*Unit	0.025	1.00
System*Day	0.025	1.00
Unit*Day	0.025	1.00



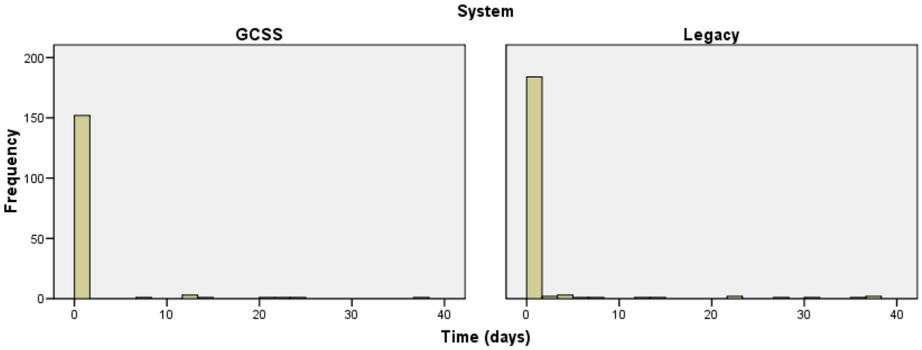
		Day				
System	**Unit**	Monday	Tuesday	Wednesday	Thursday	Friday
Legacy	M29024	10	10	10	10	10
	M29025	0	0	0	0	0
	M29026	10	10	10	10	10
	M29030	10	10	10	10	10
	M29040	10	10	10	10	10
GCSS	M29024	0	0	0	0	0
	M29025	10	10	7	7	7
	M29026	6	10	2	7	4
	M29030	6	10	10	10	5
	M29040	10	10	10	10	10

Design

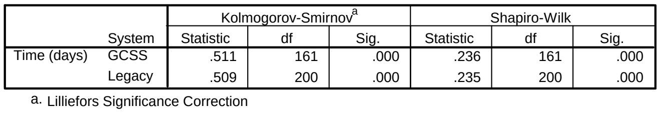
- Unbalanced
- Units available differed for GCSS and Legacy

Distribution

- Data does not conform to a normal distribution



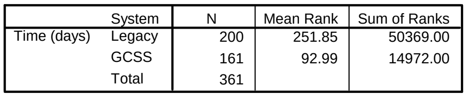
Tests of Normality



		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	System	Statistic	df	Sig.	Statistic	df	Sig.
Time (days)	GCSS	.511	161	.000	.236	161	.000
	Legacy	.509	200	.000	.235	200	.000

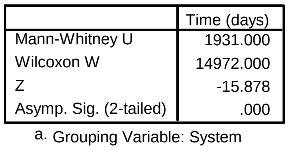
a. Lilliefors Significance Correction

Ranks



	System	N	Mean Rank	Sum of Ranks
Time (days)	Legacy	200	251.85	50369.00
	GCSS	161	92.99	14972.00
	Total	361		

Test Statistics[a]



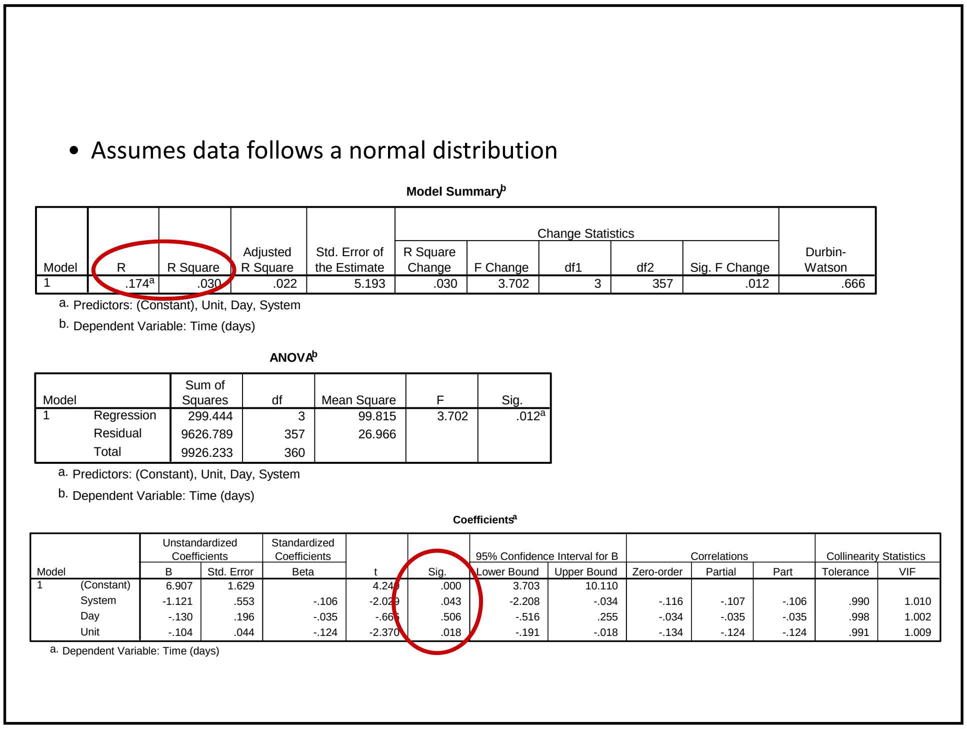
	Time (days)
Mann-Whitney U	1931.000
Wilcoxon W	14972.000
Z	-15.878
Asymp. Sig. (2-tailed)	.000

a. Grouping Variable: System

Effect Size

$$r = \frac{Z}{\sqrt{N}} = \frac{-15.878}{\sqrt{361}} = -0.836$$

- Mann-Whitney Test
  - 2-sample nonparametric test to compare means
  - Based on ranked data
- Since P-value < 0.00 at an α=0.2, the mean time to initial supply status was significantly shorter for GCSS than for the Legacy system, with an effect size of -0.836



- Assumes data follows a normal distribution

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	.174[a]	.030	.022	5.193	.030	3.702	3	357	.012	.666

a. Predictors: (Constant), Unit, Day, System

b. Dependent Variable: Time (days)

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	299.444	3	99.815	3.702	.012[a]
	Residual	9626.789	357	26.966		
	Total	9926.233	360			

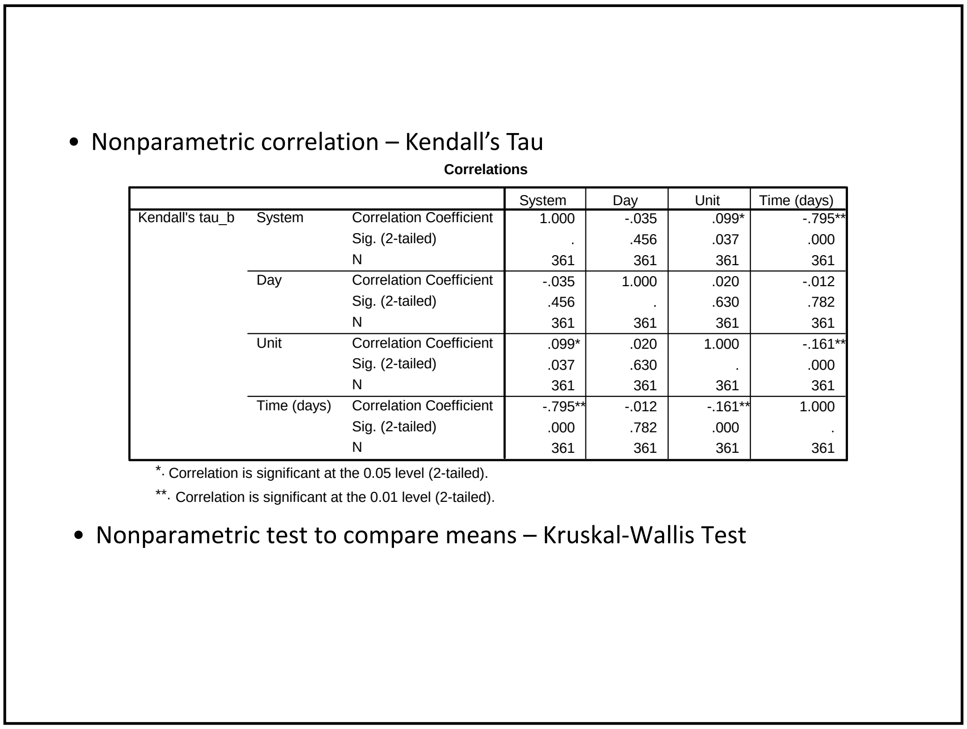
a. Predictors: (Constant), Unit, Day, System

b. Dependent Variable: Time (days)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B		Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	6.907	1.629		4.240	.000	3.703	10.110					
	System	-1.121	.553	-.106	-2.029	.043	-2.208	-.034	-.116	-.107	-.106	.990	1.010
	Day	-.130	.196	-.035	-.665	.506	-.516	.255	-.034	-.035	-.035	.998	1.002
	Unit	-.104	.044	-.124	-2.370	.018	-.191	-.018	-.134	-.124	-.124	.991	1.009

a. Dependent Variable: Time (days)



- Nonparametric correlation – Kendall's Tau

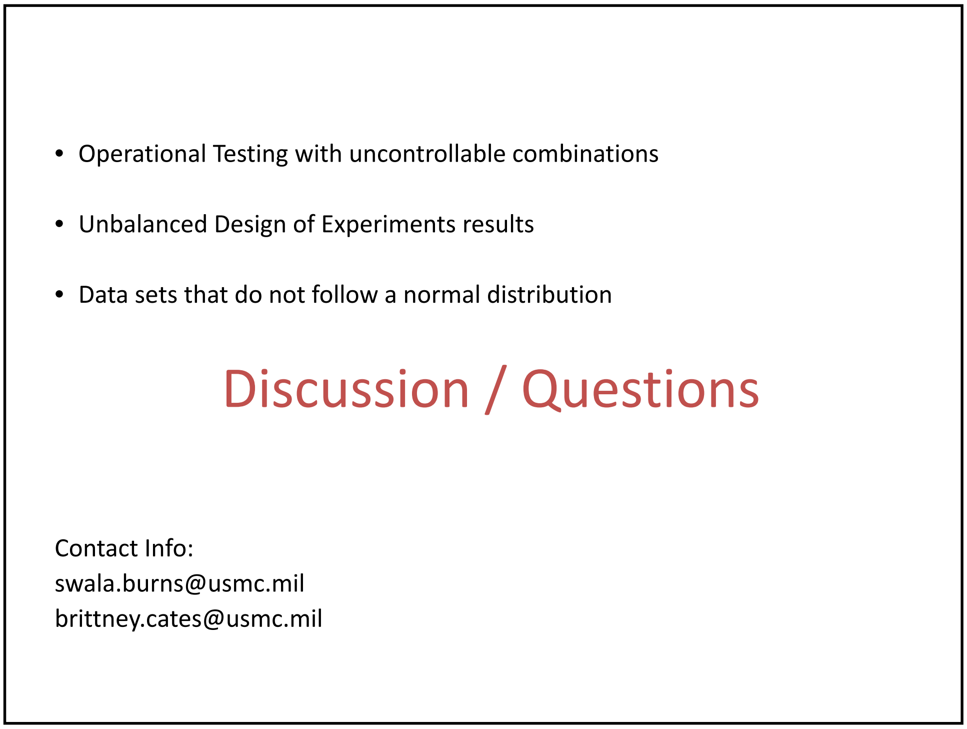
Correlations

			System	Day	Unit	Time (days)
Kendall's tau_b	System	Correlation Coefficient	1.000	-.035	.099*	-.795**
		Sig. (2-tailed)	.	.456	.037	.000
		N	361	361	361	361
	Day	Correlation Coefficient	-.035	1.000	.020	-.012
		Sig. (2-tailed)	.456	.	.630	.782
		N	361	361	361	361
	Unit	Correlation Coefficient	.099*	.020	1.000	-.161**
		Sig. (2-tailed)	.037	.630	.	.000
		N	361	361	361	361
	Time (days)	Correlation Coefficient	-.795**	-.012	-.161**	1.000
		Sig. (2-tailed)	.000	.782	.000	.
		N	361	361	361	361

*. Correlation is significant at the 0.05 level (2-tailed).

**. Correlation is significant at the 0.01 level (2-tailed).

- Nonparametric test to compare means – Kruskal-Wallis Test



- Operational Testing with uncontrollable combinations
- Unbalanced Design of Experiments results
- Data sets that do not follow a normal distribution

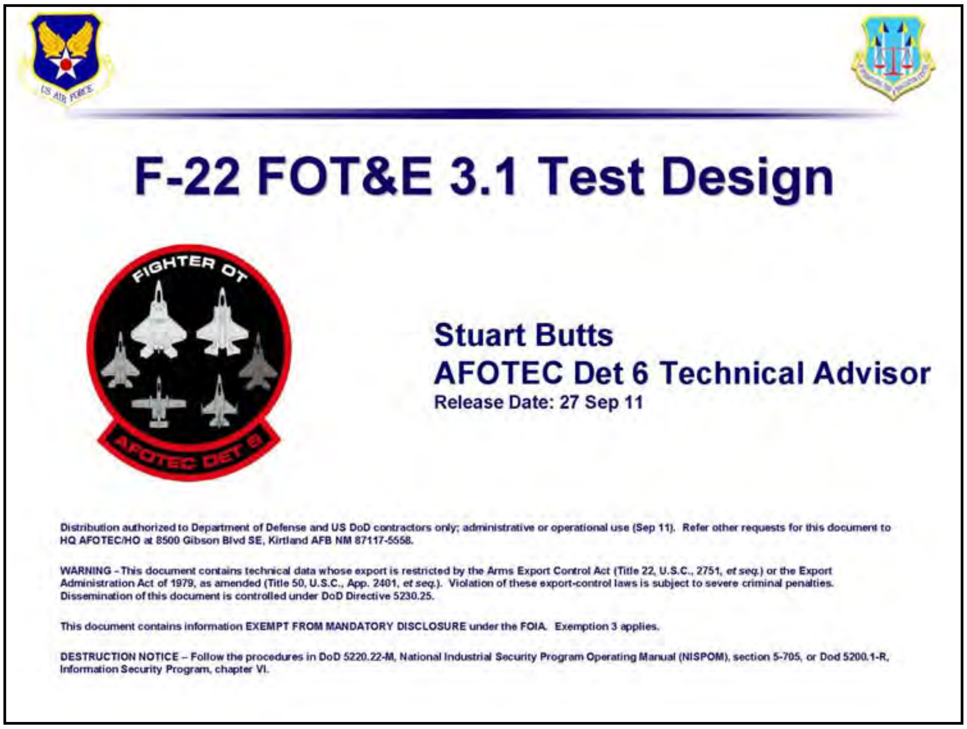
Discussion / Questions

Contact Info:
swala.burns@usmc.mil
brittney.cates@usmc.mil

Appendix 3-3.
F-22 FOT&E 3.1 Test Design

AFOTEC

AIR FORCE OPERATIONAL TEST AND EVALUATION CENTER


FIGHTER OT
AFOTEC DET 6

F-22 FOT&E 3.1 Test Design

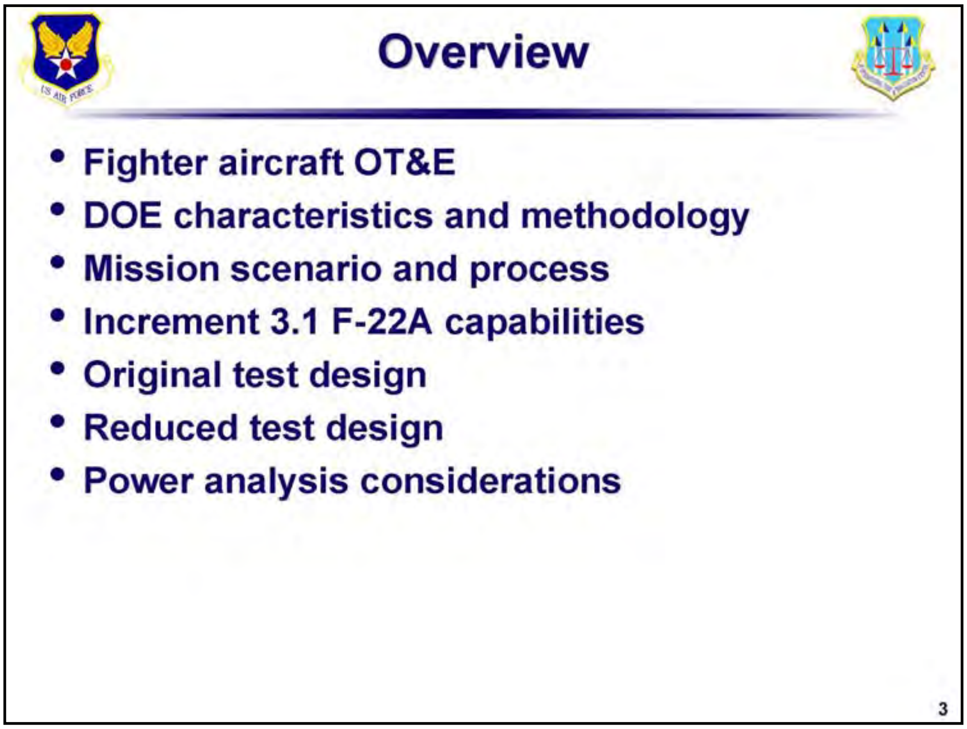
Stuart Butts
AFOTEC Det 6 Technical Advisor
Release Date: 27 Sep 11

Distribution authorized to Department of Defense and US DoD contractors only; administrative or operational use (Sep 11). Refer other requests for this document to HQ AFOTEC/HO at 8500 Gibson Blvd SE, Kirtland AFB NM 87117-5558.

WARNING - This document contains technical data whose export is restricted by the Arms Export Control Act (Title 22, U.S.C., 2751, *et seq.*) or the Export Administration Act of 1979, as amended (Title 50, U.S.C., App. 2401, *et seq.*). Violation of these export-control laws is subject to severe criminal penalties. Dissemination of this document is controlled under DoD Directive 5230.25.

This document contains information EXEMPT FROM MANDATORY DISCLOSURE under the FOIA. Exemption 3 applies.

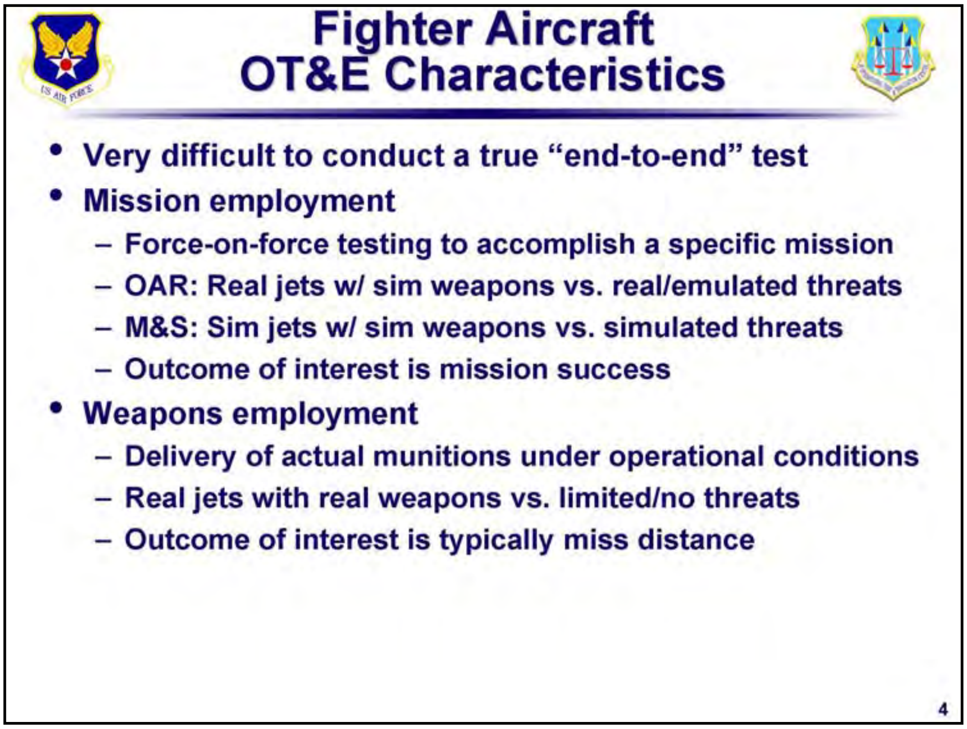
DESTRUCTION NOTICE – Follow the procedures in DoD 5220.22-M, National Industrial Security Program Operating Manual (NISPOM), section 5-705, or Dod 5200.1-R, Information Security Program, chapter VI.



Overview

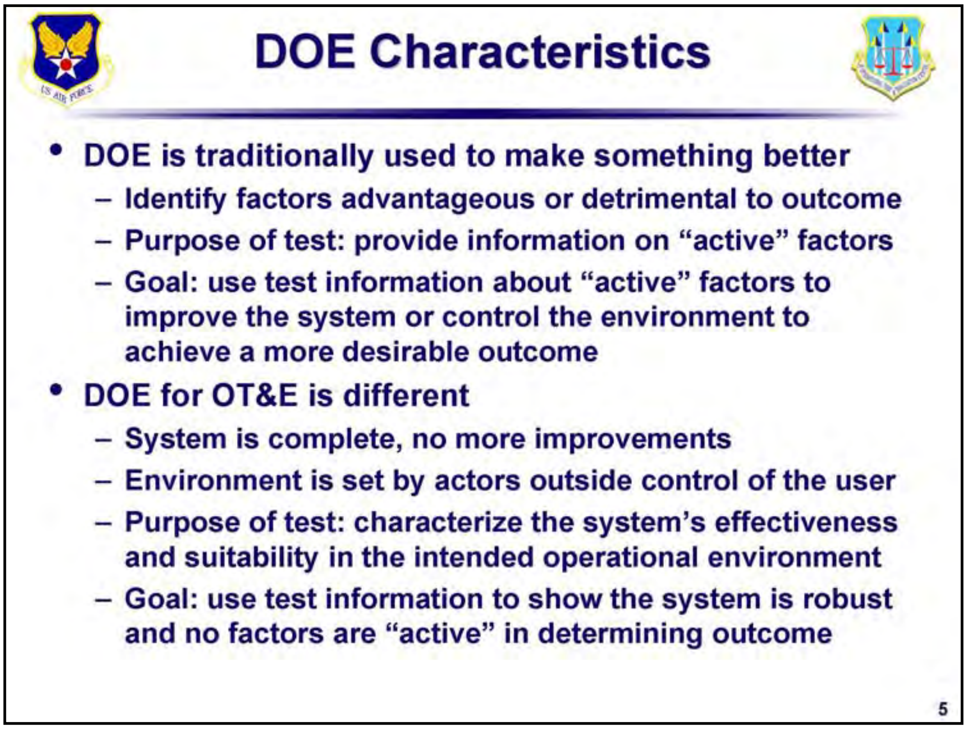
- Fighter aircraft OT&E
- DOE characteristics and methodology
- Mission scenario and process
- Increment 3.1 F-22A capabilities
- Original test design
- Reduced test design
- Power analysis considerations

3



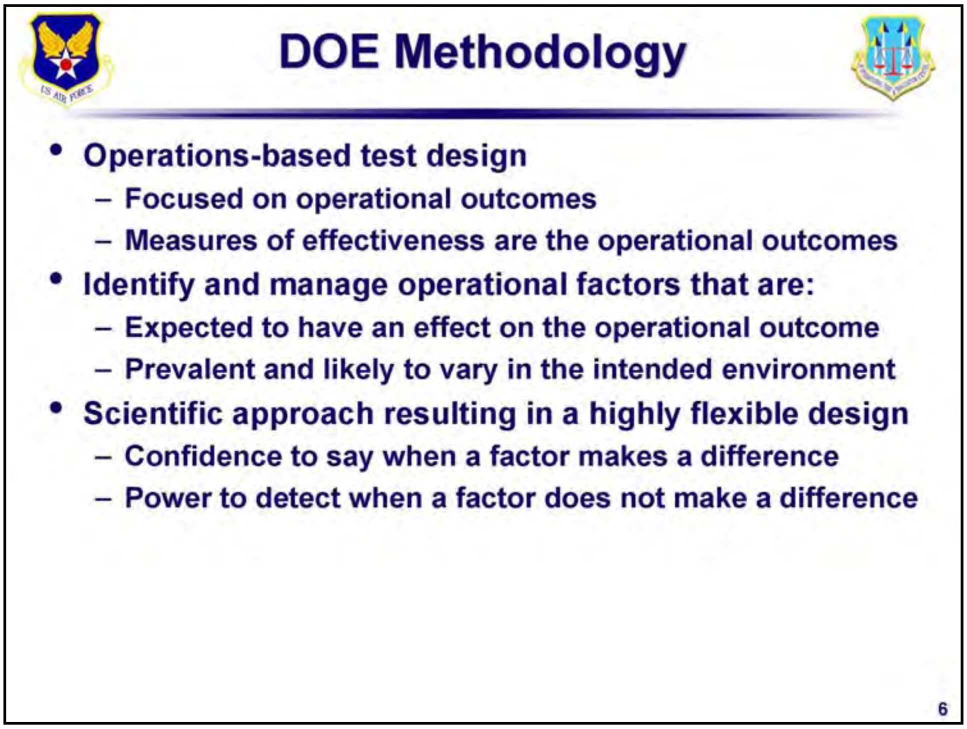
Fighter Aircraft OT&E Characteristics

- Very difficult to conduct a true “end-to-end” test
- Mission employment
  - Force-on-force testing to accomplish a specific mission
  - OAR: Real jets w/ sim weapons vs. real/emulated threats
  - M&S: Sim jets w/ sim weapons vs. simulated threats
  - Outcome of interest is mission success
- Weapons employment
  - Delivery of actual munitions under operational conditions
  - Real jets with real weapons vs. limited/no threats
  - Outcome of interest is typically miss distance



DOE Characteristics

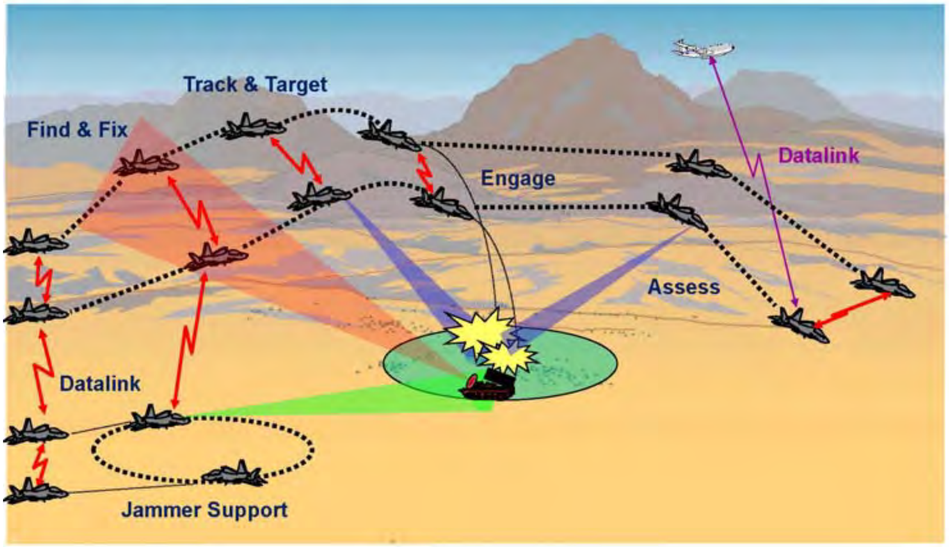
- DOE is traditionally used to make something better
  - Identify factors advantageous or detrimental to outcome
  - Purpose of test: provide information on “active” factors
  - Goal: use test information about “active” factors to improve the system or control the environment to achieve a more desirable outcome
- DOE for OT&E is different
  - System is complete, no more improvements
  - Environment is set by actors outside control of the user
  - Purpose of test: characterize the system’s effectiveness and suitability in the intended operational environment
  - Goal: use test information to show the system is robust and no factors are “active” in determining outcome

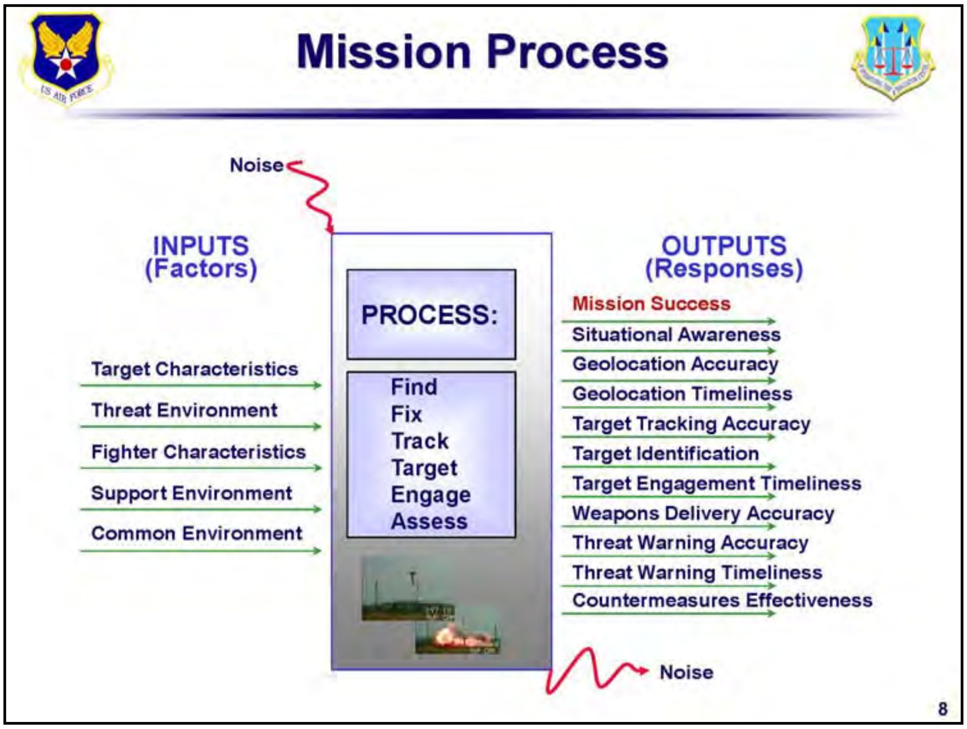


DOE Methodology

- Operations-based test design
  - Focused on operational outcomes
  - Measures of effectiveness are the operational outcomes
- Identify and manage operational factors that are:
  - Expected to have an effect on the operational outcome
  - Prevalent and likely to vary in the intended environment
- Scientific approach resulting in a highly flexible design
  - Confidence to say when a factor makes a difference
  - Power to detect when a factor does not make a difference

Mission Scenario





Mission Process

Noise

INPUTS
(Factors)

- Target Characteristics
- Threat Environment
- Fighter Characteristics
- Support Environment
- Common Environment

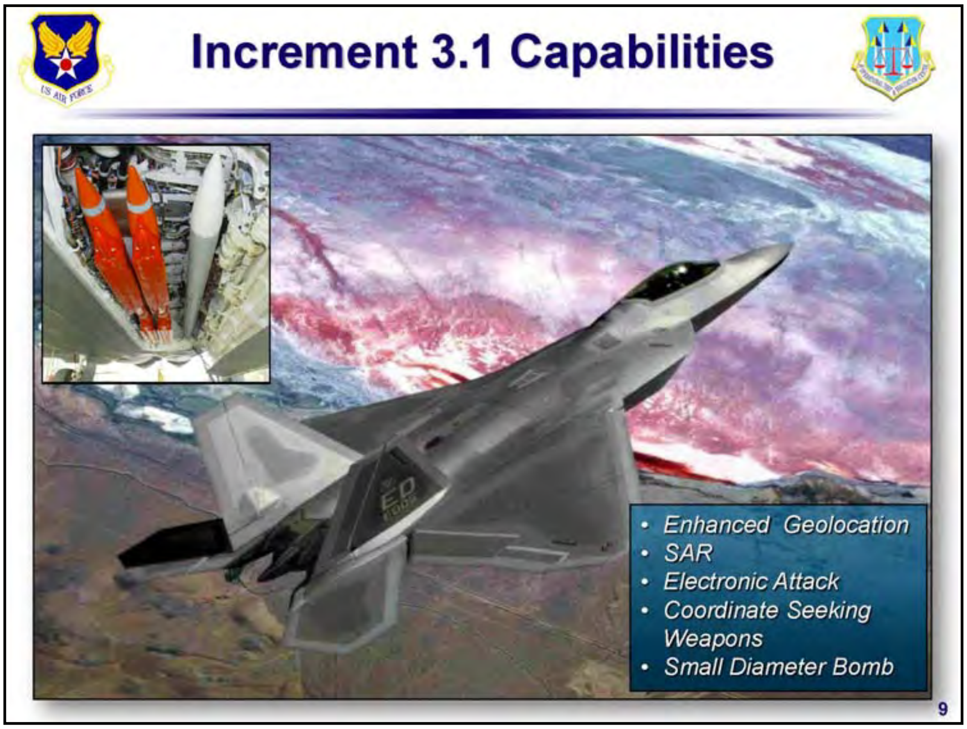
PROCESS:

Find
Fix
Track
Target
Engage
Assess

OUTPUTS
(Responses)

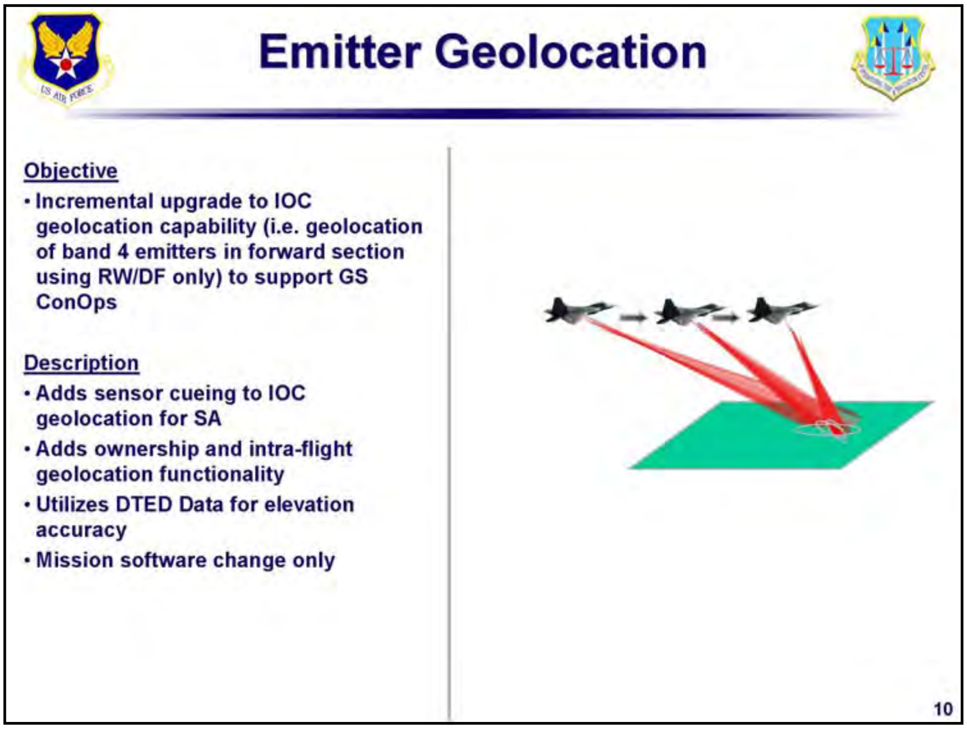
- Mission Success
- Situational Awareness
- Geolocation Accuracy
- Geolocation Timeliness
- Target Tracking Accuracy
- Target Identification
- Target Engagement Timeliness
- Weapons Delivery Accuracy
- Threat Warning Accuracy
- Threat Warning Timeliness
- Countermeasures Effectiveness

Noise



Increment 3.1 Capabilities

- *Enhanced Geolocation*
- *SAR*
- *Electronic Attack*
- *Coordinate Seeking Weapons*
- *Small Diameter Bomb*



Emitter Geolocation

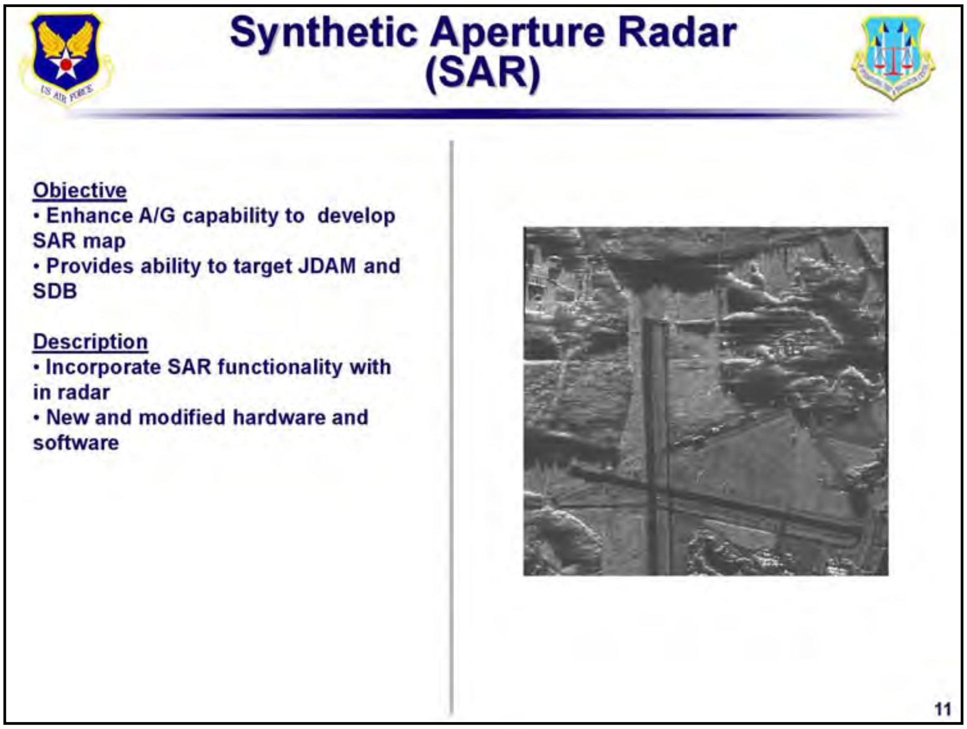
Objective

- Incremental upgrade to IOC geolocation capability (i.e. geolocation of band 4 emitters in forward section using RW/DF only) to support GS ConOps

Description

- Adds sensor cueing to IOC geolocation for SA
- Adds ownership and intra-flight geolocation functionality
- Utilizes DTED Data for elevation accuracy
- Mission software change only

10



Synthetic Aperture Radar (SAR)

Objective

• Enhance A/G capability to develop SAR map

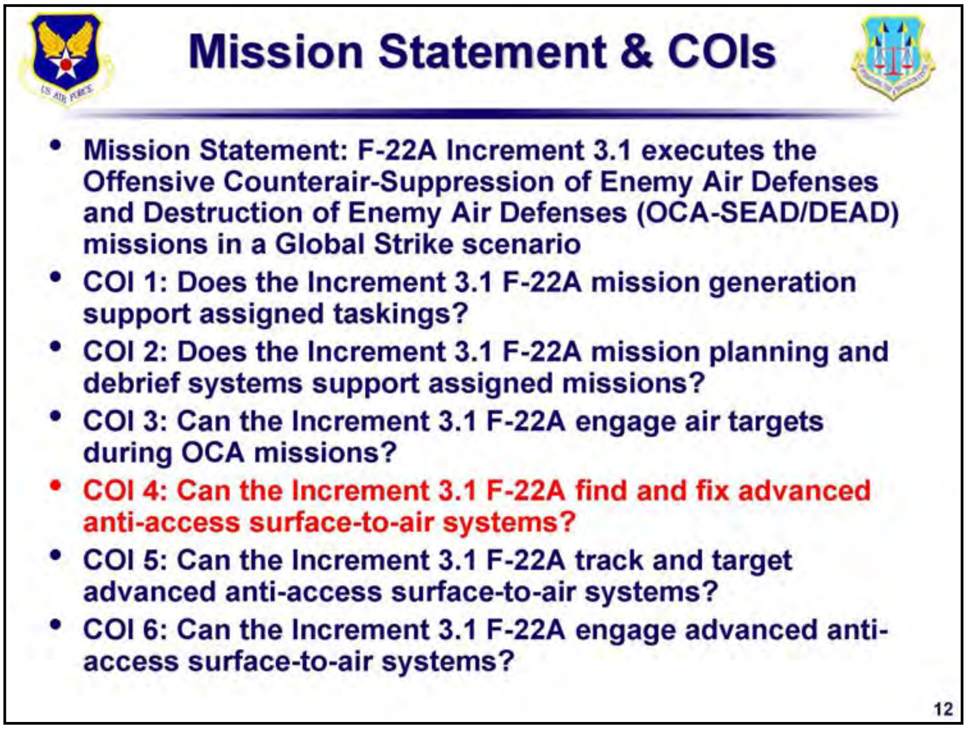
• Provides ability to target JDAM and SDB

Description

• Incorporate SAR functionality with in radar

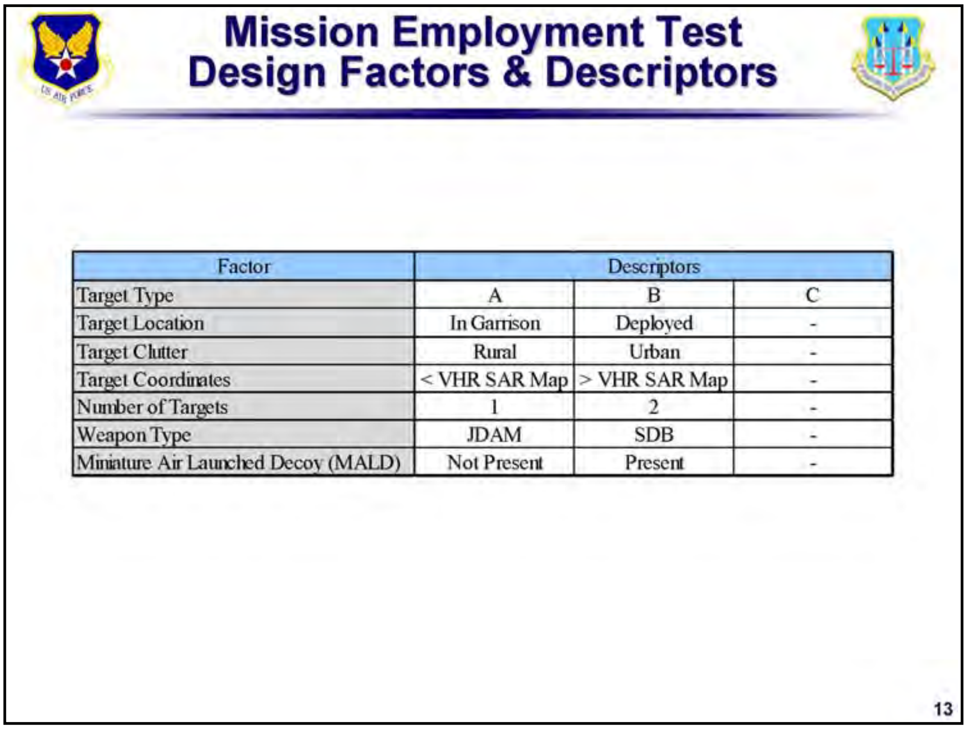
• New and modified hardware and software

11



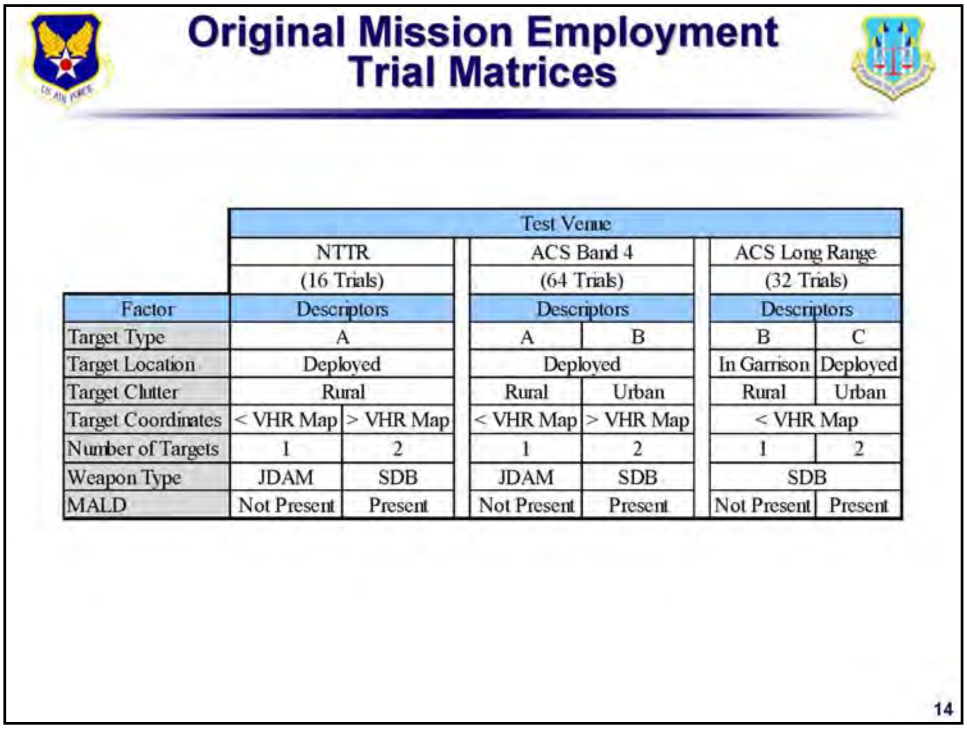
Mission Statement & COIs

- Mission Statement: F-22A Increment 3.1 executes the Offensive Counterair-Suppression of Enemy Air Defenses and Destruction of Enemy Air Defenses (OCA-SEAD/DEAD) missions in a Global Strike scenario
- COI 1: Does the Increment 3.1 F-22A mission generation support assigned taskings?
- COI 2: Does the Increment 3.1 F-22A mission planning and debrief systems support assigned missions?
- COI 3: Can the Increment 3.1 F-22A engage air targets during OCA missions?
- COI 4: Can the Increment 3.1 F-22A find and fix advanced anti-access surface-to-air systems?
- COI 5: Can the Increment 3.1 F-22A track and target advanced anti-access surface-to-air systems?
- COI 6: Can the Increment 3.1 F-22A engage advanced anti-access surface-to-air systems?



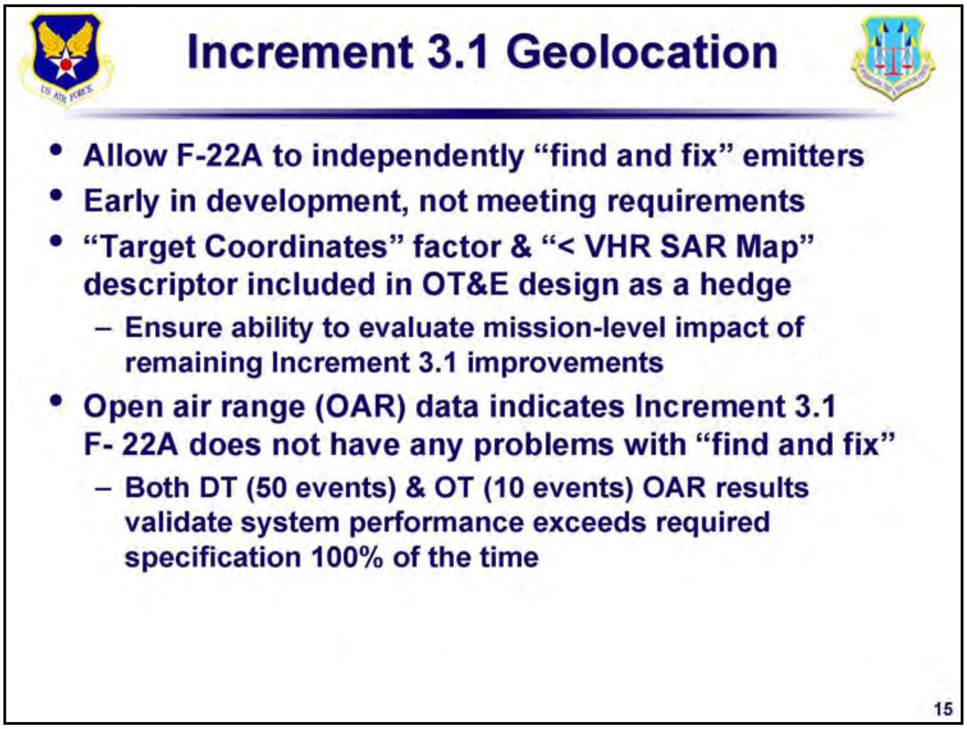
Mission Employment Test Design Factors & Descriptors

Factor	Descriptors		
Target Type	A	B	C
Target Location	In Garrison	Deployed	-
Target Clutter	Rural	Urban	-
Target Coordinates	< VHR SAR Map	> VHR SAR Map	-
Number of Targets	1	2	-
Weapon Type	JDAM	SDB	-
Miniature Air Launched Decoy (MALD)	Not Present	Present	-



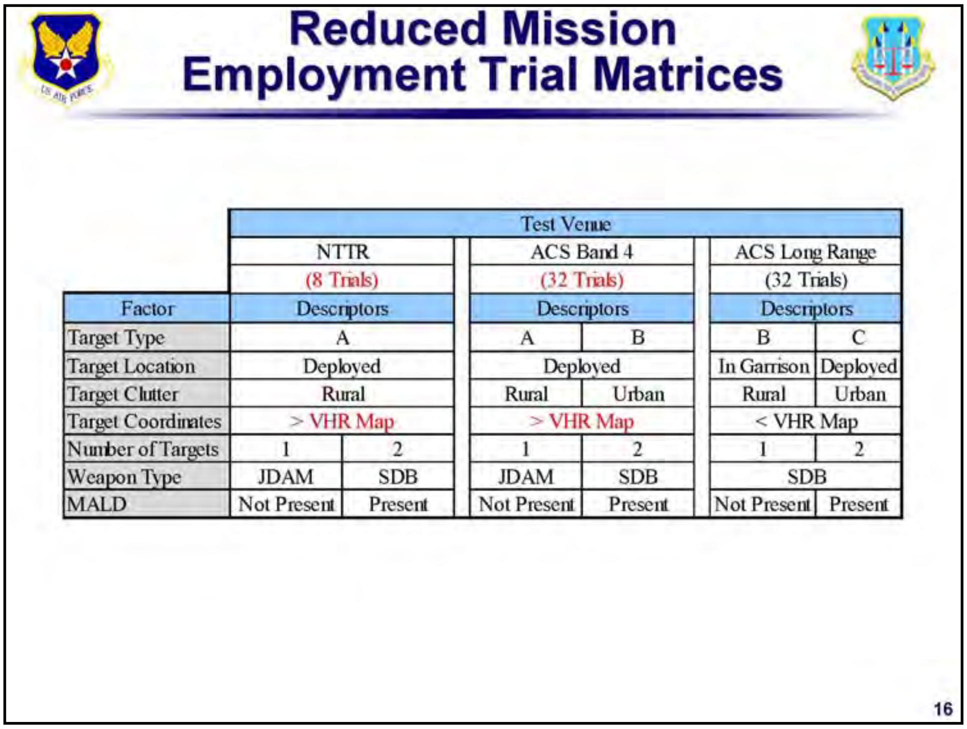
Original Mission Employment Trial Matrices

	Test Venue					
	NTTR		ACS Band 4		ACS Long Range	
	(16 Trials)		(64 Trials)		(32 Trials)	
Factor	Descriptors		Descriptors		Descriptors	
Target Type	A		A	B	B	C
Target Location	Deployed		Deployed		In Garrison	Deployed
Target Clutter	Rural		Rural	Urban	Rural	Urban
Target Coordinates	< VHR Map	> VHR Map	< VHR Map	> VHR Map	< VHR Map	
Number of Targets	1	2	1	2	1	2
Weapon Type	JDAM	SDB	JDAM	SDB	SDB	
MALD	Not Present	Present	Not Present	Present	Not Present	Present



Increment 3.1 Geolocation

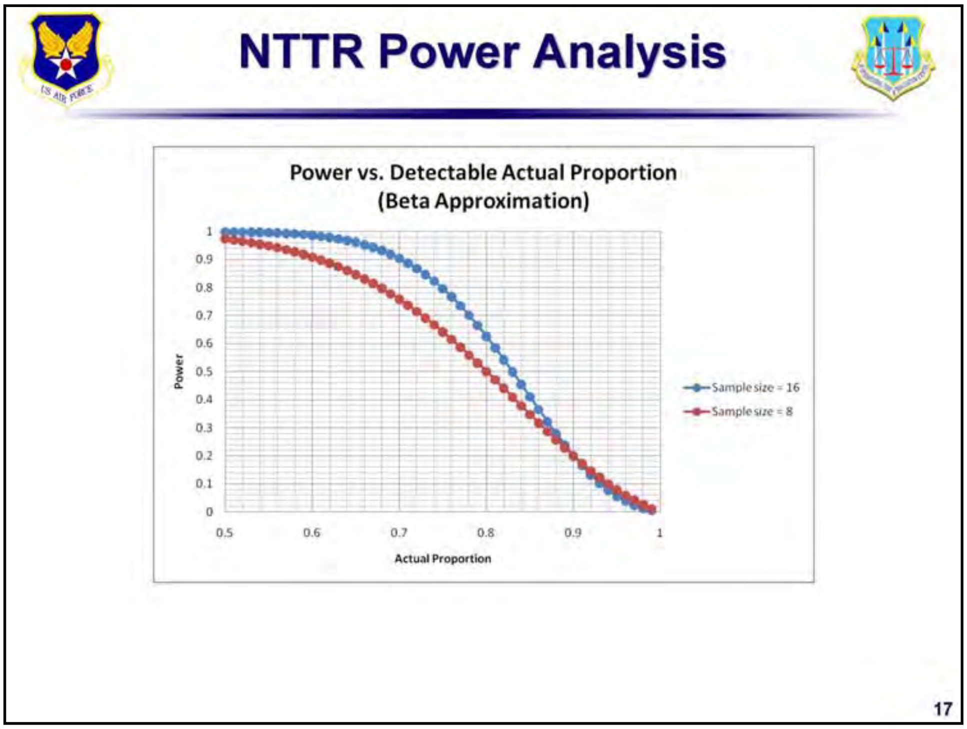
- Allow F-22A to independently "find and fix" emitters
- Early in development, not meeting requirements
- "Target Coordinates" factor & "< VHR SAR Map" descriptor included in OT&E design as a hedge
  - Ensure ability to evaluate mission-level impact of remaining Increment 3.1 improvements
- Open air range (OAR) data indicates Increment 3.1 F- 22A does not have any problems with "find and fix"
  - Both DT (50 events) & OT (10 events) OAR results validate system performance exceeds required specification 100% of the time



Reduced Mission Employment Trial Matrices

	Test Venue					
	NTTR		ACS Band 4		ACS Long Range	
	(8 Trials)		(32 Trials)		(32 Trials)	
Factor	Descriptors		Descriptors		Descriptors	
Target Type	A		A	B	B	C
Target Location	Deployed		Deployed		In Garrison	Deployed
Target Clutter	Rural		Rural	Urban	Rural	Urban
Target Coordinates	> VHR Map		> VHR Map		< VHR Map	
Number of Targets	1	2	1	2	1	2
Weapon Type	JDAM	SDB	JDAM	SDB	SDB	
MALD	Not Present	Present	Not Present	Present	Not Present	Present

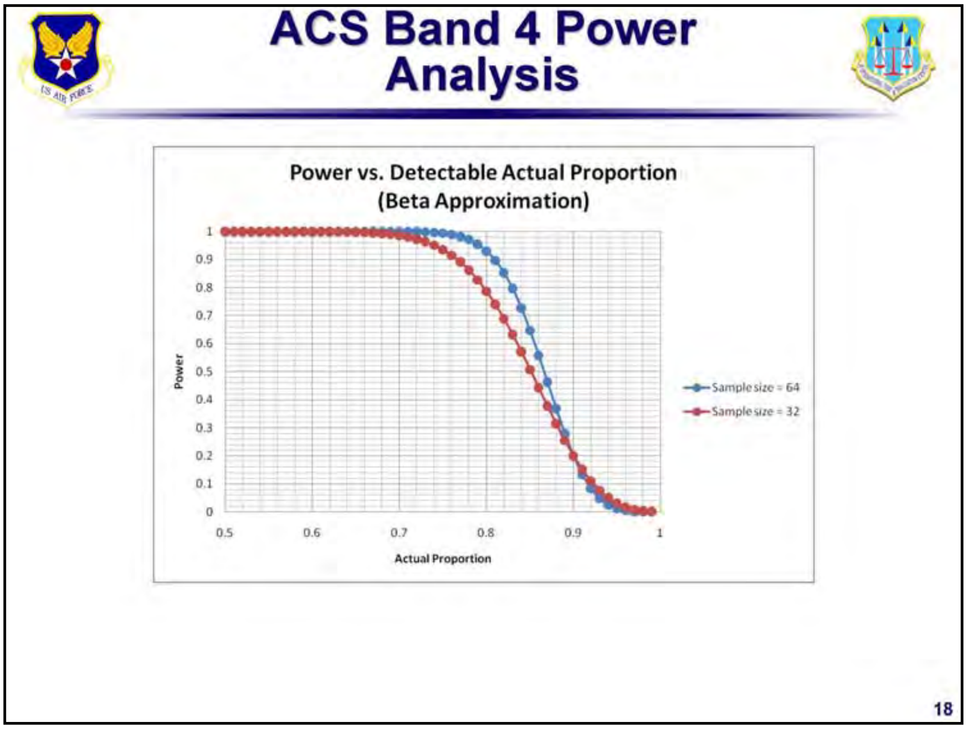
16



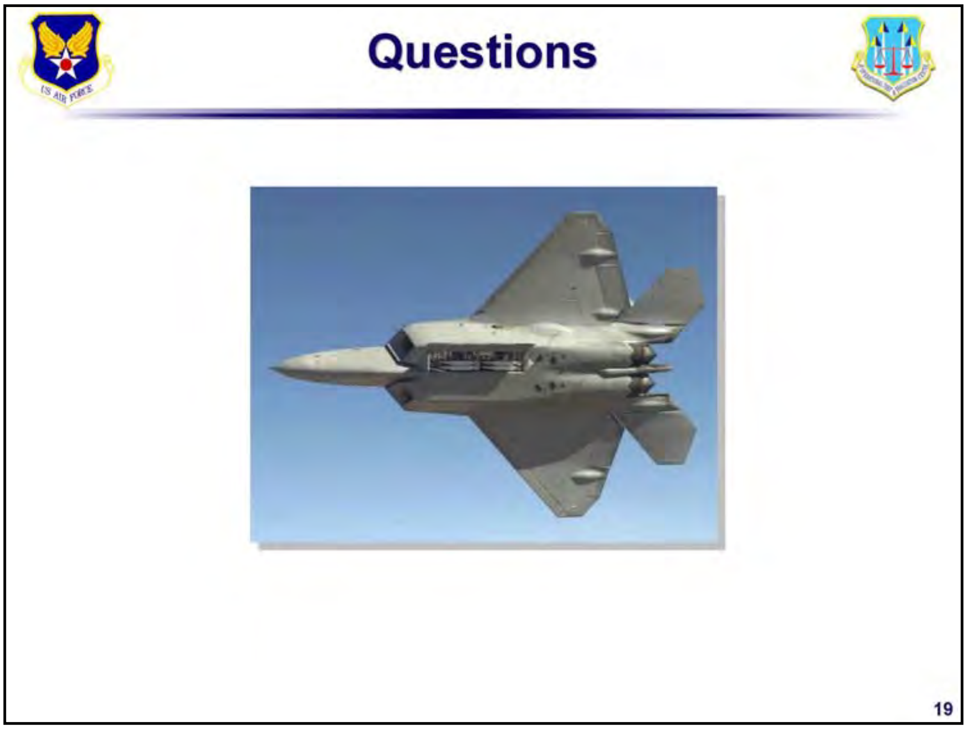
NTTR Power Analysis

Power vs. Detectable Actual Proportion (Beta Approximation)

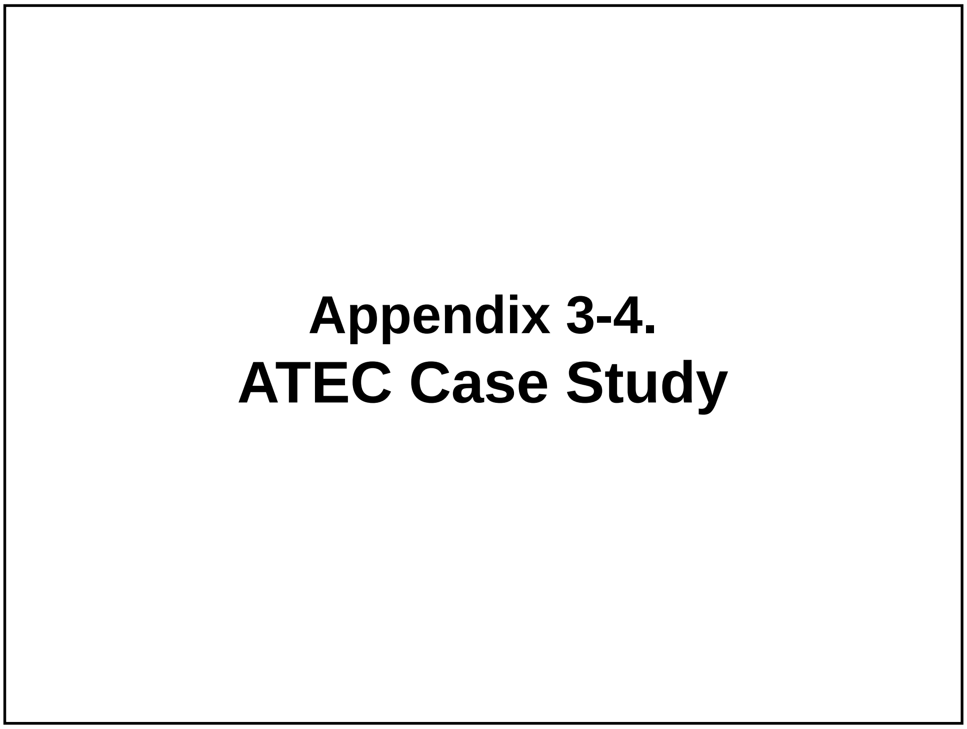
17



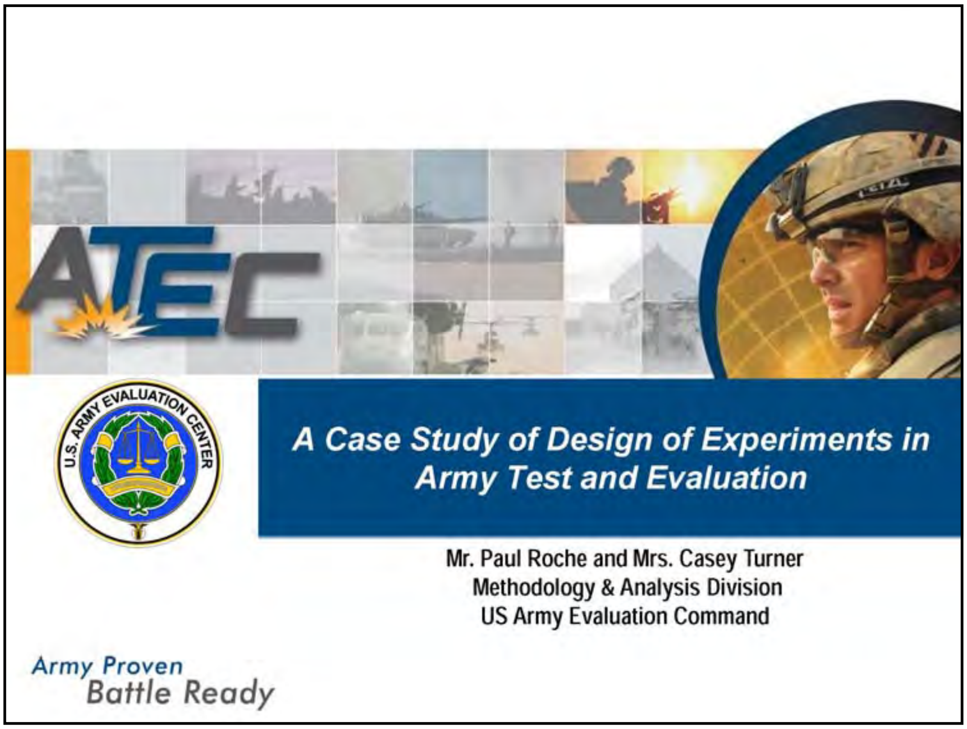
ACS Band 4 Power Analysis



Questions



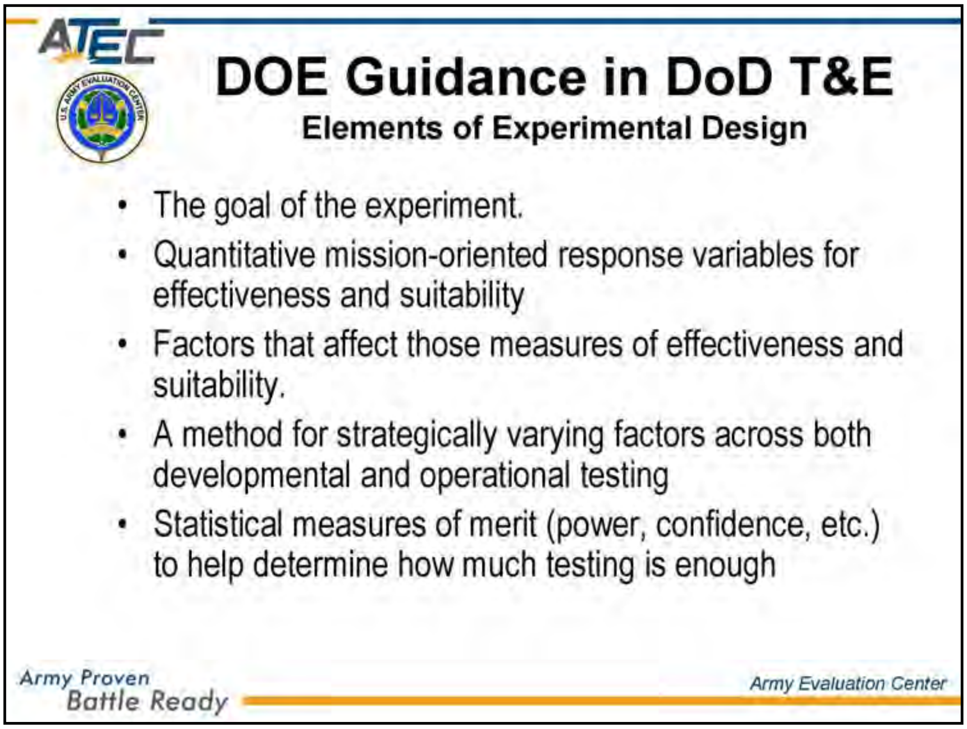
Appendix 3-4.
ATEC Case Study



A Case Study of Design of Experiments in Army Test and Evaluation

Mr. Paul Roche and Mrs. Casey Turner
Methodology & Analysis Division
US Army Evaluation Command

Army Proven
Battle Ready

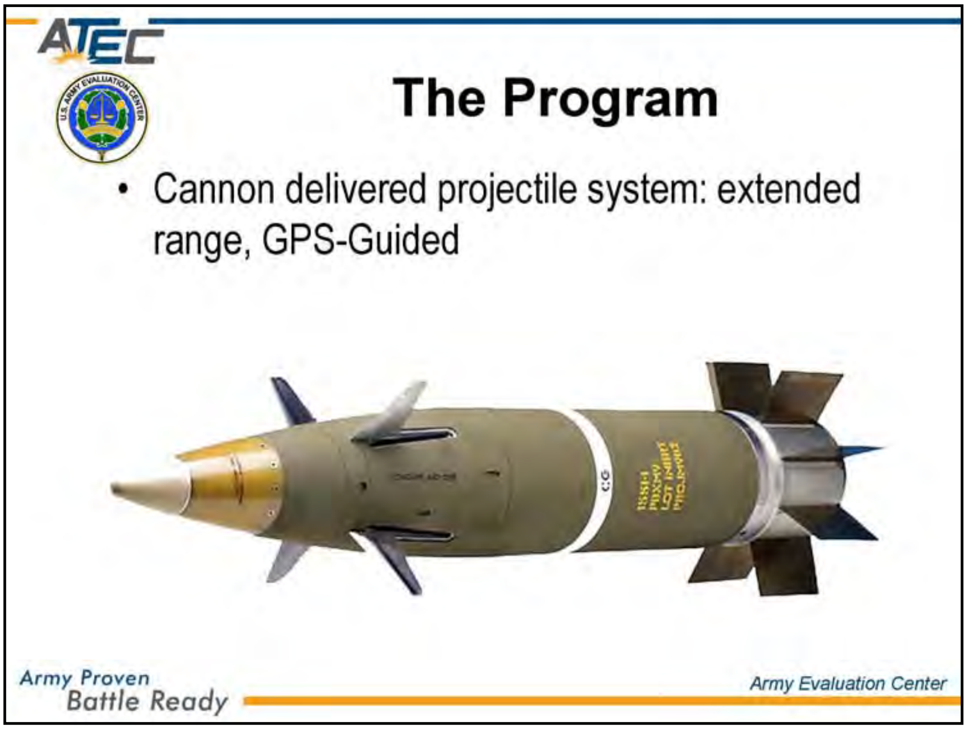


ATEC

DOE Guidance in DoD T&E

Elements of Experimental Design

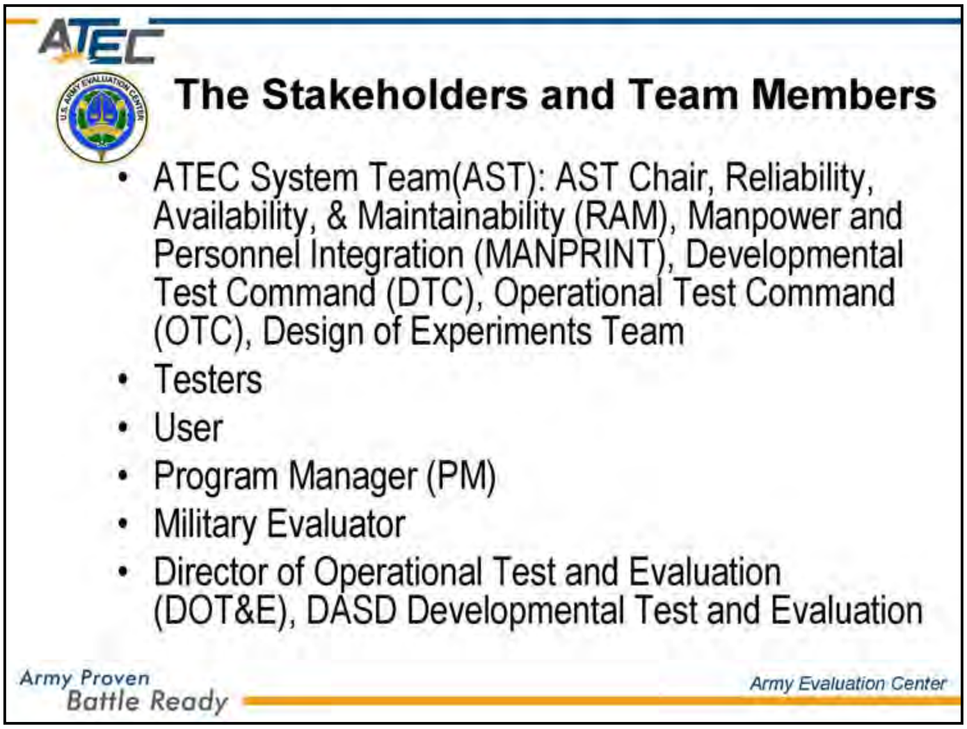
- The goal of the experiment.
- Quantitative mission-oriented response variables for effectiveness and suitability
- Factors that affect those measures of effectiveness and suitability.
- A method for strategically varying factors across both developmental and operational testing
- Statistical measures of merit (power, confidence, etc.) to help determine how much testing is enough



ATEC

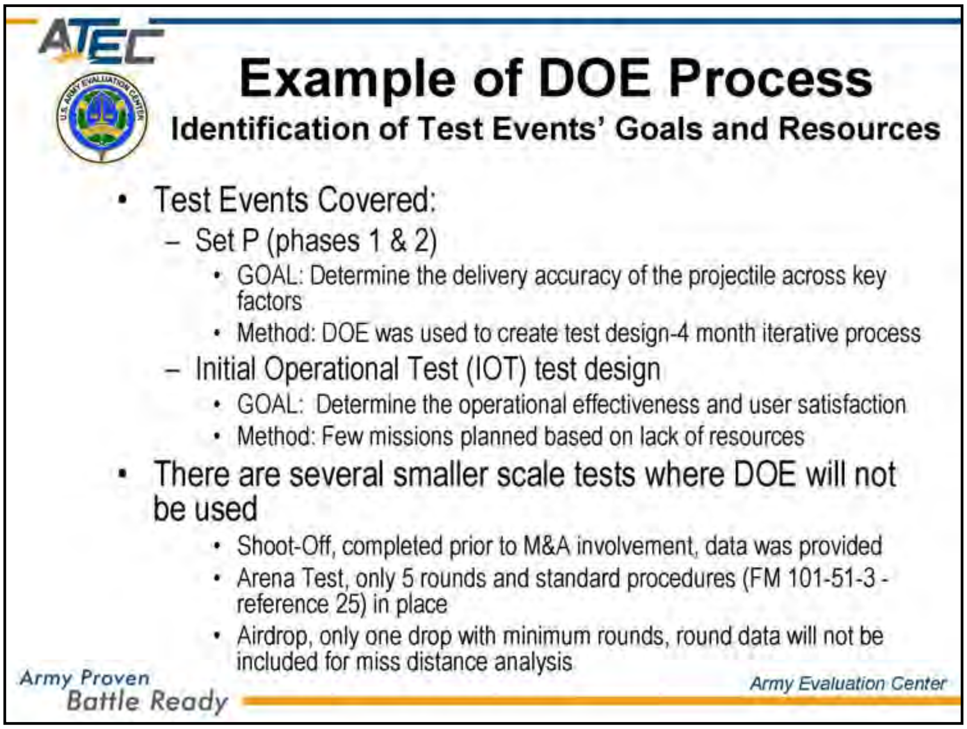
The Program

- Cannon delivered projectile system: extended range, GPS-Guided



The Stakeholders and Team Members

- ATEC System Team(AST): AST Chair, Reliability, Availability, & Maintainability (RAM), Manpower and Personnel Integration (MANPRINT), Developmental Test Command (DTC), Operational Test Command (OTC), Design of Experiments Team
- Testers
- User
- Program Manager (PM)
- Military Evaluator
- Director of Operational Test and Evaluation (DOT&E), DASD Developmental Test and Evaluation



Example of DOE Process

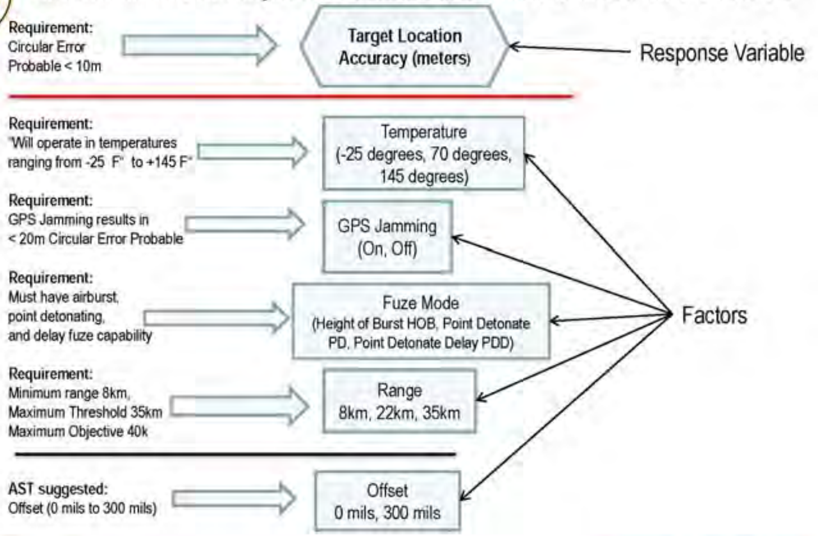
Identification of Test Events' Goals and Resources

- Test Events Covered:
  - Set P (phases 1 & 2)
    - GOAL: Determine the delivery accuracy of the projectile across key factors
    - Method: DOE was used to create test design-4 month iterative process
  - Initial Operational Test (IOT) test design
    - GOAL: Determine the operational effectiveness and user satisfaction
    - Method: Few missions planned based on lack of resources
- There are several smaller scale tests where DOE will not be used
    - Shoot-Off, completed prior to M&A involvement, data was provided
    - Arena Test, only 5 rounds and standard procedures (FM 101-51-3 - reference 25) in place
    - Airdrop, only one drop with minimum rounds, round data will not be included for miss distance analysis

ATEC

Example of DOE Process

Identification of response variable and initial factors for P1 and P2



Army Proven
Battle Ready

Army Evaluation Center

ATEC

Example of DOE Process

Initial Test Matrix Proposal for Set P

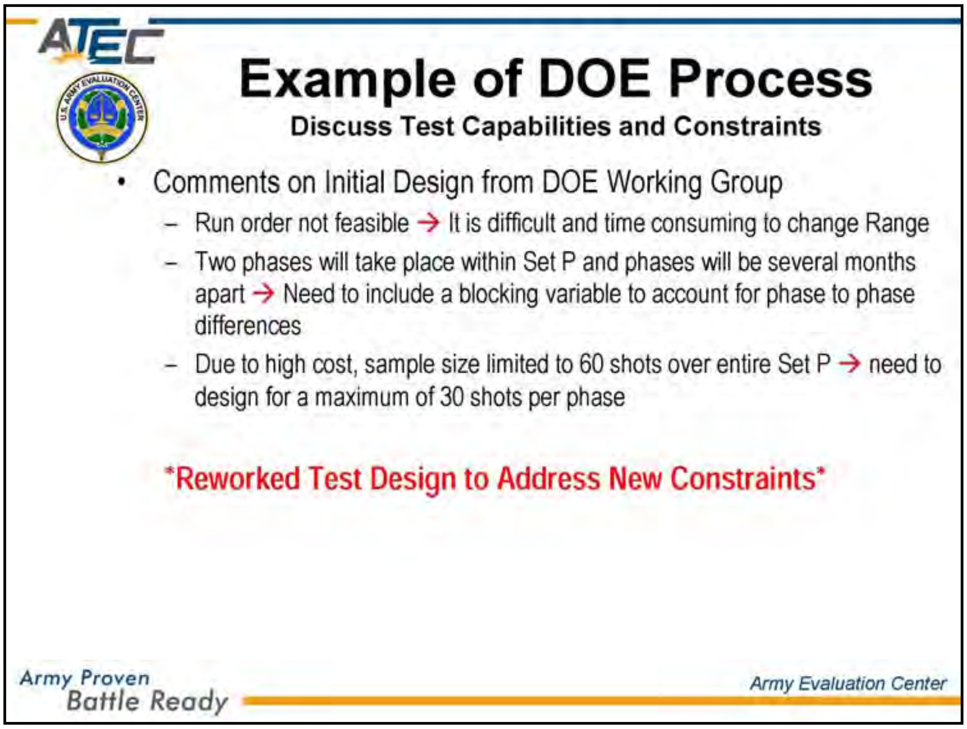
- Proposed three designs with varied sample sizes, estimation capabilities, and risks
  - 16 run main effects model
  - 31 run main effects and some two way interactions
  - 50 run main effects and all two way interactions

Balancing Cost, Schedule, and Risk

Ex Initial Test Matrix for 31 Shots

Range	GPS Jamming	Temperature	Fuze Modes	Offset
35km	On	70 degrees F	PD	0 mils
8km	On	-45 degrees F	HOB	0 mils
35km	Off	-45 degrees F	PDD	150 mils
8km	Off	70 degrees F	PDD	300 mils
35km	Off	-45 degrees F	PD	0 mils
35km	On	145 degrees F	HOB	300 mils
35km	On	70 degrees F	PDD	150 mils
8km	On	145 degrees F	PDD	0 mils
22km	On	-45 degrees F	PDD	300 mils
22km	Off	145 degrees F	PDD	0 mils
22km	On	145 degrees F	PD	150 mils
8km	Off	145 degrees F	PD	300 mils
8km	Off	145 degrees F	HOB	150 mils
22km	Off	70 degrees F	HOB	150 mils
35km	On	145 degrees F	HOB	300 mils
8km	On	-45 degrees F	PD	150 mils
8km	Off	-45 degrees F	HOB	0 mils
22km	Off	-45 degrees F	HOB	0 mils
35km	Off	-45 degrees F	PD	0 mils
22km	On	145 degrees F	PD	0 mils
8km	Off	145 degrees F	PDD	0 mils
35km	On	145 degrees F	PDD	0 mils
35km	Off	-45 degrees F	HOB	150 mils
22km	Off	70 degrees F	HOB	150 mils
35km	Off	70 degrees F	HOB	150 mils
35km	Off	70 degrees F	PD	150 mils
22km	Off	-45 degrees F	PDD	150 mils
22km	Off	-45 degrees F	PD	300 mils
22km	Off	70 degrees F	PDD	300 mils
35km	Off	70 degrees F	PDD	300 mils
35km	Off	145 degrees F	PDD	300 mils

Army Proven
Battle Ready



ATEC

Example of DOE Process

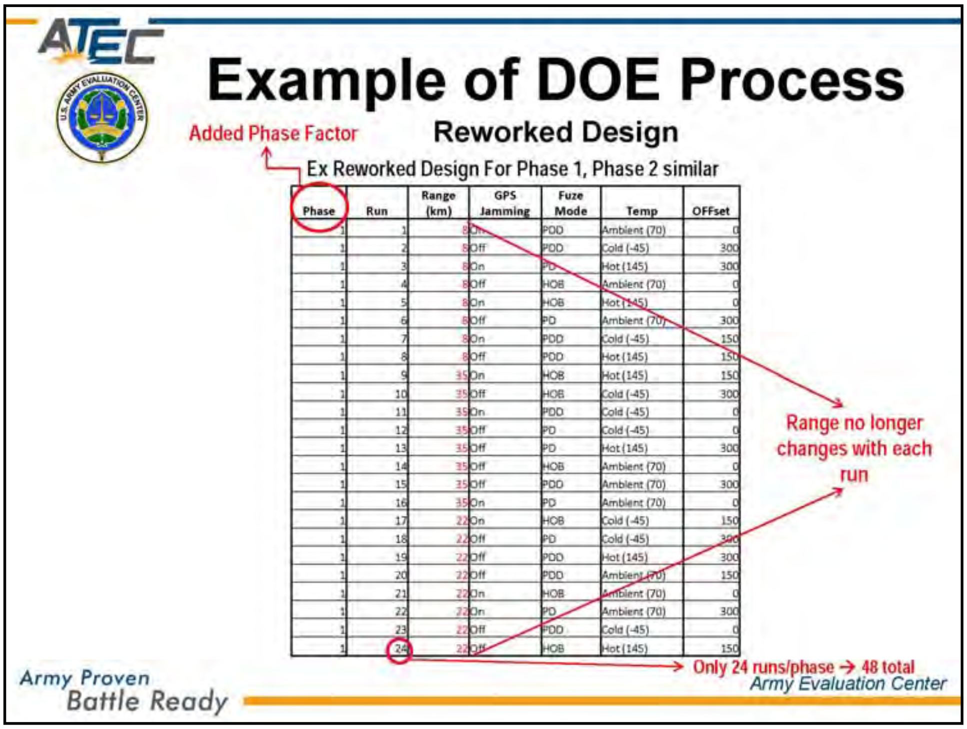
Discuss Test Capabilities and Constraints

- Comments on Initial Design from DOE Working Group
  - Run order not feasible → It is difficult and time consuming to change Range
  - Two phases will take place within Set P and phases will be several months apart → Need to include a blocking variable to account for phase to phase differences
  - Due to high cost, sample size limited to 60 shots over entire Set P → need to design for a maximum of 30 shots per phase

Reworked Test Design to Address New Constraints

Army Proven Battle Ready

Army Evaluation Center



ATEC

Example of DOE Process

Reworked Design

Added Phase Factor

Ex Reworked Design For Phase 1, Phase 2 similar

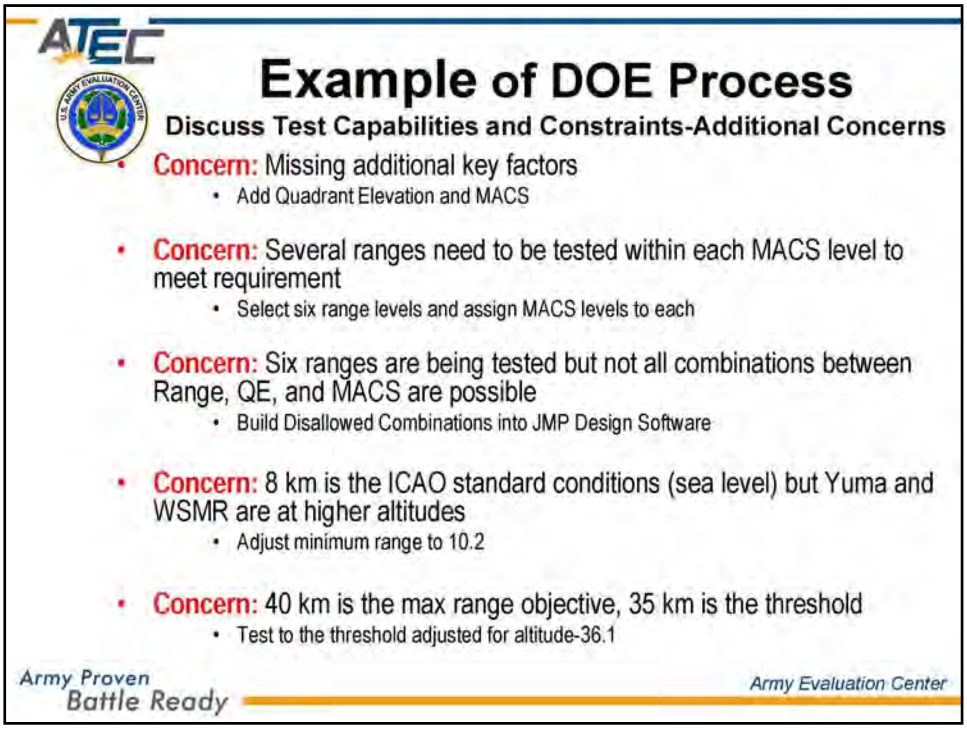
Phase	Run	Range (km)	GPS Jamming	Fuze Mode	Temp	OFFset
1	1	8	On	PDD	Ambient (70)	0
1	2	8	Off	PDD	Cold (-45)	300
1	3	8	On	PD	Hot (145)	300
1	4	8	Off	HOB	Ambient (70)	0
1	5	8	On	HOB	Hot (145)	0
1	6	8	Off	PD	Ambient (70)	300
1	7	8	On	PDD	Cold (-45)	150
1	8	8	Off	PDD	Hot (145)	150
1	9	35	On	HOB	Hot (145)	150
1	10	35	Off	HOB	Cold (-45)	300
1	11	35	On	PDD	Cold (-45)	0
1	12	35	Off	PD	Cold (-45)	0
1	13	35	Off	PD	Hot (145)	300
1	14	35	Off	HOB	Ambient (70)	0
1	15	35	Off	PDD	Ambient (70)	300
1	16	35	On	PD	Ambient (70)	0
1	17	22	On	HOB	Cold (-45)	150
1	18	22	Off	PD	Cold (-45)	300
1	19	22	Off	PDD	Hot (145)	300
1	20	22	Off	PDD	Ambient (70)	150
1	21	22	On	HOB	Ambient (70)	0
1	22	22	On	PD	Ambient (70)	300
1	23	22	Off	PDD	Cold (-45)	0
1	24	22	Off	HOB	Hot (145)	150

Range no longer changes with each run

Only 24 runs/phase → 48 total

Army Proven Battle Ready

Army Evaluation Center



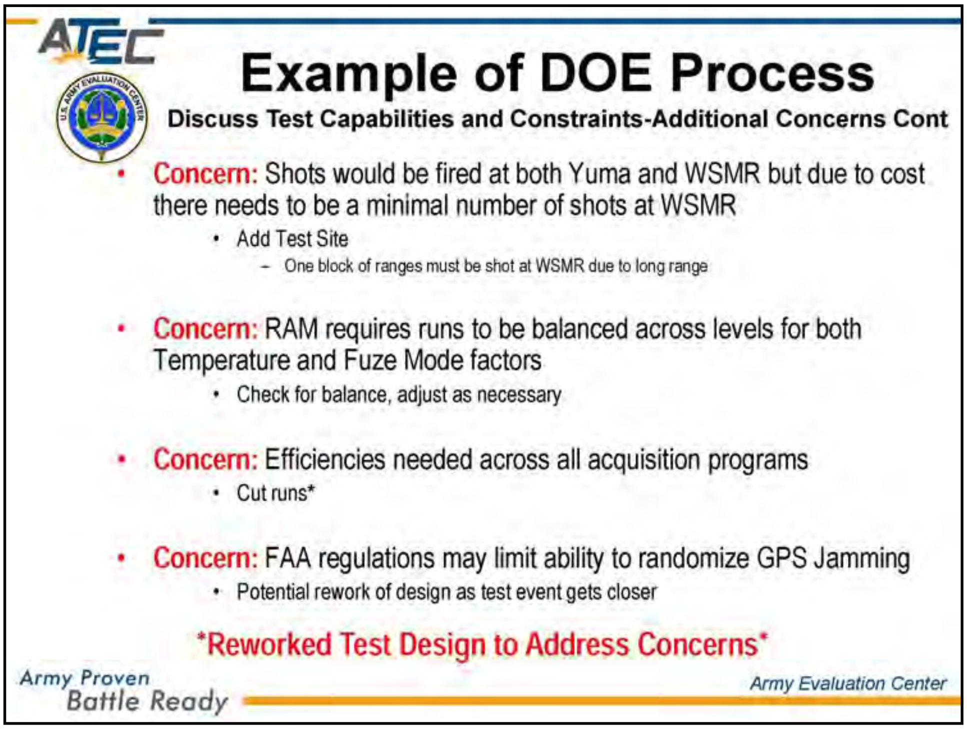
Example of DOE Process

Discuss Test Capabilities and Constraints-Additional Concerns

- Concern: Missing additional key factors
  - Add Quadrant Elevation and MACS
- Concern: Several ranges need to be tested within each MACS level to meet requirement
  - Select six range levels and assign MACS levels to each
- Concern: Six ranges are being tested but not all combinations between Range, QE, and MACS are possible
  - Build Disallowed Combinations into JMP Design Software
- Concern: 8 km is the ICAO standard conditions (sea level) but Yuma and WSMR are at higher altitudes
  - Adjust minimum range to 10.2
- Concern: 40 km is the max range objective, 35 km is the threshold
  - Test to the threshold adjusted for altitude-36.1

Army Proven
Battle Ready

Army Evaluation Center



Example of DOE Process

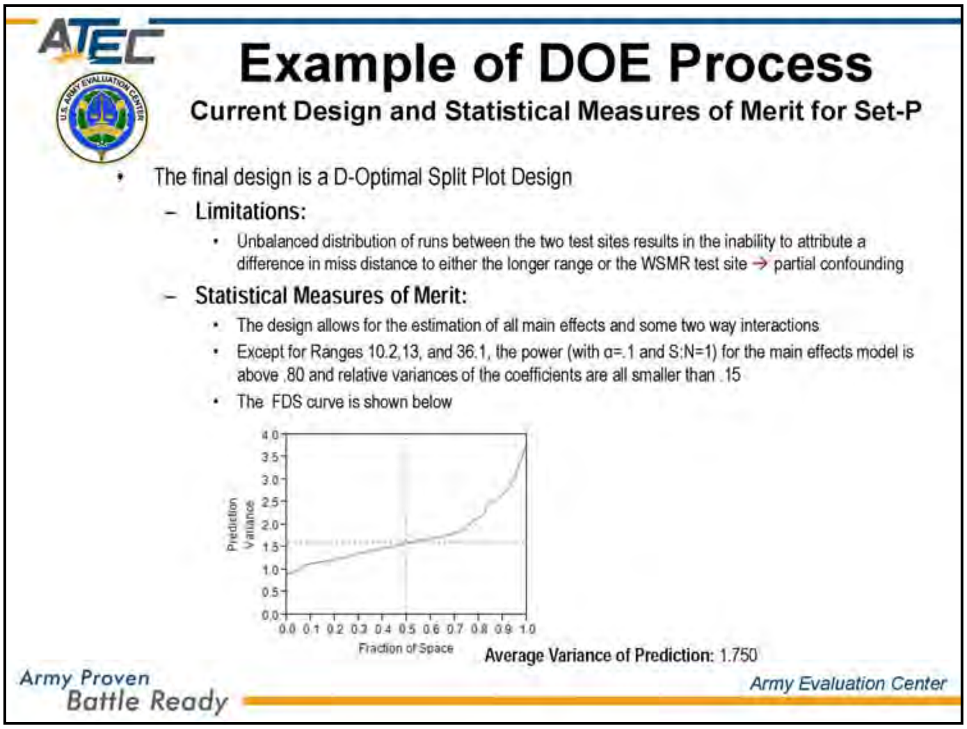
Discuss Test Capabilities and Constraints-Additional Concerns Cont

- Concern: Shots would be fired at both Yuma and WSMR but due to cost there needs to be a minimal number of shots at WSMR
  - Add Test Site
    - One block of ranges must be shot at WSMR due to long range
- Concern: RAM requires runs to be balanced across levels for both Temperature and Fuze Mode factors
  - Check for balance, adjust as necessary
- Concern: Efficiencies needed across all acquisition programs
  - Cut runs*
- Concern: FAA regulations may limit ability to randomize GPS Jamming
  - Potential rework of design as test event gets closer

Reworked Test Design to Address Concerns

Army Proven
Battle Ready

Army Evaluation Center



Example of DOE Process

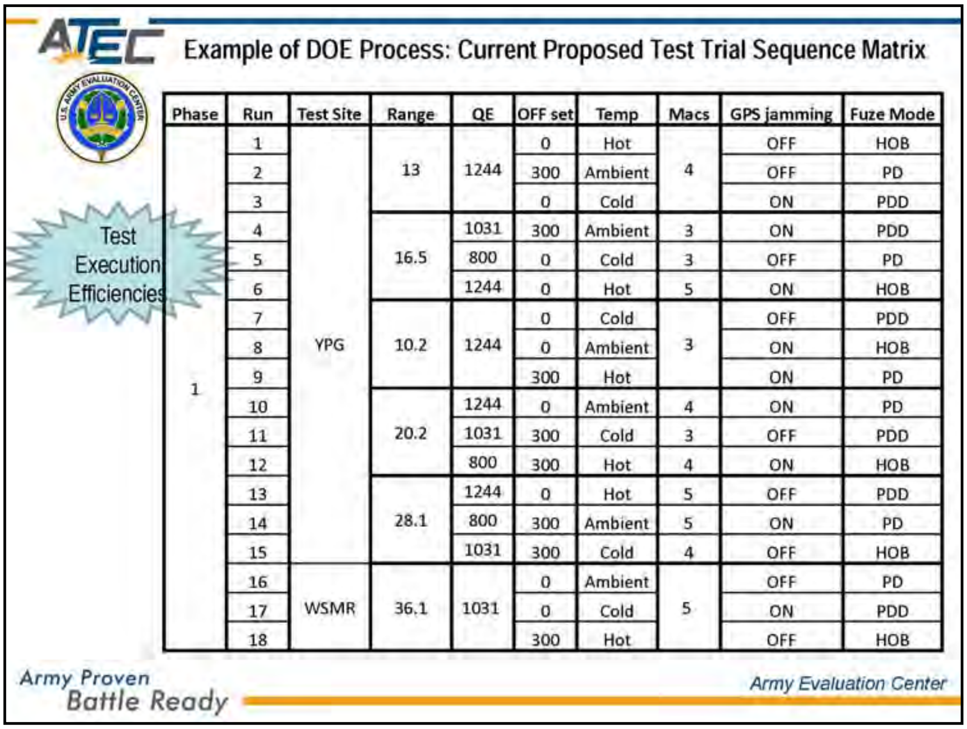
Current Design and Statistical Measures of Merit for Set-P

- The final design is a D-Optimal Split Plot Design
  - Limitations:
    - Unbalanced distribution of runs between the two test sites results in the inability to attribute a difference in miss distance to either the longer range or the WSMR test site → partial confounding
  - Statistical Measures of Merit:
    - The design allows for the estimation of all main effects and some two way interactions
    - Except for Ranges 10.2,13, and 36.1, the power (with α=.1 and S:N=1) for the main effects model is above .80 and relative variances of the coefficients are all smaller than .15
    - The FDS curve is shown below

Average Variance of Prediction: 1.750

Army Proven
Battle Ready

Army Evaluation Center



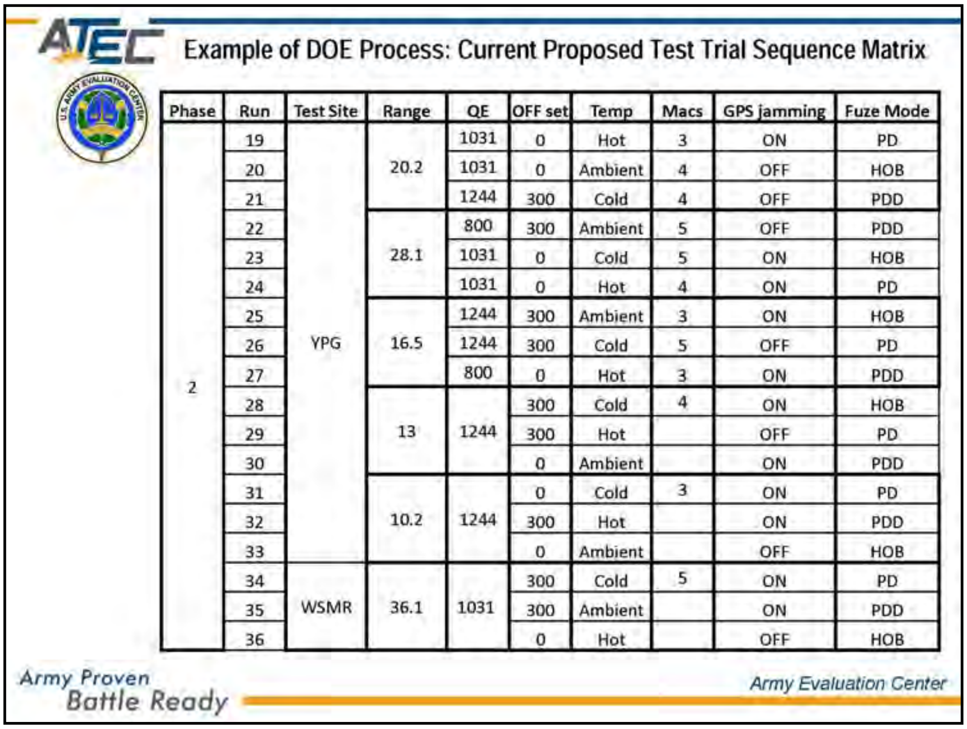
Example of DOE Process: Current Proposed Test Trial Sequence Matrix

Test Execution Efficiencies

Phase	Run	Test Site	Range	QE	OFF set	Temp	Macs	GPS jamming	Fuze Mode
1	1	YPG	13	1244	0	Hot	4	OFF	HOB
	2				300	Ambient		OFF	PD
	3				0	Cold		ON	PDD
	4		16.5	1031	300	Ambient	3	ON	PDD
	5			800	0	Cold	3	OFF	PD
	6			1244	0	Hot	5	ON	HOB
	7		10.2	1244	0	Cold	3	OFF	PDD
	8				0	Ambient		ON	HOB
	9				300	Hot		ON	PD
	10		20.2	1244	0	Ambient	4	ON	PD
	11			1031	300	Cold	3	OFF	PDD
	12			800	300	Hot	4	ON	HOB
	13		28.1	1244	0	Hot	5	OFF	PDD
	14			800	300	Ambient	5	ON	PD
	15			1031	300	Cold	4	OFF	HOB
	16	WSMR	36.1	1031	0	Ambient	5	OFF	PD
	17				0	Cold		ON	PDD
	18				300	Hot		OFF	HOB

Army Proven
Battle Ready

Army Evaluation Center

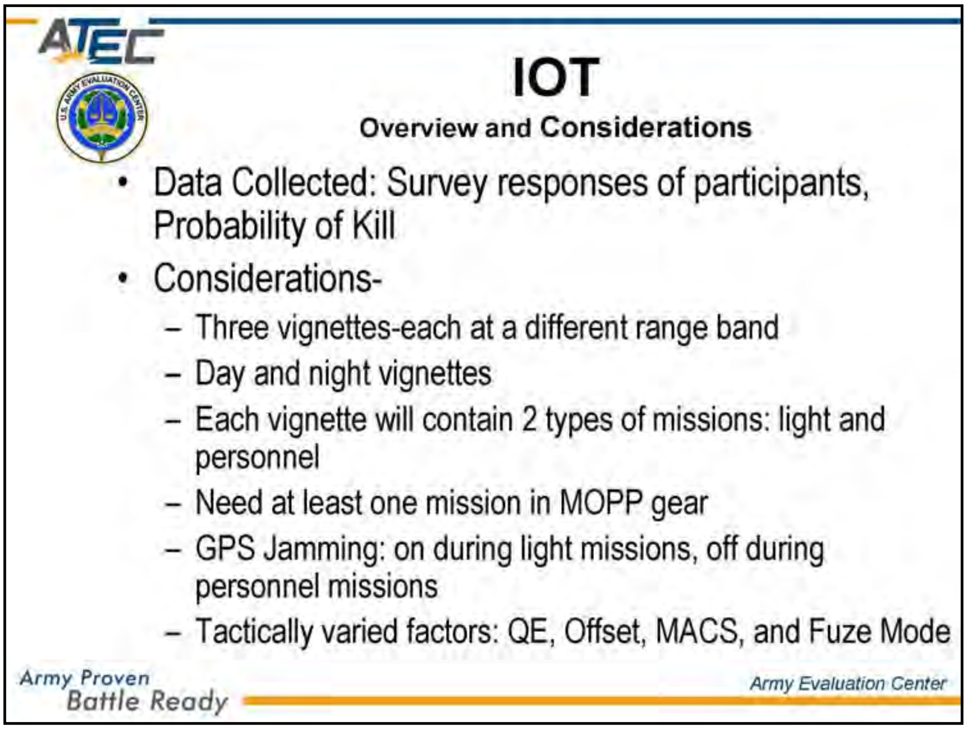


Example of DOE Process: Current Proposed Test Trial Sequence Matrix

Phase	Run	Test Site	Range	QE	OFF set	Temp	Macs	GPS jamming	Fuze Mode
2	19	YPG	20.2	1031	0	Hot	3	ON	PD
	20			1031	0	Ambient	4	OFF	HOB
	21			1244	300	Cold	4	OFF	PDD
	22		28.1	800	300	Ambient	5	OFF	PDD
	23			1031	0	Cold	5	ON	HOB
	24			1031	0	Hot	4	ON	PD
	25		16.5	1244	300	Ambient	3	ON	HOB
	26			1244	300	Cold	5	OFF	PD
	27			800	0	Hot	3	ON	PDD
	28		13	1244	300	Cold	4	ON	HOB
	29				300	Hot		OFF	PD
	30				0	Ambient		ON	PDD
	31		10.2	1244	0	Cold	3	ON	PD
	32				300	Hot		ON	PDD
	33				0	Ambient		OFF	HOB
	34	WSMR	36.1	1031	300	Cold	5	ON	PD
	35				300	Ambient		ON	PDD
	36				0	Hot		OFF	HOB

Army Proven
Battle Ready

Army Evaluation Center



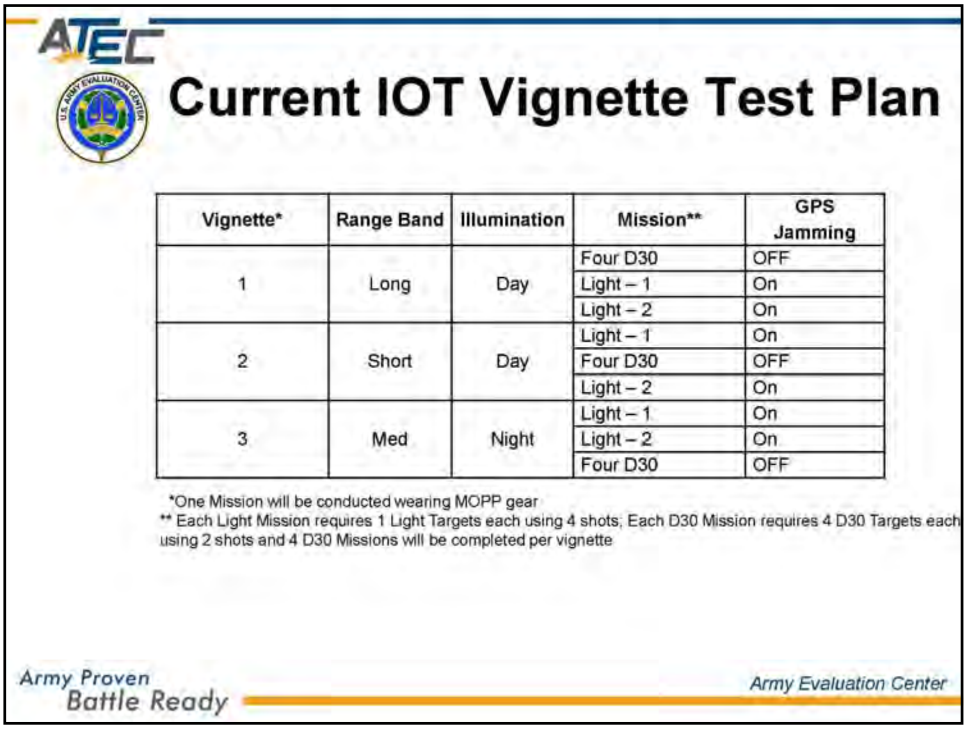
IOT

Overview and Considerations

- Data Collected: Survey responses of participants, Probability of Kill
- Considerations-
  - Three vignettes-each at a different range band
  - Day and night vignettes
  - Each vignette will contain 2 types of missions: light and personnel
  - Need at least one mission in MOPP gear
  - GPS Jamming: on during light missions, off during personnel missions
  - Tactically varied factors: QE, Offset, MACS, and Fuze Mode

Army Proven
Battle Ready

Army Evaluation Center



ATEC

Current IOT Vignette Test Plan

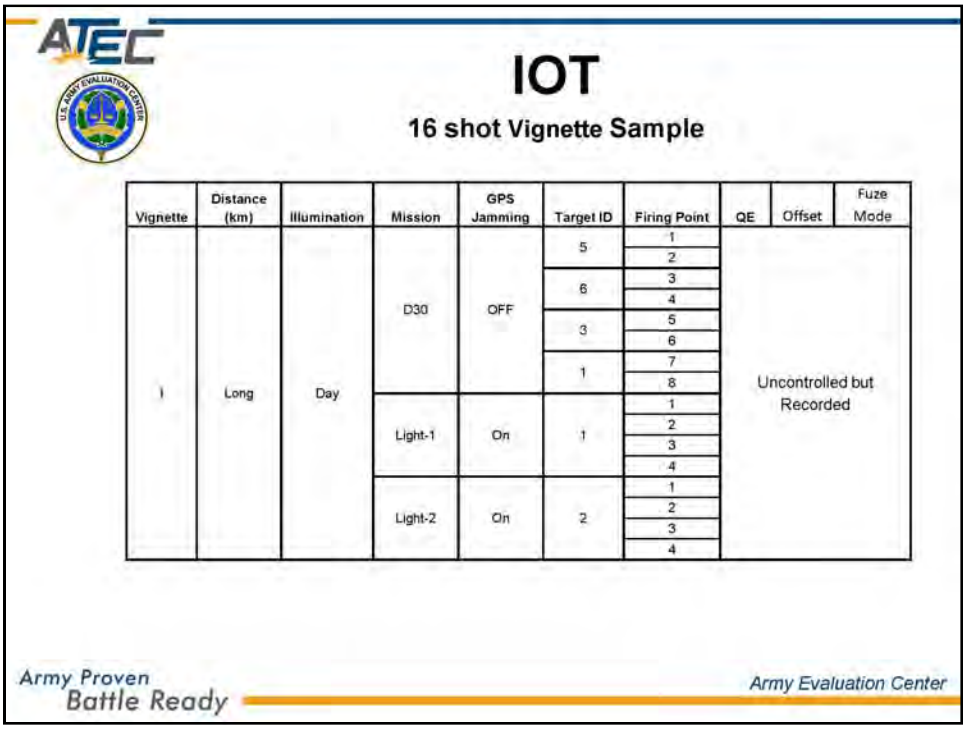
Vignette*	Range Band	Illumination	Mission**	GPS Jamming
1	Long	Day	Four D30	OFF
			Light – 1	On
			Light – 2	On
2	Short	Day	Light – 1	On
			Four D30	OFF
			Light – 2	On
3	Med	Night	Light – 1	On
			Light – 2	On
			Four D30	OFF

*One Mission will be conducted wearing MOPP gear

** Each Light Mission requires 1 Light Targets each using 4 shots; Each D30 Mission requires 4 D30 Targets each using 2 shots and 4 D30 Missions will be completed per vignette

Army Proven
Battle Ready

Army Evaluation Center



ATEC

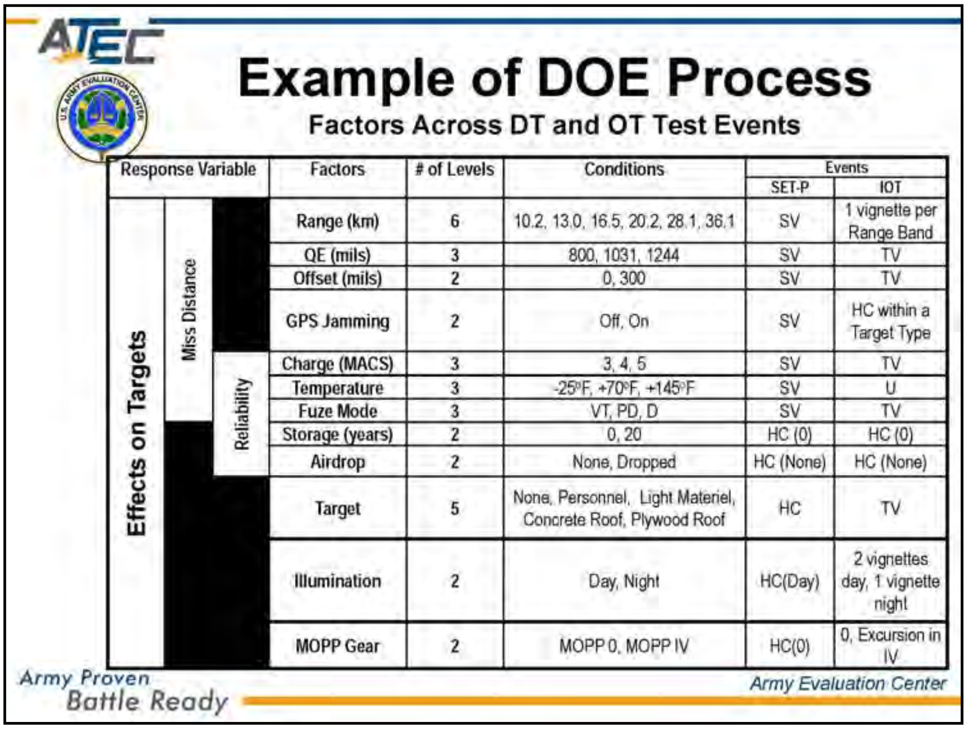
IOT

16 shot Vignette Sample

Vignette	Distance (km)	Illumination	Mission	GPS Jamming	Target ID	Firing Point	QE	Offset	Fuze Mode
1	Long	Day	D30	OFF	5	1	Uncontrolled but Recorded		
						2			
					6	3			
						4			
					3	5			
						6			
					1	7			
						8			
			Light-1	On	1	1			
						2			
						3			
						4			
			Light-2	On	2	1			
						2			
						3			
						4			

Army Proven
Battle Ready

Army Evaluation Center



ATEC

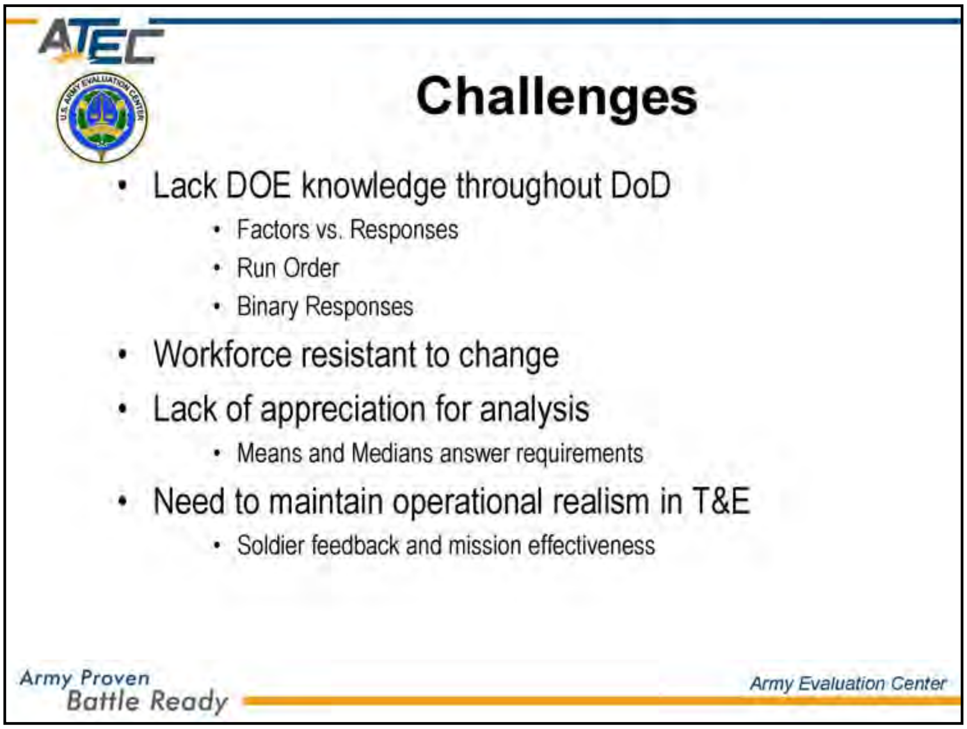
Example of DOE Process

Factors Across DT and OT Test Events

Response Variable	Factors	# of Levels	Conditions	Events	
				SET-P	IOT
Effects on Targets Miss Distance	Range (km)	6	10.2, 13.0, 16.5, 20.2, 28.1, 36.1	SV	1 vignette per Range Band
	QE (mils)	3	800, 1031, 1244	SV	TV
	Offset (mils)	2	0, 300	SV	TV
	GPS Jamming	2	Off, On	SV	HC within a Target Type
Reliability	Charge (MACS)	3	3, 4, 5	SV	TV
	Temperature	3	-25ºF, +70ºF, +145ºF	SV	U
	Fuze Mode	3	VT, PD, D	SV	TV
	Storage (years)	2	0, 20	HC (0)	HC (0)
	Airdrop	2	None, Dropped	HC (None)	HC (None)
	Target	5	None, Personnel, Light Materiel, Concrete Roof, Plywood Roof	HC	TV
	Illumination	2	Day, Night	HC(Day)	2 vignettes day, 1 vignette night
	MOPP Gear	2	MOPP 0, MOPP IV	HC(0)	0, Excursion in IV

Army Proven
Battle Ready

Army Evaluation Center



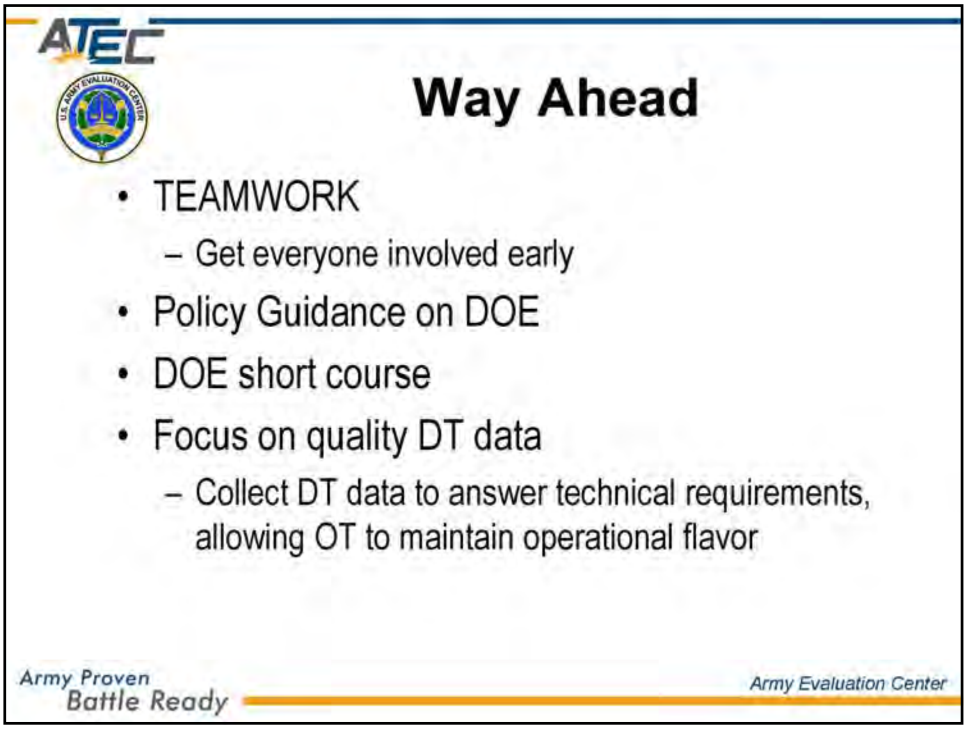
ATEC

Challenges

- Lack DOE knowledge throughout DoD
  - Factors vs. Responses
  - Run Order
  - Binary Responses
- Workforce resistant to change
- Lack of appreciation for analysis
  - Means and Medians answer requirements
- Need to maintain operational realism in T&E
  - Soldier feedback and mission effectiveness

Army Proven
Battle Ready

Army Evaluation Center



ATEC

Way Ahead

- TEAMWORK
  - Get everyone involved early
- Policy Guidance on DOE
- DOE short course
- Focus on quality DT data
  - Collect DT data to answer technical requirements, allowing OT to maintain operational flavor

Army Proven
Battle Ready

Army Evaluation Center

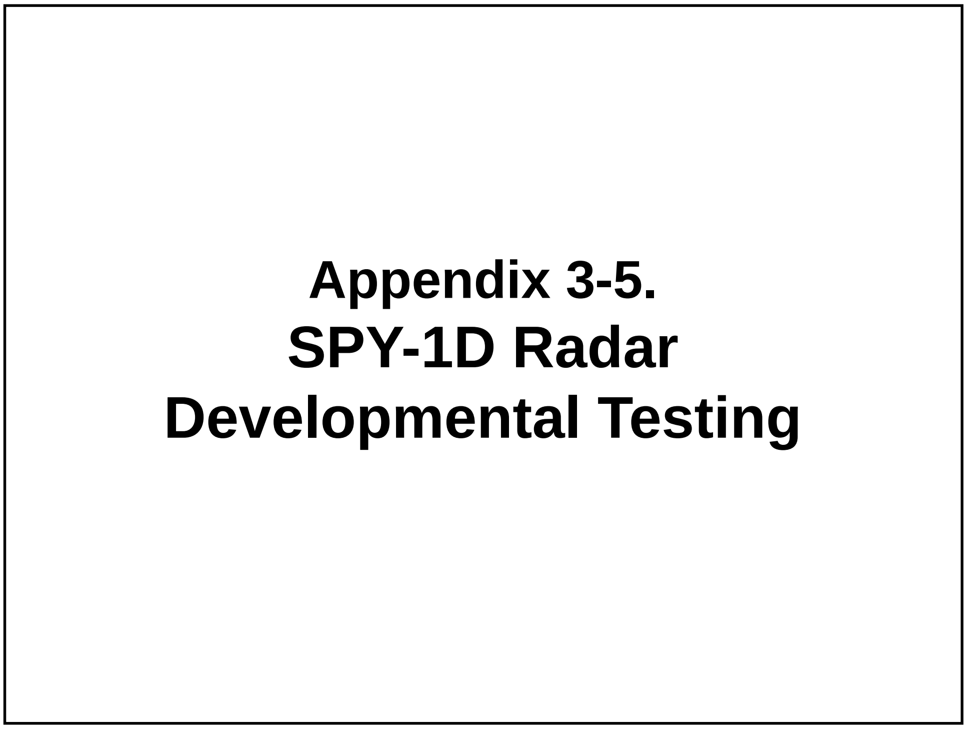
ATEC

QUESTIONS?

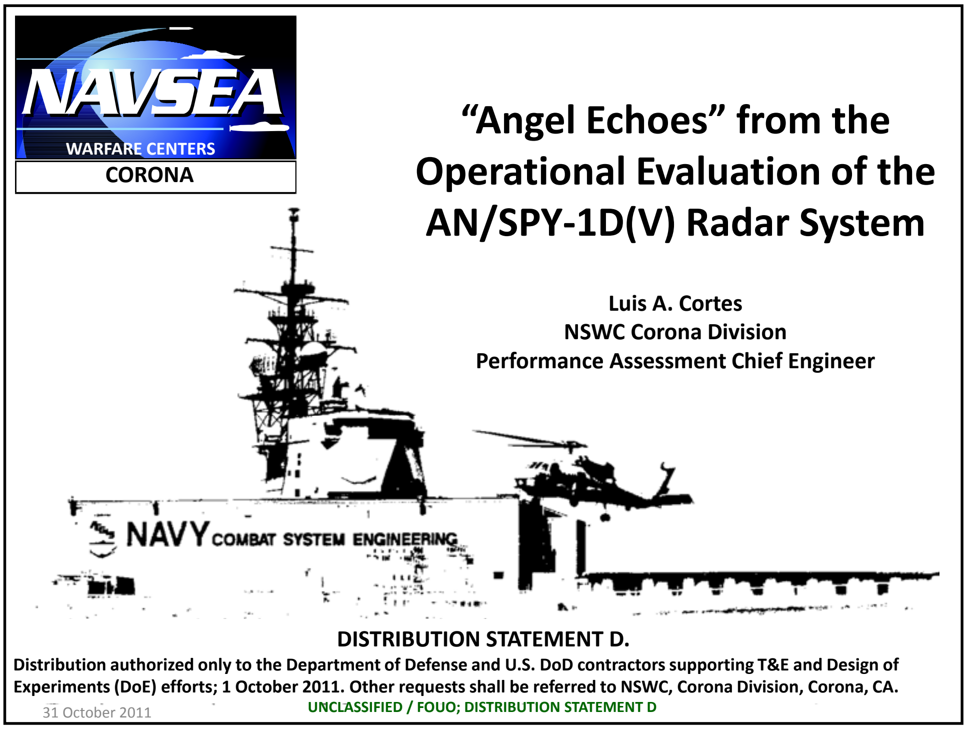
Army Proven
Battle Ready

Army Evaluation Center

This page intentionally left blank.



Appendix 3-5.
SPY-1D Radar
Developmental Testing



NAVSEA
WARFARE CENTERS
CORONA
NAVY COMBAT SYSTEM ENGINEERING

"Angel Echoes" from the Operational Evaluation of the AN/SPY-1D(V) Radar System

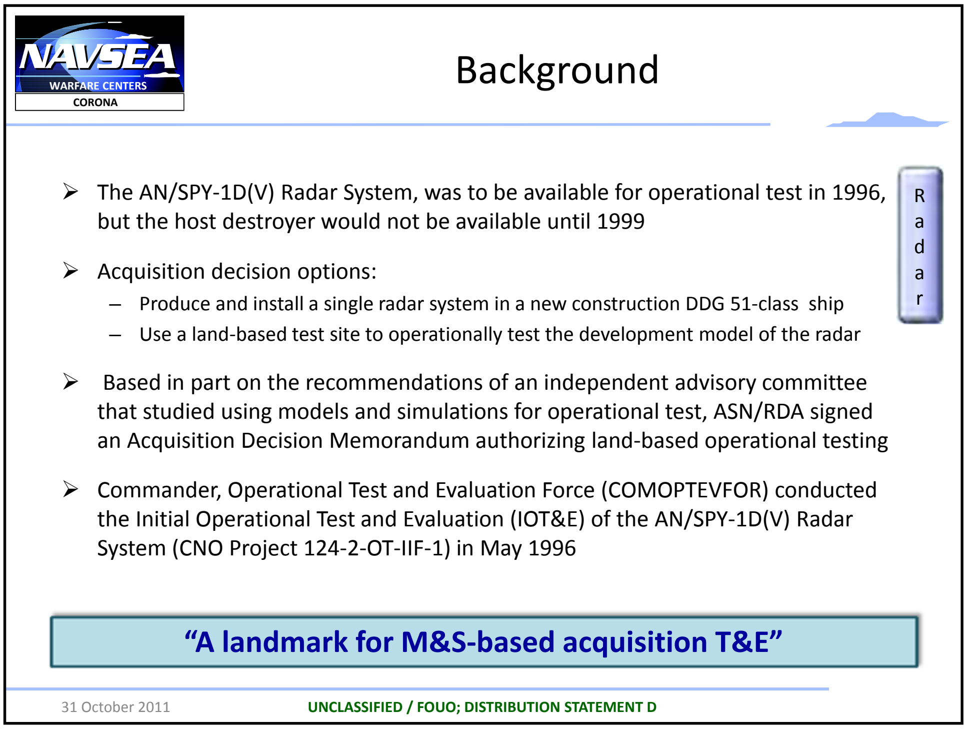
Luis A. Cortes
NSWC Corona Division
Performance Assessment Chief Engineer

DISTRIBUTION STATEMENT D.

Distribution authorized only to the Department of Defense and U.S. DoD contractors supporting T&E and Design of Experiments (DoE) efforts; 1 October 2011. Other requests shall be referred to NSWC, Corona Division, Corona, CA.

31 October 2011

UNCLASSIFIED / FOUO; DISTRIBUTION STATEMENT D



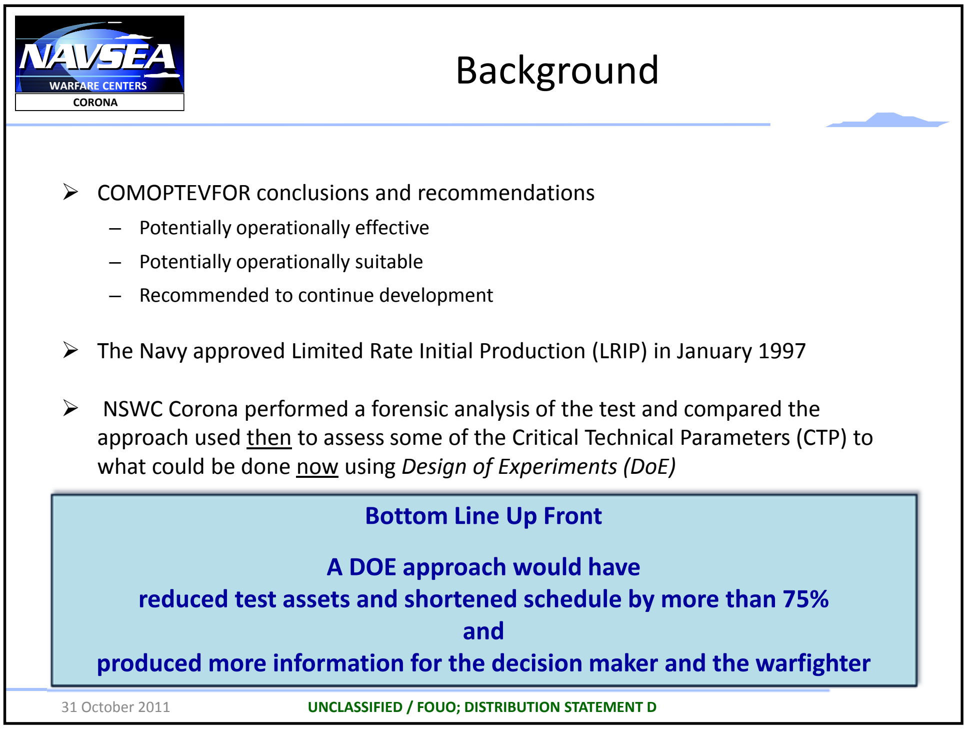
NAVSEA WARFARE CENTERS CORONA

Background

Radar

- The AN/SPY-1D(V) Radar System, was to be available for operational test in 1996, but the host destroyer would not be available until 1999
- Acquisition decision options:
  - Produce and install a single radar system in a new construction DDG 51-class ship
  - Use a land-based test site to operationally test the development model of the radar
- Based in part on the recommendations of an independent advisory committee that studied using models and simulations for operational test, ASN/RDA signed an Acquisition Decision Memorandum authorizing land-based operational testing
- Commander, Operational Test and Evaluation Force (COMOPTEVFOR) conducted the Initial Operational Test and Evaluation (IOT&E) of the AN/SPY-1D(V) Radar System (CNO Project 124-2-OT-IIF-1) in May 1996

“A landmark for M&S-based acquisition T&E”



NAVSEA WARFARE CENTERS CORONA

Background

- COMOPTEVFOR conclusions and recommendations
  - Potentially operationally effective
  - Potentially operationally suitable
  - Recommended to continue development
- The Navy approved Limited Rate Initial Production (LRIP) in January 1997
- NSWC Corona performed a forensic analysis of the test and compared the approach used then to assess some of the Critical Technical Parameters (CTP) to what could be done now using *Design of Experiments (DoE)*

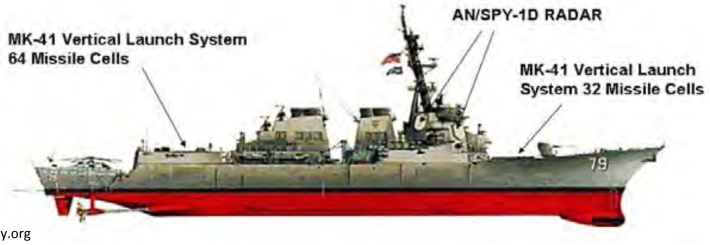
Bottom Line Up Front

A DOE approach would have
reduced test assets and shortened schedule by more than 75%
and
produced more information for the decision maker and the warfighter

NAVSEA WARFARE CENTERS CORONA

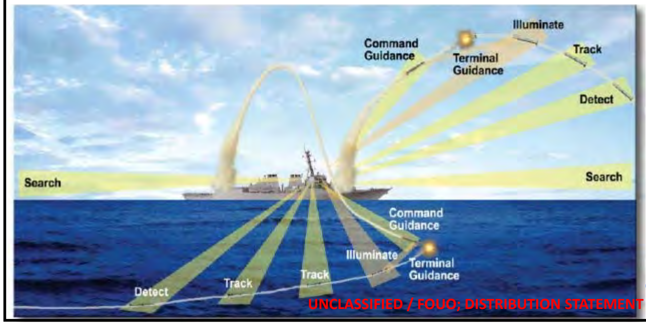
AN/SPY-1D(V) Radar System

USS OSCAR AUSTIN (DDG 79) - First Flight IIA, Commissioned August 2000



Back

Ref: Global Security.org



AN/SPY-1D(V)

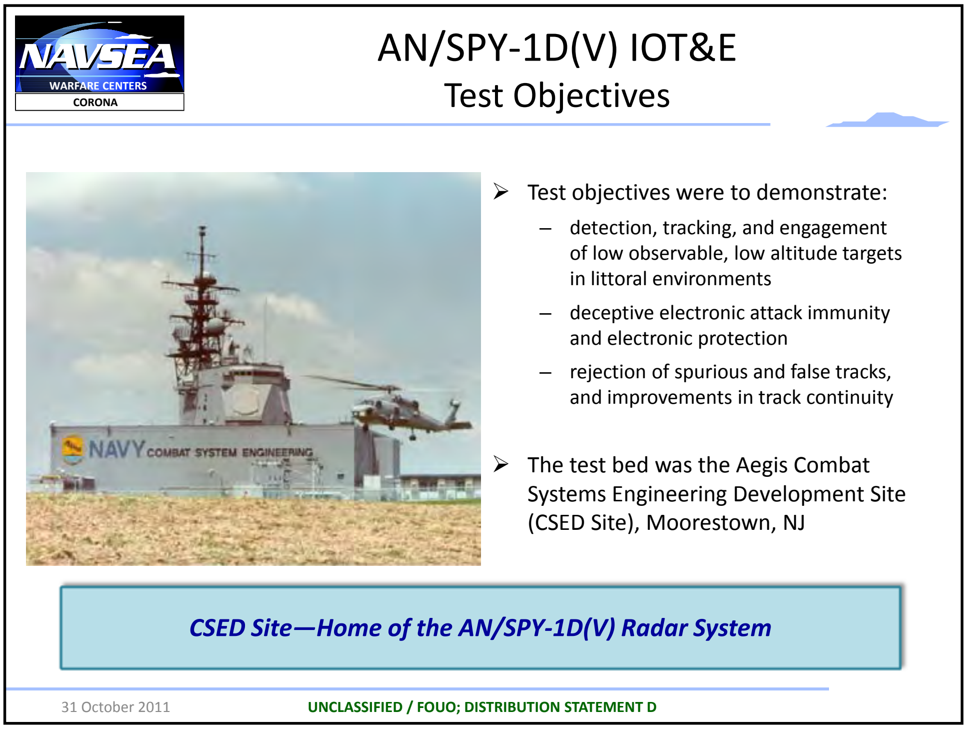
Type	3D Air-search
Frequency	S band
Range	100+ nm
Azimuth	0-360°
Elevation	Horizon-Zenith
Power	6 MW

UNCLASSIFIED / FOUO; DISTRIBUTION STATEMENT D

NAVSEA WARFARE CENTERS CORONA

Discussion Topics

- AN/SPY-1D(V) IOT&E
  - Background
  - Test Objectives
  - Test Site Limitations
  - Models and Simulation
- T&E Approach (Then)
  - Test Planning
  - Test Execution
  - Analysis and Assessment (TEMP Detection Requirements)
- DoE Approach (Now)
  - Test Design
  - Test Execution
  - Analysis and Assessment (TEMP Detection Requirements)
- Summary and Conclusions


NAVSEA
WARFARE CENTERS
CORONA
NAVY COMBAT SYSTEM ENGINEERING

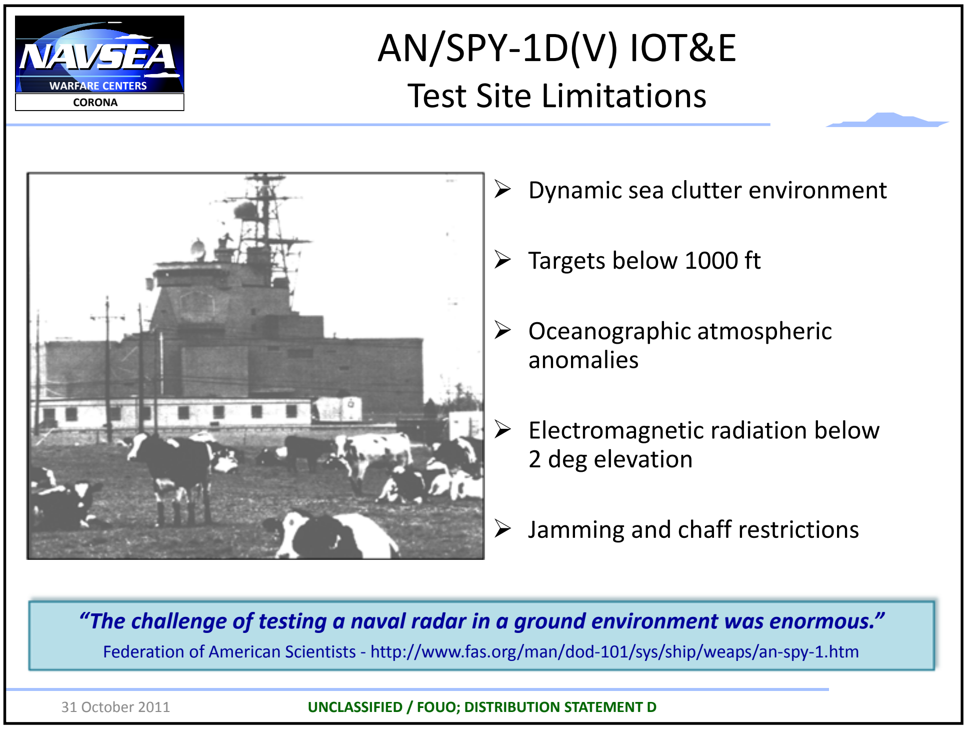
AN/SPY-1D(V) IOT&E

Test Objectives

- Test objectives were to demonstrate:
  - detection, tracking, and engagement of low observable, low altitude targets in littoral environments
  - deceptive electronic attack immunity and electronic protection
  - rejection of spurious and false tracks, and improvements in track continuity
- The test bed was the Aegis Combat Systems Engineering Development Site (CSED Site), Moorestown, NJ

CSED Site—Home of the AN/SPY-1D(V) Radar System

31 October 2011 UNCLASSIFIED / FOUO; DISTRIBUTION STATEMENT D


NAVSEA
WARFARE CENTERS
CORONA

AN/SPY-1D(V) IOT&E

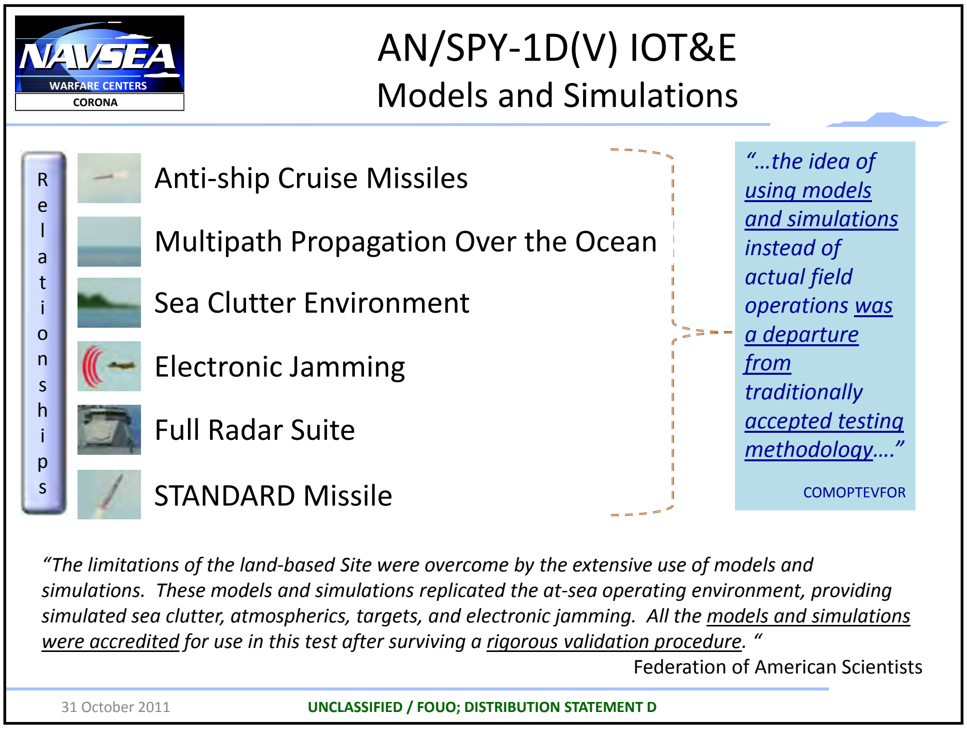
Test Site Limitations

- Dynamic sea clutter environment
- Targets below 1000 ft
- Oceanographic atmospheric anomalies
- Electromagnetic radiation below 2 deg elevation
- Jamming and chaff restrictions

"The challenge of testing a naval radar in a ground environment was enormous."

Federation of American Scientists - http://www.fas.org/man/dod-101/sys/ship/weaps/an-spy-1.htm

31 October 2011 UNCLASSIFIED / FOUO; DISTRIBUTION STATEMENT D



NAVSEA WARFARE CENTERS CORONA

AN/SPY-1D(V) IOT&E
Models and Simulations

Relationships

- Anti-ship Cruise Missiles
- Multipath Propagation Over the Ocean
- Sea Clutter Environment
- Electronic Jamming
- Full Radar Suite
- STANDARD Missile

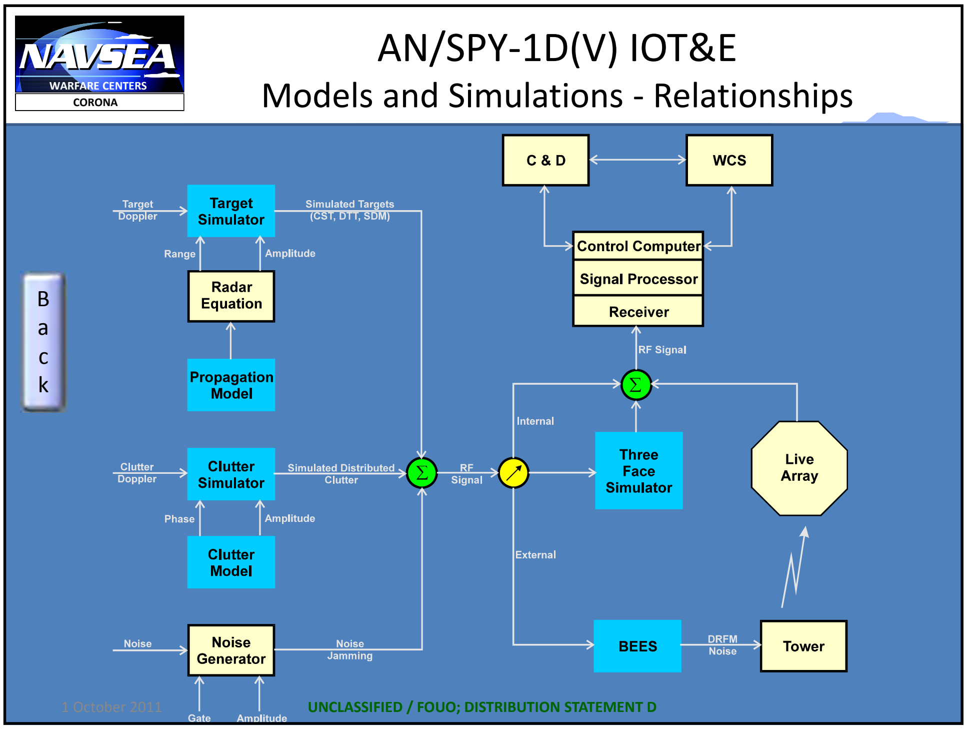
"…the idea of using models and simulations instead of actual field operations was a departure from traditionally accepted testing methodology…."

COMOPTEVFOR

"The limitations of the land-based Site were overcome by the extensive use of models and simulations. These models and simulations replicated the at-sea operating environment, providing simulated sea clutter, atmospherics, targets, and electronic jamming. All the models and simulations were accredited for use in this test after surviving a rigorous validation procedure. "

Federation of American Scientists

31 October 2011 UNCLASSIFIED / FOUO; DISTRIBUTION STATEMENT D



NAVSEA WARFARE CENTERS CORONA

AN/SPY-1D(V) IOT&E
Models and Simulations - Relationships

Back

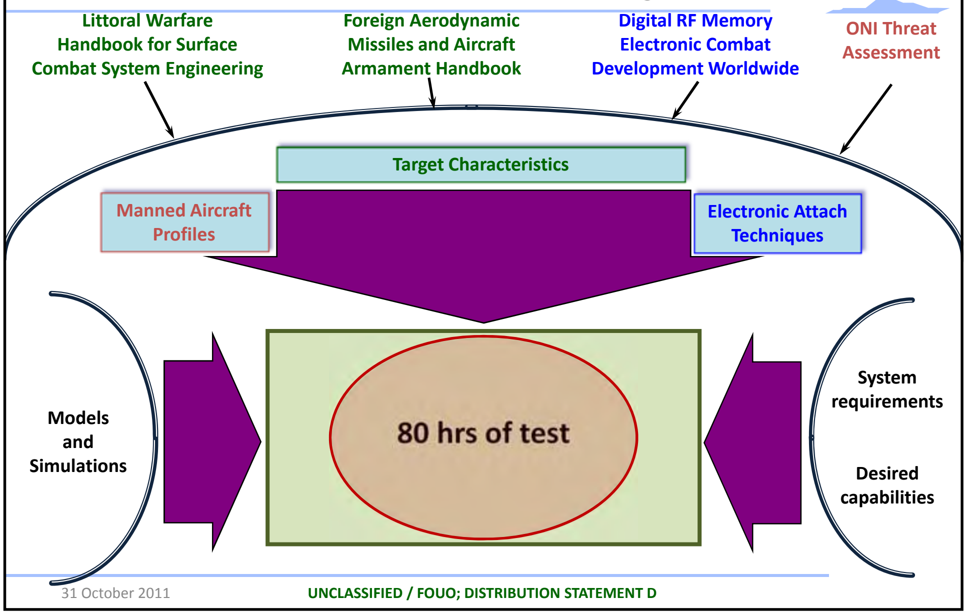
1 October 2011 UNCLASSIFIED / FOUO; DISTRIBUTION STATEMENT D

Virtual Prototype

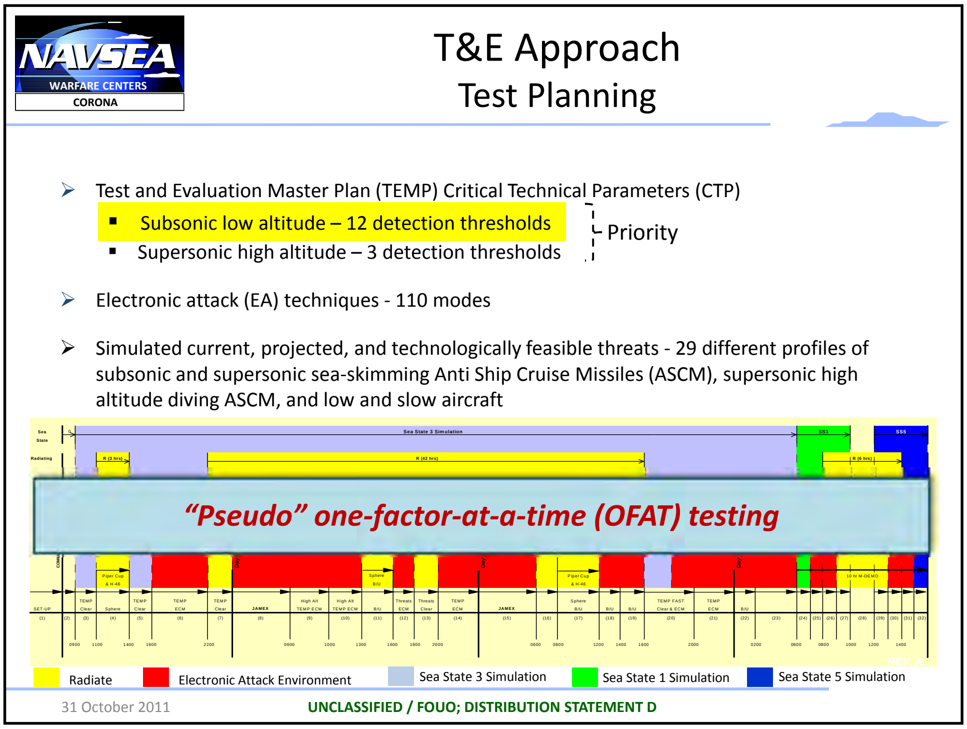
NAVSEA WARFARE CENTERS CORONA

T&E Approach

Test Planning



Littoral Warfare Handbook for Surface Combat System Engineering
Foreign Aerodynamic Missiles and Aircraft Armament Handbook
Digital RF Memory Electronic Combat Development Worldwide
ONI Threat Assessment
Target Characteristics
Manned Aircraft Profiles
Electronic Attach Techniques
Models and Simulations
80 hrs of test
System requirements
Desired capabilities



T&E Approach

Test Planning

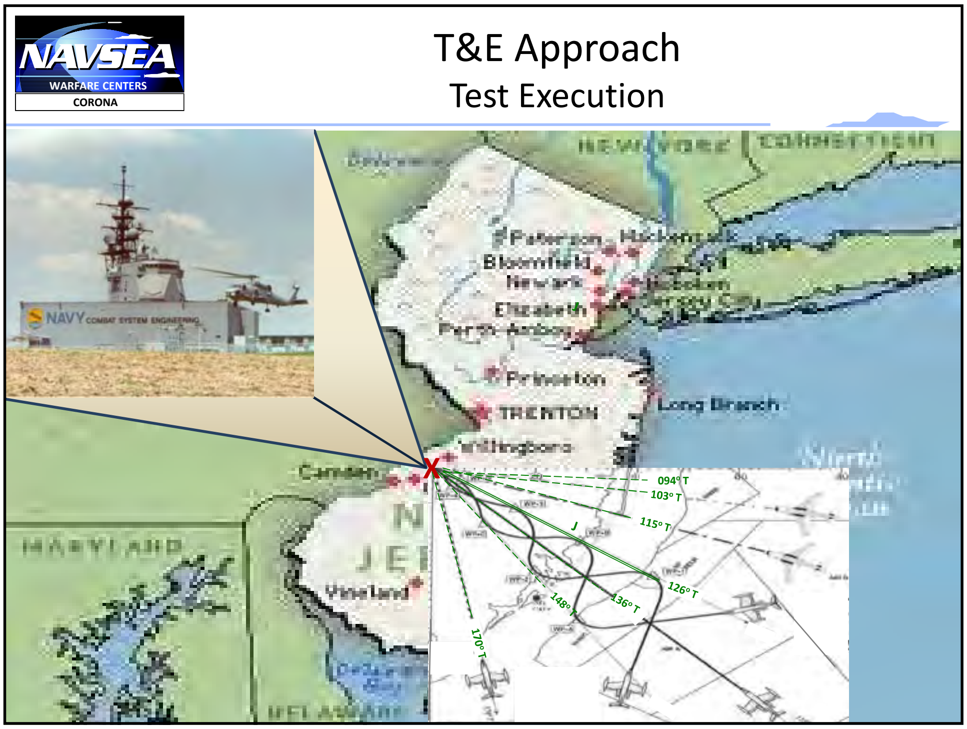
- Test and Evaluation Master Plan (TEMP) Critical Technical Parameters (CTP)
  - Subsonic low altitude – 12 detection thresholds
  - Supersonic high altitude – 3 detection thresholds

  Priority
- Electronic attack (EA) techniques - 110 modes
- Simulated current, projected, and technologically feasible threats - 29 different profiles of subsonic and supersonic sea-skimming Anti Ship Cruise Missiles (ASCM), supersonic high altitude diving ASCM, and low and slow aircraft

"Pseudo" one-factor-at-a-time (OFAT) testing

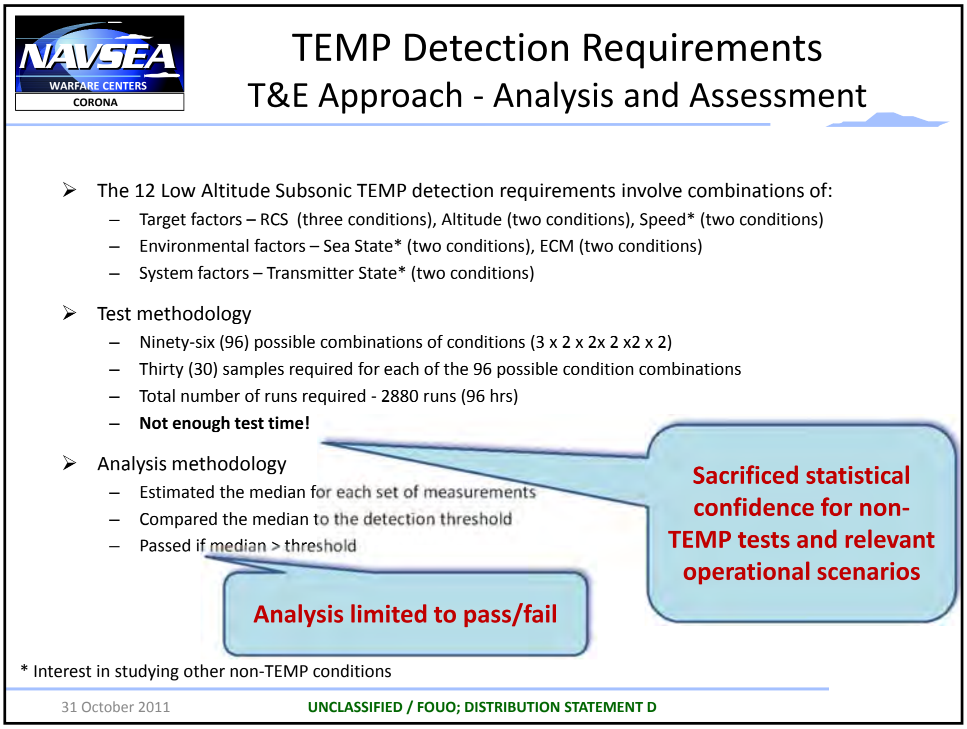
🟨 Radiate 🟥 Electronic Attack Environment ▨ Sea State 3 Simulation 🟩 Sea State 1 Simulation 🟦 Sea State 5 Simulation

31 October 2011 UNCLASSIFIED / FOUO; DISTRIBUTION STATEMENT D



T&E Approach

Test Execution



TEMP Detection Requirements
T&E Approach - Analysis and Assessment

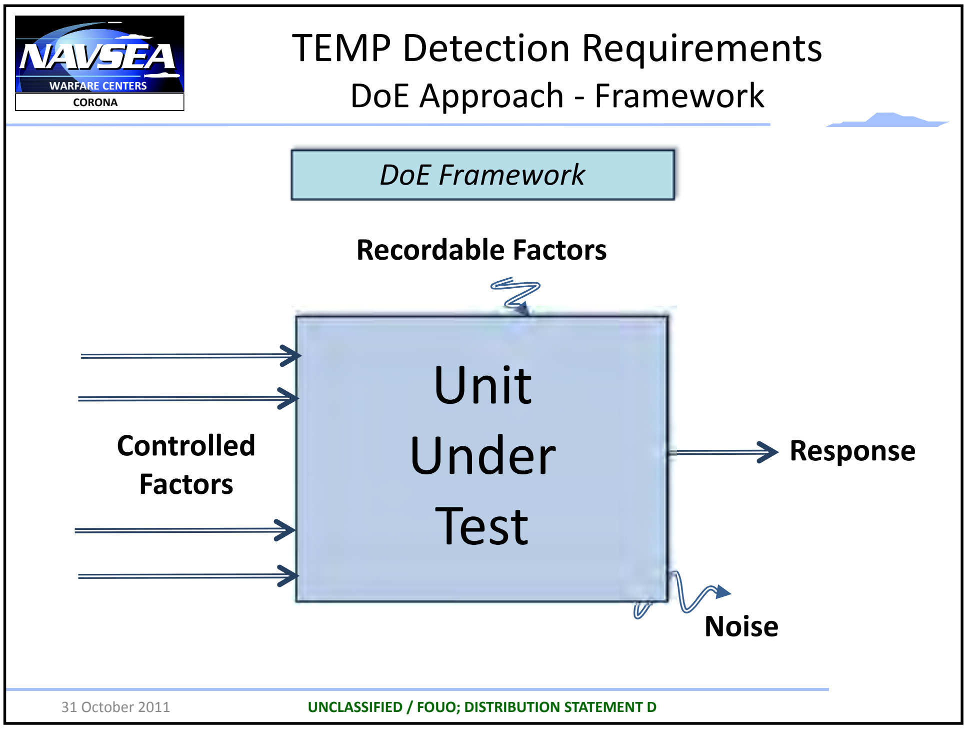
- The 12 Low Altitude Subsonic TEMP detection requirements involve combinations of:
  - Target factors – RCS (three conditions), Altitude (two conditions), Speed* (two conditions)
  - Environmental factors – Sea State* (two conditions), ECM (two conditions)
  - System factors – Transmitter State* (two conditions)
- Test methodology
  - Ninety-six (96) possible combinations of conditions (3 x 2 x 2x 2 x2 x 2)
  - Thirty (30) samples required for each of the 96 possible condition combinations
  - Total number of runs required - 2880 runs (96 hrs)
  - **Not enough test time!**
- Analysis methodology
  - Estimated the median for each set of measurements
  - Compared the median to the detection threshold
  - Passed if median > threshold

Sacrificed statistical confidence for non-TEMP tests and relevant operational scenarios

Analysis limited to pass/fail

* Interest in studying other non-TEMP conditions

31 October 2011 UNCLASSIFIED / FOUO; DISTRIBUTION STATEMENT D



TEMP Detection Requirements
DoE Approach - Framework

DoE Framework

Recordable Factors

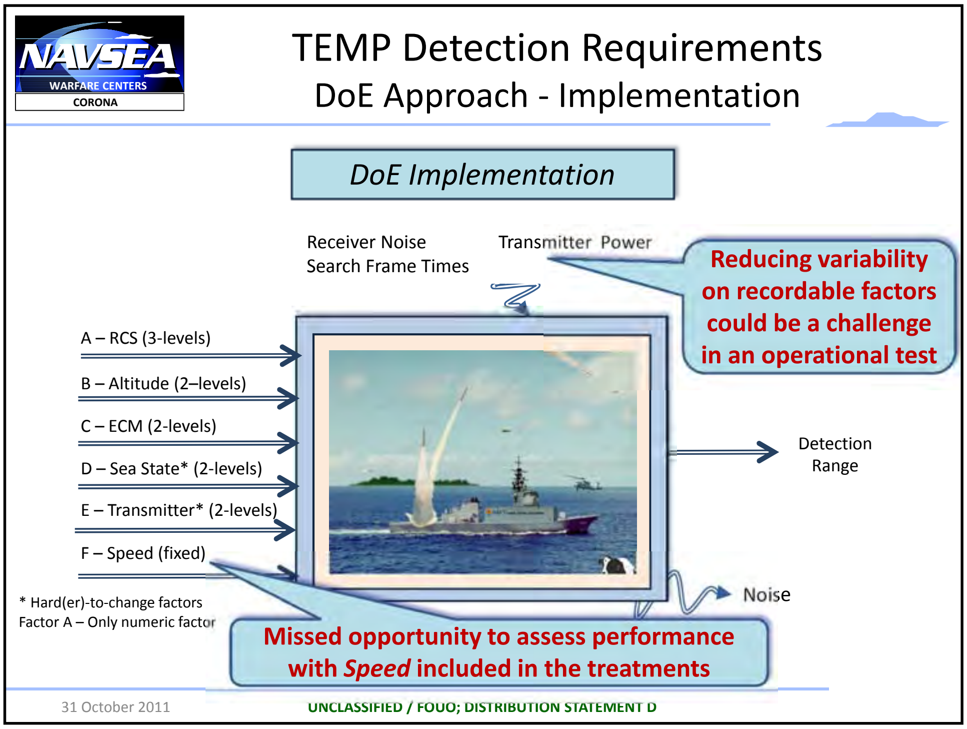
Controlled Factors

Unit Under Test

Response

Noise

31 October 2011 UNCLASSIFIED / FOUO; DISTRIBUTION STATEMENT D



NAVSEA WARFARE CENTERS CORONA

TEMP Detection Requirements

DoE Approach - Implementation

DoE Implementation

Receiver Noise
Search Frame Times

Transmitter Power

Reducing variability on recordable factors could be a challenge in an operational test

A – RCS (3-levels)

B – Altitude (2–levels)

C – ECM (2-levels)

D – Sea State* (2-levels)

E – Transmitter* (2-levels)

F – Speed (fixed)

Detection Range

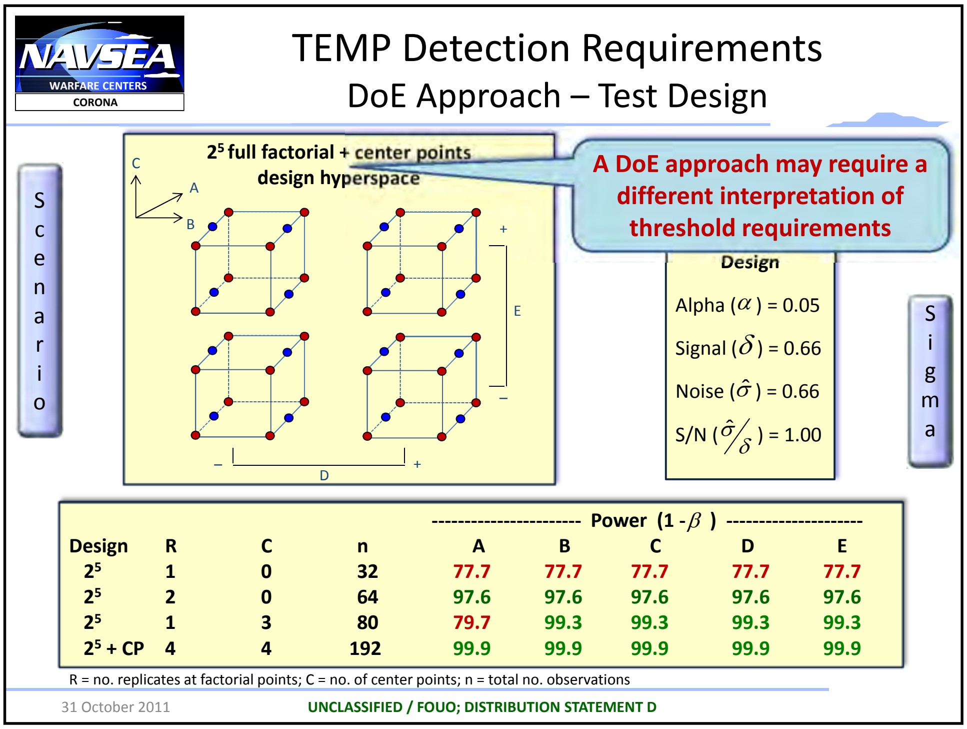
Noise

* Hard(er)-to-change factors
Factor A – Only numeric factor

Missed opportunity to assess performance with *Speed* included in the treatments

31 October 2011

UNCLASSIFIED / FOUO; DISTRIBUTION STATEMENT D



NAVSEA WARFARE CENTERS CORONA

TEMP Detection Requirements

DoE Approach – Test Design

Scenario

2^5 full factorial + center points design hyperspace

A DoE approach may require a different interpretation of threshold requirements

Design

Alpha (α) = 0.05

Signal (δ) = 0.66

Noise ($\hat{\sigma}$) = 0.66

S/N ($\hat{\sigma}/\delta$) = 1.00

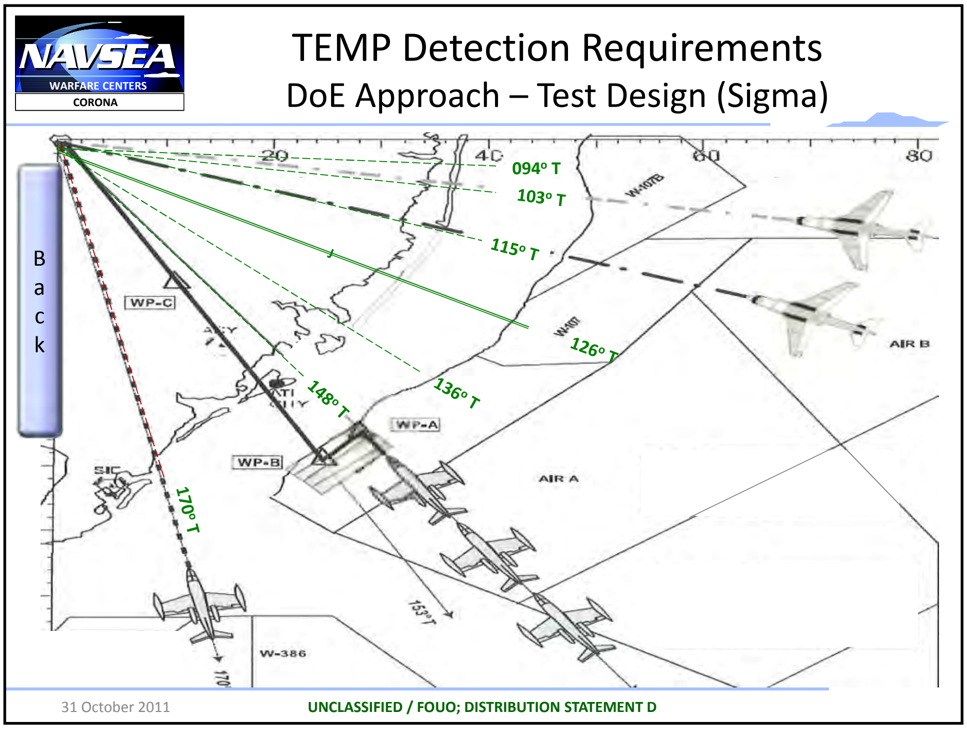
Sigma

				----------------------	Power	(1 - β)	----------------------	
Design	**R**	**C**	**n**	**A**	**B**	**C**	**D**	**E**
2^5	**1**	**0**	**32**	**77.7**	**77.7**	**77.7**	**77.7**	**77.7**
2^5	**2**	**0**	**64**	**97.6**	**97.6**	**97.6**	**97.6**	**97.6**
2^5	**1**	**3**	**80**	**79.7**	**99.3**	**99.3**	**99.3**	**99.3**
2^5 + CP	**4**	**4**	**192**	**99.9**	**99.9**	**99.9**	**99.9**	**99.9**

R = no. replicates at factorial points; C = no. of center points; n = total no. observations

31 October 2011

UNCLASSIFIED / FOUO; DISTRIBUTION STATEMENT D

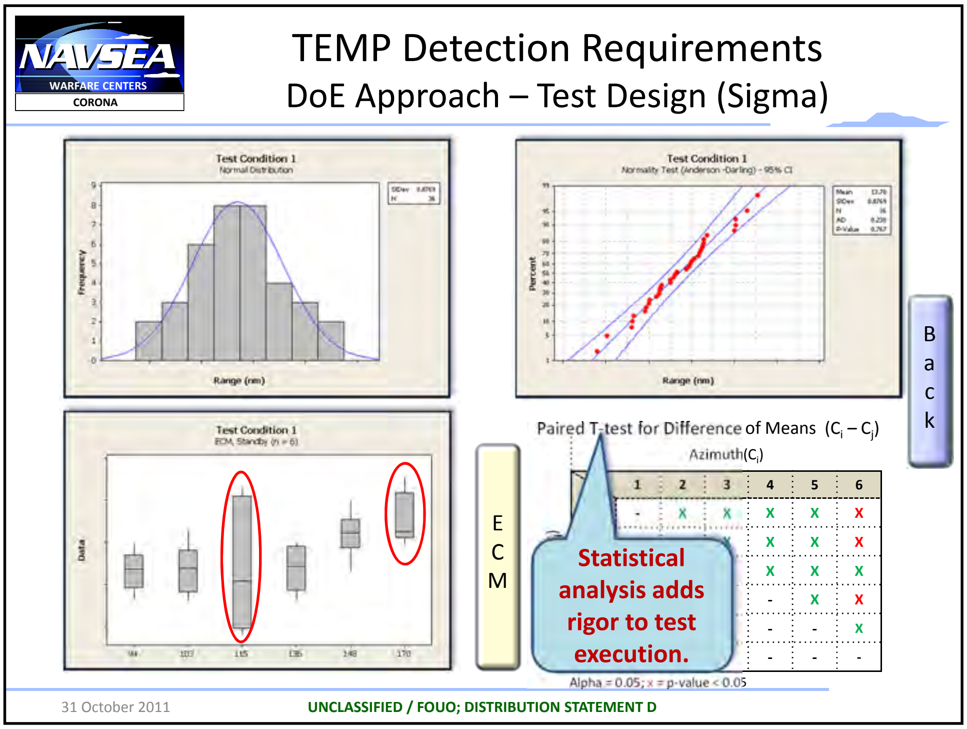


NAVSEA WARFARE CENTERS CORONA

TEMP Detection Requirements

DoE Approach – Test Design (Sigma)

31 October 2011 UNCLASSIFIED / FOUO; DISTRIBUTION STATEMENT D



NAVSEA WARFARE CENTERS CORONA

TEMP Detection Requirements

DoE Approach – Test Design (Sigma)

Paired T-test for Difference of Means ($C_i – C_j$)

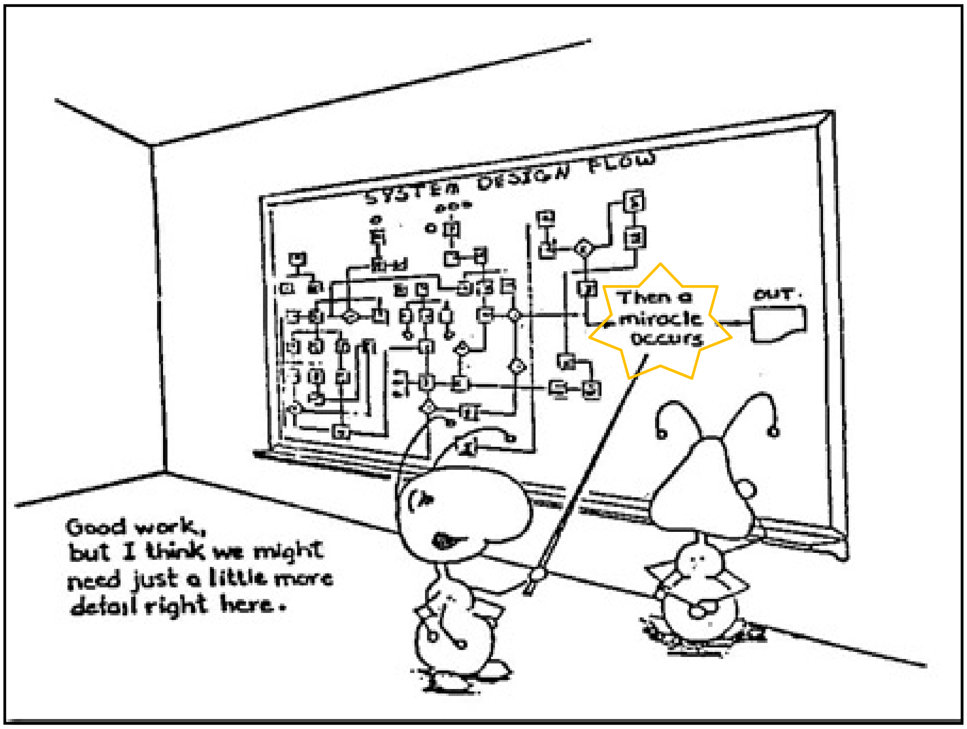
Azimuth(C_i)

	1	2	3	4	5	6
	-	X	X	X	X	X
				X	X	X
				X	X	X
				-	X	X
				-	-	X
				-	-	-

Statistical analysis adds rigor to test execution.

Alpha = 0.05; x = p-value < 0.05

31 October 2011 UNCLASSIFIED / FOUO; DISTRIBUTION STATEMENT D


SYSTEM DESIGN FLOW
Then a miracle occurs
OUT
Good work, but I think we might need just a little more detail right here.

NAVSEA WARFARE CENTERS CORONA

TEMP Detection Requirements

DoE Approach – Analysis

Design

2^5 Full Factorial

1 Replicate (32 Runs)

Diagnostics

Completely Randomized Design

Analysis of variance table [Partial sum of squares - Type III]

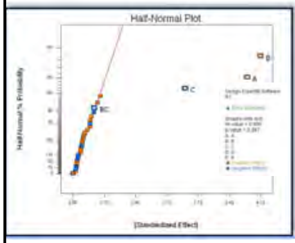
Source	Sum of Squares	df	Mean Square	F Value	p-value Prob > F	
Model	313.94	4	78.48	125.56	< 0.0001	significant
A-RCS	*121.29*	*1*	*121.29*	*194.05*	*< 0.0001*	
B-Alt						
C-ECM						
BC						
Residual						
Cor Total						

Sea State (Factor D) has no significant effect on detection range. The clutter model and clutter simulator are suspect.

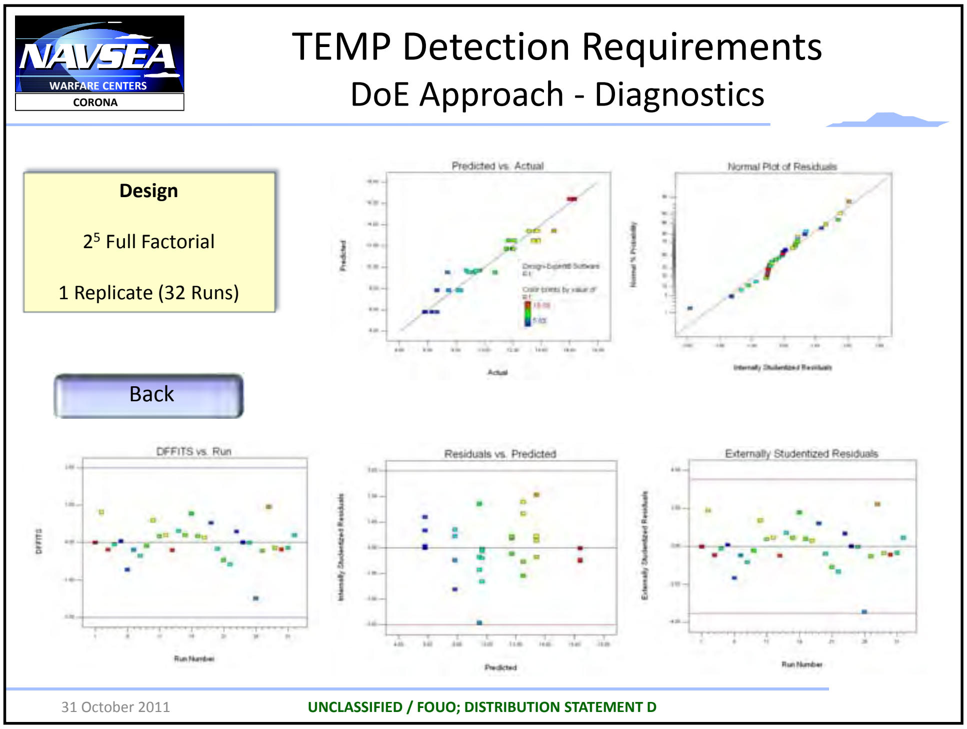
Reduced Empirical Model (Coded Factors)

$$R = I + 1.95A + 2.10B - 1.25C - 0.24BC$$

$R^2 = 0.9490$ Adj. $R^2 = 0.9414$ Pred. $R^2 = 0.9283$ Adeq. Precision = 33.9



Analysis via DoE yields an empirical detection model that is useful for tactical decision aids, training, and performance assessment.



TEMP Detection Requirements
DoE Approach - Diagnostics

Design

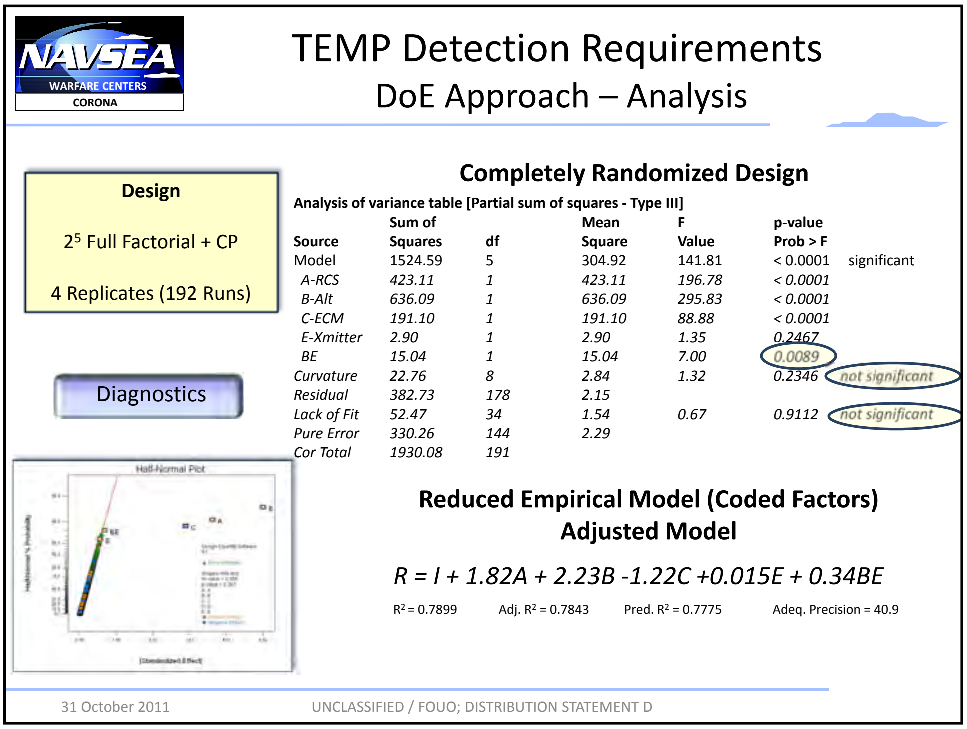
2^5 Full Factorial

1 Replicate (32 Runs)

Back

31 October 2011

UNCLASSIFIED / FOUO; DISTRIBUTION STATEMENT D


NAVSEA
WARFARE CENTERS
CORONA
Half-Normal Plot
Half-Normal % Probability
|Standardized Effect|
A
B
C
E
BE

TEMP Detection Requirements
DoE Approach – Analysis

Design

2^5 Full Factorial + CP

4 Replicates (192 Runs)

Diagnostics

Completely Randomized Design

Analysis of variance table [Partial sum of squares - Type III]

Source	**Sum of Squares**	**df**	**Mean Square**	**F Value**	**p-value Prob > F**	
Model	1524.59	5	304.92	141.81	< 0.0001	significant
A-RCS	*423.11*	*1*	*423.11*	*196.78*	*< 0.0001*	
B-Alt	*636.09*	*1*	*636.09*	*295.83*	*< 0.0001*	
C-ECM	*191.10*	*1*	*191.10*	*88.88*	*< 0.0001*	
E-Xmitter	*2.90*	*1*	*2.90*	*1.35*	*0.2467*	
BE	*15.04*	*1*	*15.04*	*7.00*	*0.0089*	
Curvature	*22.76*	*8*	*2.84*	*1.32*	*0.2346*	*not significant*
Residual	*382.73*	*178*	*2.15*			
Lack of Fit	*52.47*	*34*	*1.54*	*0.67*	*0.9112*	*not significant*
Pure Error	*330.26*	*144*	*2.29*			
Cor Total	*1930.08*	*191*				

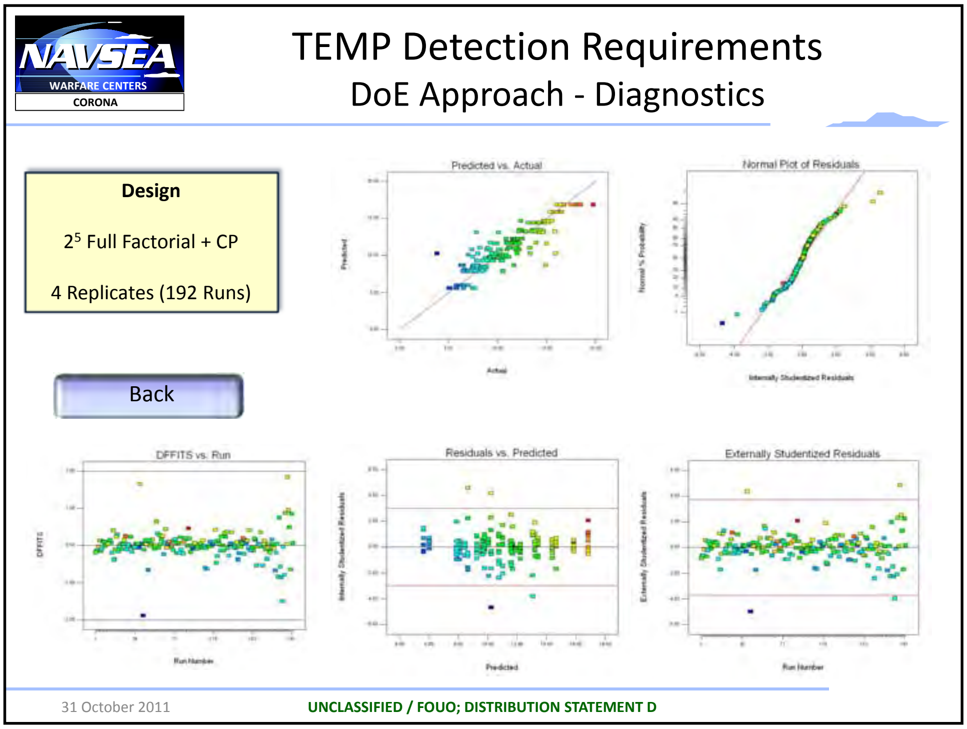
Reduced Empirical Model (Coded Factors)
Adjusted Model

$$R = I + 1.82A + 2.23B - 1.22C + 0.015E + 0.34BE$$

$R^2 = 0.7899$ Adj. $R^2 = 0.7843$ Pred. $R^2 = 0.7775$ Adeq. Precision = 40.9

31 October 2011

UNCLASSIFIED / FOUO; DISTRIBUTION STATEMENT D



TEMP Detection Requirements

DoE Approach - Diagnostics

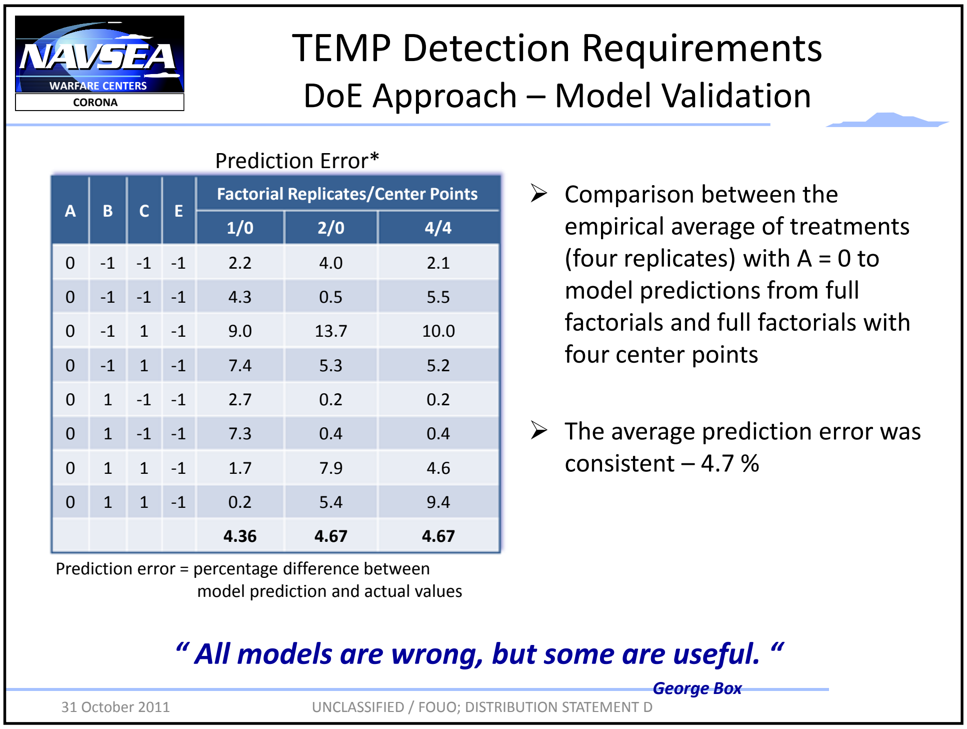
Design

2^5 Full Factorial + CP

4 Replicates (192 Runs)

Back

31 October 2011 UNCLASSIFIED / FOUO; DISTRIBUTION STATEMENT D



TEMP Detection Requirements

DoE Approach – Model Validation

Prediction Error*

A	B	C	E	Factorial Replicates/Center Points		
				1/0	2/0	4/4
0	-1	-1	-1	2.2	4.0	2.1
0	-1	-1	-1	4.3	0.5	5.5
0	-1	1	-1	9.0	13.7	10.0
0	-1	1	-1	7.4	5.3	5.2
0	1	-1	-1	2.7	0.2	0.2
0	1	-1	-1	7.3	0.4	0.4
0	1	1	-1	1.7	7.9	4.6
0	1	1	-1	0.2	5.4	9.4
				4.36	**4.67**	**4.67**

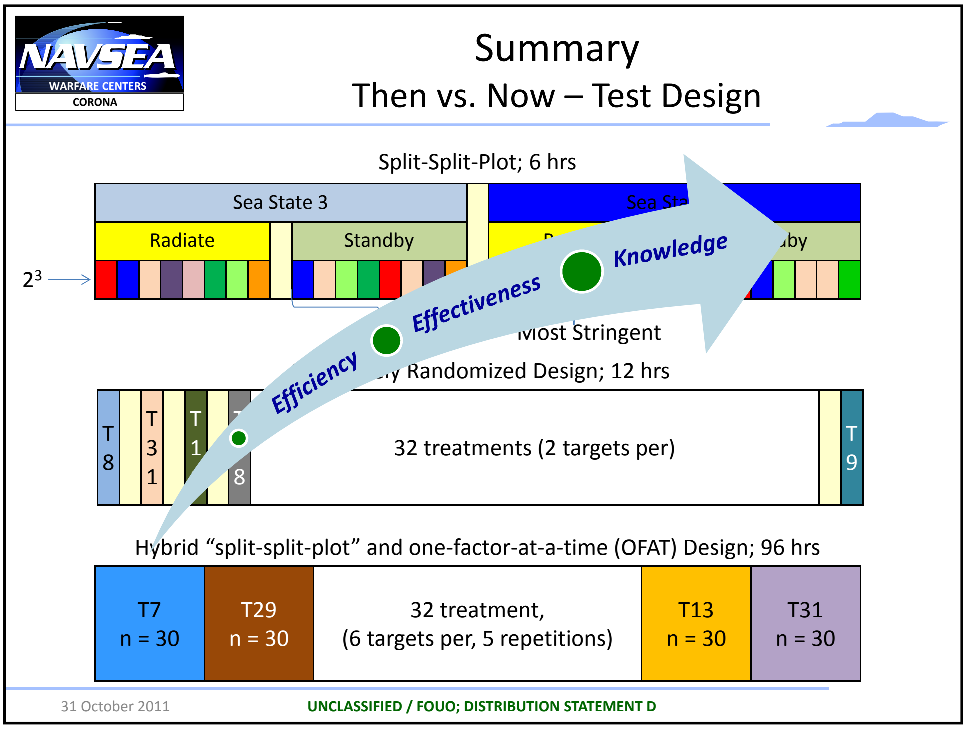
Prediction error = percentage difference between model prediction and actual values

- Comparison between the empirical average of treatments (four replicates) with A = 0 to model predictions from full factorials and full factorials with four center points
- The average prediction error was consistent – 4.7 %

" All models are wrong, but some are useful. "

George Box

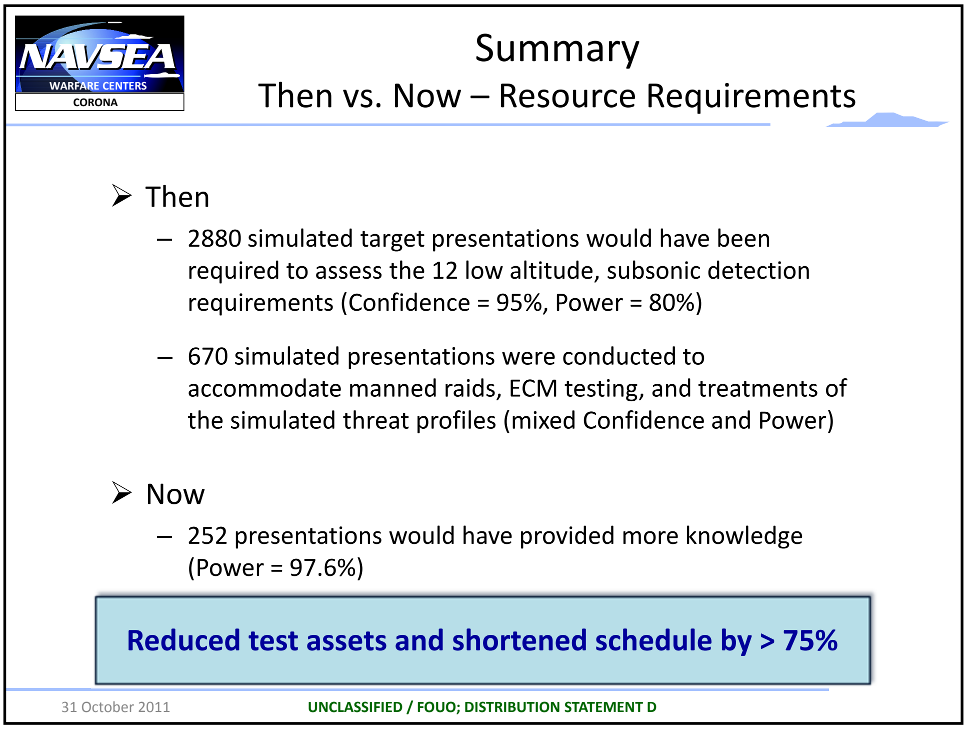
31 October 2011 UNCLASSIFIED / FOUO; DISTRIBUTION STATEMENT D



NAVSEA WARFARE CENTERS CORONA

Summary
Then vs. Now – Test Design

Split-Split-Plot; 6 hrs
Sea State 3
Sea St
Radiate
Standby
R
ıby
2^3
Most Stringent
ly Randomized Design; 12 hrs
T 8
T 3 1
T 1
8
32 treatments (2 targets per)
T 9
Efficiency
Effectiveness
Knowledge
Hybrid "split-split-plot" and one-factor-at-a-time (OFAT) Design; 96 hrs
T7 n = 30
T29 n = 30
32 treatment, (6 targets per, 5 repetitions)
T13 n = 30
T31 n = 30

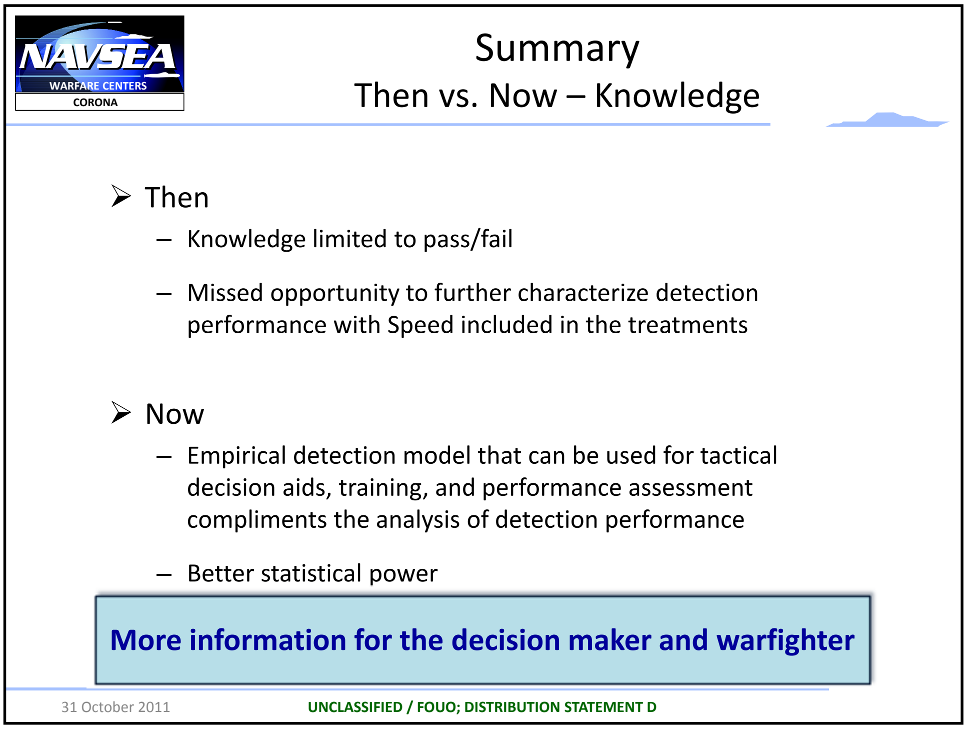


NAVSEA WARFARE CENTERS CORONA

Summary
Then vs. Now – Resource Requirements

- Then
  - 2880 simulated target presentations would have been required to assess the 12 low altitude, subsonic detection requirements (Confidence = 95%, Power = 80%)
  - 670 simulated presentations were conducted to accommodate manned raids, ECM testing, and treatments of the simulated threat profiles (mixed Confidence and Power)
- Now
  - 252 presentations would have provided more knowledge (Power = 97.6%)

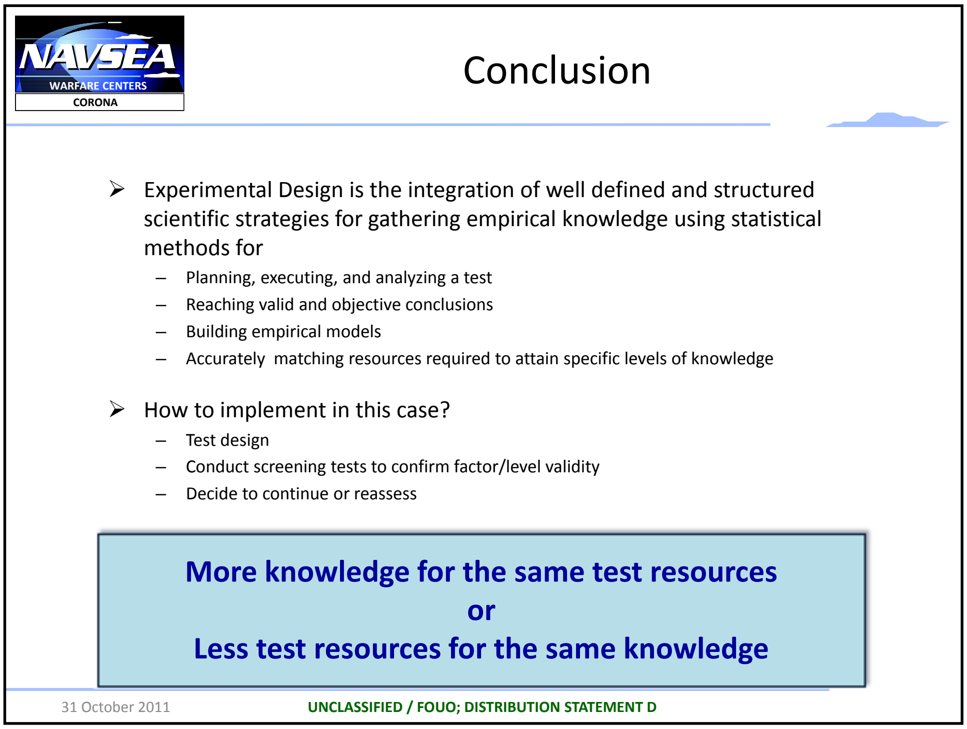
Reduced test assets and shortened schedule by > 75%



Summary
Then vs. Now – Knowledge

- Then
  - Knowledge limited to pass/fail
  - Missed opportunity to further characterize detection performance with Speed included in the treatments
- Now
  - Empirical detection model that can be used for tactical decision aids, training, and performance assessment compliments the analysis of detection performance
  - Better statistical power

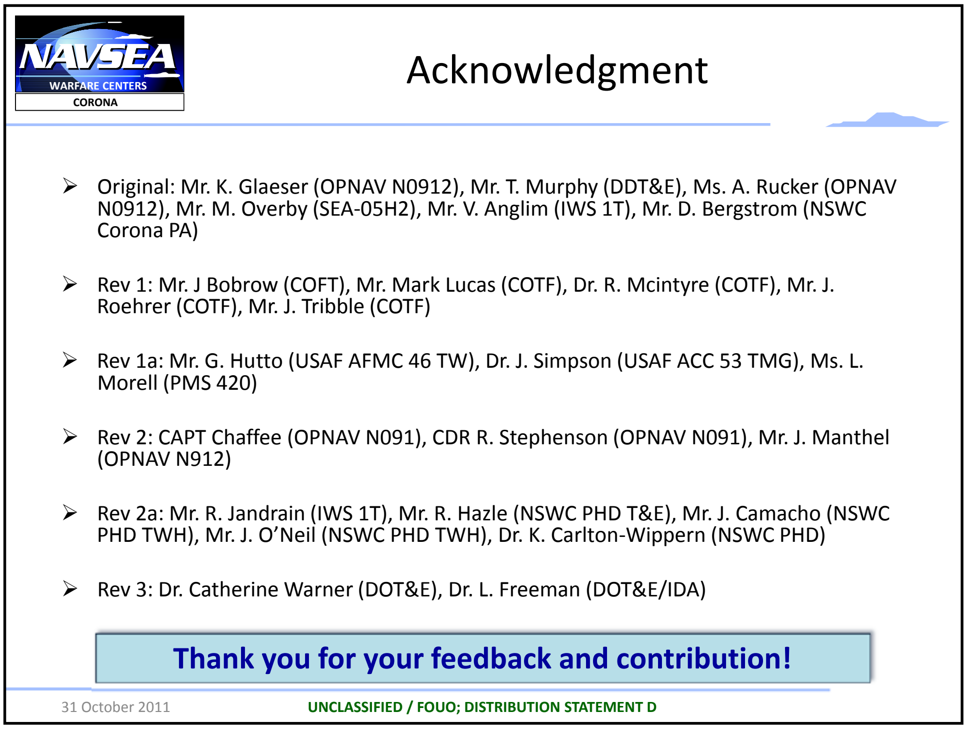
More information for the decision maker and warfighter



Conclusion

- Experimental Design is the integration of well defined and structured scientific strategies for gathering empirical knowledge using statistical methods for
  - Planning, executing, and analyzing a test
  - Reaching valid and objective conclusions
  - Building empirical models
  - Accurately matching resources required to attain specific levels of knowledge
- How to implement in this case?
  - Test design
  - Conduct screening tests to confirm factor/level validity
  - Decide to continue or reassess

More knowledge for the same test resources
or
Less test resources for the same knowledge



Acknowledgment

- Original: Mr. K. Glaeser (OPNAV N0912), Mr. T. Murphy (DDT&E), Ms. A. Rucker (OPNAV N0912), Mr. M. Overby (SEA-05H2), Mr. V. Anglim (IWS 1T), Mr. D. Bergstrom (NSWC Corona PA)
- Rev 1: Mr. J Bobrow (COFT), Mr. Mark Lucas (COTF), Dr. R. Mcintyre (COTF), Mr. J. Roehrer (COTF), Mr. J. Tribble (COTF)
- Rev 1a: Mr. G. Hutto (USAF AFMC 46 TW), Dr. J. Simpson (USAF ACC 53 TMG), Ms. L. Morell (PMS 420)
- Rev 2: CAPT Chaffee (OPNAV N091), CDR R. Stephenson (OPNAV N091), Mr. J. Manthel (OPNAV N912)
- Rev 2a: Mr. R. Jandrain (IWS 1T), Mr. R. Hazle (NSWC PHD T&E), Mr. J. Camacho (NSWC PHD TWH), Mr. J. O'Neil (NSWC PHD TWH), Dr. K. Carlton-Wippern (NSWC PHD)
- Rev 3: Dr. Catherine Warner (DOT&E), Dr. L. Freeman (DOT&E/IDA)

Thank you for your feedback and contribution!

Appendix 4
IDA Background Case Studies

4-1. DOE in TEMPs, T&E Concepts, Test Plans, and BLRIPS

4-2. Joint Chemical Agent Detector (JCAD) Test Design

4-3. Mobile Gun System (MGS) Case Study

4-4. Apache Block III

4-5. Integrated Defensive Electronic Countermeasures (IDECM)

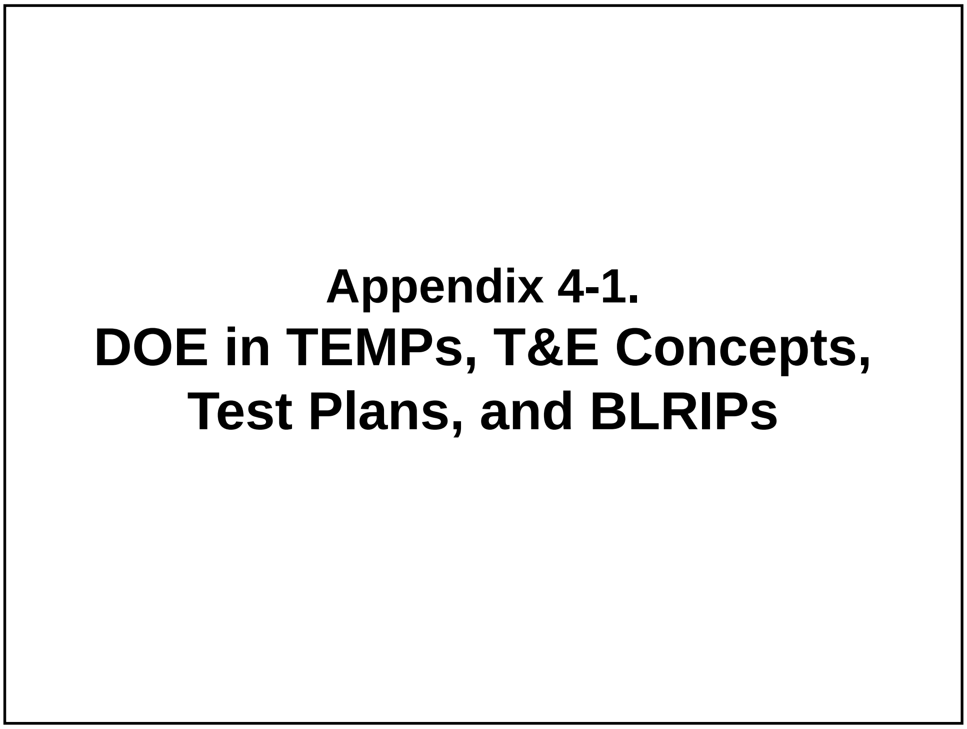
4-6. Censored Data Analysis Briefing

4-7. Excalibur Logistic Regression

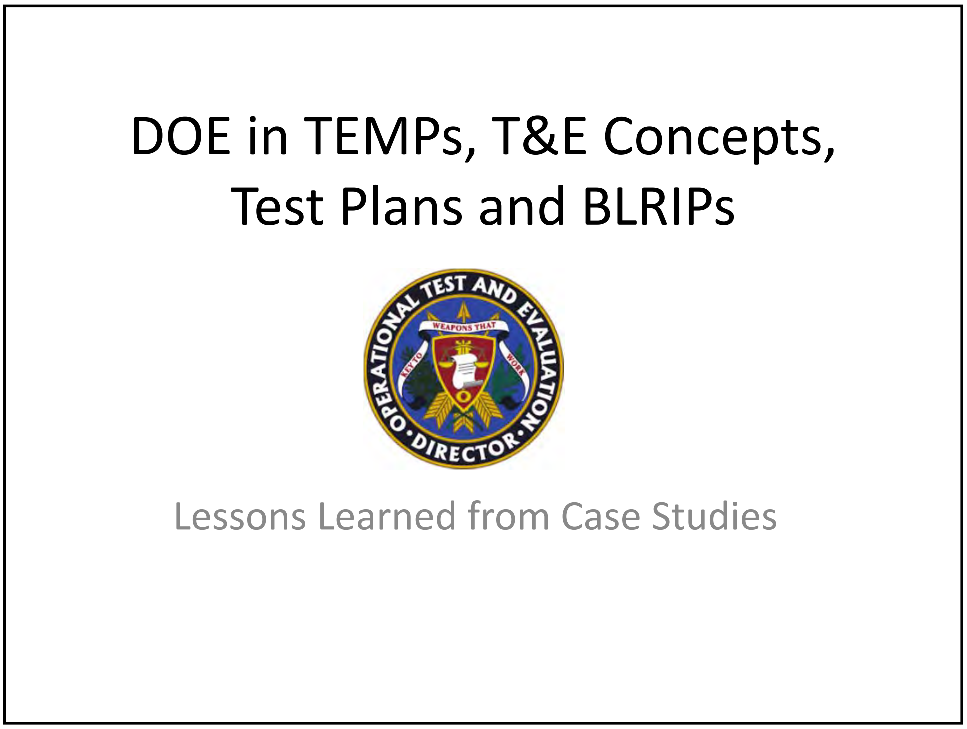
4-8. Stryker Reliability Case Study

4-9. Survey Case Study – Measuring Workload and Operator Latency: Command and Control Dynamic Targeting Cell

This page intentionally left blank.



Appendix 4-1.
DOE in TEMPs, T&E Concepts, Test Plans, and BLRIPs



DOE in TEMPs, T&E Concepts, Test Plans and BLRIPs

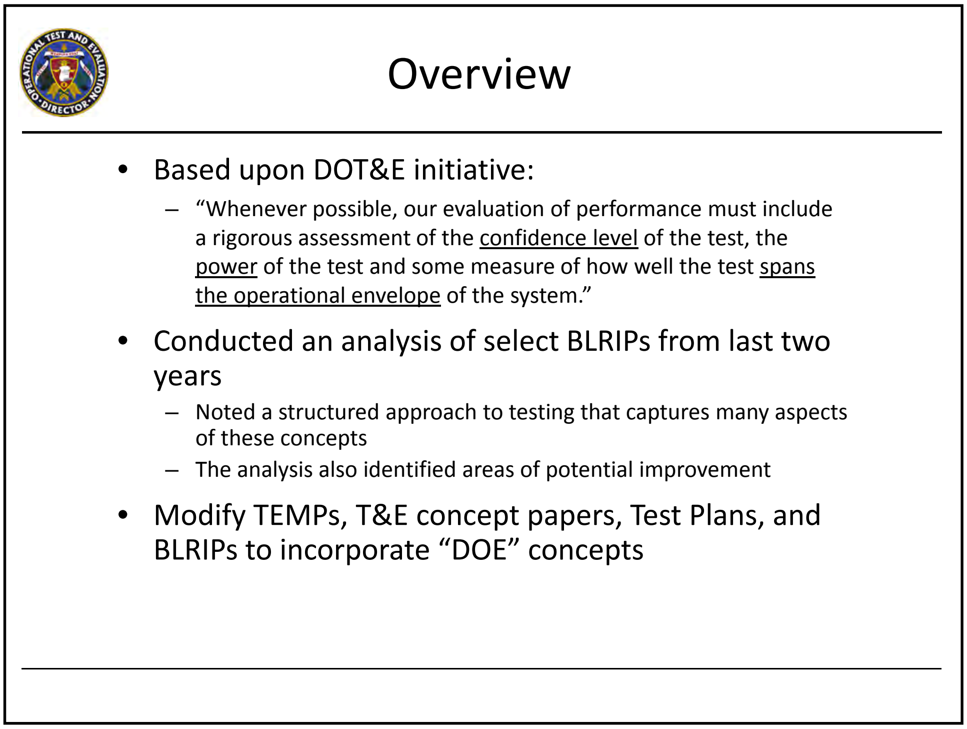
Lessons Learned from Case Studies

Purpose

- Discuss lessons learned from past tests
- Illustrate how DOE thinking can be applied to TEMPs, Test Plans, and other documents

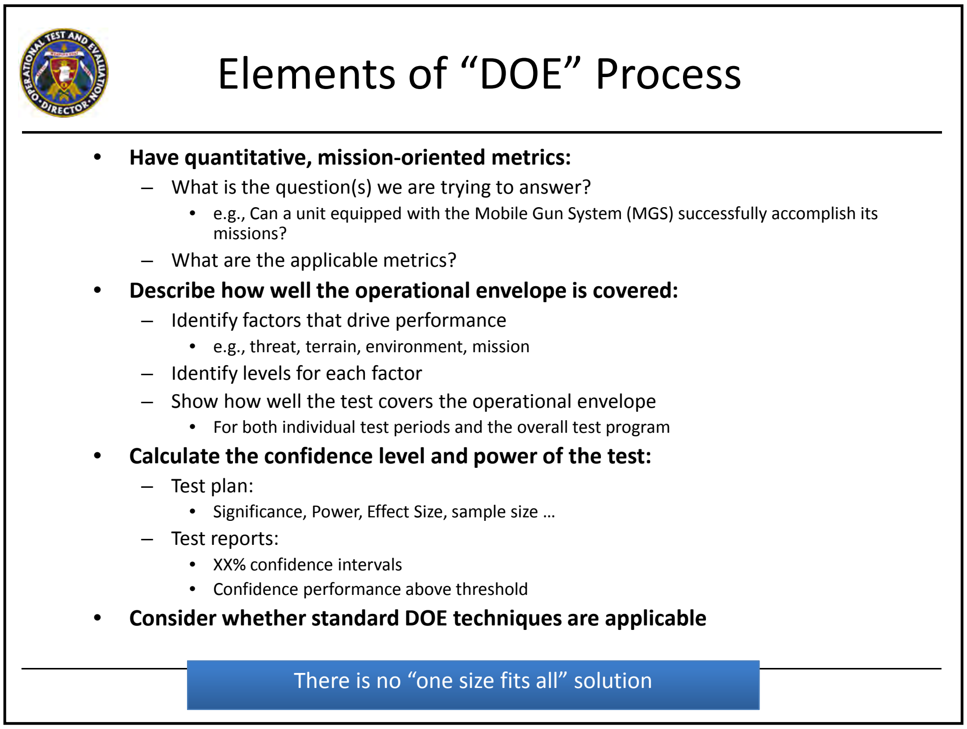
Outline

- Overview
- Elements of “DOE” Process
- Examples
  - Quantitative, Mission-Oriented Metrics
  - Coverage of Operational Envelope
  - Confidence and Power of Test
- Summary



Overview

- Based upon DOT&E initiative:
  - "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."
- Conducted an analysis of select BLRIPs from last two years
  - Noted a structured approach to testing that captures many aspects of these concepts
  - The analysis also identified areas of potential improvement
- Modify TEMPs, T&E concept papers, Test Plans, and BLRIPs to incorporate "DOE" concepts



Elements of "DOE" Process

- **Have quantitative, mission-oriented metrics:**
  - What is the question(s) we are trying to answer?
    - e.g., Can a unit equipped with the Mobile Gun System (MGS) successfully accomplish its missions?
  - What are the applicable metrics?
- **Describe how well the operational envelope is covered:**
  - Identify factors that drive performance
    - e.g., threat, terrain, environment, mission
  - Identify levels for each factor
  - Show how well the test covers the operational envelope
    - For both individual test periods and the overall test program
- **Calculate the confidence level and power of the test:**
  - Test plan:
    - Significance, Power, Effect Size, sample size …
  - Test reports:
    - XX% confidence intervals
    - Confidence performance above threshold
- **Consider whether standard DOE techniques are applicable**

There is no "one size fits all" solution

Quantitative, Mission-Oriented Metrics

Mission-Oriented Metrics

- Case studies identified several areas for potential improvement
- Ensure metrics and KPPs are measureable and testable
  - As defined, many are not, e.g., “The Mobile Gun System (MGS) primary armament must defeat a standard infantry bunker and create an opening in a double reinforced concrete wall, through which infantry can pass.”
- Mission-oriented metrics frequently do not have thresholds
  - Consider whether they should have a threshold
- Is the standard “at least as good as (or better than) the legacy system?”
  - Do you have quantitative data on the legacy system?

Look at metrics during JCIDS process

Surveys

- Surveys frequently have been qualitative and poorly designed
- There is a science behind survey design; use it
  - Be quantitative (e.g., Likert scale)
- During analysis, watch for discrepancies between numerical scores and written comments

Be careful with surveys

Coverage of the Operational Envelope

Coverage of the Operational Envelope

- 1st Step: Identify factors & levels of interest
  - In case studies, factors and levels of interest were sometimes specified; other times they were added in retrospective study.
- 2nd Step: Determine breadth of coverage of operational envelope
  - Tools illustrated in following examples: cross-tabular matrices, continuous plots, other graphical representations
  - These are examples, do not restrict yourself to these alone
- Power analysis can help determine if test design is sufficient
  - Next section of brief

Mobile Gun System (MGS)
Coverage of Operational Envelope

4 Factors: Mission Type, Terrain Type, Threat Level & Illumination

		Mission	Attack				Defend				Stability and Support				
Illum	OPFOR	Terrain	Urban	Mixed	Forest	Desert	Urban	Mixed	Forest	Desert	Urban	Mixed	Forest	Desert	
Day	Low		1	1		▨			🟩	▨	🟩		🟩	▨	2
Day	Med		1				1								2
Day	High		1				1	3							5
Night	Low					▨			🟩	▨	🟩	2	🟩	▨	2
Night	Med		2												2
Night	High			2			1								3
			5	3		▨	3	3	🟩	▨	🟩	2	🟩	▨	16

Weather: as it occurred; not controlled

Key

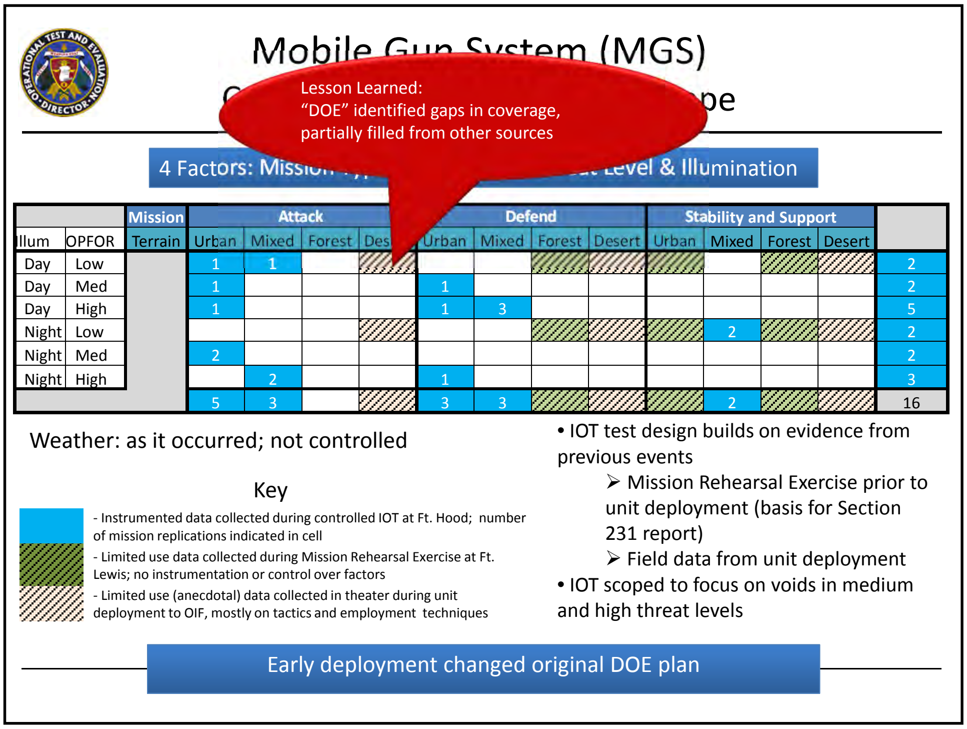
🟦 - Instrumented data collected during controlled IOT at Ft. Hood; number of mission replications indicated in cell

🟩 - Limited use data collected during Mission Rehearsal Exercise at Ft. Lewis; no instrumentation or control over factors

▨ - Limited use (anecdotal) data collected in theater during unit deployment to OIF, mostly on tactics and employment techniques

- IOT test design builds on evidence from previous events
  - ➢ Mission Rehearsal Exercise prior to unit deployment (basis for Section 231 report)
  - ➢ Field data from unit deployment
- IOT scoped to focus on voids in medium and high threat levels

Early deployment changed original DOE plan



Mobile Gun System (MGS)

C pe

Lesson Learned:
“DOE” identified gaps in coverage,
partially filled from other sources

4 Factors: Missio Level & Illumination

		Mission	Attack				Defend				Stability and Support				
Illum	OPFOR	Terrain	Urban	Mixed	Forest	Des	Urban	Mixed	Forest	Desert	Urban	Mixed	Forest	Desert	
Day	Low		1	1		▨			🟩	▨	🟩		🟩	▨	2
Day	Med		1				1								2
Day	High		1				1	3							5
Night	Low					▨			🟩	▨	🟩	2	🟩	▨	2
Night	Med		2												2
Night	High			2			1								3
			5	3		▨	3	3	🟩	▨	🟩	2	🟩	▨	16

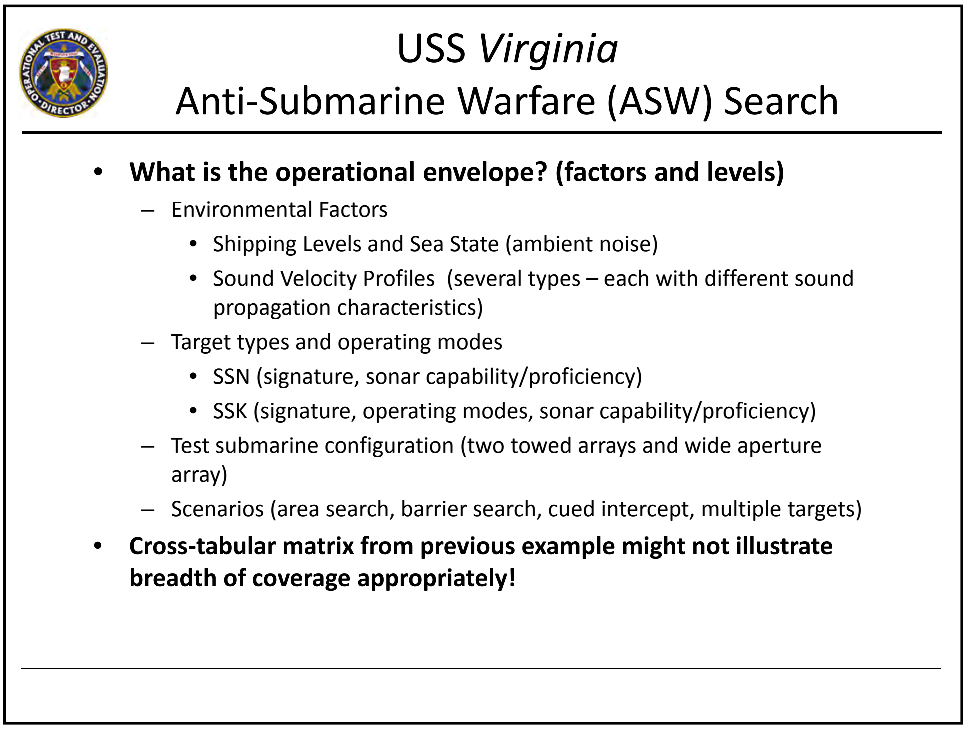
Weather: as it occurred; not controlled

Key

🟦	- Instrumented data collected during controlled IOT at Ft. Hood; number of mission replications indicated in cell
🟩	- Limited use data collected during Mission Rehearsal Exercise at Ft. Lewis; no instrumentation or control over factors
▨	- Limited use (anecdotal) data collected in theater during unit deployment to OIF, mostly on tactics and employment techniques

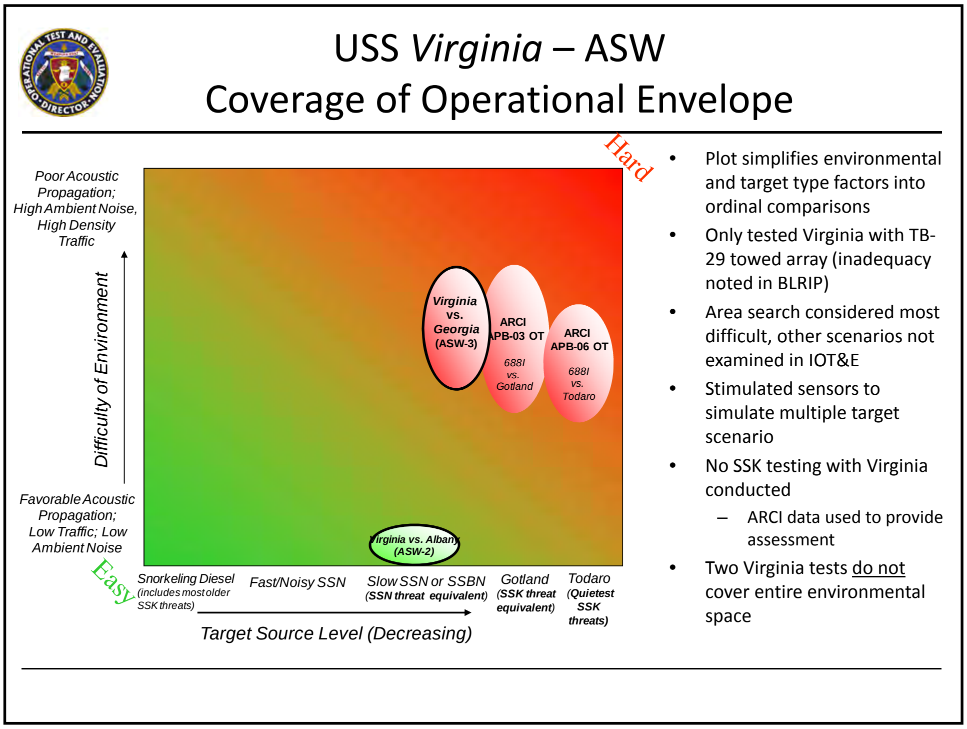
- IOT test design builds on evidence from previous events
  - ➢ Mission Rehearsal Exercise prior to unit deployment (basis for Section 231 report)
  - ➢ Field data from unit deployment
- IOT scoped to focus on voids in medium and high threat levels

Early deployment changed original DOE plan



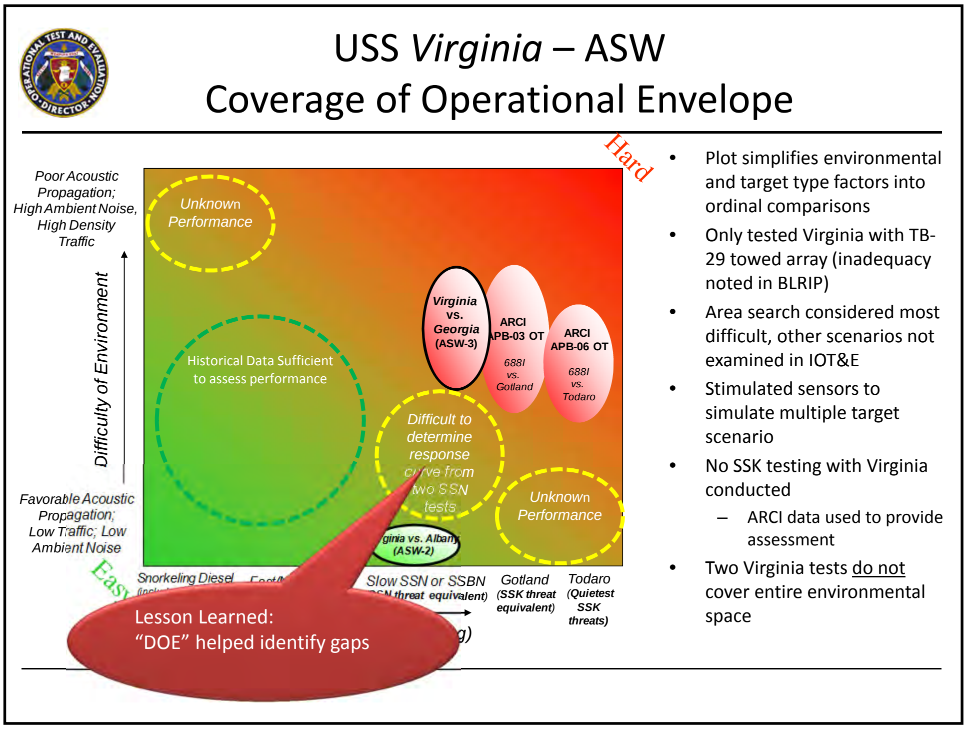
USS *Virginia*
Anti-Submarine Warfare (ASW) Search

- **What is the operational envelope? (factors and levels)**
  - Environmental Factors
    - Shipping Levels and Sea State (ambient noise)
    - Sound Velocity Profiles (several types – each with different sound propagation characteristics)
  - Target types and operating modes
    - SSN (signature, sonar capability/proficiency)
    - SSK (signature, operating modes, sonar capability/proficiency)
  - Test submarine configuration (two towed arrays and wide aperture array)
  - Scenarios (area search, barrier search, cued intercept, multiple targets)
- **Cross-tabular matrix from previous example might not illustrate breadth of coverage appropriately!**



USS *Virginia* – ASW Coverage of Operational Envelope

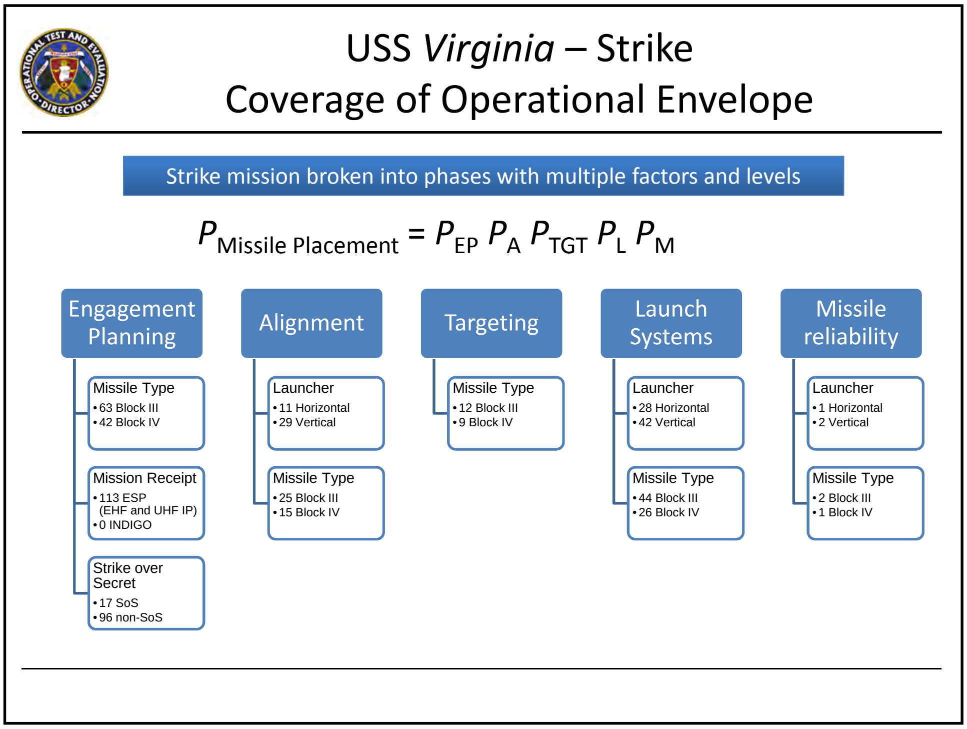
- Plot simplifies environmental and target type factors into ordinal comparisons
- Only tested Virginia with TB-29 towed array (inadequacy noted in BLRIP)
- Area search considered most difficult, other scenarios not examined in IOT&E
- Stimulated sensors to simulate multiple target scenario
- No SSK testing with Virginia conducted
    - ARCI data used to provide assessment
- Two Virginia tests <u>do not</u> cover entire environmental space



USS *Virginia* – ASW Coverage of Operational Envelope

Lesson Learned:
"DOE" helped identify gaps

- Plot simplifies environmental and target type factors into ordinal comparisons
- Only tested Virginia with TB-29 towed array (inadequacy noted in BLRIP)
- Area search considered most difficult, other scenarios not examined in IOT&E
- Stimulated sensors to simulate multiple target scenario
- No SSK testing with Virginia conducted
    - ARCI data used to provide assessment
- Two Virginia tests <u>do not</u> cover entire environmental space



USS *Virginia* – Strike Coverage of Operational Envelope

Strike mission broken into phases with multiple factors and levels

$$P_{\text{Missile Placement}} = P_{\text{EP}}\, P_{\text{A}}\, P_{\text{TGT}}\, P_{\text{L}}\, P_{\text{M}}$$

Engagement Planning

- Missile Type
  - 63 Block III
  - 42 Block IV
- Mission Receipt
  - 113 ESP (EHF and UHF IP)
  - 0 INDIGO
- Strike over Secret
  - 17 SoS
  - 96 non-SoS

Alignment

- Launcher
  - 11 Horizontal
  - 29 Vertical
- Missile Type
  - 25 Block III
  - 15 Block IV

Targeting

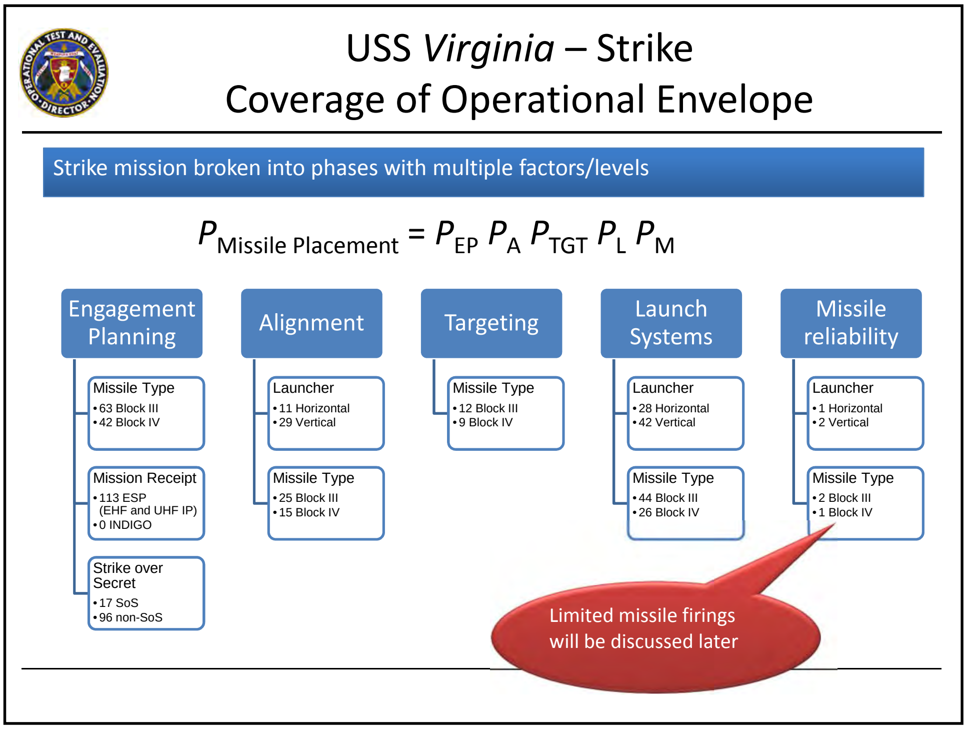
- Missile Type
  - 12 Block III
  - 9 Block IV

Launch Systems

- Launcher
  - 28 Horizontal
  - 42 Vertical
- Missile Type
  - 44 Block III
  - 26 Block IV

Missile reliability

- Launcher
  - 1 Horizontal
  - 2 Vertical
- Missile Type
  - 2 Block III
  - 1 Block IV



USS *Virginia* – Strike Coverage of Operational Envelope

Strike mission broken into phases with multiple factors/levels

$$P_{\text{Missile Placement}} = P_{\text{EP}}\, P_{\text{A}}\, P_{\text{TGT}}\, P_{\text{L}}\, P_{\text{M}}$$

Engagement Planning

- Missile Type
  - 63 Block III
  - 42 Block IV
- Mission Receipt
  - 113 ESP (EHF and UHF IP)
  - 0 INDIGO
- Strike over Secret
  - 17 SoS
  - 96 non-SoS

Alignment

- Launcher
  - 11 Horizontal
  - 29 Vertical
- Missile Type
  - 25 Block III
  - 15 Block IV

Targeting

- Missile Type
  - 12 Block III
  - 9 Block IV

Launch Systems

- Launcher
  - 28 Horizontal
  - 42 Vertical
- Missile Type
  - 44 Block III
  - 26 Block IV

Missile reliability

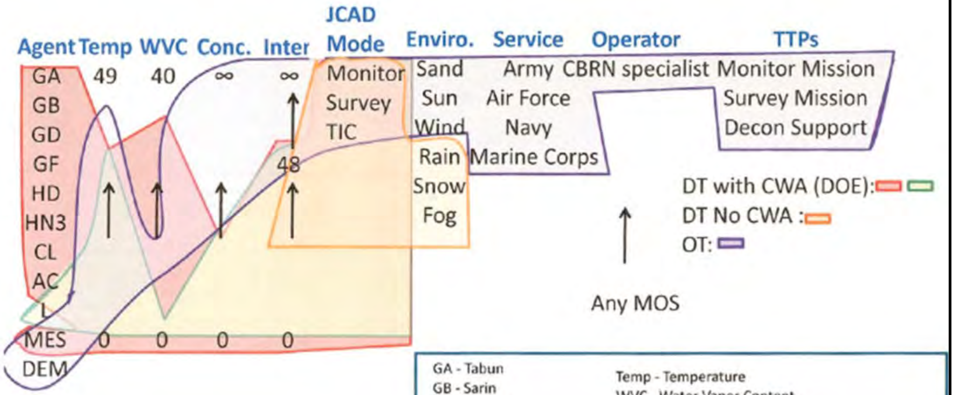
- Launcher
  - 1 Horizontal
  - 2 Vertical
- Missile Type
  - 2 Block III
  - 1 Block IV

Limited missile firings will be discussed later

Joint Chemical Agent Detector (JCAD)

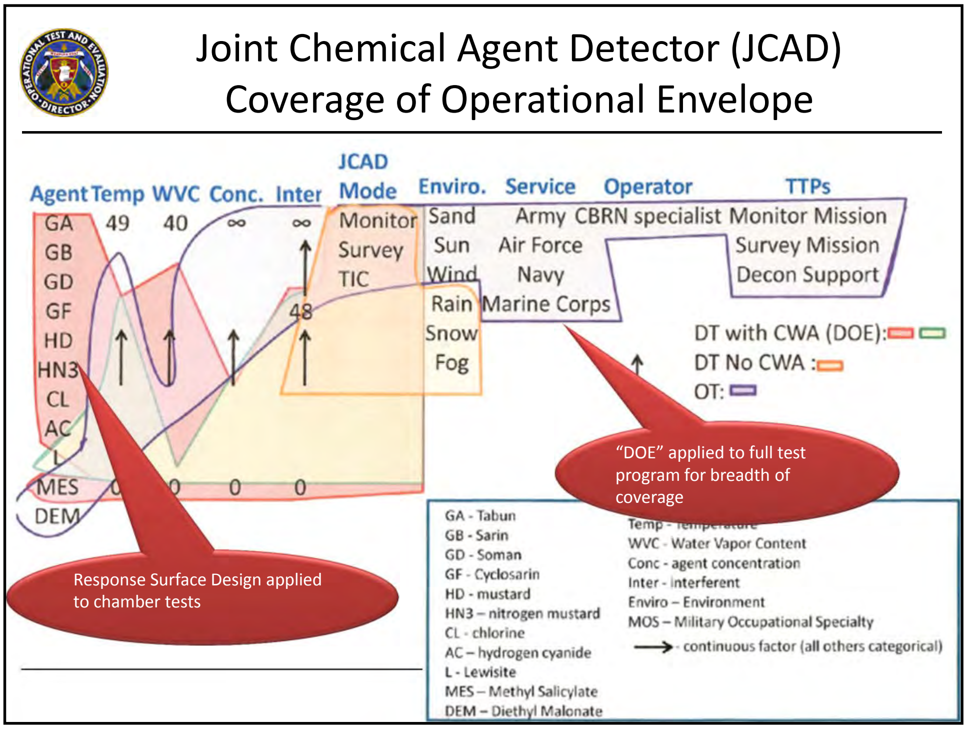
- What is the operational envelope? (factors and levels)
  - Agent (9 agents and 2 simulants)
  - Temperature, water vapor concentration, agent concentration, interferent (continuous)
  - Environment (sand, sun, wind, rain, snow, fog)
  - Service (Army, Air Force, Navy, Marine Corps)
  - JCAD Mode (Monitor, Survey, TIC)
  - Operator (Any MOS to CBRN Specialist)
  - TTP (Monitor Mission, Survey Mission, Decon Support)

Joint Chemical Agent Detector (JCAD) Coverage of Operational Envelope



GA - Tabun
GB - Sarin
GD - Soman
GF - Cyclosarin
HD - mustard
HN3 – nitrogen mustard
CL - chlorine
AC – hydrogen cyanide
L - Lewisite
MES – Methyl Salicylate
DEM – Diethyl Malonate

Temp - Temperature
WVC - Water Vapor Content
Conc - agent concentration
Inter - interferent
Enviro – Environment
MOS – Military Occupational Specialty
→ - continuous factor (all others categorical)



Joint Chemical Agent Detector (JCAD)
Coverage of Operational Envelope

Agent | Temp | WVC | Conc. | Inter | JCAD Mode | Enviro. | Service | Operator | TTPs

Agent: GA, GB, GD, GF, HD, HN3, CL, AC, L, MES, DEM

Temp: 49 … 0

WVC: 40 … 0

Conc.: ∞ … 0

Inter: ∞ … 48 … 0

JCAD Mode: Monitor, Survey, TIC

Enviro.: Sand, Sun, Wind, Rain, Snow, Fog

Service: Army, Air Force, Navy, Marine Corps

Operator: CBRN specialist

TTPs: Monitor Mission, Survey Mission, Decon Support

DT with CWA (DOE): 🟥 🟩

DT No CWA : 🟧

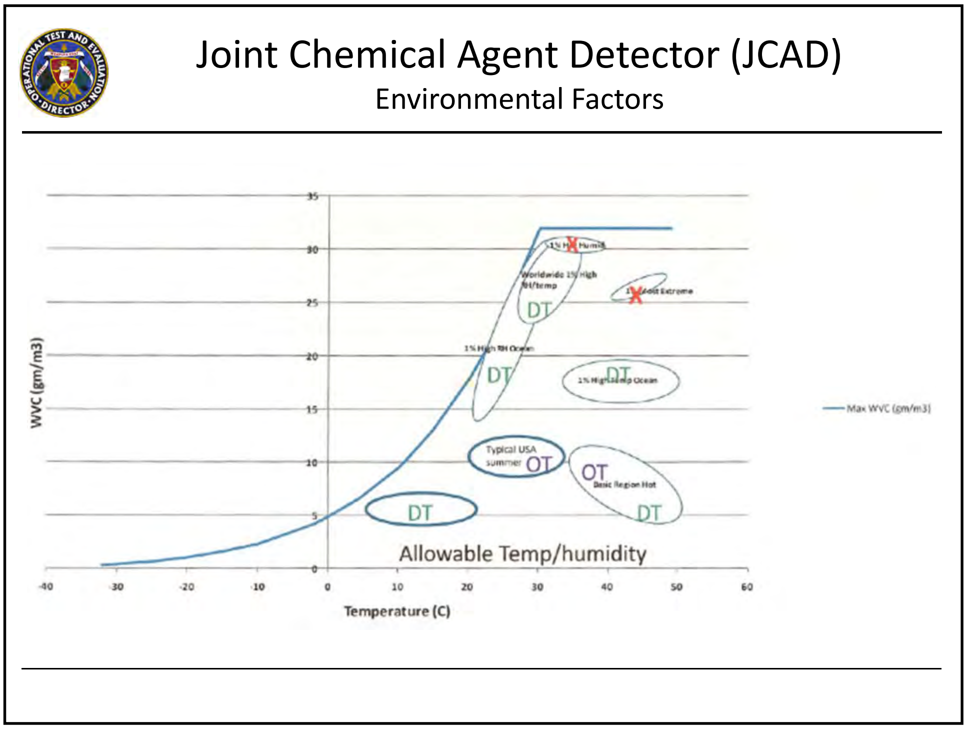
OT: 🟪

"DOE" applied to full test program for breadth of coverage

Response Surface Design applied to chamber tests

GA - Tabun
GB - Sarin
GD - Soman
GF - Cyclosarin
HD - mustard
HN3 – nitrogen mustard
CL - chlorine
AC – hydrogen cyanide
L - Lewisite
MES – Methyl Salicylate
DEM – Diethyl Malonate

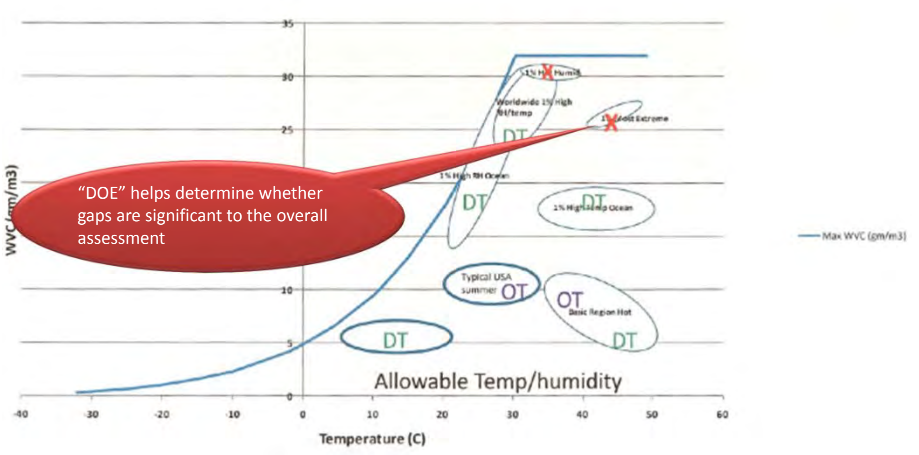
Temp - Temperature
WVC - Water Vapor Content
Conc - agent concentration
Inter - interferent
Enviro – Environment
MOS – Military Occupational Specialty
⟶ continuous factor (all others categorical)



Joint Chemical Agent Detector (JCAD)
Environmental Factors

Joint Chemical Agent Detector (JCAD)

Environmental Factors



"DOE" helps determine whether gaps are significant to the overall assessment

Confidence and Power of Test

Confidence and Power of Test

- Test Planning vs. Test Reporting
- Test Planning
  - What confidence level is needed?
  - Construct power of test – does the test have a high probability of detecting important differences?
- Test Reporting
  - Provide confidence intervals for all results.
  - Provide confidence above threshold when required.

Joint Chemical Agent Detector (JCAD) Power of Test

- Power Analysis for JCAD Chamber Test
  - DT Testing
  - Statistical Response Surface Design (I-Optimal)
  - High power test plan

Factor	S:N* = 0.5	S:N = 1.0	S:N = 2.0
Temperature	32.0%	84.7%	99.9%
Water Vapor Content (WVC)	42.1%	94.1%	99.9%
Concentration	46.5%	96.3%	99.9%

*S:N – signal-to-noise ratio, goal detectable difference as a ratio to the design standard deviation

Mobile Gun System (MGS) Power of Test

- Original Test Plan
  (Sample Size = 22)

Factor	S:N* = 0.5	S:N = 1.0	S:N = 2.0
Mission Type	7.7%	16.6%	54.1%
Terrain Type	17.0%	51.3%	97.8%
Threat Level	9.4%	24.4%	75.5%
Illumin.	15.9%	47.9%	96.7%

- DOE Interrupted by Deployment
  (Sample Size = 16)

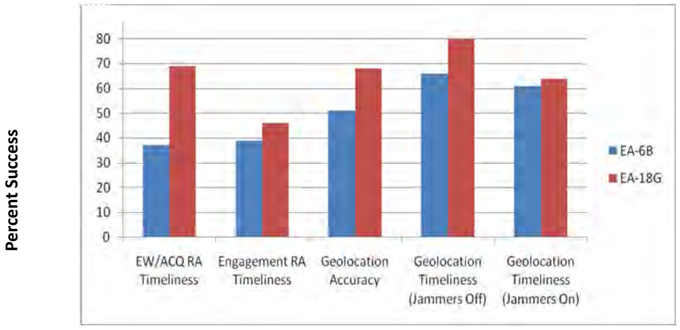
Factor	S:N = 0.5	S:N = 1.0	S:N = 2.0
Mission Type	5.7%	8.1%	18.3%
Terrain Type	10.6%	28.0%	78.2%
Threat Level	6.4%	10.9%	31.2%
Illumin.	10.1%	26.0%	74.3%

*S:N – signal-to-noise ratio, goal detectable difference as a ratio to the design standard deviation

Lesson Learned: smaller sample size decreases power

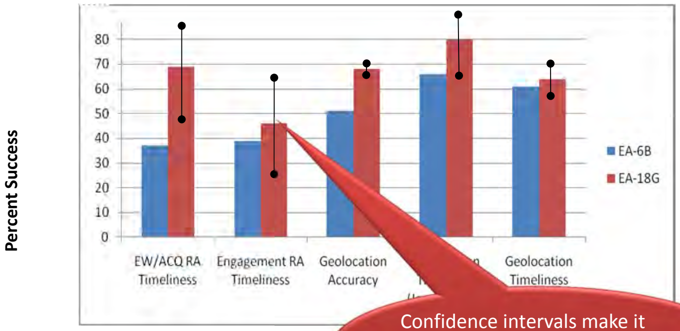
EA-18G/EA-6B Comparison Confidence Intervals

Figure from DOT&E EA-18G BLRIP



EA-18G/EA-6B Comparison Confidence Intervals

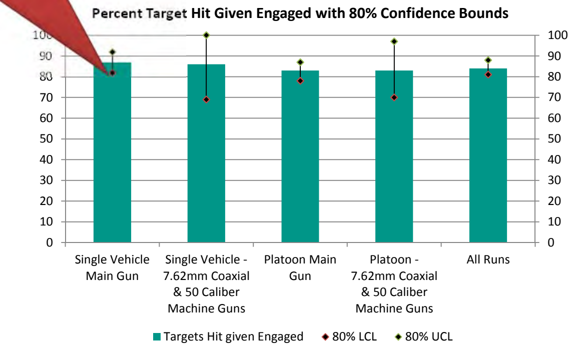
Figure from DOT&E EA-18G BLRIP

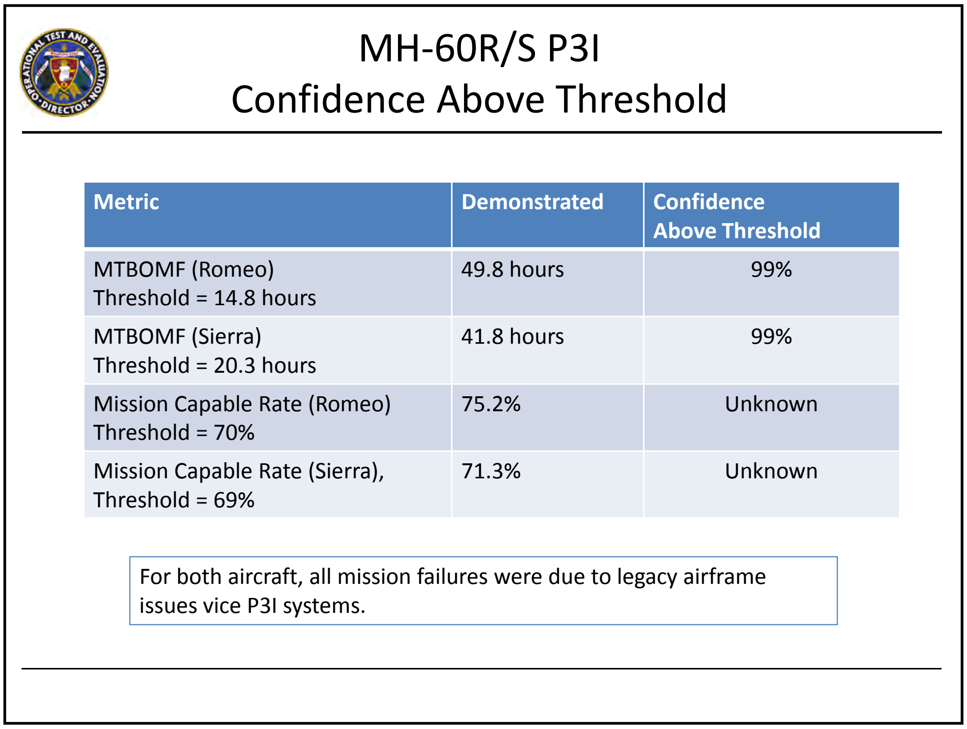


Confidence intervals make it clear that performance is comparable

Mobile Gun System (MGS) Confidence Intervals

Even without a threshold, confidence intervals quantifies how well the metric was measured

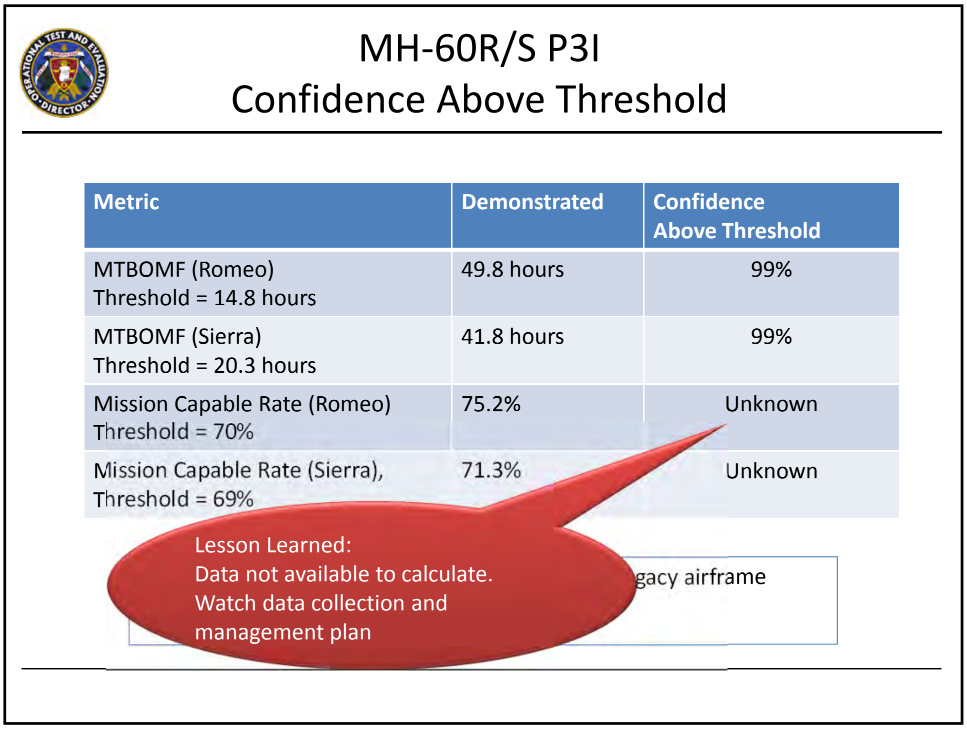




MH-60R/S P3I
Confidence Above Threshold

Metric	Demonstrated	Confidence Above Threshold
MTBOMF (Romeo) Threshold = 14.8 hours	49.8 hours	99%
MTBOMF (Sierra) Threshold = 20.3 hours	41.8 hours	99%
Mission Capable Rate (Romeo) Threshold = 70%	75.2%	Unknown
Mission Capable Rate (Sierra), Threshold = 69%	71.3%	Unknown

For both aircraft, all mission failures were due to legacy airframe issues vice P3I systems.

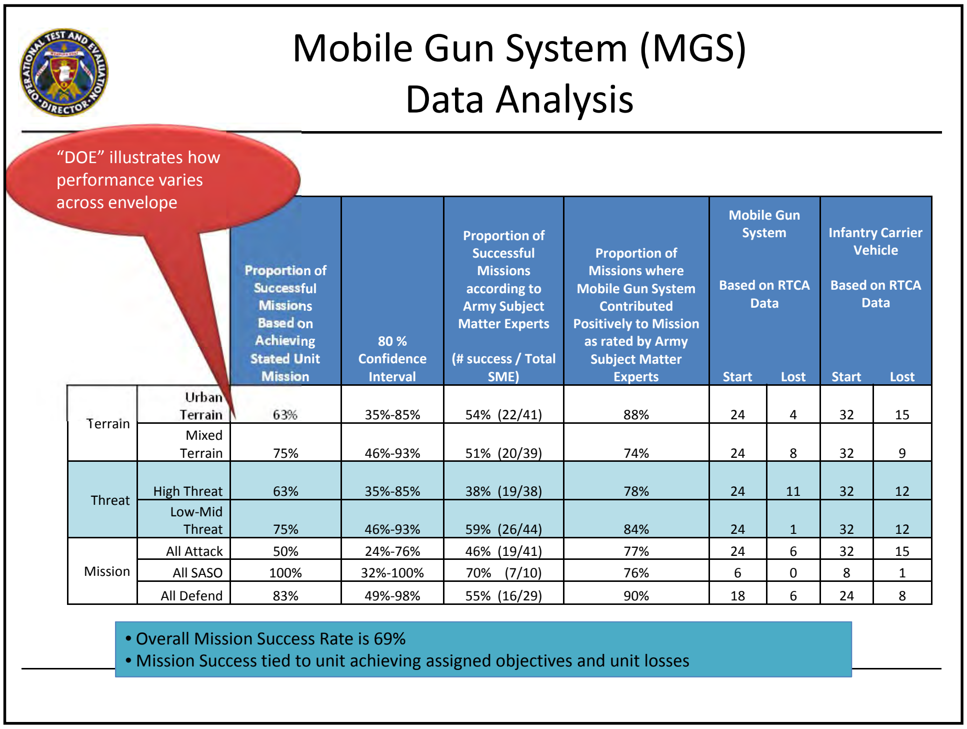


MH-60R/S P3I
Confidence Above Threshold

Metric	Demonstrated	Confidence Above Threshold
MTBOMF (Romeo) Threshold = 14.8 hours	49.8 hours	99%
MTBOMF (Sierra) Threshold = 20.3 hours	41.8 hours	99%
Mission Capable Rate (Romeo) Threshold = 70%	75.2%	Unknown
Mission Capable Rate (Sierra), Threshold = 69%	71.3%	Unknown

Lesson Learned:
Data not available to calculate.
Watch data collection and
management plan

gacy airframe

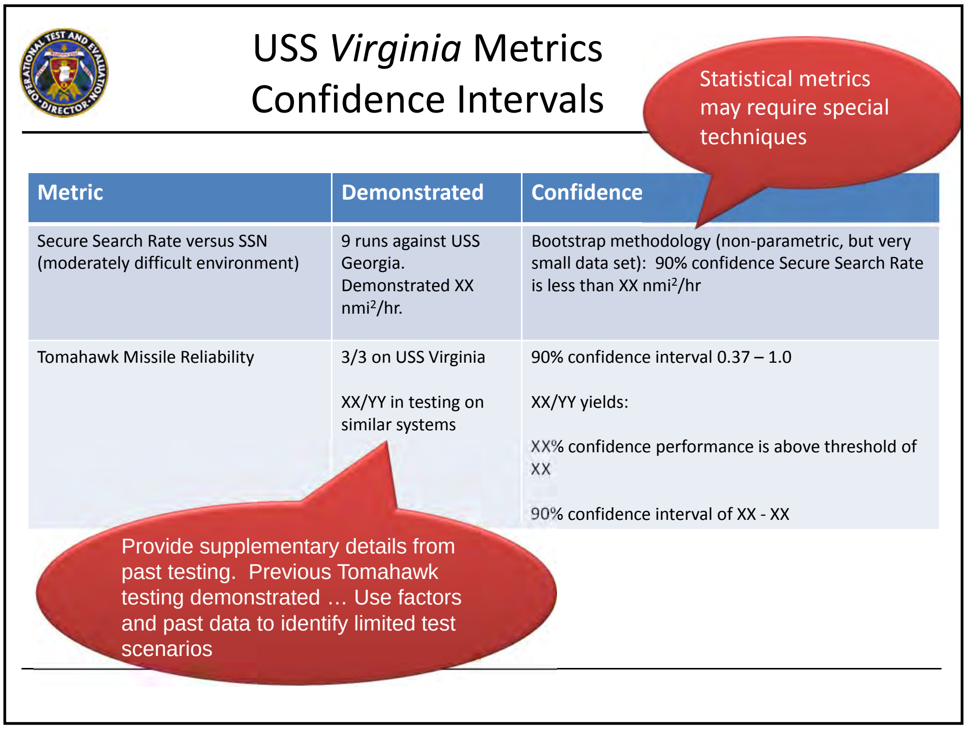


Mobile Gun System (MGS) Data Analysis

"DOE" illustrates how performance varies across envelope

		Proportion of Successful Missions Based on Achieving Stated Unit Mission	80 % Confidence Interval	Proportion of Successful Missions according to Army Subject Matter Experts (# success / Total SME)	Proportion of Missions where Mobile Gun System Contributed Positively to Mission as rated by Army Subject Matter Experts	Mobile Gun System Based on RTCA Data		Infantry Carrier Vehicle Based on RTCA Data	
						Start	Lost	Start	Lost
Terrain	Urban Terrain	63%	35%-85%	54% (22/41)	88%	24	4	32	15
	Mixed Terrain	75%	46%-93%	51% (20/39)	74%	24	8	32	9
Threat	High Threat	63%	35%-85%	38% (19/38)	78%	24	11	32	12
	Low-Mid Threat	75%	46%-93%	59% (26/44)	84%	24	1	32	12
Mission	All Attack	50%	24%-76%	46% (19/41)	77%	24	6	32	15
	All SASO	100%	32%-100%	70% (7/10)	76%	6	0	8	1
	All Defend	83%	49%-98%	55% (16/29)	90%	18	6	24	8

- Overall Mission Success Rate is 69%
- Mission Success tied to unit achieving assigned objectives and unit losses



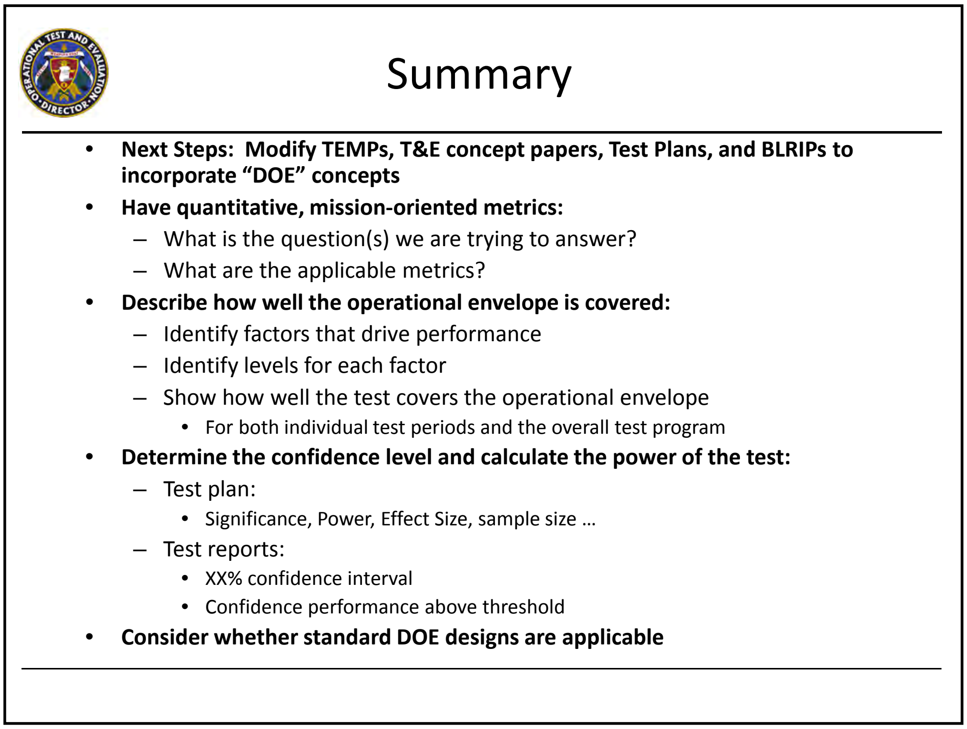
USS *Virginia* Metrics Confidence Intervals

Statistical metrics may require special techniques

Metric	Demonstrated	Confidence
Secure Search Rate versus SSN (moderately difficult environment)	9 runs against USS Georgia. Demonstrated XX nmi^2/hr.	Bootstrap methodology (non-parametric, but very small data set): 90% confidence Secure Search Rate is less than XX nmi^2/hr
Tomahawk Missile Reliability	3/3 on USS Virginia XX/YY in testing on similar systems	90% confidence interval 0.37 – 1.0 XX/YY yields: XX% confidence performance is above threshold of XX 90% confidence interval of XX - XX

Provide supplementary details from past testing. Previous Tomahawk testing demonstrated … Use factors and past data to identify limited test scenarios

Summary



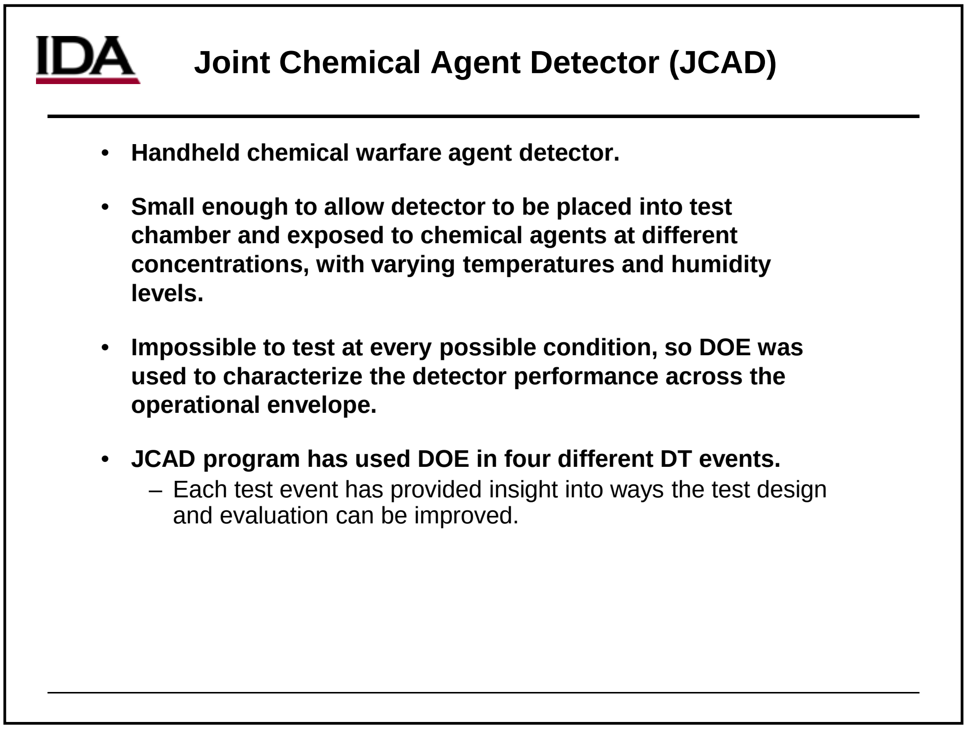
Summary

- **Next Steps: Modify TEMPs, T&E concept papers, Test Plans, and BLRIPs to incorporate “DOE” concepts**
- **Have quantitative, mission-oriented metrics:**
  - What is the question(s) we are trying to answer?
  - What are the applicable metrics?
- **Describe how well the operational envelope is covered:**
  - Identify factors that drive performance
  - Identify levels for each factor
  - Show how well the test covers the operational envelope
    - For both individual test periods and the overall test program
- **Determine the confidence level and calculate the power of the test:**
  - Test plan:
    - Significance, Power, Effect Size, sample size …
  - Test reports:
    - XX% confidence interval
    - Confidence performance above threshold
- **Consider whether standard DOE designs are applicable**

Appendix 4-2.
Joint Chemical Agent Detector (JCAD) Test Design

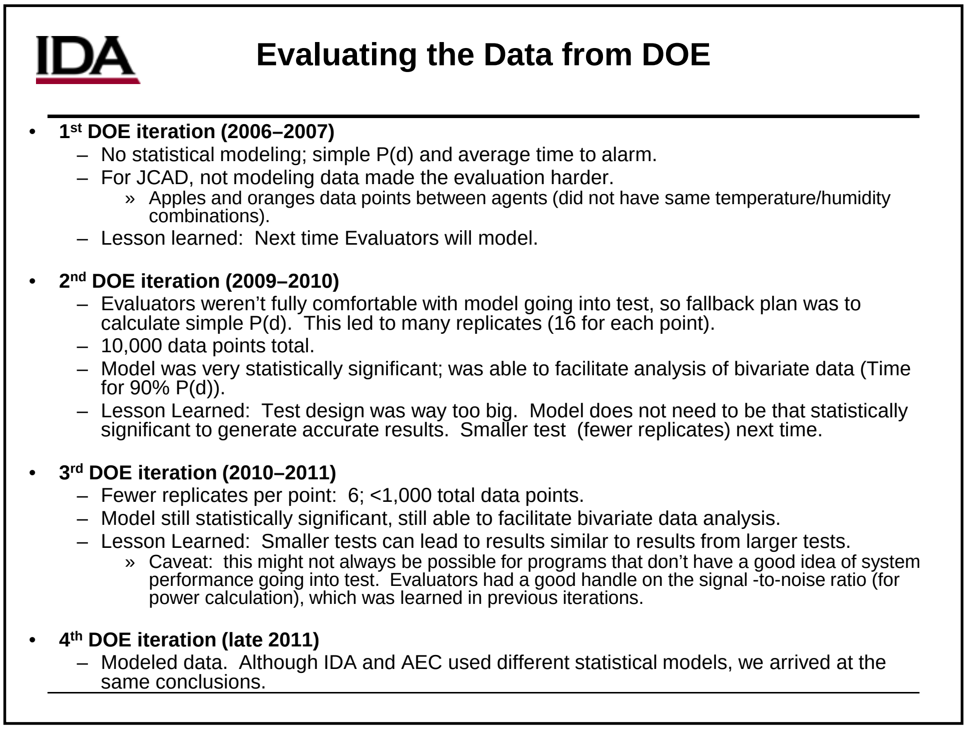
Case Study:
Joint Chemical Agent Detector DOE Analysis

IDA



Joint Chemical Agent Detector (JCAD)

- **Handheld chemical warfare agent detector.**
- **Small enough to allow detector to be placed into test chamber and exposed to chemical agents at different concentrations, with varying temperatures and humidity levels.**
- **Impossible to test at every possible condition, so DOE was used to characterize the detector performance across the operational envelope.**
- **JCAD program has used DOE in four different DT events.**
  - Each test event has provided insight into ways the test design and evaluation can be improved.



Evaluating the Data from DOE

- **1st DOE iteration (2006–2007)**
  - No statistical modeling; simple P(d) and average time to alarm.
  - For JCAD, not modeling data made the evaluation harder.
    - » Apples and oranges data points between agents (did not have same temperature/humidity combinations).
  - Lesson learned: Next time Evaluators will model.
- **2nd DOE iteration (2009–2010)**
  - Evaluators weren't fully comfortable with model going into test, so fallback plan was to calculate simple P(d). This led to many replicates (16 for each point).
  - 10,000 data points total.
  - Model was very statistically significant; was able to facilitate analysis of bivariate data (Time for 90% P(d)).
  - Lesson Learned: Test design was way too big. Model does not need to be that statistically significant to generate accurate results. Smaller test (fewer replicates) next time.
- **3rd DOE iteration (2010–2011)**
  - Fewer replicates per point: 6; <1,000 total data points.
  - Model still statistically significant, still able to facilitate bivariate data analysis.
  - Lesson Learned: Smaller tests can lead to results similar to results from larger tests.
    - » Caveat: this might not always be possible for programs that don't have a good idea of system performance going into test. Evaluators had a good handle on the signal -to-noise ratio (for power calculation), which was learned in previous iterations.
- **4th DOE iteration (late 2011)**
  - Modeled data. Although IDA and AEC used different statistical models, we arrived at the same conclusions.

JCAD DOE Overview

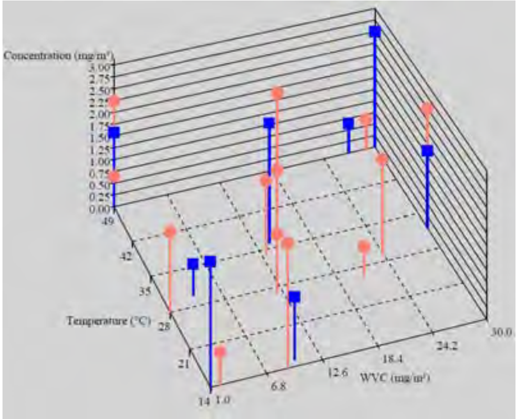
	FACTORS				
LEVELS	Temperature	Relative Humidity	Agent Concentration *	Detector Mode	Detector Type
	49°C	100%	High	Monitor	Legacy Detector
	↑	↑	↑	Survey	JCAD
	5°C	5%	Low		

** Range different for each agent*

<u>Response Variables (user requirement):</u>

Probability of Detection
Time to Alarm
Time to Reset after Alarm

Space-filling Model of JCAD DOE Matrix Points



Test Limitations:

- *Chamber can't go below 5°C or above 80% relative humidity (30 mg/m^3 water vapor content).*
- *Time and money constraints, and chamber limitations prevent randomization (one agent at a time, from low to high temperature).*

JCAD DOE Evaluation Without Modeling

- 1st DOE evaluation
- Standard statistical methods, no modeling

Agent	Conc (mg/m³)	Environmental Conditions			JCAD			
		Temp (ºC)	RH (%)	WVC (g/m³)	# Detects/ # Valid DO	P(D) (%)	80% LCB	Avg Time to Alarm (mm:ss)
R	B6	8	49	4	19/22	86.4	76.3	4:06
		36	69	29	30/30	100	94.8	0:10
T	B8	8	49	4	18/18	100	91.4	0:06
		36	69	29	24/24	100	93.5	0:00**
U	B0	8	49	4	19/24	79.2	69.1	1:44
		36	69	29	24/24	100	93.5	0:16
	K0	20	74	13	24/24	100	93.5	1:34
W	B0	8	49	4	16/20	80.0	68.7	6:16
		36	69	29	18/18	100	91.4	1:43
Q	A5	8	0	0	0/24	0	0	N/A
		36	0	0	1/18	5.6	1.2	6:36
	< K0	8	0	0	15/18	83.3	71.5	1:23
	K0	36	0	0	18/18	100	91.4	0:44
F	L0	20	74	13	24/24	100	93.5	0:48
		36	0	0	30/30	100	94.8	1:49
	T25	20	74	13	18/18	100	91.4	0:29
		36	69	29	17/18	94.4	84.3	0:06
L	L0	36	0	0	35/35	100	95.5	0:08
	T25	8	0	0	18/18	100	91.4	0:13
		36	0	0	18/18	100	91.4	0:04

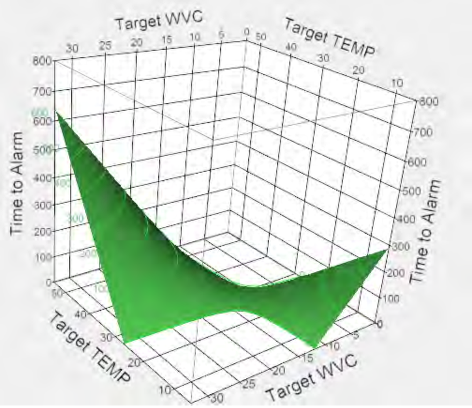
Which condition do you pick to determine effectiveness?

Which metric do you pick to determine effectiveness?

- *Test designed using DOE methods, but evaluated using standard statistics*
- ***User requirement : 90% P(d) within 30 seconds***
- *How do you evaluate against the requirement with this analysis?*

JCAD DOE Evaluation With Modeling – Many Ways of Depicting Same Data

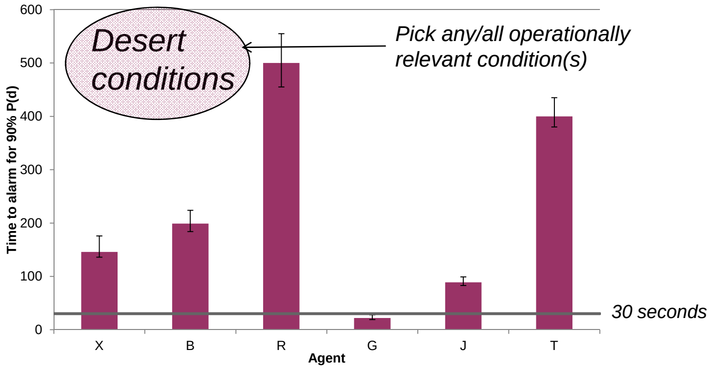
- *Example: Generate Response Surface Curve*
- *Displays performance over the entire operational envelope*



Performance at a Specified Concentration (e.g., user requirement)

JCAD DOE Evaluation With Modeling – Cont'd

Example: Reliability/Survivability Modeling



Desert conditions

Pick any/all operationally relevant condition(s)

30 seconds

- *Bivariate data analysis – Time to achieve a specified Probability of Detection*
- *Allows the evaluator to pick any condition of interest, even if the system wasn't specifically tested in that condition.*
- *Allows a direct comparison to the requirement with 95% confidence*
- *Can generate modeling results within minutes despite very large amounts of data (> 10,000 individual records).*

JCAD DOE Evaluation with Modeling – Cont'd

- *Generate model equation to estimate performance at specified conditions*
- *Compare performance between detectors/modes*

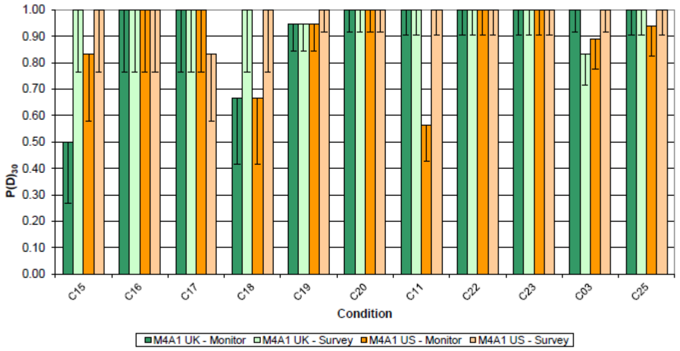


Figure 3-1. Agent C P(D) Within 30 Seconds.

IDA

Determining Test Adequacy: Chemical Agent Detector

- **Goal: Determine the probability of detection within one minute**
  - Threshold is least 90% within one minute
- **Metric (response variables) :**
  - Detect (Yes/No)
  - Detection time (seconds)
- **Factors to consider:**
  - Temperature, water vapor concentration, agent concentration, agent type
- **Notional test design: Full factorial (2^4)**

DOE Matrix

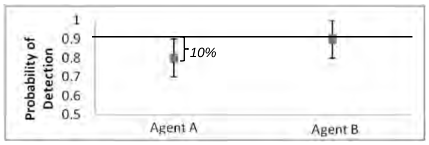
Agent Type	Agent Concentration	Low Temperature		High Temperature		Agent Type	Agent Concentration	Low Temperature		High Temperature	
		Low WVC	High WVC	Low WVC	High WVC			Low WVC	High WVC	Low WVC	High WVC
A	Low	?	?	?	?	B	Low	?	?	?	?
	High	?	?	?	?		High	?	?	?	?

What sample size is do we need to determine probability of detection?

IDA

Adequate Test Resources

- **Goal: Determine an adequate sample size to determine a 10% change in probability of detection across the operational envelope?**
  - For example, for each agent type can we conclude we meet the requirement?

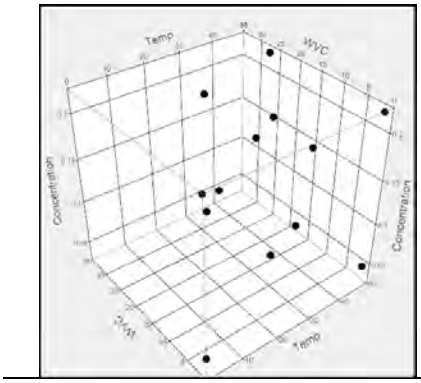


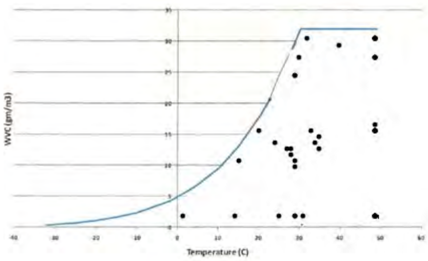
- **Assumptions:**
  - Detectable difference = 10%
  - 90% Confidence Level, 80% Power
- **Results:**
  - Binomial response (detect/non-detect):
    - » 14 replications of full factorial (**224 total test points**)
  - Continuous response (time until detection):
    - » 5 replications of full factorial **(80 total test points**) – 65% reduction in test cases!

This example results in a 65% reduction in test cases!

Example Analysis: Chemical Agent Detector

- **Design points from Chemical Agent Test are shown below**
  - Employed an optimal design methodology
  - Responses times are hypothetical
  - What is the implication in test analysis?





Chemical Agent Detector Results

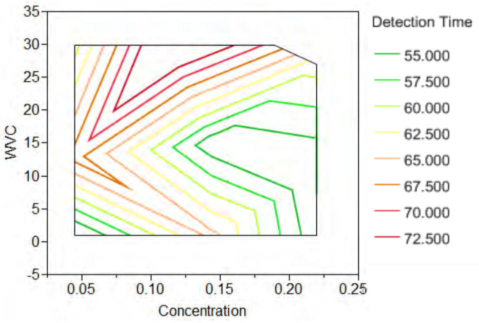
(notional analysis – not based on actual data)

- **Data determine significant factors:**

Factor	Model Coefficient Estimate	Standard Error	F-Ratio	P-value
Temperature	-7.07	1.30	29.7	<0.001
Water Vapor Content	5.13	1.06	23.6	<0.001
Agent Concentration	5.13	2.01	96.5	<0.001
Agent Type	N/A	N/A	4.34	<0.001

- **Allows for understanding of performance across the operational envelope.**

- **Note: All results are for Illustration only**

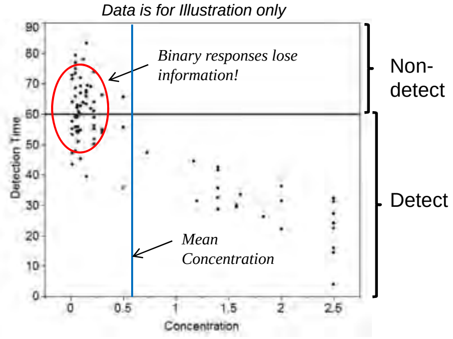


IDA

Chemical Agent Detector Results

- Estimate the probability of detection at 60 seconds at the mean concentration
- Detection times and detect/non-detect information recorded
- Binary analysis results in **300% increase** in confidence interval width

Data is for Illustration only



Response	Probability of Detection within 60 seconds at mean	Lower 90% Confidence Bound	Upper 90% Confidence Bound	Confidence Interval Width
Binary (Detect: Yes/No)	83.5%	60.5%	94.4%	33.9%
Continuous (Time)	91.0%	86.3%	94.5%	8.2%

IDA

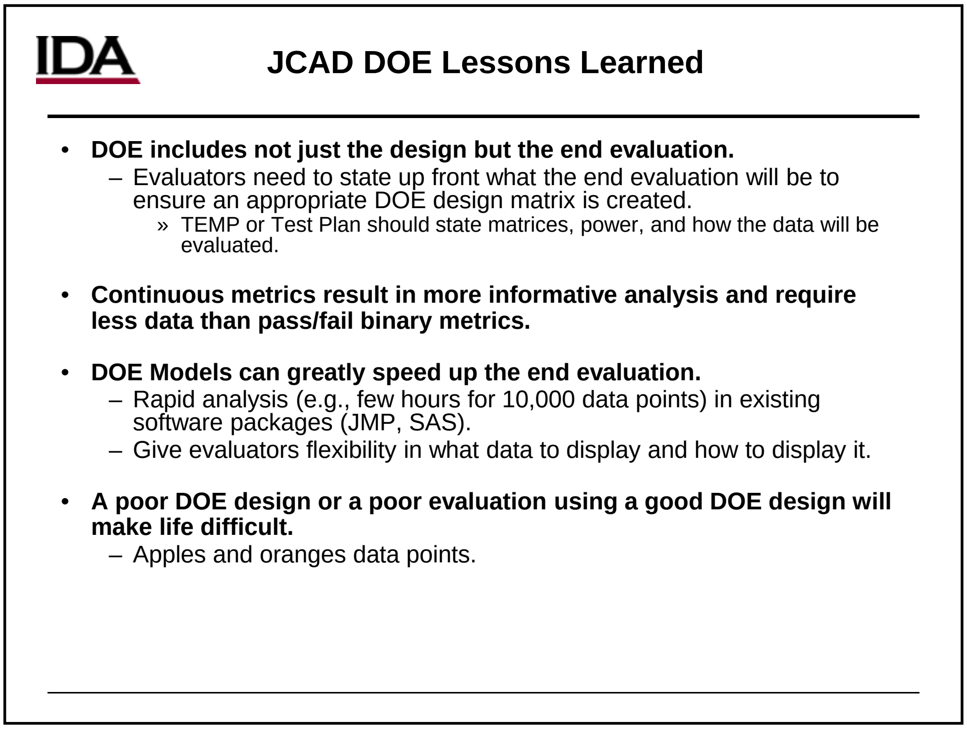
JCAD DOE Pros and Cons

Pros

- **Rapid analysis of large data sets**
- **Flexibility to display data multiple different ways**
- **Allowed for direct comparison to requirement**
- **Could analyze performance in any potential operating environment, even if we didn't specifically test that condition**

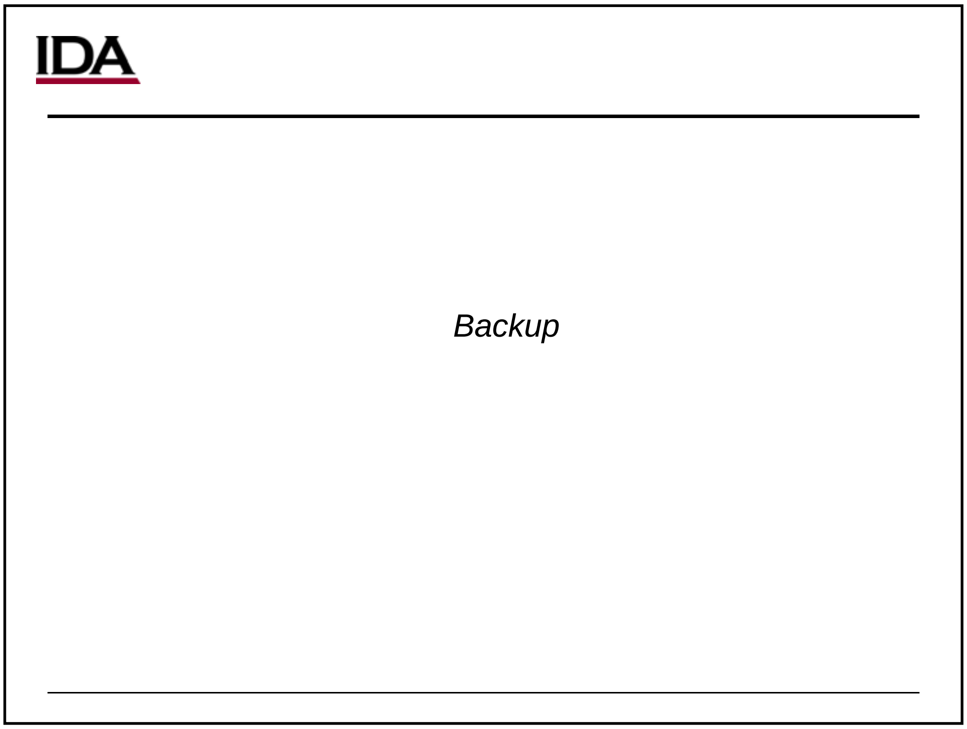
Cons

- **Test community (including PM) needed to buy into using modeled data to evaluate against the requirement**
- **If not modeling data, analysis becomes very difficult (apples and oranges)**

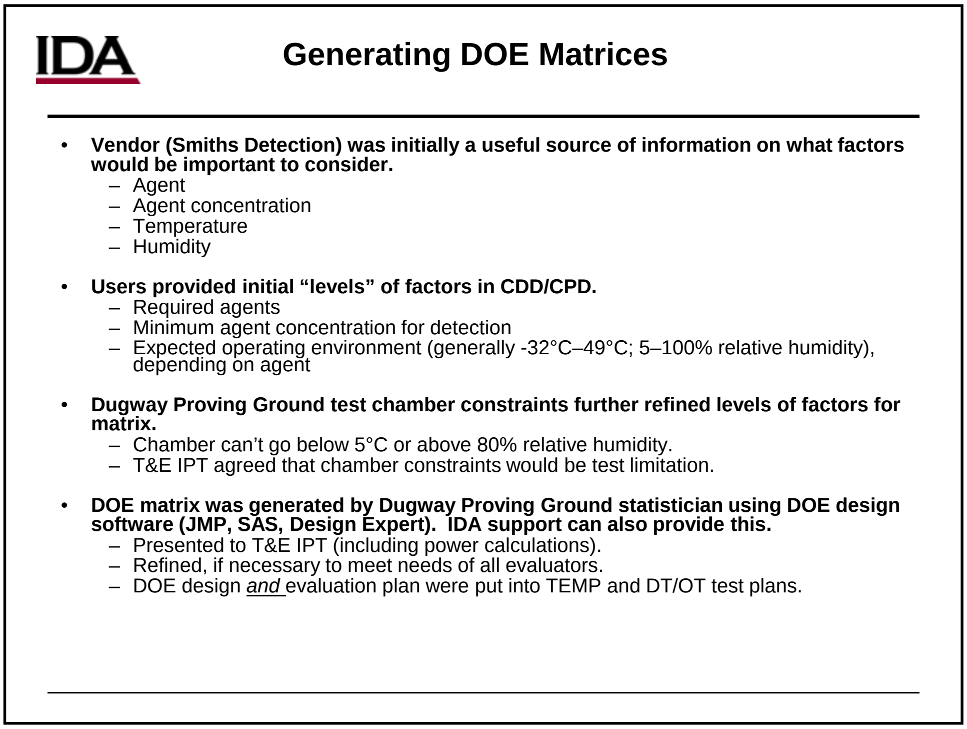


JCAD DOE Lessons Learned

- **DOE includes not just the design but the end evaluation.**
  - Evaluators need to state up front what the end evaluation will be to ensure an appropriate DOE design matrix is created.
    - » TEMP or Test Plan should state matrices, power, and how the data will be evaluated.
- **Continuous metrics result in more informative analysis and require less data than pass/fail binary metrics.**
- **DOE Models can greatly speed up the end evaluation.**
  - Rapid analysis (e.g., few hours for 10,000 data points) in existing software packages (JMP, SAS).
  - Give evaluators flexibility in what data to display and how to display it.
- **A poor DOE design or a poor evaluation using a good DOE design will make life difficult.**
  - Apples and oranges data points.



Backup



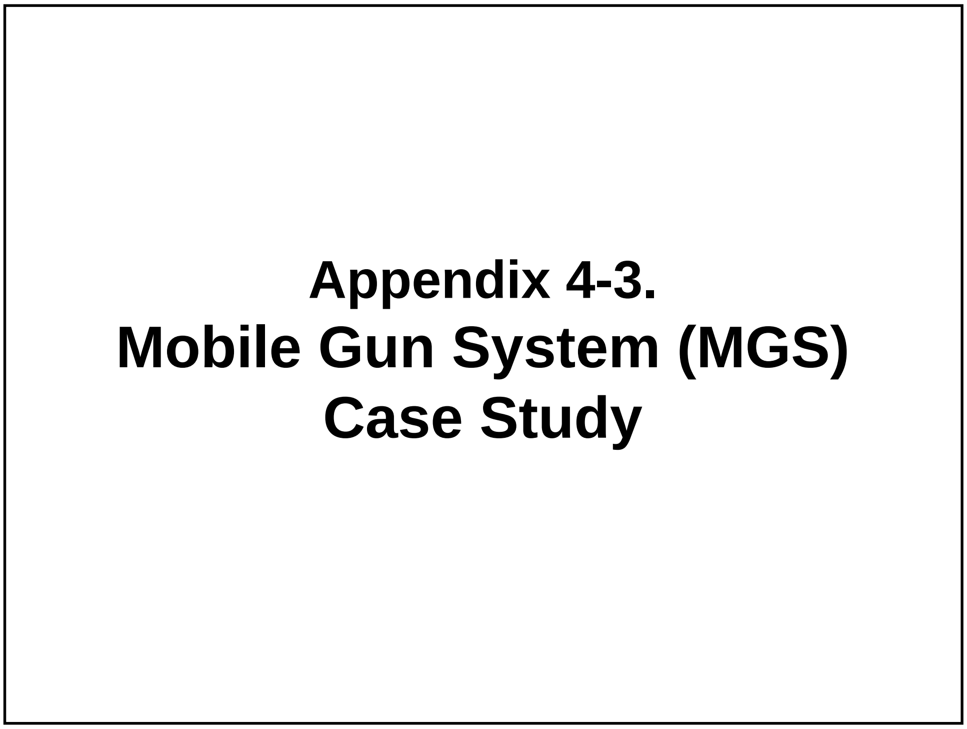
Generating DOE Matrices

- **Vendor (Smiths Detection) was initially a useful source of information on what factors would be important to consider.**
  - Agent
  - Agent concentration
  - Temperature
  - Humidity

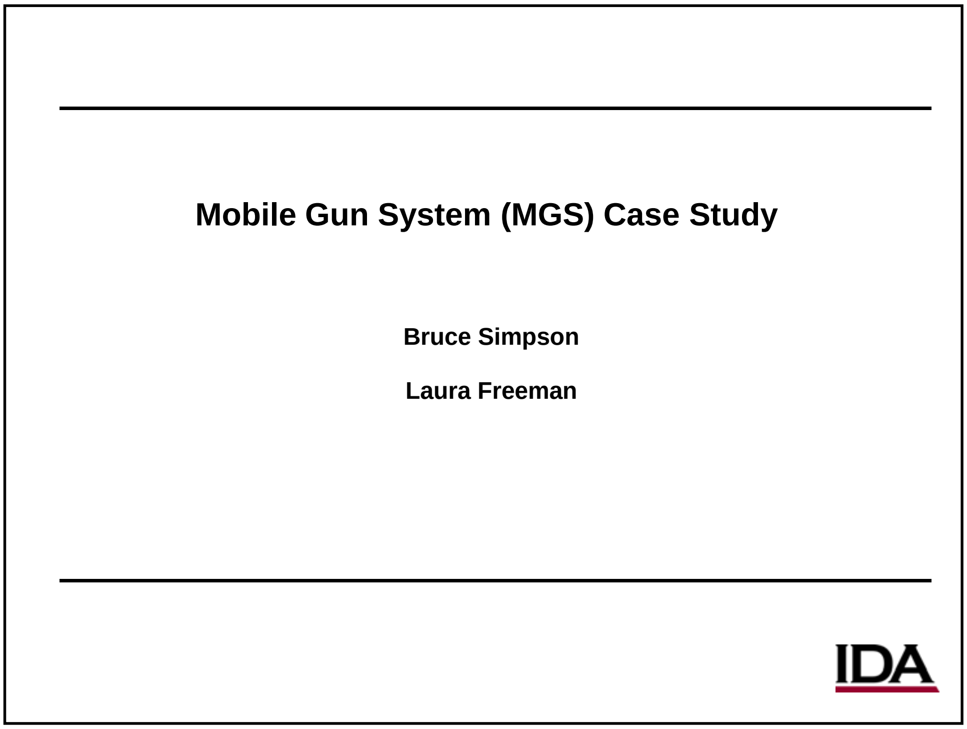
- **Users provided initial “levels” of factors in CDD/CPD.**
  - Required agents
  - Minimum agent concentration for detection
  - Expected operating environment (generally -32°C–49°C; 5–100% relative humidity), depending on agent

- **Dugway Proving Ground test chamber constraints further refined levels of factors for matrix.**
  - Chamber can’t go below 5°C or above 80% relative humidity.
  - T&E IPT agreed that chamber constraints would be test limitation.

- **DOE matrix was generated by Dugway Proving Ground statistician using DOE design software (JMP, SAS, Design Expert). IDA support can also provide this.**
  - Presented to T&E IPT (including power calculations).
  - Refined, if necessary to meet needs of all evaluators.
  - DOE design *and* evaluation plan were put into TEMP and DT/OT test plans.



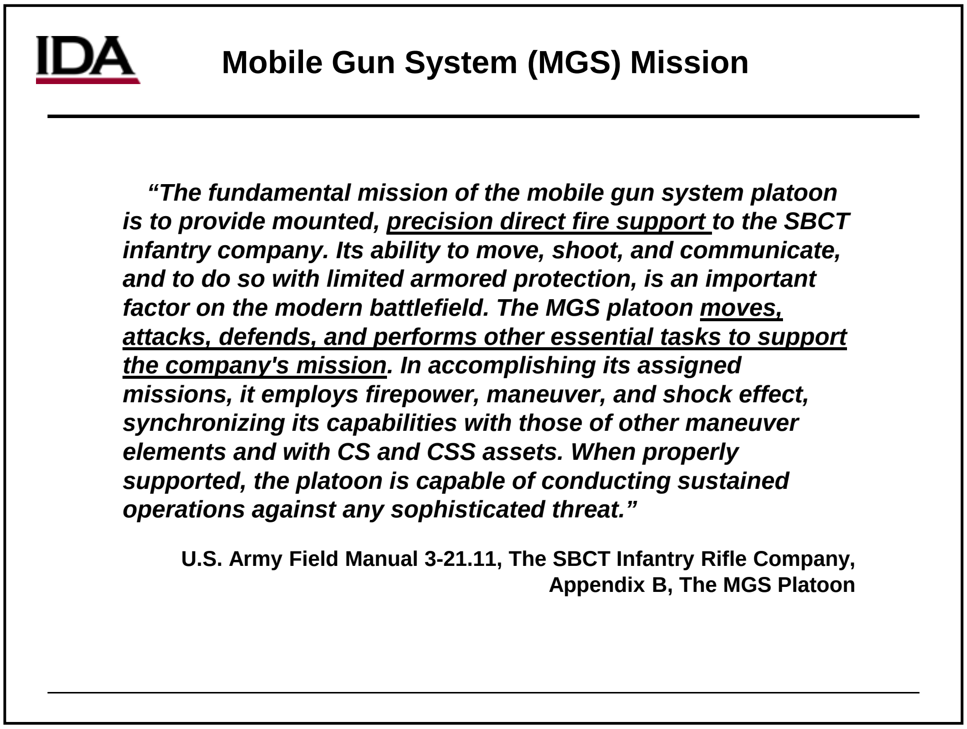
Appendix 4-3.
Mobile Gun System (MGS) Case Study



Mobile Gun System (MGS) Case Study

Bruce Simpson

Laura Freeman

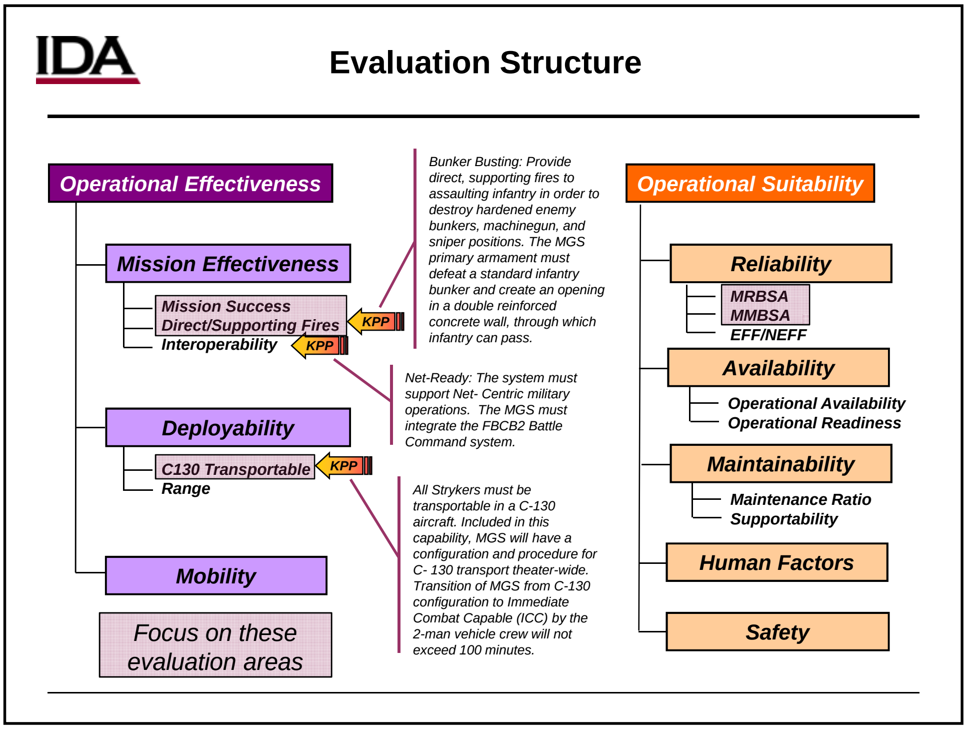


IDA

Mobile Gun System (MGS) Mission

> ***"The fundamental mission of the mobile gun system platoon is to provide mounted, <u>precision direct fire support</u> to the SBCT infantry company. Its ability to move, shoot, and communicate, and to do so with limited armored protection, is an important factor on the modern battlefield. The MGS platoon <u>moves, attacks, defends, and performs other essential tasks to support the company's mission</u>. In accomplishing its assigned missions, it employs firepower, maneuver, and shock effect, synchronizing its capabilities with those of other maneuver elements and with CS and CSS assets. When properly supported, the platoon is capable of conducting sustained operations against any sophisticated threat."***

U.S. Army Field Manual 3-21.11, The SBCT Infantry Rifle Company,
Appendix B, The MGS Platoon



IDA

Evaluation Structure

Operational Effectiveness

- **Mission Effectiveness**
  - Mission Success
  - Direct/Supporting Fires KPP
  - Interoperability KPP
- **Deployability**
  - C130 Transportable KPP
  - Range
- **Mobility**

Focus on these evaluation areas

Bunker Busting: Provide direct, supporting fires to assaulting infantry in order to destroy hardened enemy bunkers, machinegun, and sniper positions. The MGS primary armament must defeat a standard infantry bunker and create an opening in a double reinforced concrete wall, through which infantry can pass.

Net-Ready: The system must support Net- Centric military operations. The MGS must integrate the FBCB2 Battle Command system.

All Strykers must be transportable in a C-130 aircraft. Included in this capability, MGS will have a configuration and procedure for C- 130 transport theater-wide. Transition of MGS from C-130 configuration to Immediate Combat Capable (ICC) by the 2-man vehicle crew will not exceed 100 minutes.

Operational Suitability

- **Reliability**
  - MRBSA
  - MMBSA
  - EFF/NEFF
- **Availability**
  - Operational Availability
  - Operational Readiness
- **Maintainability**
  - Maintenance Ratio
  - Supportability
- **Human Factors**
- **Safety**

Design Factors

- **Mission Success-Can a unit equipped with the MGS successfully accomplish its missions**
  - Mission Type: Attack, Defend, Stability and Support Operations
  - Terrain Type: Urban, Mixed, Forest, Desert
  - Threat Level (OPFOR): Low, Medium, High
  - Illumination: Day, Night
  - Weather: Clear, Rain, Snow, Fog, Wind
- **Direct/Supporting Fires (Gunnery)**
  - Weapon System: Main gun, coaxial machine gun, 0.50 cal. machine gun
  - Weapon Sight: Primary (Day), Primary (Thermal), Auxiliary
  - Engagement Type: Offensive (Moving), Defensive (Stationary)
  - Target Type
    - » Moving, Stationary
    - » Tank, Armored Personnel Carrier, Bunker/Building, Troops
  - Range to target
  - Single Vehicle, Platoon
- **C-130 Transportability**
  - Add-on armor
  - Crew Training
  - Availability of Materiel Handling Equipment
- **Reliability**
  - Chassis: 1,000 Mean Miles Between System Aborts (MMBSA)
  - Mission Equipment Package (MEP): 81 Mean Rounds Between System Aborts (MRBSA)
  - Terrain Conditions (Operational Mode Summary/Mission Profile [OMS/MP])
    - » Trail/Cross Country
    - » Secondary Road
    - » Primary Road

Mission Design Factors

		Mission	Attack				Defend				Stability and Support				
Illum	OPFOR	Terrain	Urban	Mixed	Forest	Desert	Urban	Mixed	Forest	Desert	Urban	Mixed	Forest	Desert	
Day	Low		1	1		▨			🟩	▨	🟩		🟩	▨	2
Day	Med		1				1								2
Day	High		1				1	3							5
Night	Low					▨			🟩	▨	🟩	2	🟩	▨	2
Night	Med		2												2
Night	High			2			1								3
			5	3		▨	3	3	🟩	▨	🟩	2	🟩	▨	16

Weather: as it occurred; not controlled

Key

🟦	- Instrumented data collected during controlled IOT at Ft. Hood; number of mission replications indicated in cell
🟩	- Limited use data collected during Mission Rehearsal Exercise at Ft. Lewis; no instrumentation or control over factors
▨	- Limited use (anecdotal) data collected in theater during unit deployment to OIF, mostly on tactics and employment techniques

- *IOT test design builds on evidence from previous events*
  - ➢ *Mission Rehearsal Exercise prior to unit deployment (basis for Section 231 report)*
  - ➢ *Field data from unit deployment*
- *IOT scoped to focus on voids in medium and high threat levels*

Impact of Design of Experiments

- **Case Study: Mobile Gun System Design Comparison**

	Executed Cases in IOT&E	DOE I - Factorial Design	DOE II – Optimal Design (large)	DOE III – Optimal Design (small)
Factors & Levels	4 factors: Mission Type (3), Terrain Type (4), Treat Level (3), Illumination (2)			
Total Tests	16	72	36	16
Confidence	Set to the same level across all 4 designs: Confidence = 80%			
Power	0% - 53.1%	88.5% - 99.8%	64.7% - 93.4%	36.8% - 71.7%

- **The case study suggests that 16 runs may not be adequate to span the operational battle space with high power and confidence.**
- **The DOE optimal design is a more powerful allocation of the 16 tests than the case based design.**
- **DOE allows us to understand what we are giving up**
  - In the case of MGS, the system was deployed early which altered the original test plan.

MGS Design Comparison

Case Based Design Executed in IOT&E

		Mission	Attack				Defend				Stability and Support				
Illum	OPFOR	Terrain	Urban	Mixed	Forest	Desert	Urban	Mixed	Forest	Desert	Urban	Mixed	Forest	Desert	
Day	Low		1	1											2
Day	Med		1				1								2
Day	High		1				1	3							5
Night	Low											2			2
Night	Med		2												2
Night	High			2			1								3
			5	3			3	3				2			16

Statistical D-Optimal Design

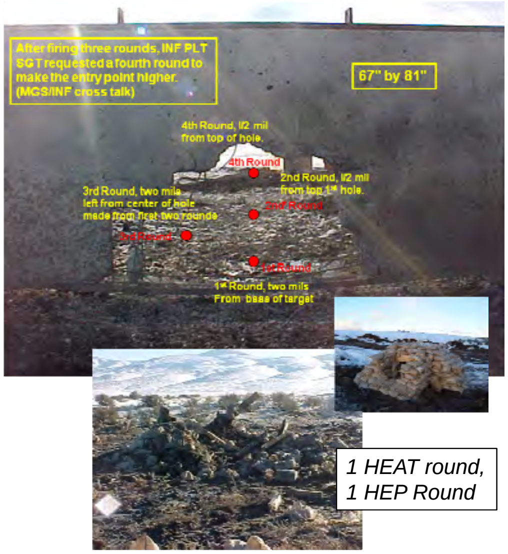
		Mission	Attack				Defend				Stability and Support				
Illum	OPFOR	Terrain	Urban	Mixed	Forest	Desert	Urban	Mixed	Forest	Desert	Urban	Mixed	Forest	Desert	
Day	Low						1		1					1	3
Day	Med			1		1									2
Day	High		1						1			1			3
Night	Low			1	1										2
Night	Med						1	1					1		3
Night	High					1				1	1				3
			1	2	1	2	2	1	2	1	1	1	1	1	16

Mission Success Results

		Proportion of Successful Missions Based on Achieving Stated Unit Mission	80 % Confidence Interval	Proportion of Successful Missions according to Army Subject Matter Experts (# success / Total SME)	Proportion of Missions where Mobile Gun System Contributed Positively to Mission as rated by Army Subject Matter Experts	Mobile Gun System Based on RTCA Data		Infantry Carrier Vehicle Based on RTCA Data	
						Start	Lost	Start	Lost
Terrain	Urban Terrain	63%	35%-85%	54% (22/41)	88%	24	4	32	15
	Mixed Terrain	75%	46%-93%	51% (20/39)	74%	24	8	32	9
Threat	High Threat	63%	35%-85%	38% (19/38)	78%	24	11	32	12
	Low-Mid Threat	75%	46%-93%	59% (26/44)	84%	24	1	32	12
Mission	All Attack	50%	24%-76%	46% (19/41)	77%	24	6	32	15
	All SASO	100%	32%-100%	70% (7/10)	76%	6	0	8	1
	All Defend	83%	49%-98%	55% (16/29)	90%	18	6	24	8

- *Overall Mission Success Rate is 69% (p-value=0.105)*
- *Mission Success tied to unit achieving assigned objectives and unit losses*
- *No confidence interval on Subject Matter Expert ratings*

Bunker Busting/Wall Breach KPP

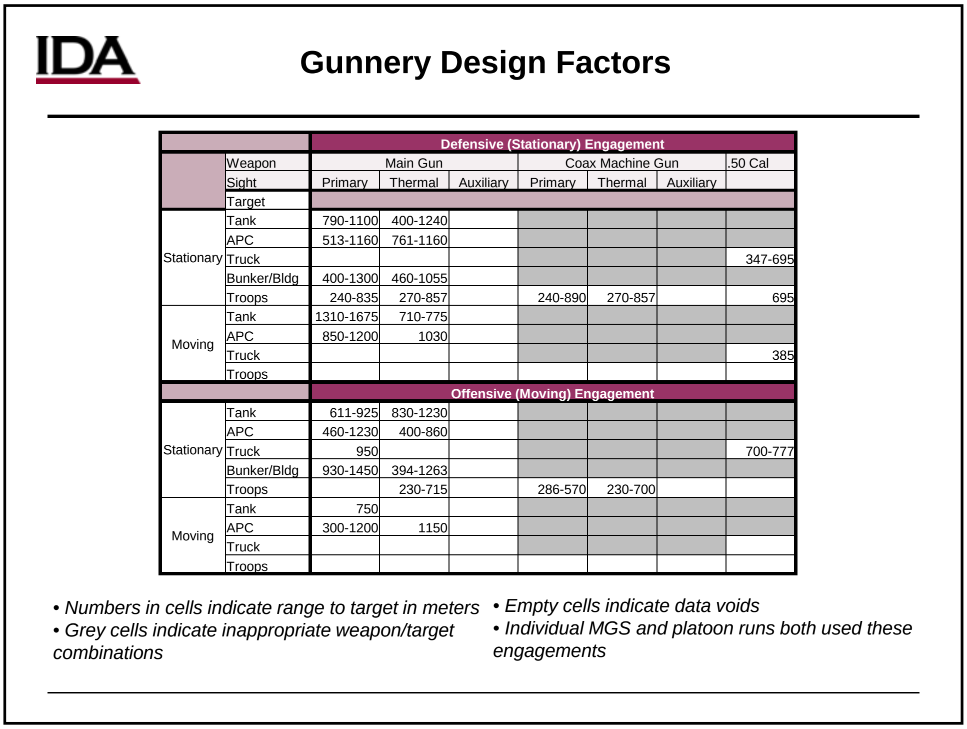

After firing three rounds, INF PLT SGT requested a fourth round to make the entry point higher. (MGS/INF cross talk)
67" by 81"
4th Round, l/2 mil from top of hole.
4th Round
2nd Round, l/2 mil from top 1st hole.
2nd Round
3rd Round, two mils left from center of hole made from first two rounds
3rd Round
1st Round
1st Round, two mils From base of target

1 HEAT round, 1 HEP Round

KPP demonstrated at Force Development Exercise, Yakima Training Center, WA Feb. 2004

Run	Height (inches)	Width (inches)	Rounds HEP
1	68	60	3
2	51	56	3
3	71	41	3
4	67	81	4
5	52	47	3
6	60	51	4
7	57	74	3
8	50	60	4

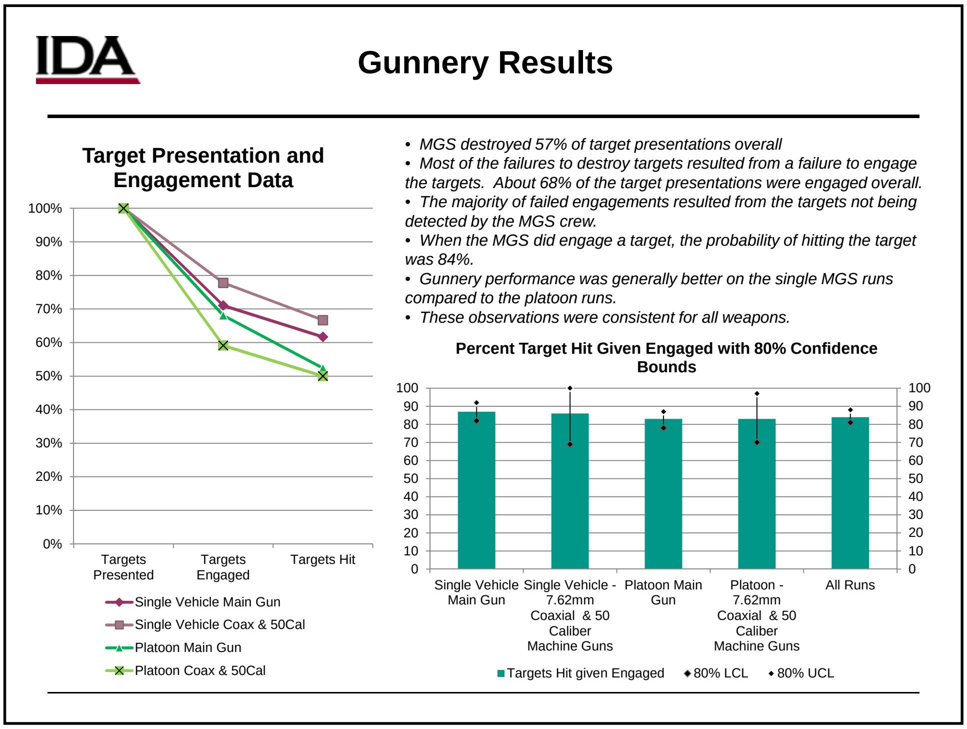
At 8 of 8 attempts, the system has an 80% LCB of 75% probability of breaching a concrete wall in 3-4 rounds, as demonstrated in the FDE



Gunnery Design Factors

	Weapon	Main Gun			Coax Machine Gun			.50 Cal
	Sight	Primary	Thermal	Auxiliary	Primary	Thermal	Auxiliary	
	Target							
		Defensive (Stationary) Engagement						
Stationary	Tank	790-1100	400-1240					
	APC	513-1160	761-1160					
	Truck							347-695
	Bunker/Bldg	400-1300	460-1055					
	Troops	240-835	270-857		240-890	270-857		695
Moving	Tank	1310-1675	710-775					
	APC	850-1200	1030					
	Truck							385
	Troops							
		Offensive (Moving) Engagement						
Stationary	Tank	611-925	830-1230					
	APC	460-1230	400-860					
	Truck	950						700-777
	Bunker/Bldg	930-1450	394-1263					
	Troops		230-715		286-570	230-700		
Moving	Tank	750						
	APC	300-1200	1150					
	Truck							
	Troops							

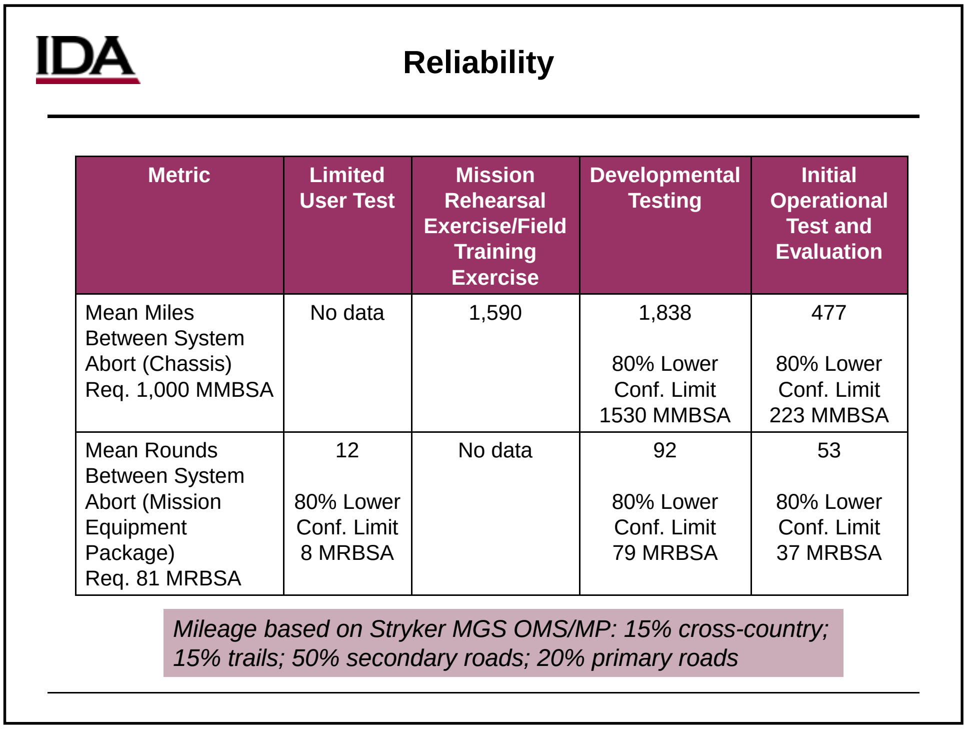
- *Numbers in cells indicate range to target in meters*
- *Grey cells indicate inappropriate weapon/target combinations*
- *Empty cells indicate data voids*
- *Individual MGS and platoon runs both used these engagements*



Gunnery Results

- *MGS destroyed 57% of target presentations overall*
- *Most of the failures to destroy targets resulted from a failure to engage the targets. About 68% of the target presentations were engaged overall.*
- *The majority of failed engagements resulted from the targets not being detected by the MGS crew.*
- *When the MGS did engage a target, the probability of hitting the target was 84%.*
- *Gunnery performance was generally better on the single MGS runs compared to the platoon runs.*
- *These observations were consistent for all weapons.*

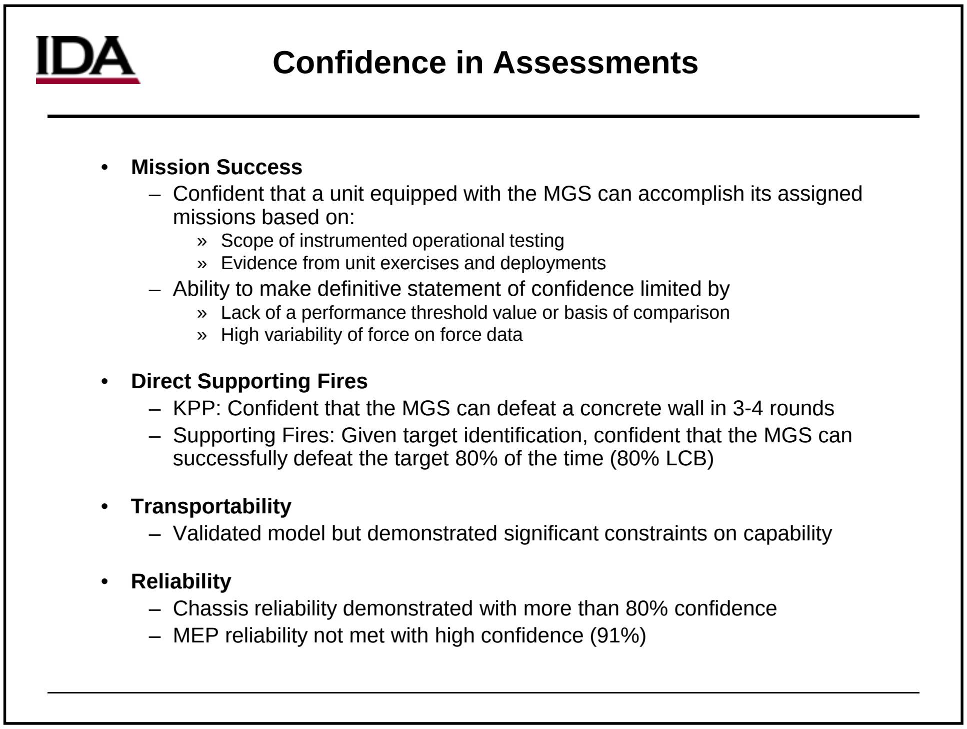
Percent Target Hit Given Engaged with 80% Confidence Bounds
100
90
80
70
60
50
40
30
20
10
0
Single Vehicle Main Gun
Single Vehicle - 7.62mm Coaxial & 50 Caliber Machine Guns
Platoon Main Gun
Platoon - 7.62mm Coaxial & 50 Caliber Machine Guns
All Runs
Targets Hit given Engaged
80% LCL
80% UCL



Reliability

Metric	Limited User Test	Mission Rehearsal Exercise/Field Training Exercise	Developmental Testing	Initial Operational Test and Evaluation
Mean Miles Between System Abort (Chassis) Req. 1,000 MMBSA	No data	1,590	1,838 80% Lower Conf. Limit 1530 MMBSA	477 80% Lower Conf. Limit 223 MMBSA
Mean Rounds Between System Abort (Mission Equipment Package) Req. 81 MRBSA	12 80% Lower Conf. Limit 8 MRBSA	No data	92 80% Lower Conf. Limit 79 MRBSA	53 80% Lower Conf. Limit 37 MRBSA

Mileage based on Stryker MGS OMS/MP: 15% cross-country; 15% trails; 50% secondary roads; 20% primary roads



Confidence in Assessments

- **Mission Success**
  - Confident that a unit equipped with the MGS can accomplish its assigned missions based on:
    - Scope of instrumented operational testing
    - Evidence from unit exercises and deployments
  - Ability to make definitive statement of confidence limited by
    - Lack of a performance threshold value or basis of comparison
    - High variability of force on force data
- **Direct Supporting Fires**
  - KPP: Confident that the MGS can defeat a concrete wall in 3-4 rounds
  - Supporting Fires: Given target identification, confident that the MGS can successfully defeat the target 80% of the time (80% LCB)
- **Transportability**
  - Validated model but demonstrated significant constraints on capability
- **Reliability**
  - Chassis reliability demonstrated with more than 80% confidence
  - MEP reliability not met with high confidence (91%)

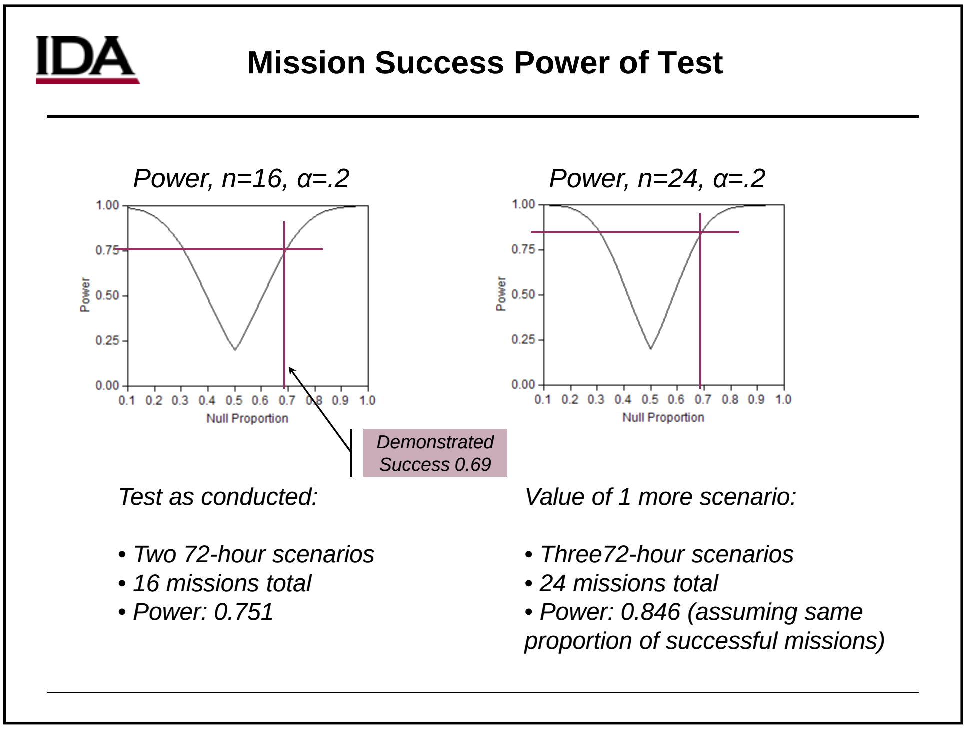
IDA

DOE Lessons from MGS

- **Force on force exercises contain far more sources of variability than can be controlled**
  - Underlying distributions of battlefield phenomena not well understood
  - Human decision making limits repeatability
- **DOE-like structured analysis can define the operational envelope and inform testing**
  - Mission space
  - Gunnery performance
- **Operational Effectiveness and Operational Suitability are frequently multi-dimensional**
  - DOE can be used on individual sub-elements
  - Roll-up of several sub-elements makes a numerical assessment of the overall “power of the test” difficult
- **Can be used to allocate test resources based on other evidence**
  - Using data from training or operational events to focus IOT
  - Using previous test results for reliability to focus IOT

IDA

BACKUP



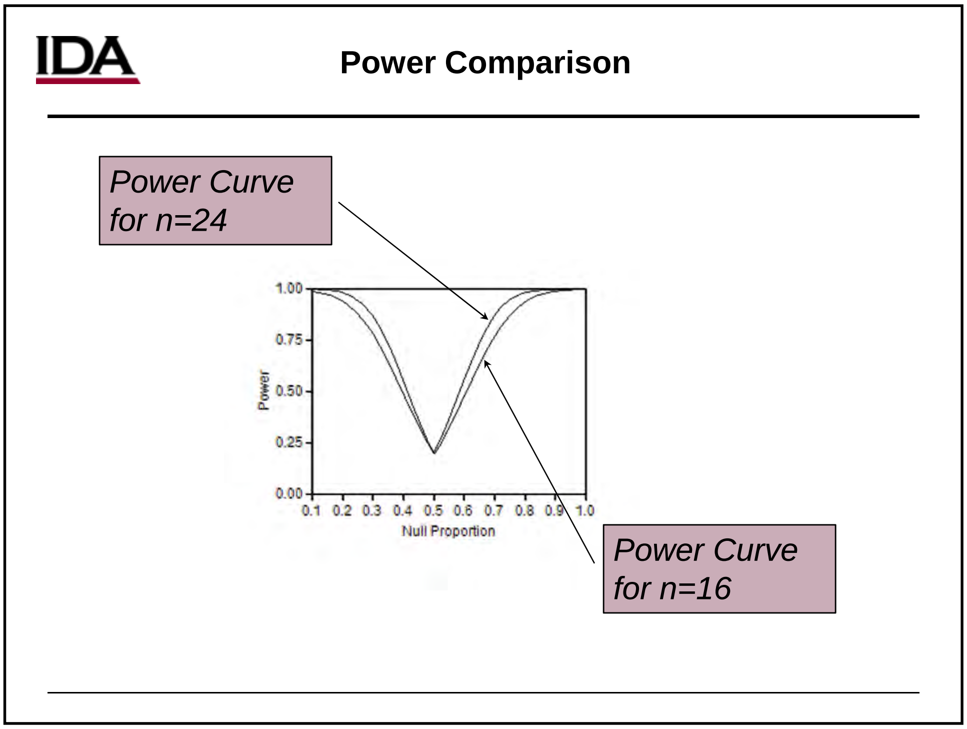
Mission Success Power of Test

Test as conducted:

- *Two 72-hour scenarios*
- *16 missions total*
- *Power: 0.751*

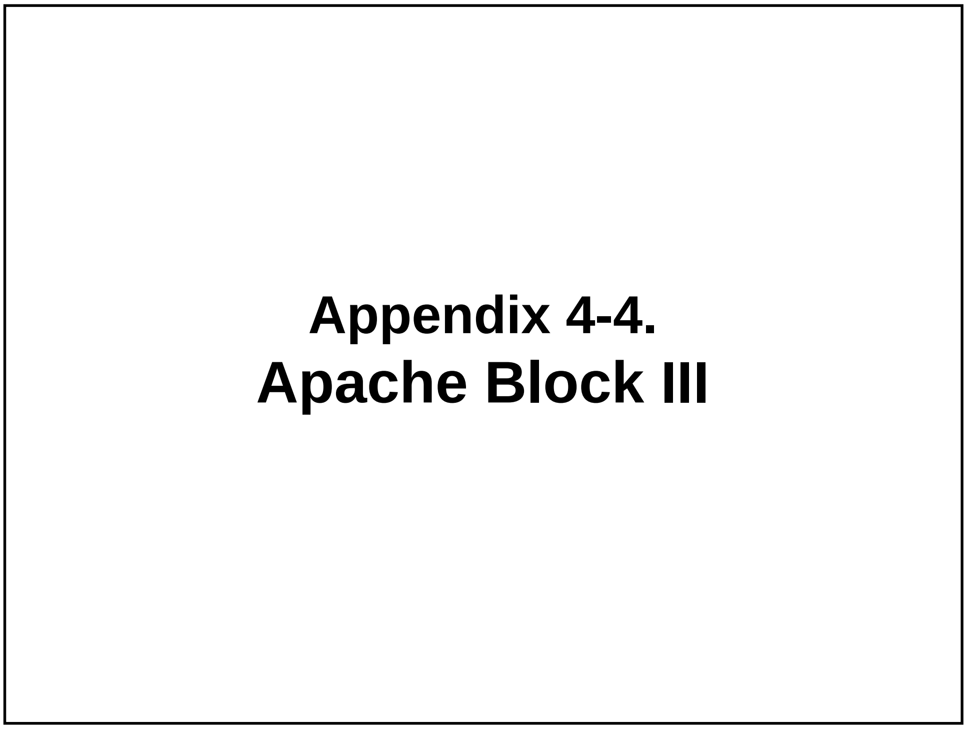
Value of 1 more scenario:

- *Three72-hour scenarios*
- *24 missions total*
- *Power: 0.846 (assuming same proportion of successful missions)*

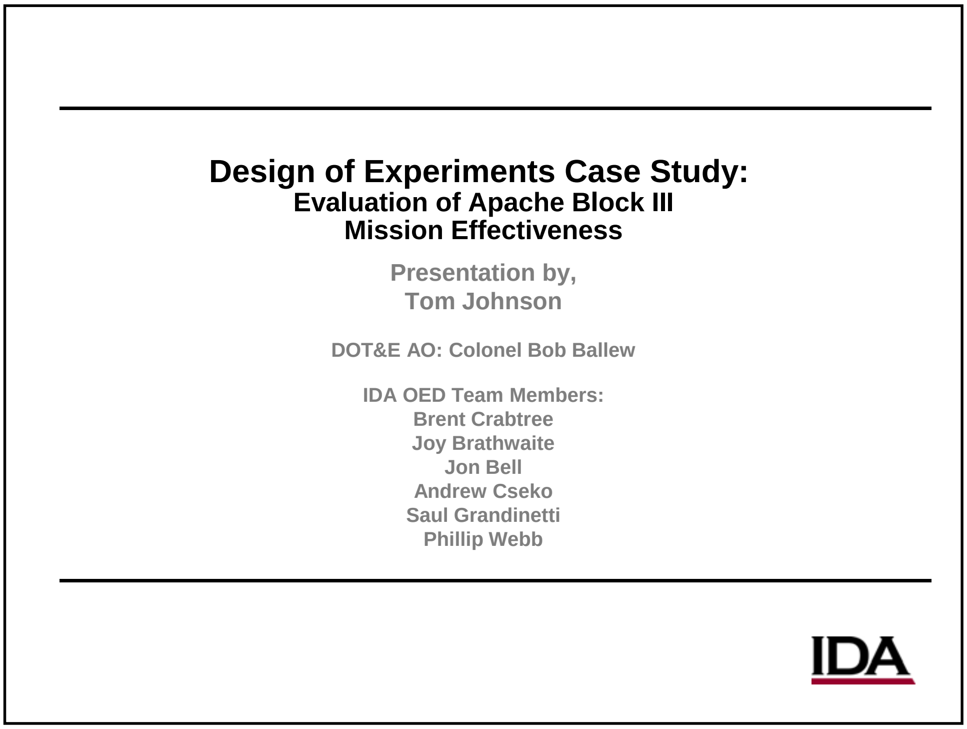


Power Comparison

This page intentionally left blank.



Appendix 4-4.
Apache Block III



Design of Experiments Case Study:
Evaluation of Apache Block III
Mission Effectiveness

Presentation by,
Tom Johnson

DOT&E AO: Colonel Bob Ballew

IDA OED Team Members:
Brent Crabtree
Joy Brathwaite
Jon Bell
Andrew Cseko
Saul Grandinetti
Phillip Webb

Case Study Outline

- **Apache Block III Background Information**
- **Purpose of Experiment / Response Variable**
- **Factors and Levels**
- **Experimental Design**
  - Sample size, Model Form, Power
- **Analysis / Conclusions**
  - ANOVA, Results Plots
- **Lessons Learned / Future Testing**

"By failing to prepare, you are preparing to fail."
– Benjamin Franklin

Apache Block III (AB3)

- **AB3 is an updated version of the AH-64D attack helicopter**
  - Will modernize the entire fleet of 690 aircraft
- **New AB3 Lot 1 features:**

Improved Drive System:
New Transmission Design
Increases Power
Increases Performance

AB3 Avionics:
Expands Communication Options
Adds Instrument Flight Capability
Eliminates Obsolescence
Expands Processing Capability

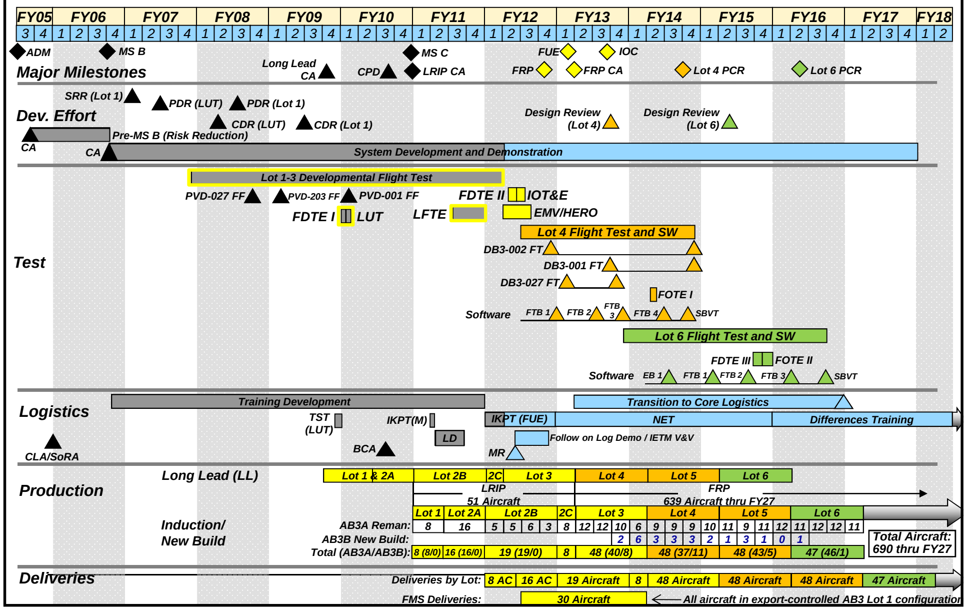
Improved Helmet and Display Sight System:
Avoids Obsolescence

UAS Interoperability:
Receives UAS Video
Controls UAS Sensors
Repositions UAS Air Vehicle

Radar Electronics Unit:
Enhances Radar Processing Capacity
Replaces Obsolescent Components

- **Lot 4 is scheduled for operational testing in 2014. Lot 6 is scheduled for 2015.**

Apache Block III IOT&E Background



	Lot 1	Lot 2A	Lot 2B				2C	Lot 3				Lot 4				Lot 5				Lot 6			
AB3A Reman:	8	16	5	5	6	3	8	12	12	10	6	9	9	9	10	11	9	11	12	11	12	12	11
AB3B New Build:										2	6	3	3	3	2	1	3	1	0	1			
Total (AB3A/AB3B):	8 (8/0)	16 (16/0)	19 (19/0)				8	48 (40/8)				48 (37/11)				48 (43/5)				47 (46/1)			

Three AB3 IOT&E Configurations

Common to all Apache Block III (AB3) Aircraft

Block III Drive Train

- *6K 95 Hover performance*
- *Increased payload*

Block III Avionics

- *Weapon/display processor upgrades*
- *Enhanced display electronics unit*

Integrated Communications

- *Multi-band ARC-231 radios*
- *SATCOM*
- *Blue Force Tracker*

Instrument Flight Capability (IFC)

1. An AB3 with no mast-mounted assembly (Slick)

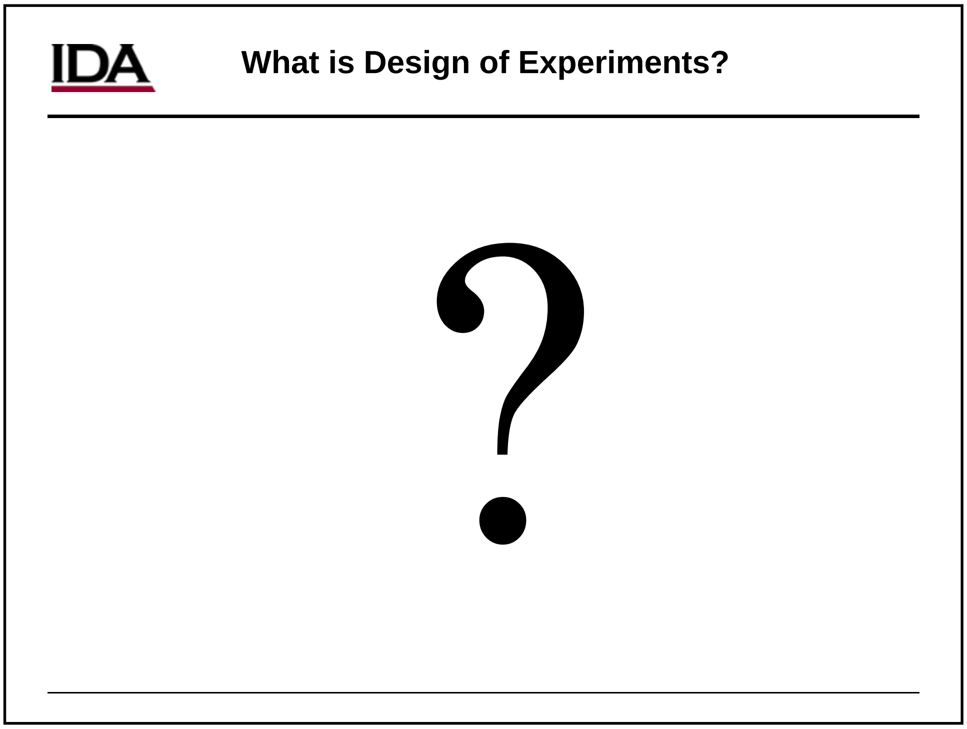
2. Fire Control Radar (FCR) Aircraft

- *Provides legacy FCR functionality with new hardware and software*
- *Faster processor with potential for enhanced FCR range at Lot 6*

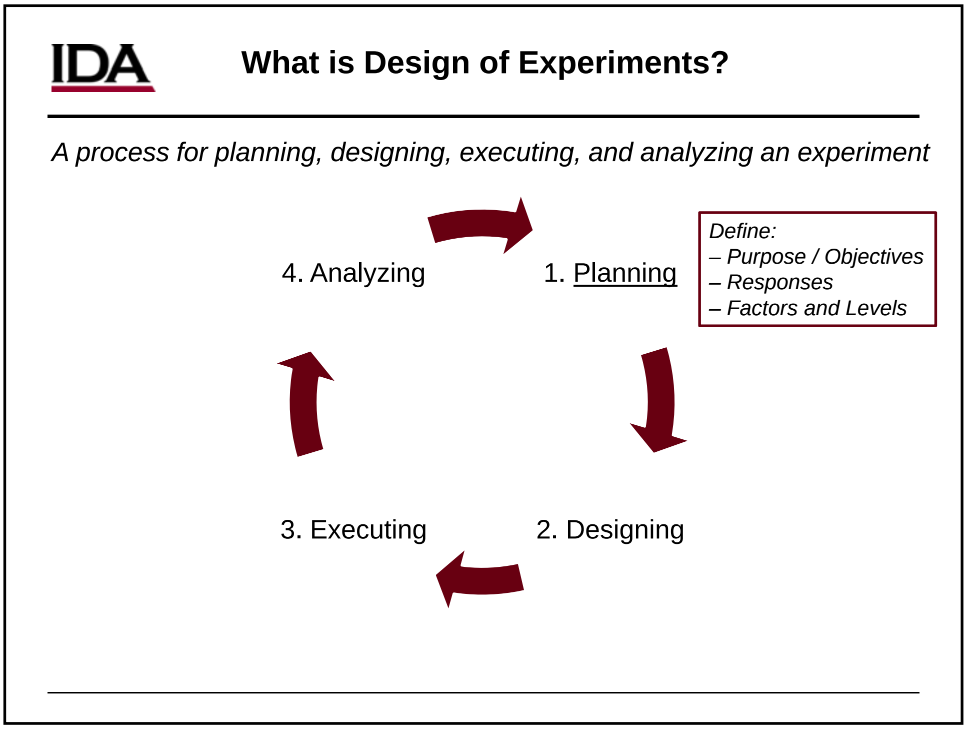
3. UAS TCDL Assembly (UTA) Aircraft

- *Enables Manned Unmanned (MUM) Teaming*
  - *UTA exercises Unmanned Aircraft System (UAS) Control with Tactical Common Data Link (TCDL)*
  - *Levels of Interoperability (LOI)*
    - *LOI 1 = Voice communication with UAS operator*
    - *LOI 2 = Receive UAS video in UTA aircraft*
    - *LOI 3 = Control UAS sensor from UTA aircraft*
    - *LOI 4 = Control UAS flight from UTA aircraft*

New AB3 Configuration



What is Design of Experiments?

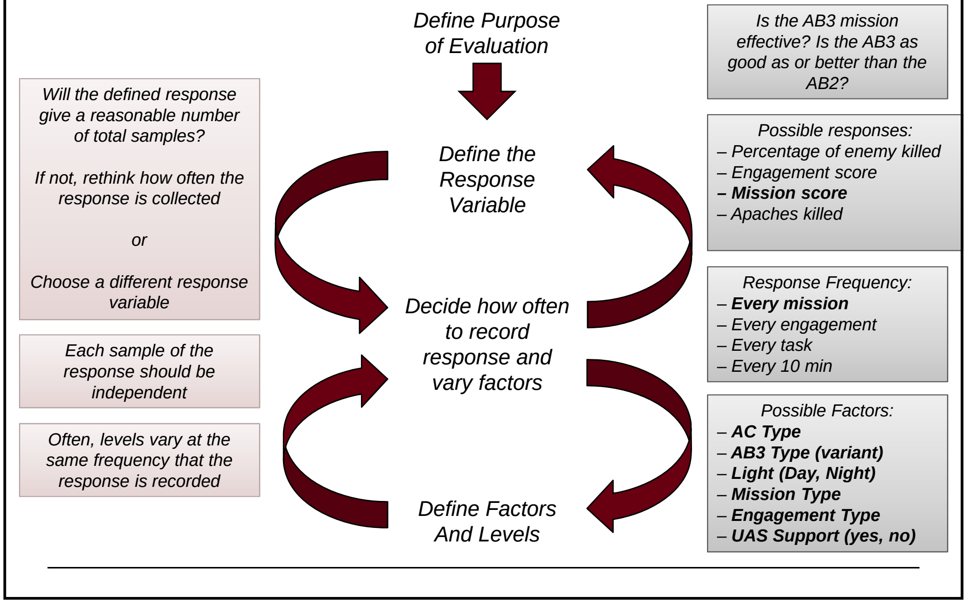


What is Design of Experiments?

A process for planning, designing, executing, and analyzing an experiment

Define:
- *Purpose / Objectives*
- *Responses*
- *Factors and Levels*

DOE Planning



Define Purpose of Evaluation

Is the AB3 mission effective? Is the AB3 as good as or better than the AB2?

Define the Response Variable

Will the defined response give a reasonable number of total samples?

If not, rethink how often the response is collected

or

Choose a different response variable

Possible responses:
- *Percentage of enemy killed*
- *Engagement score*
- ***Mission score***
- *Apaches killed*

Decide how often to record response and vary factors

Each sample of the response should be independent

Often, levels vary at the same frequency that the response is recorded

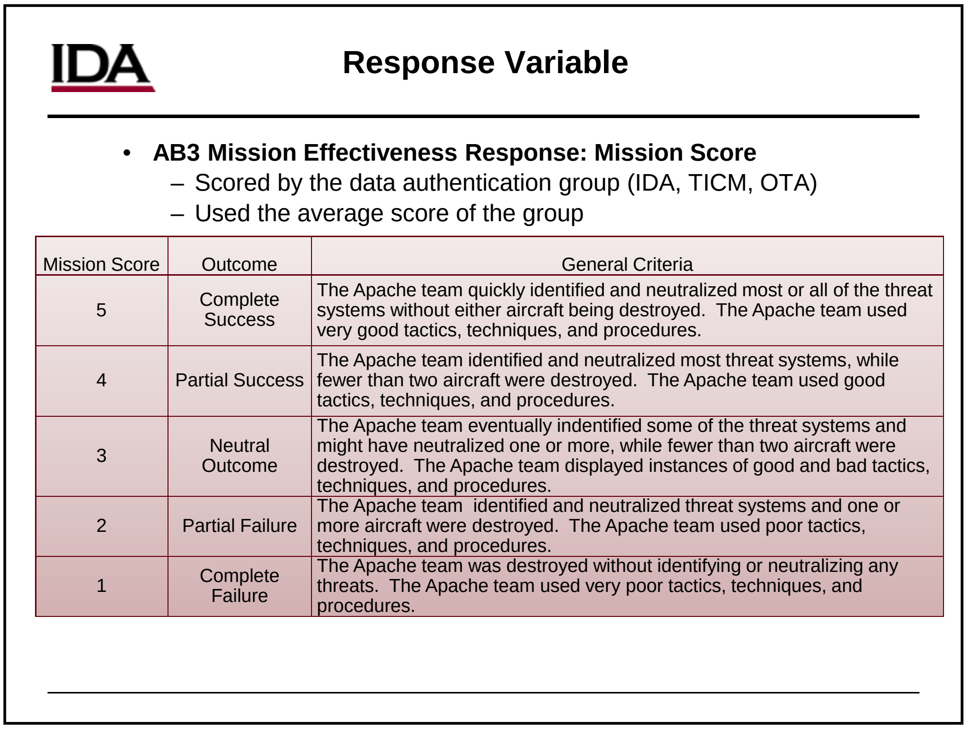
Response Frequency:
- ***Every mission***
- *Every engagement*
- *Every task*
- *Every 10 min*

Define Factors And Levels

Possible Factors:
- ***AC Type***
- ***AB3 Type (variant)***
- ***Light (Day, Night)***
- ***Mission Type***
- ***Engagement Type***
- ***UAS Support (yes, no)***

Purpose of AB3 DOE Evaluation

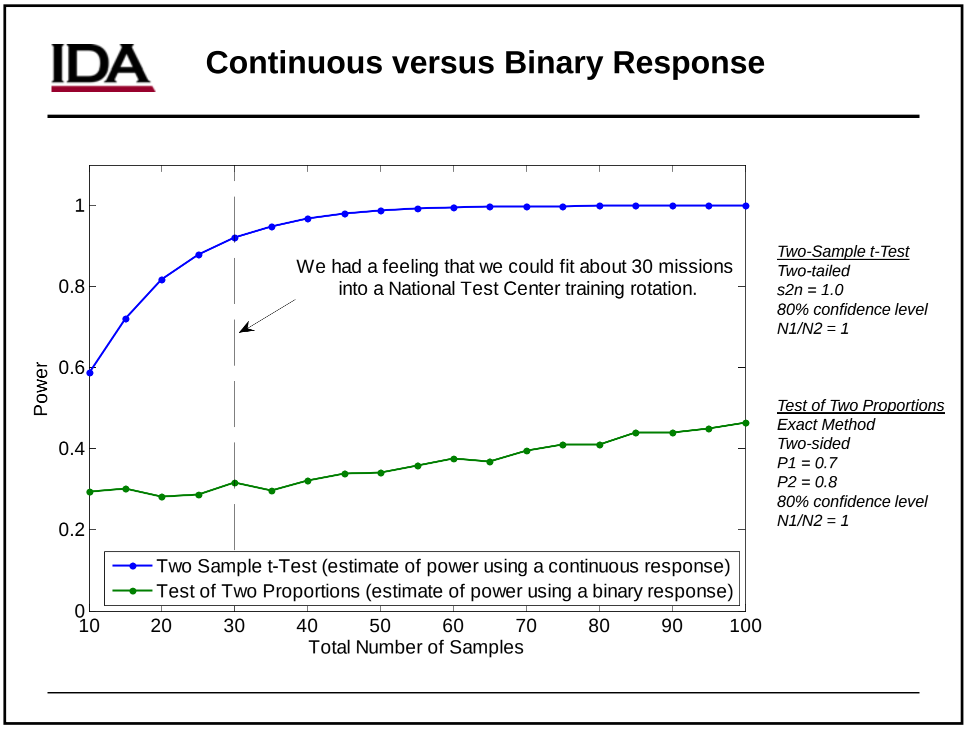
- **Purpose of AB3 Mission Effectiveness Evaluation**
  1. Is the AB3 mission effective?
     - » What does that mean? There can be many interpretations.
     - » Define with no jargon: Can the AB3 complete missions in a timely manner without being killed?
     - » Under what conditions?
  2. Is it as good as or better than the AB2?
     - » This question requires a comparative test
     - » Is this question too vague?
- **Must be operationally realistic**
  - » Do not let the statistical design take precedence over operational realism
  - » Statistics are meaningless if the test is not operational
- **Consider reliability requirements as well**
  - – Testing conditions should span the entire operational envelope to provide a comprehensive reliability assessment



Response Variable

- **AB3 Mission Effectiveness Response: Mission Score**
  - Scored by the data authentication group (IDA, TICM, OTA)
  - Used the average score of the group

Mission Score	Outcome	General Criteria
5	Complete Success	The Apache team quickly identified and neutralized most or all of the threat systems without either aircraft being destroyed. The Apache team used very good tactics, techniques, and procedures.
4	Partial Success	The Apache team identified and neutralized most threat systems, while fewer than two aircraft were destroyed. The Apache team used good tactics, techniques, and procedures.
3	Neutral Outcome	The Apache team eventually indentified some of the threat systems and might have neutralized one or more, while fewer than two aircraft were destroyed. The Apache team displayed instances of good and bad tactics, techniques, and procedures.
2	Partial Failure	The Apache team identified and neutralized threat systems and one or more aircraft were destroyed. The Apache team used poor tactics, techniques, and procedures.
1	Complete Failure	The Apache team was destroyed without identifying or neutralizing any threats. The Apache team used very poor tactics, techniques, and procedures.



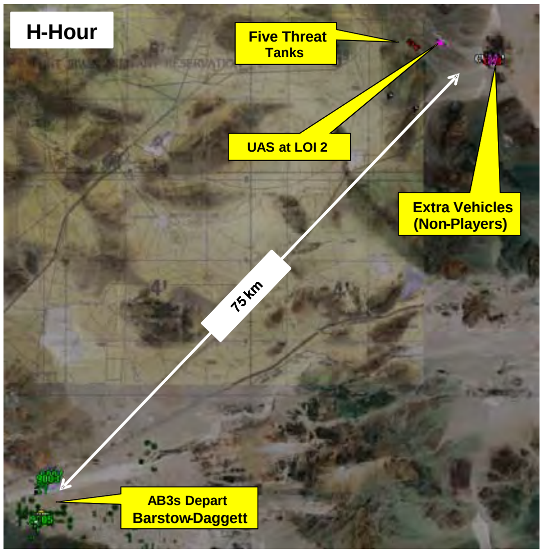
Continuous versus Binary Response

We had a feeling that we could fit about 30 missions into a National Test Center training rotation.

Two-Sample t-Test
Two-tailed
s2n = 1.0
80% confidence level
N1/N2 = 1

Test of Two Proportions
Exact Method
Two-sided
P1 = 0.7
P2 = 0.8
80% confidence level
N1/N2 = 1

Typical Mission during AB3 IOT&E

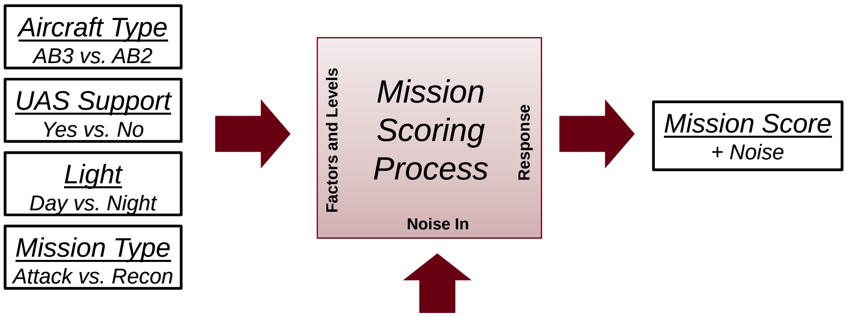


H-Hour
- UAS arrives on station; begins to build SA
- AB3 attempts link with UAS while at airfield
- LOI 2 (receive UAS video) to observe targets
- LOI 3 (control UAS sensor) to mark and store targets (if possible)

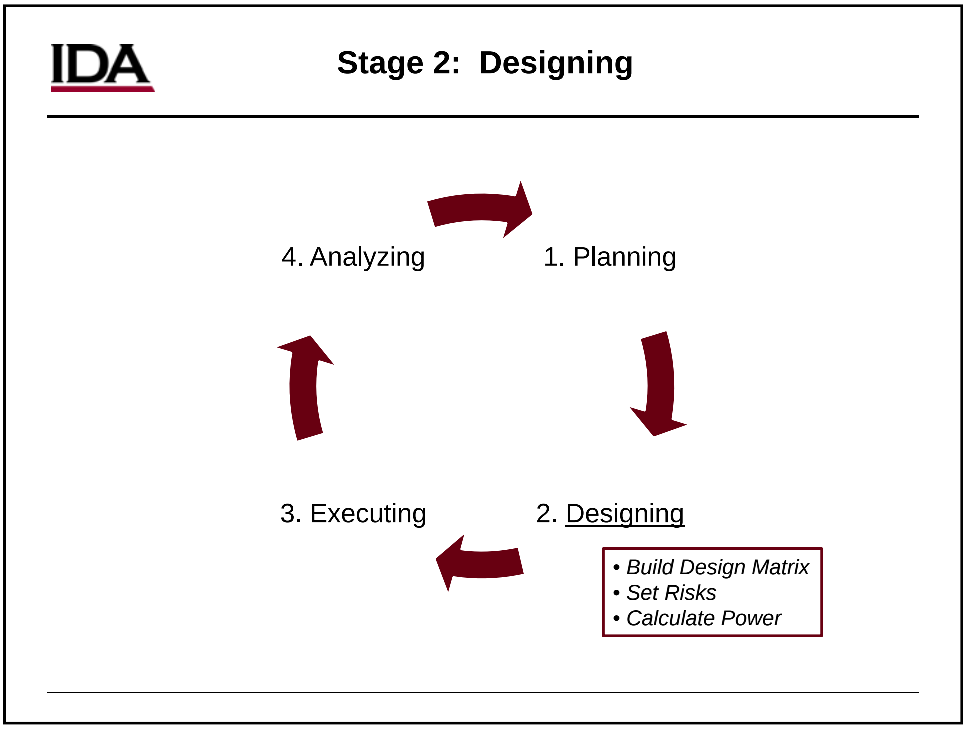
H-Hour to H + 0:30
- AB3 in LOI 2 (observe threat systems) while enroute

H + 0:30 to H + 1:00
- AB3 relocates to appropriate firing position
- AB3 engages OPFOR with UAS in LOI 2 or LOI 3 (AB3 pilot's choice)

Defining Factors and Levels



- *Instrumentation Problems*
- *Poor communications with OC*
- *Weather*
- *Apache pilot skills*
- *UAS pilot skills*
- *UAS/AB3 teaming guidelines*
- *Fatigue*



Stage 2: Designing

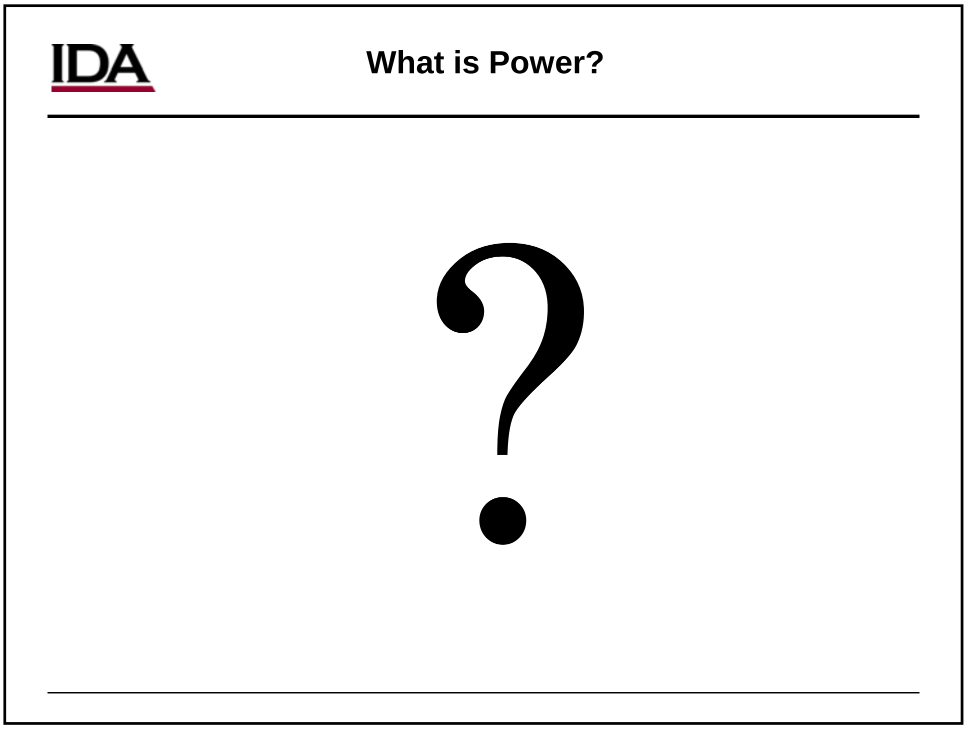
4. Analyzing

1. Planning

3. Executing

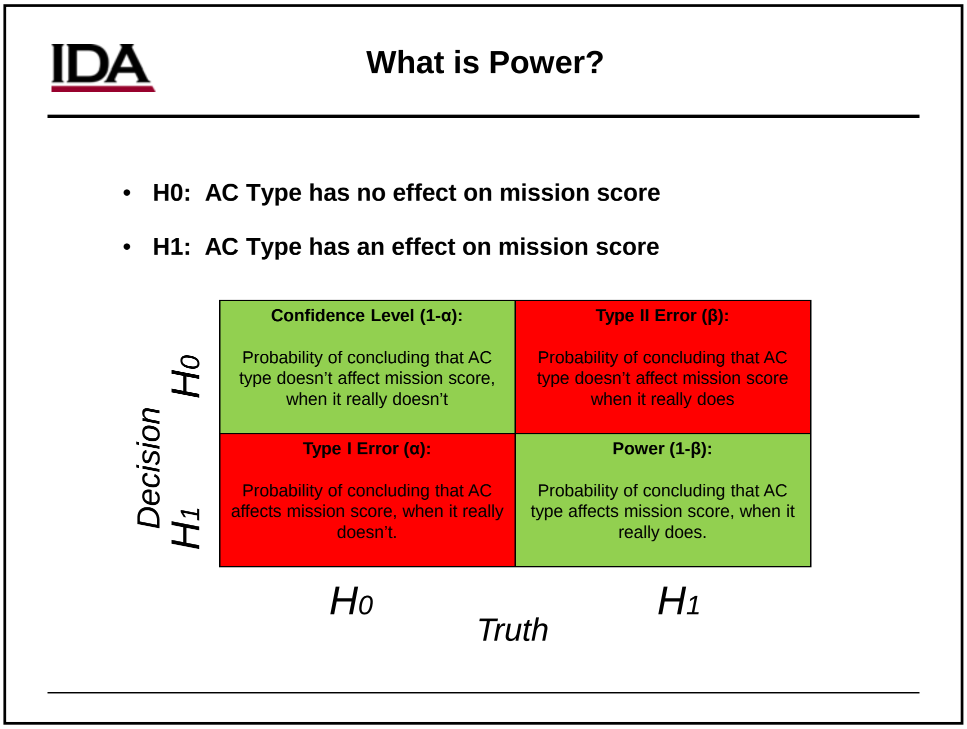
2. Designing

- *Build Design Matrix*
- *Set Risks*
- *Calculate Power*



What is Power?

?

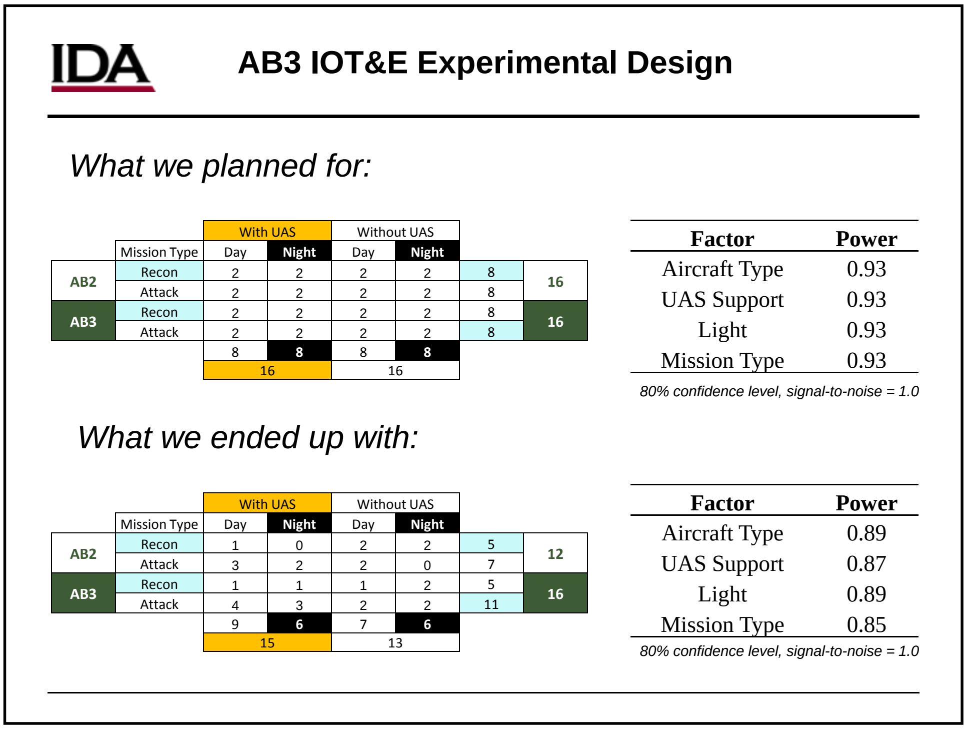


IDA

What is Power?

- **H0: AC Type has no effect on mission score**
- **H1: AC Type has an effect on mission score**

Decision	*Truth: H0*	*Truth: H1*
H0	**Confidence Level (1-α):** Probability of concluding that AC type doesn't affect mission score, when it really doesn't	**Type II Error (β):** Probability of concluding that AC type doesn't affect mission score when it really does
H1	**Type I Error (α):** Probability of concluding that AC affects mission score, when it really doesn't.	**Power (1-β):** Probability of concluding that AC type affects mission score, when it really does.



IDA

AB3 IOT&E Experimental Design

What we planned for:

		With UAS		Without UAS			
	Mission Type	Day	Night	Day	Night		
AB2	Recon	2	2	2	2	8	16
	Attack	2	2	2	2	8	
AB3	Recon	2	2	2	2	8	16
	Attack	2	2	2	2	8	
		8	8	8	8		
		16		16			

Factor	Power
Aircraft Type	0.93
UAS Support	0.93
Light	0.93
Mission Type	0.93

80% confidence level, signal-to-noise = 1.0

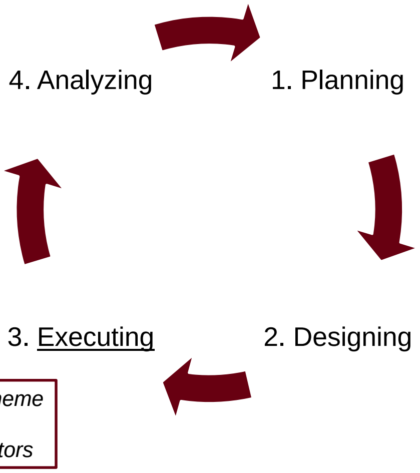
What we ended up with:

		With UAS		Without UAS			
	Mission Type	Day	Night	Day	Night		
AB2	Recon	1	0	2	2	5	12
	Attack	3	2	2	0	7	
AB3	Recon	1	1	1	2	5	16
	Attack	4	3	2	2	11	
		9	6	7	6		
		15		13			

Factor	Power
Aircraft Type	0.89
UAS Support	0.87
Light	0.89
Mission Type	0.85

80% confidence level, signal-to-noise = 1.0

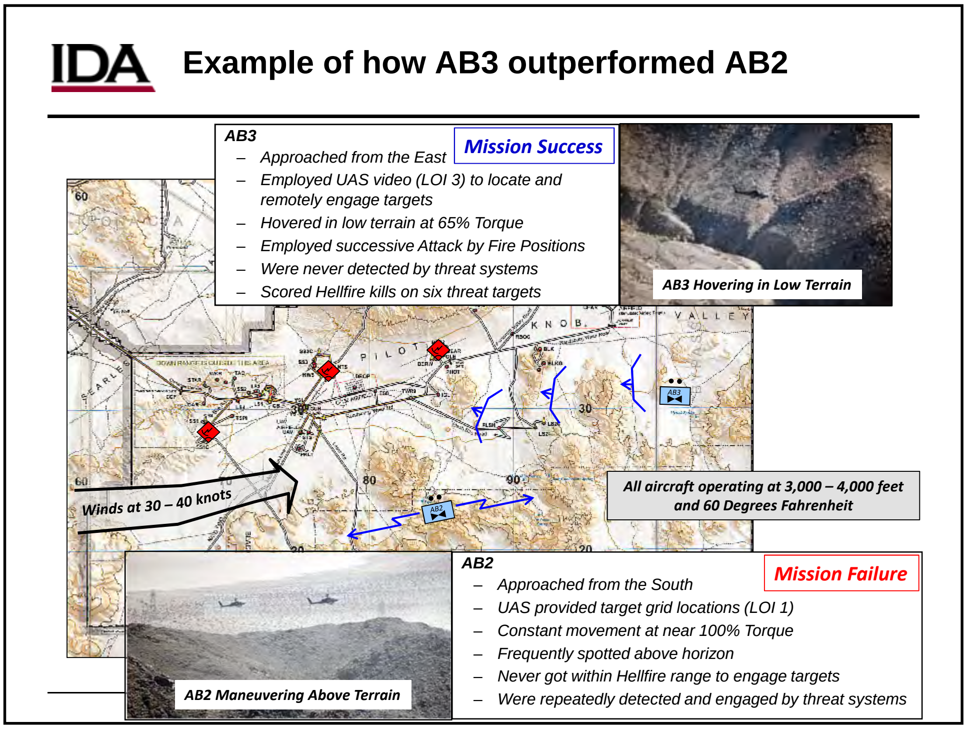
Stage 3: Executing



- *Randomization Scheme*
- *Blocking*
- *Hard to change factors*

IOT&E Execution

																	MARCH																															APRIL													
13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	1	2	3	4	5	6	7	8	9	10	11	12	13	14
NTC ROTATION 12-04																					NTC ROTATION 12-05 (3-3 HBCT)																																												
3	4	5	6	7	8	9	10	11	12	13	14			REGEN							RSOI					1	2	3	4	5	6	7	8	9	10	11	12	13	14			REGEN							WHITE WEEK							RSOI					
Gray Eagle Support											NAT	NAT	NAT	NAT	NAT	NAT	NAT	NAT	NAT	UAS	UAS	WXX		UAS	UAS	UAS		UAS	UAS	UAS		WXX	WXX	WXX	UAS	UAS	UAS	UAS					UAS	UAS	UAS				UAS	UAS	WXX	WXX					UAS		WXX	WXX	
Apache Flights									🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩		🟩	🟩	🟩		🟩	🟩	🟩		🟩	🟩	🟩	🟩	🟩	🟩	🟩					🟩	🟩	🟩				🟩	🟩	🟩	🟩				🟩	🟩		WXX	WXX	
AB3 TRAINING (MESA, AZ)					1/1 ARB Pilots Travel to Barstow-Daggett				Local Orientation AB3 Flight Training Unit Training					PT1	PT2	Local Orientation AB3 Flight Training Unit Training							OTRR3a	FDTE Night				FDTE Day			OTRR3b	IOT&E Missions Day and Night										GUNNERY TRAINING						GUNNERY							CHINA LAKE			CHINA LAKE		
Total Days at Barstow-Daggett																																																												
									1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53



Example of how AB3 outperformed AB2

AB3 — *Mission Success*

- *Approached from the East*
- *Employed UAS video (LOI 3) to locate and remotely engage targets*
- *Hovered in low terrain at 65% Torque*
- *Employed successive Attack by Fire Positions*
- *Were never detected by threat systems*
- *Scored Hellfire kills on six threat targets*

AB3 Hovering in Low Terrain

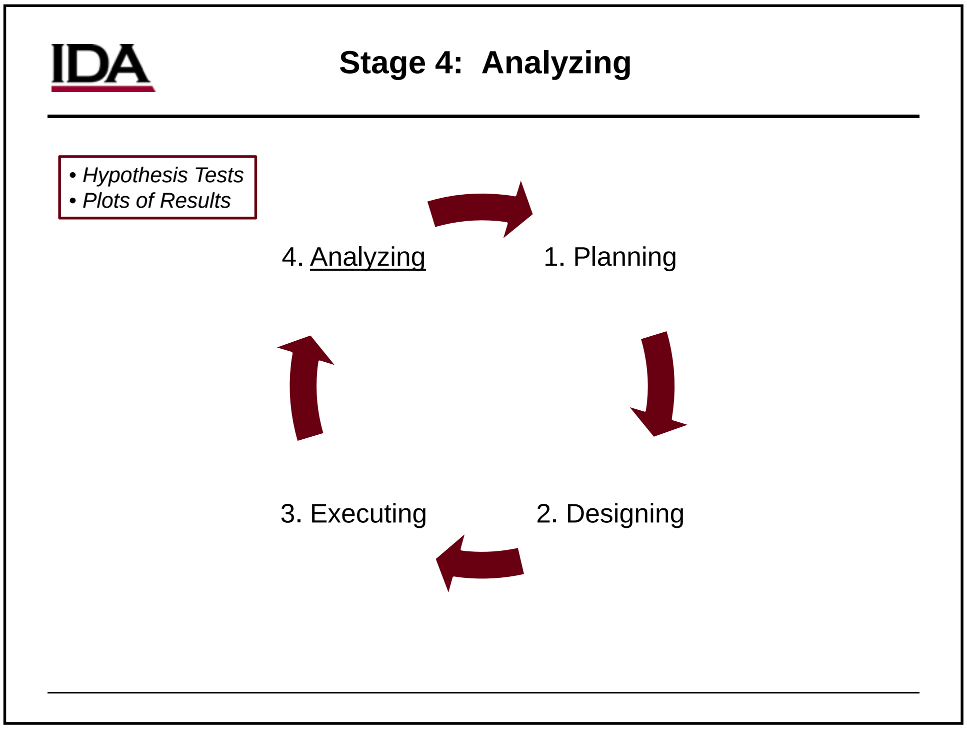
Winds at 30 – 40 knots

All aircraft operating at 3,000 – 4,000 feet and 60 Degrees Fahrenheit

AB2 Maneuvering Above Terrain

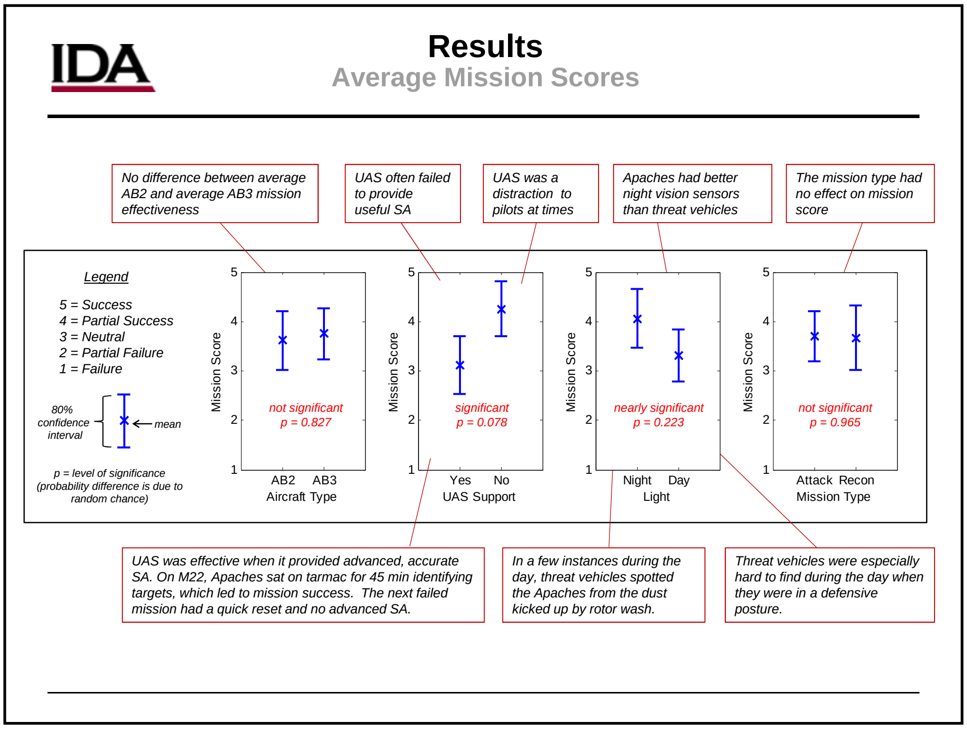
AB2 — *Mission Failure*

- *Approached from the South*
- *UAS provided target grid locations (LOI 1)*
- *Constant movement at near 100% Torque*
- *Frequently spotted above horizon*
- *Never got within Hellfire range to engage targets*
- *Were repeatedly detected and engaged by threat systems*



Stage 4: Analyzing

- *Hypothesis Tests*
- *Plots of Results*



IDA

Results

Average Mission Scores

No difference between average AB2 and average AB3 mission effectiveness

UAS often failed to provide useful SA

UAS was a distraction to pilots at times

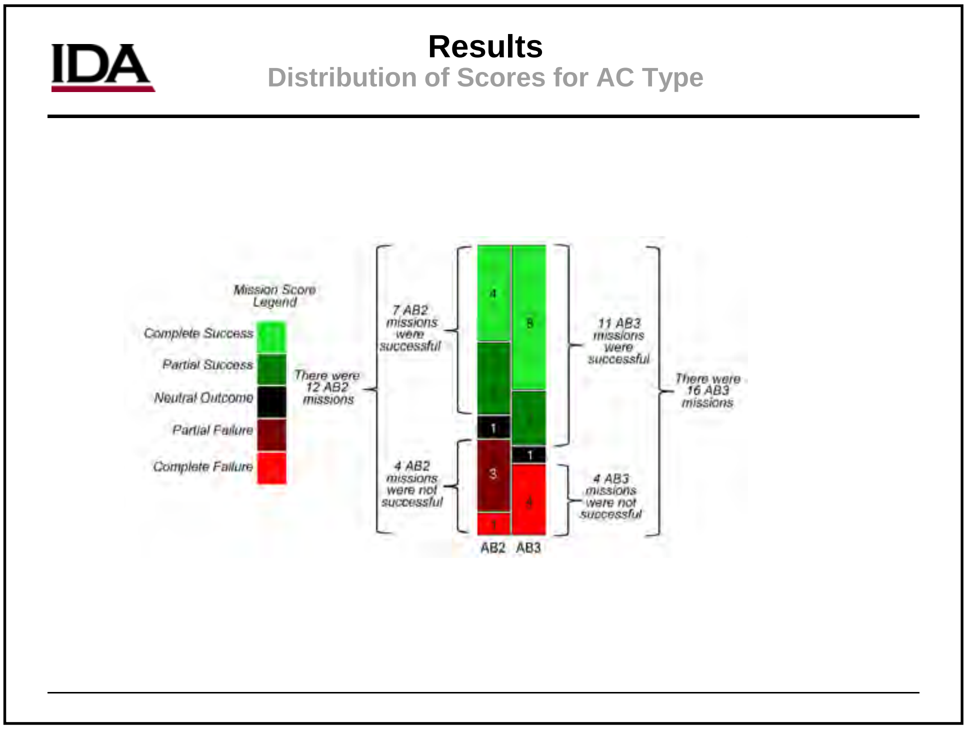
Apaches had better night vision sensors than threat vehicles

The mission type had no effect on mission score

UAS was effective when it provided advanced, accurate SA. On M22, Apaches sat on tarmac for 45 min identifying targets, which led to mission success. The next failed mission had a quick reset and no advanced SA.

In a few instances during the day, threat vehicles spotted the Apaches from the dust kicked up by rotor wash.

Threat vehicles were especially hard to find during the day when they were in a defensive posture.



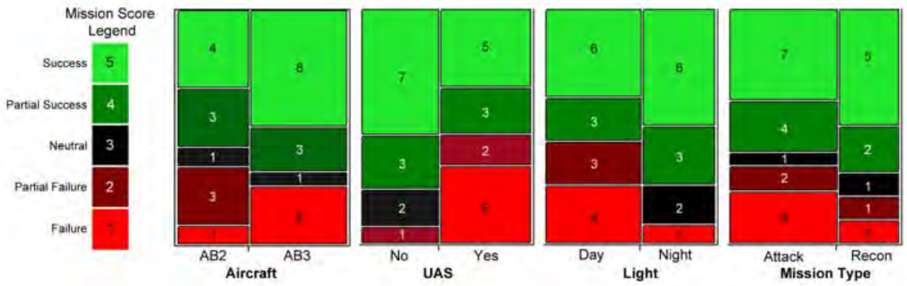
IDA

Results

Distribution of Scores for AC Type

Results

Distribution of Scores for all Factors



AB2 scores tended to be more neutral, while AB3 had more complete successes and failures. Half of the AB3 missions were complete successes. Six of 8 of the AB3 complete successes were attack missions

Five of 15 missions with UAS were complete failures. There were no missions without UAS that were complete failures. Four of the 5 complete failures with UAS support were AB3 missions.

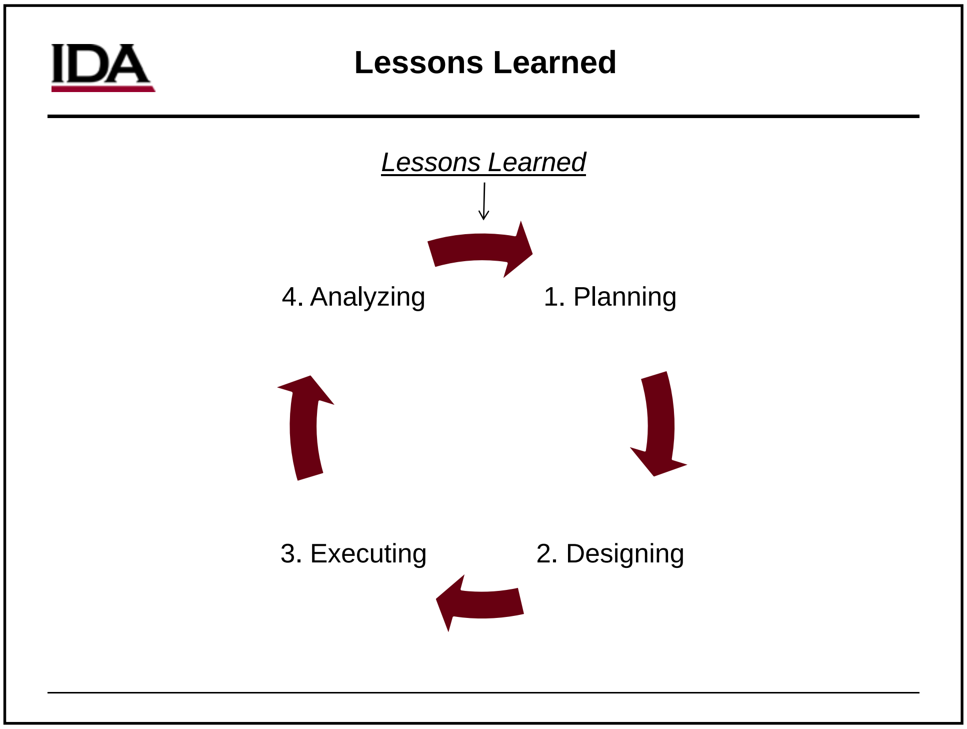
There was only one night mission that was worse than neutral. Half of the day missions were worse than neutral. All of the complete failures during the day had UAS support.

The distributions of scores for attack and recon missions were similar. Both of the failed Recon missions occurred during the day. Five of the 6 failed attack missions occurred with UAS.

Type of	**Aircraft Type**		**UAS Support**		**Light**		**Mission Type**	
Test	ChSquare	P>ChSquare	ChSquare	P>ChSquare	ChSquare	P>ChSquare	ChiSquare	P>ChSquare
Likelihood Ratio	6.872	0.143*	10.236	0.037*	8.286	0.082*	0.964	0.915
Pearson	5.678	0.225	7.562	0.109*	6.358	0.174*	0.923	0.921

Mission Effectiveness Conclusions

- **DOE supported conclusions:**
  - AB3 has the potential to be the superior war fighter
  - When implemented properly UAS support leads to increase mission effectiveness, and leads to failure when not
  - TTPs for UAS/AB3 teaming need to be refined
- **Conclusions based on crew observations:**
  - Crews liked flight performance, speed, and power of AB3
  - Crews believed that UAS/AB3 teaming enhanced situational awareness
- **Conclusions based on specific examples:**
  - The two China Lake missions demonstrated that AB3 engine performance increased mission effectiveness



Lessons Learned

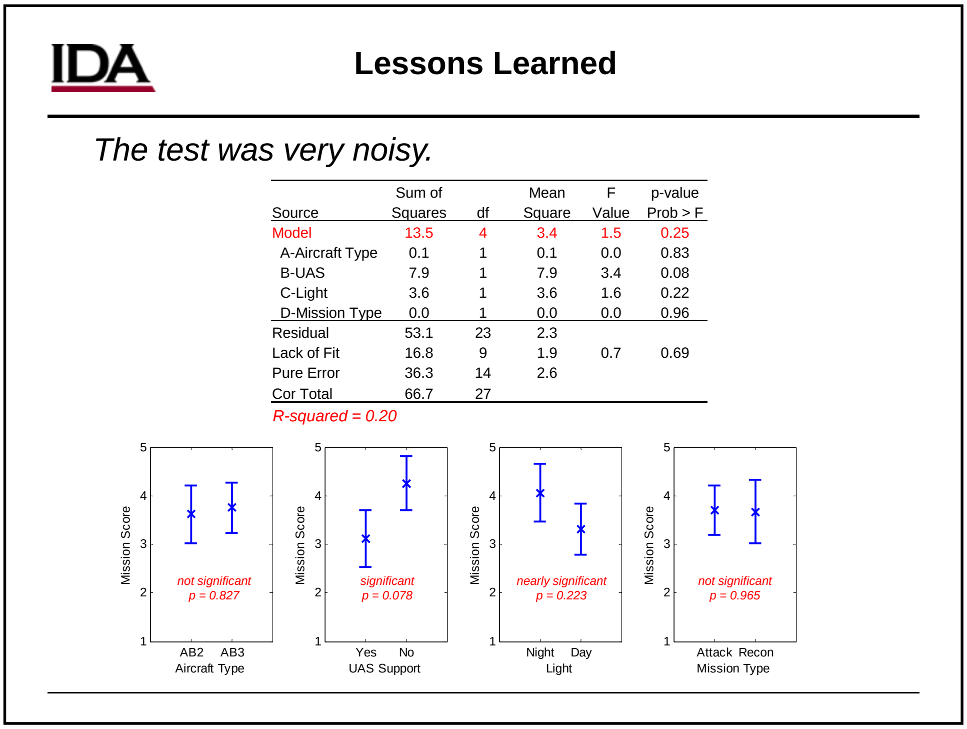
Lessons Learned

4. Analyzing

1. Planning

3. Executing

2. Designing

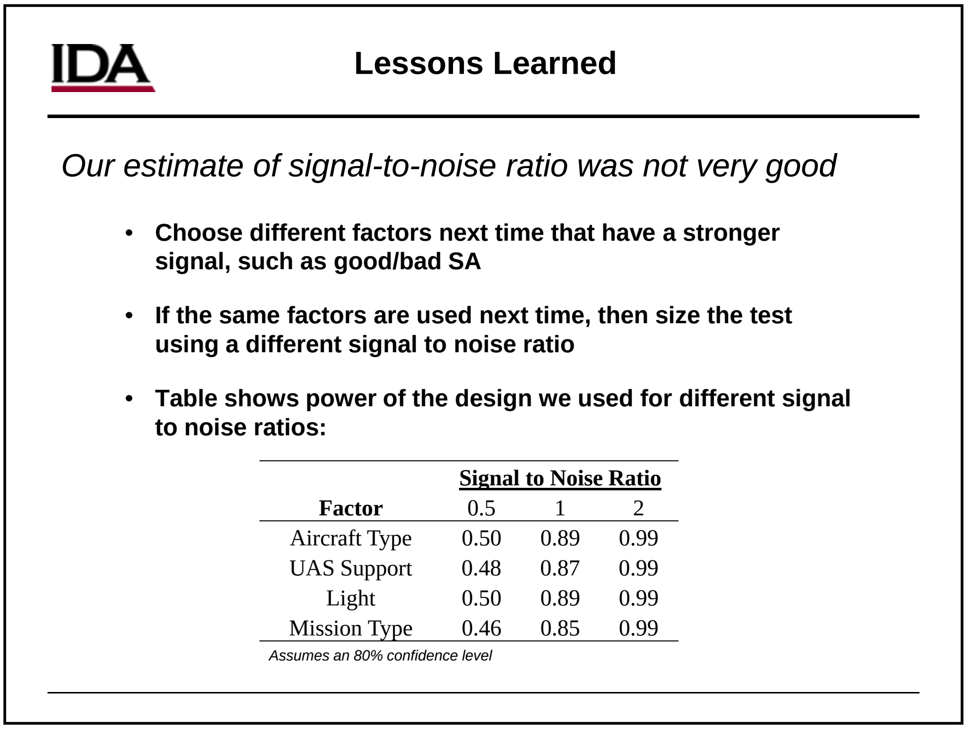


Lessons Learned

The test was very noisy.

Source	Sum of Squares	df	Mean Square	F Value	p-value Prob > F
Model	13.5	4	3.4	1.5	0.25
A-Aircraft Type	0.1	1	0.1	0.0	0.83
B-UAS	7.9	1	7.9	3.4	0.08
C-Light	3.6	1	3.6	1.6	0.22
D-Mission Type	0.0	1	0.0	0.0	0.96
Residual	53.1	23	2.3		
Lack of Fit	16.8	9	1.9	0.7	0.69
Pure Error	36.3	14	2.6		
Cor Total	66.7	27			

R-squared = 0.20



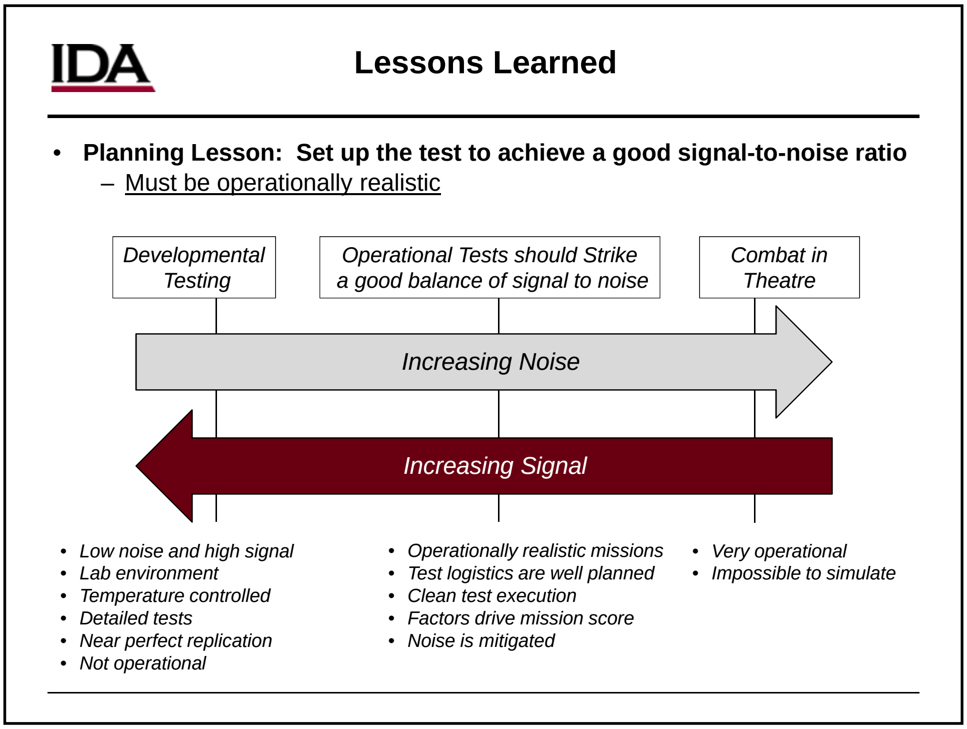
Lessons Learned

Our estimate of signal-to-noise ratio was not very good

- **Choose different factors next time that have a stronger signal, such as good/bad SA**
- **If the same factors are used next time, then size the test using a different signal to noise ratio**
- **Table shows power of the design we used for different signal to noise ratios:**

	Signal to Noise Ratio		
Factor	0.5	1	2
Aircraft Type	0.50	0.89	0.99
UAS Support	0.48	0.87	0.99
Light	0.50	0.89	0.99
Mission Type	0.46	0.85	0.99

Assumes an 80% confidence level



Lessons Learned

- **Planning Lesson: Set up the test to achieve a good signal-to-noise ratio**
  - Must be operationally realistic

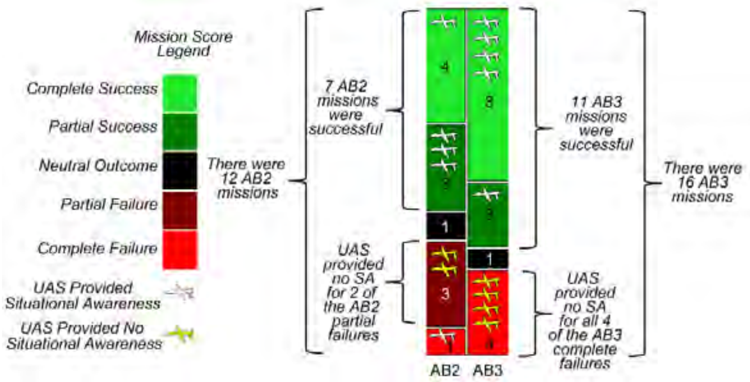
- *Low noise and high signal*
- *Lab environment*
- *Temperature controlled*
- *Detailed tests*
- *Near perfect replication*
- *Not operational*

- *Operationally realistic missions*
- *Test logistics are well planned*
- *Clean test execution*
- *Factors drive mission score*
- *Noise is mitigated*

- *Very operational*
- *Impossible to simulate*

Lessons Learned

Situational awareness drives mission score.



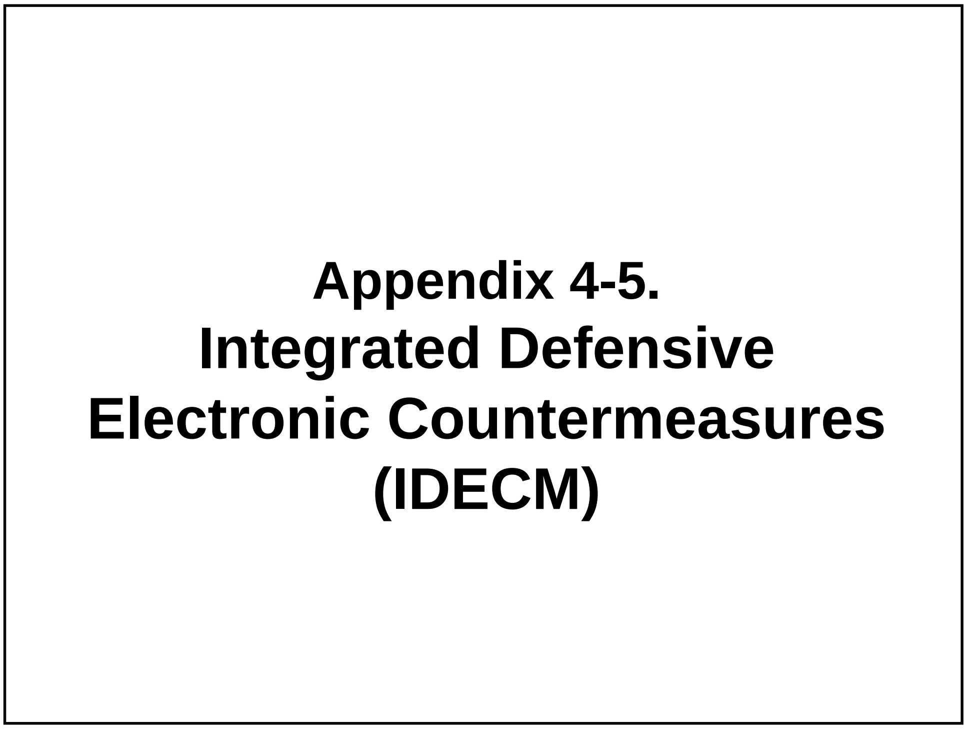
Future FOTE1 DOE Test Plan

- **Response:**
  - Mission Score
- **Factors:**
  - SA Quality (good vs. bad)
  - SA timeliness (early vs. late)
  - Threat Density (1x vs. 2x)
- **Design:**
  - Fully replicated 2^3 factorial design
  - Supports a main effects model
- **Execution:**
  - Two phases
    - Phase 1) DOE phase that is tightly controlled.
    - Phase 2) Demonstration phase.

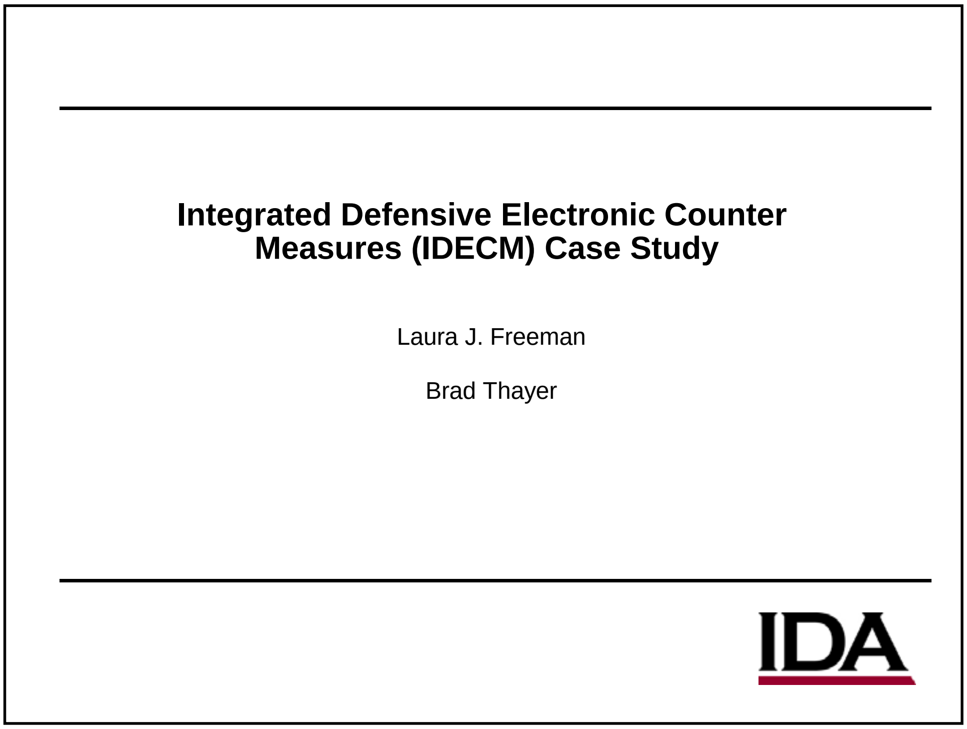
IDA

The End

This page intentionally left blank.



Appendix 4-5.
Integrated Defensive Electronic Countermeasures (IDECM)



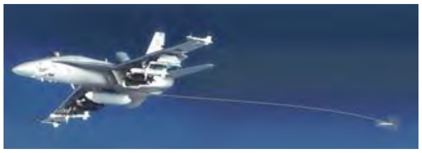
Integrated Defensive Electronic Counter Measures (IDECM) Case Study

Laura J. Freeman

Brad Thayer

Jammer Case Study

- **Goals of the Test**
  - Characterize performance of a new jammer
  - New jammer is required to be a measurable improvement over the legacy jammer
  - Screen factors for future testing
- **Response variables**
  - Reduction in lethality
  - Miss distance of missile shots



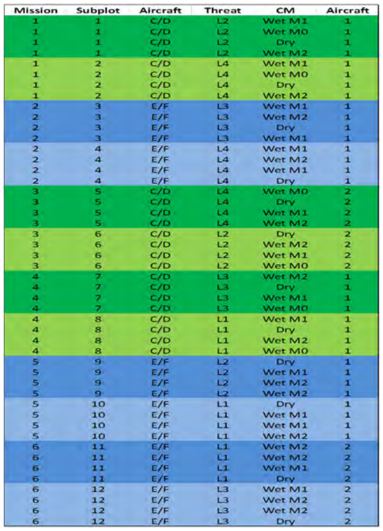
- **Factors and levels**
  - Aircraft variant: 2 variants (A1, A2)
  - Threat: 4 different type of threats (T1, T2, T3, T4)
  - Jammer type: legacy and new
  - Counter Measures: dry, wet Non-maneuvering, or wet with one of three maneuvers
  - Number of sorties per mission: 1 ship or 2 ship

Jammer Case Study: DOE Solution

- **DOE Challenges**
  - Complete randomization is not possible
    - » Each mission allowed for up to 8 potential engagements but aircraft and threat could not be easily varied from run to run
  - Disallowed combinations of factors
    - » The legacy system can only be used on one type of aircraft
    - » The legacy system will only be flown in a subset of the operational envelope
      - Dry and wet non-maneuvering
      - Single ship missions
    - » The second aircraft variant can only do a subset of the three maneuvers
  - Limited sample size
    - » 8 operational sorties
- **DOE Solution**
  - D-optimal Split-Split-Plot Design
    - » Allows for restrictions in randomization
  - Creation of new "factors"
    - » Combine original factors into allowed cases for design generation
    - » Accounts for disallowed combinations of factors

Jammer Case Study: Run Table

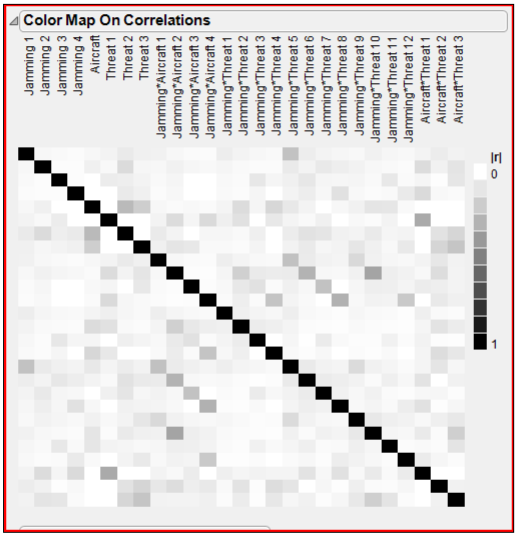
- **Design approach (a.k.a tricks of the trade)**
  - Use a generation variable to appropriately weight runs and eliminate some disallowed combinations
  - Practice counting to make sure the right number of whole-plots and sub-plots are selected
  - Customize design & import into software to check properties **on and split-plot solution are similar**
  - Split-plot requires replication of the hard-to-change factor

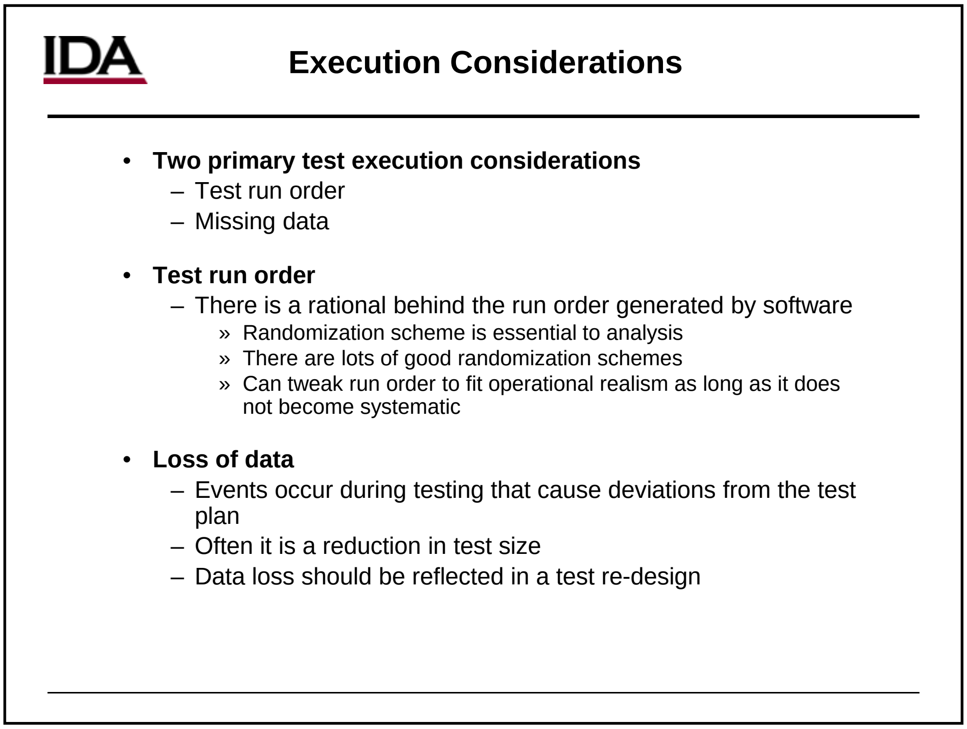


Mission	Subplot	Aircraft	Threat	CM	Aircraft
1	1	C/D	L2	Wet M1	1
1	1	C/D	L2	Wet M0	1
1	1	C/D	L2	Dry	1
1	1	C/D	L2	Wet M2	1
1	2	C/D	L4	Wet M1	1
1	2	C/D	L4	Wet M0	1
1	2	C/D	L4	Dry	1
1	2	C/D	L4	Wet M2	1
2	3	E/F	L3	Wet M1	1
2	3	E/F	L3	Wet M2	1
2	3	E/F	L3	Dry	1
2	3	E/F	L3	Wet M1	1
2	4	E/F	L4	Wet M1	1
2	4	E/F	L4	Wet M2	1
2	4	E/F	L4	Wet M1	1
2	4	E/F	L4	Dry	1
3	5	C/D	L4	Wet M0	2
3	5	C/D	L4	Dry	2
3	5	C/D	L4	Wet M1	2
3	5	C/D	L4	Wet M2	2
3	6	C/D	L2	Dry	2
3	6	C/D	L2	Wet M2	2
3	6	C/D	L2	Wet M1	2
3	6	C/D	L2	Wet M0	2
4	7	C/D	L3	Wet M2	1
4	7	C/D	L3	Dry	1
4	7	C/D	L3	Wet M1	1
4	7	C/D	L3	Wet M0	1
4	8	C/D	L1	Wet M1	1
4	8	C/D	L1	Dry	1
4	8	C/D	L1	Wet M2	1
4	8	C/D	L1	Wet M0	1
5	9	E/F	L2	Dry	1
5	9	E/F	L2	Wet M1	1
5	9	E/F	L2	Wet M2	1
5	9	E/F	L2	Wet M2	1
5	10	E/F	L1	Dry	1
5	10	E/F	L1	Wet M1	1
5	10	E/F	L1	Wet M1	1
5	10	E/F	L1	Wet M2	1
6	11	E/F	L1	Wet M2	2
6	11	E/F	L1	Wet M2	2
6	11	E/F	L1	Wet M1	2
6	11	E/F	L1	Dry	2
6	12	E/F	L3	Wet M1	2
6	12	E/F	L3	Wet M2	2
6	12	E/F	L3	Wet M2	2
6	12	E/F	L3	Dry	2

Jammer Case Study: Design Properties

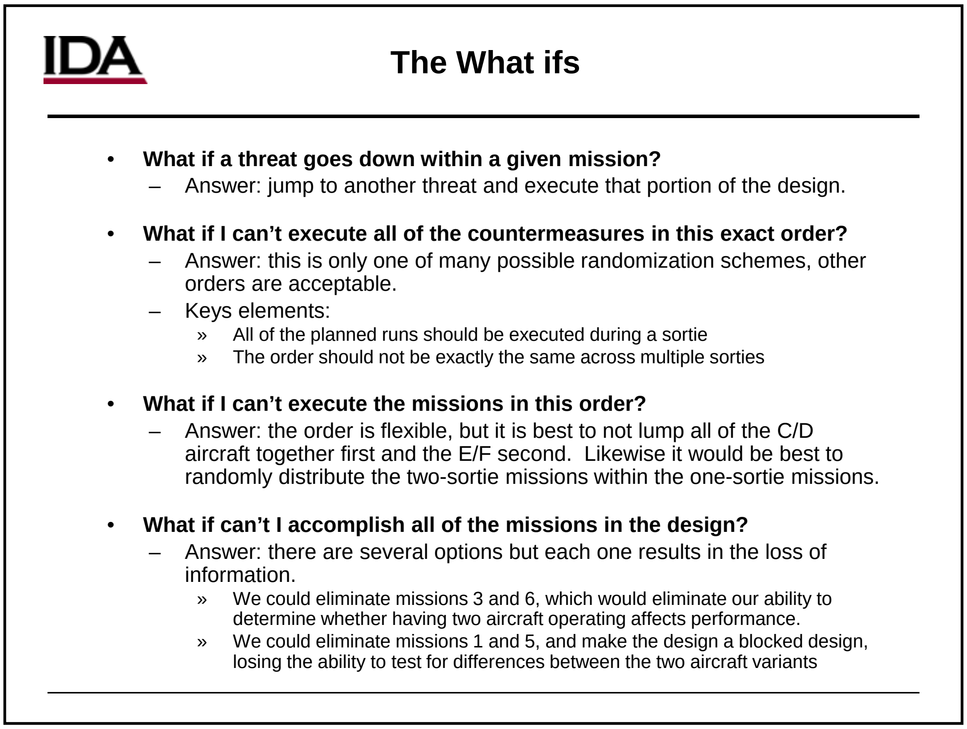
Power Numbers		
Factor	S:N = 1	S:N = 2
Aircraft	0.258	0.745
Variant	0.258	0.745
Jamming	0.975	0.999
Threat	0.388	0.844
Wingman	0.258	0.745





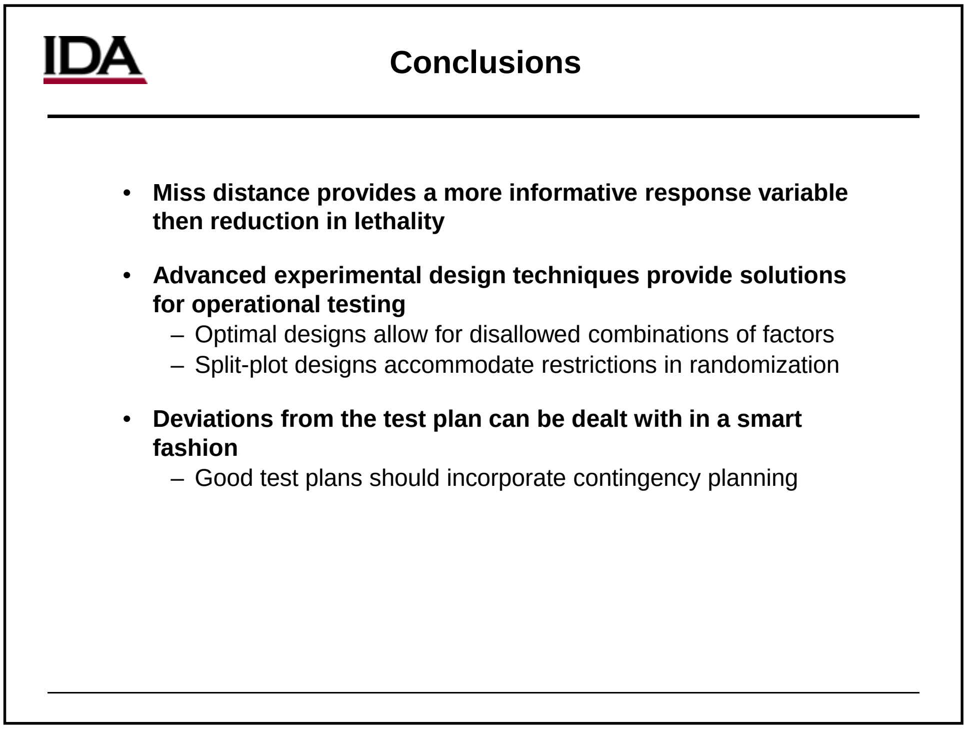
Execution Considerations

- **Two primary test execution considerations**
  - Test run order
  - Missing data
- **Test run order**
  - There is a rational behind the run order generated by software
    - Randomization scheme is essential to analysis
    - There are lots of good randomization schemes
    - Can tweak run order to fit operational realism as long as it does not become systematic
- **Loss of data**
  - Events occur during testing that cause deviations from the test plan
  - Often it is a reduction in test size
  - Data loss should be reflected in a test re-design



The What ifs

- **What if a threat goes down within a given mission?**
  - Answer: jump to another threat and execute that portion of the design.
- **What if I can’t execute all of the countermeasures in this exact order?**
  - Answer: this is only one of many possible randomization schemes, other orders are acceptable.
  - Keys elements:
    - All of the planned runs should be executed during a sortie
    - The order should not be exactly the same across multiple sorties
- **What if I can’t execute the missions in this order?**
  - Answer: the order is flexible, but it is best to not lump all of the C/D aircraft together first and the E/F second. Likewise it would be best to randomly distribute the two-sortie missions within the one-sortie missions.
- **What if can’t I accomplish all of the missions in the design?**
  - Answer: there are several options but each one results in the loss of information.
    - We could eliminate missions 3 and 6, which would eliminate our ability to determine whether having two aircraft operating affects performance.
    - We could eliminate missions 1 and 5, and make the design a blocked design, losing the ability to test for differences between the two aircraft variants

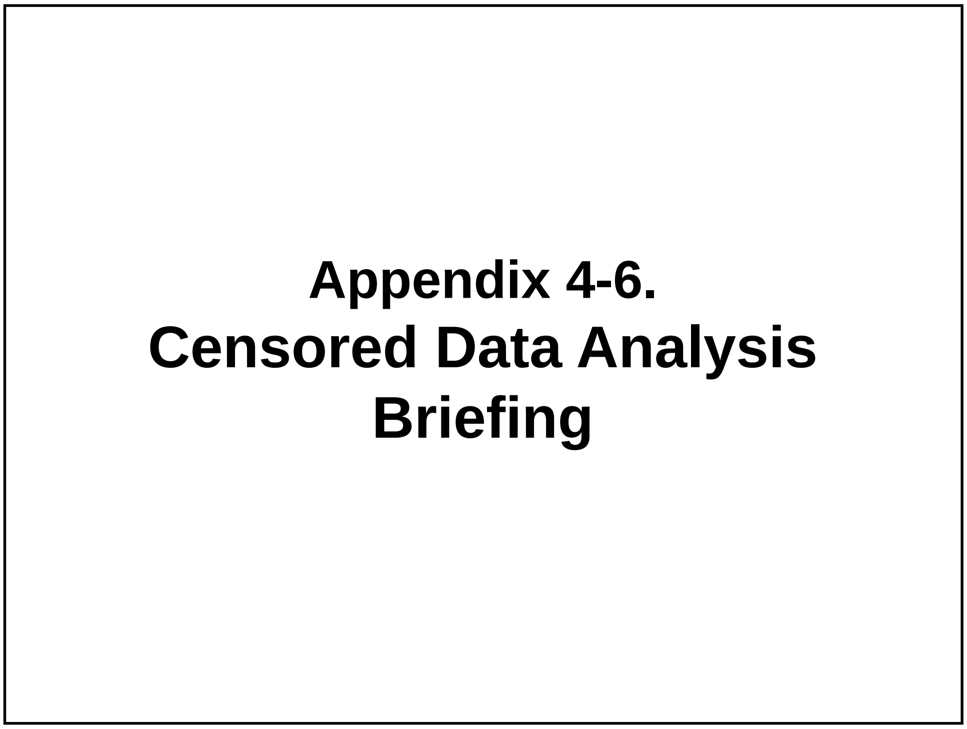


IDA

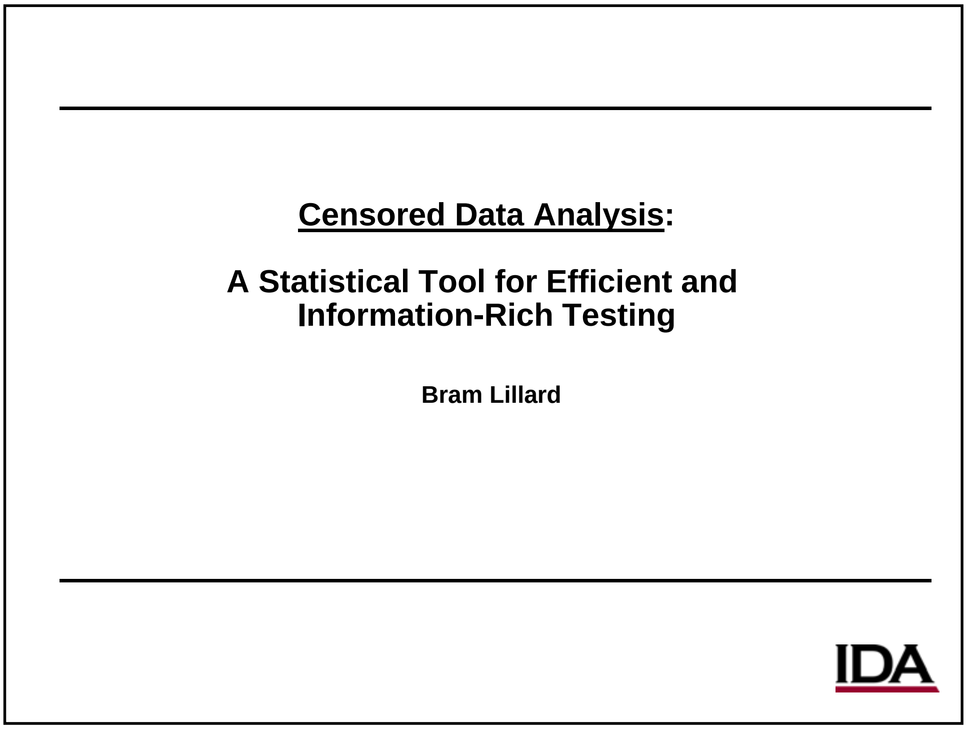
Conclusions

- **Miss distance provides a more informative response variable then reduction in lethality**
- **Advanced experimental design techniques provide solutions for operational testing**
  - Optimal designs allow for disallowed combinations of factors
  - Split-plot designs accommodate restrictions in randomization
- **Deviations from the test plan can be dealt with in a smart fashion**
  - Good test plans should incorporate contingency planning

This page intentionally left blank.



Appendix 4-6. Censored Data Analysis Briefing



Censored Data Analysis:

A Statistical Tool for Efficient and Information-Rich Testing

Bram Lillard

IDA

RECALL….

Continuous Metrics for Efficient and Effective Testing

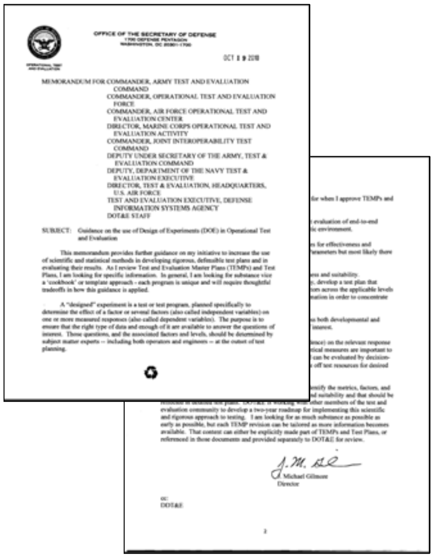
Laura J. Freeman
&
Bram Lillard

NDIA National Test and Evaluation Conference

March 15, 2012

DOT&E Guidance

Dr. Gilmore's October 19, 2010 Memo to OTAs



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

OPERATIONAL TEST AND EVALUATION

OCT 1 9 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
Parameters but most likely there

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

ss both developmental and
f interest.

dence) on the relevant response
tical measures are important to
l 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. M. Gil
J. Michael Gilmore
Director

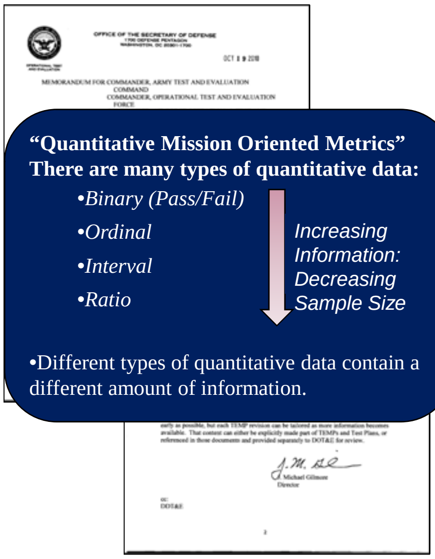
cc:
DOT&E

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 understand "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.

DOT&E Guidance

Dr. Gilmore’s October 19, 2010 Memo to OTAs



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

OCT 1 9 2010

OPERATIONAL TEST AND EVALUATION

MEMORANDUM FOR COMMANDER, ARMY TEST AND EVALUATION COMMAND
COMMANDER, OPERATIONAL TEST AND EVALUATION FORCE

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. M. Gilmore
J. Michael Gilmore
Director

cc:
DOT&E

2

“Quantitative Mission Oriented Metrics”
There are many types of quantitative data:

- *Binary (Pass/Fail)*
- *Ordinal*
- *Interval*
- *Ratio*

Increasing Information: Decreasing Sample Size

- Different types of quantitative data contain a different amount of information.

- ❑ **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 understand "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.

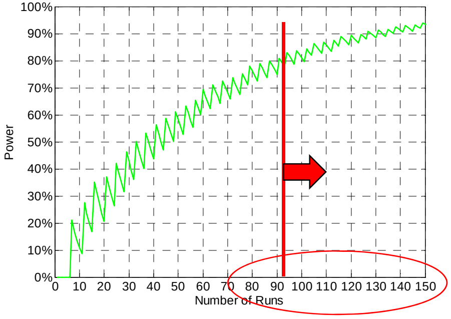
The Binomial Conundrum

- **Testing for a binary metric requires large sample sizes**

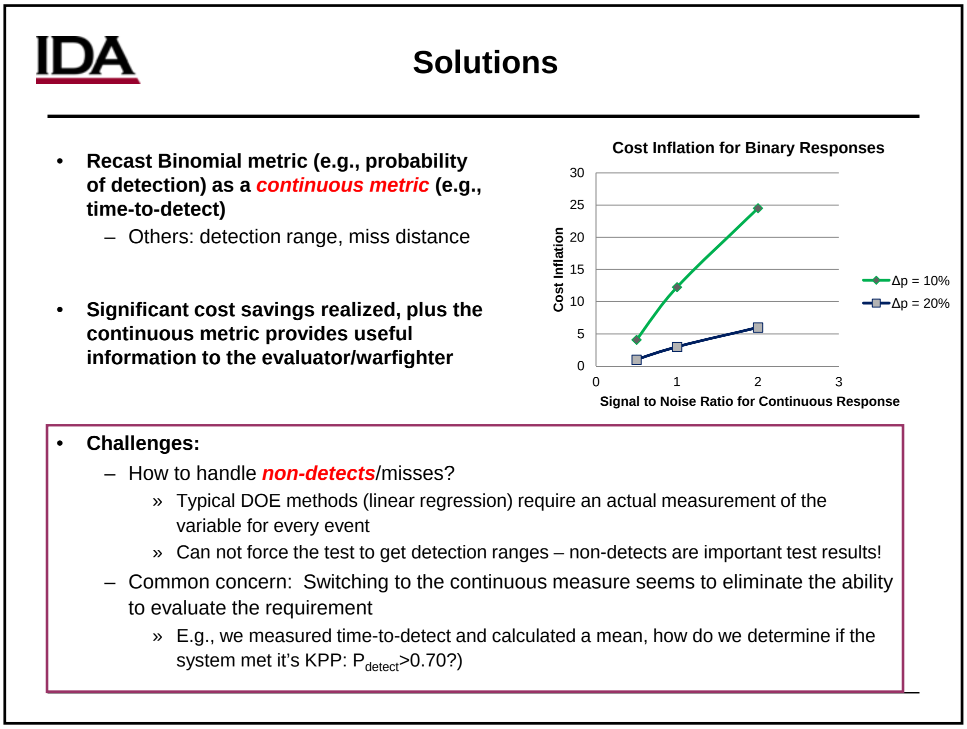
Sample Size Requirements

Sample Size	90% Confidence Interval Width (p = 0.5)	90% Confidence Interval Width (p = 0.8)
10	± 26%	± 21%
50	± 11.6%	± 9.3%
100	± 8.2%	± 6.6%
500	± 3.7%	± 2.9%

Power Calculation, 90% confidence, Performance 10% better than threshold



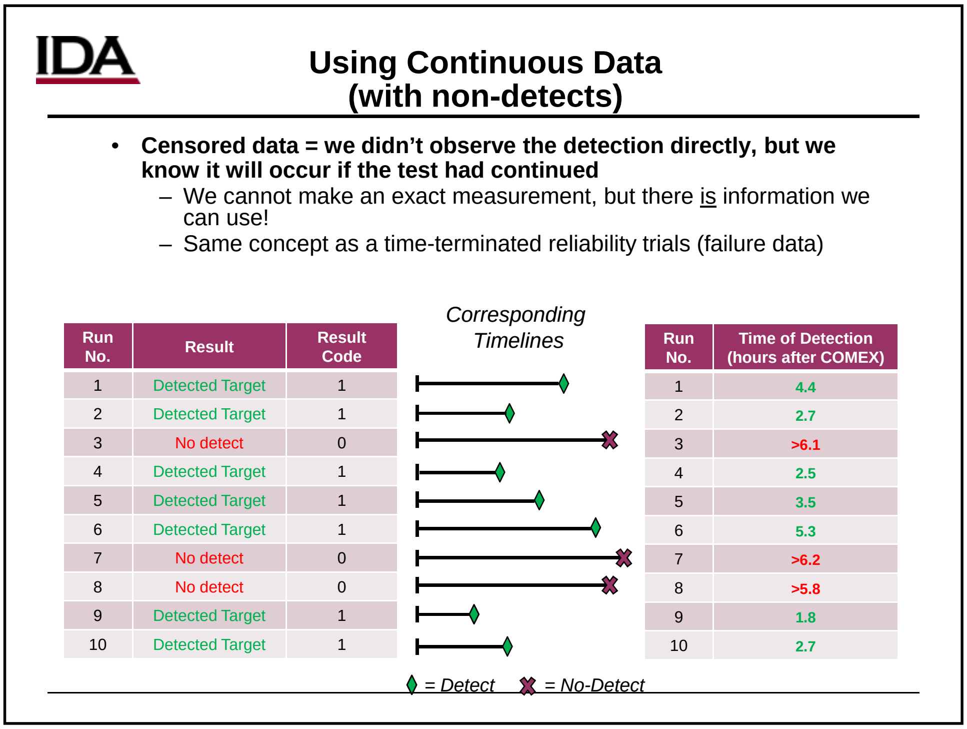
- **Difficult (impossible?) to achieve acceptable power for factor analysis unless many runs *(often >100)* can be resourced**
  - Non-starter for implementing DOE concepts (characterizing performance across multiple conditions)



Solutions

- **Recast Binomial metric (e.g., probability of detection) as a *continuous metric* (e.g., time-to-detect)**
  - Others: detection range, miss distance
- **Significant cost savings realized, plus the continuous metric provides useful information to the evaluator/warfighter**

- **Challenges:**
  - How to handle ***non-detects***/misses?
    - » Typical DOE methods (linear regression) require an actual measurement of the variable for every event
    - » Can not force the test to get detection ranges – non-detects are important test results!
  - Common concern: Switching to the continuous measure seems to eliminate the ability to evaluate the requirement
    - » E.g., we measured time-to-detect and calculated a mean, how do we determine if the system met it's KPP: P_{detect}>0.70?)



Using Continuous Data (with non-detects)

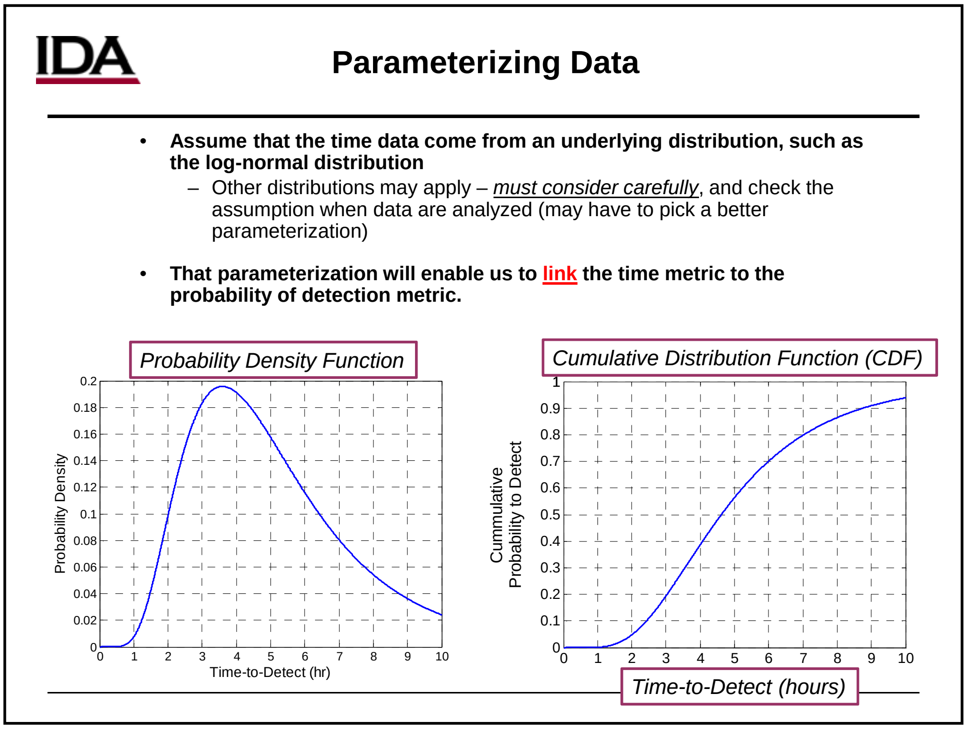
- **Censored data = we didn't observe the detection directly, but we know it will occur if the test had continued**
  - We cannot make an exact measurement, but there is information we can use!
  - Same concept as a time-terminated reliability trials (failure data)

Run No.	Result	Result Code
1	Detected Target	1
2	Detected Target	1
3	No detect	0
4	Detected Target	1
5	Detected Target	1
6	Detected Target	1
7	No detect	0
8	No detect	0
9	Detected Target	1
10	Detected Target	1

Corresponding Timelines

Run No.	Time of Detection (hours after COMEX)
1	4.4
2	2.7
3	>6.1
4	2.5
5	3.5
6	5.3
7	>6.2
8	>5.8
9	1.8
10	2.7

◆ *= Detect* ✖ *= No-Detect*



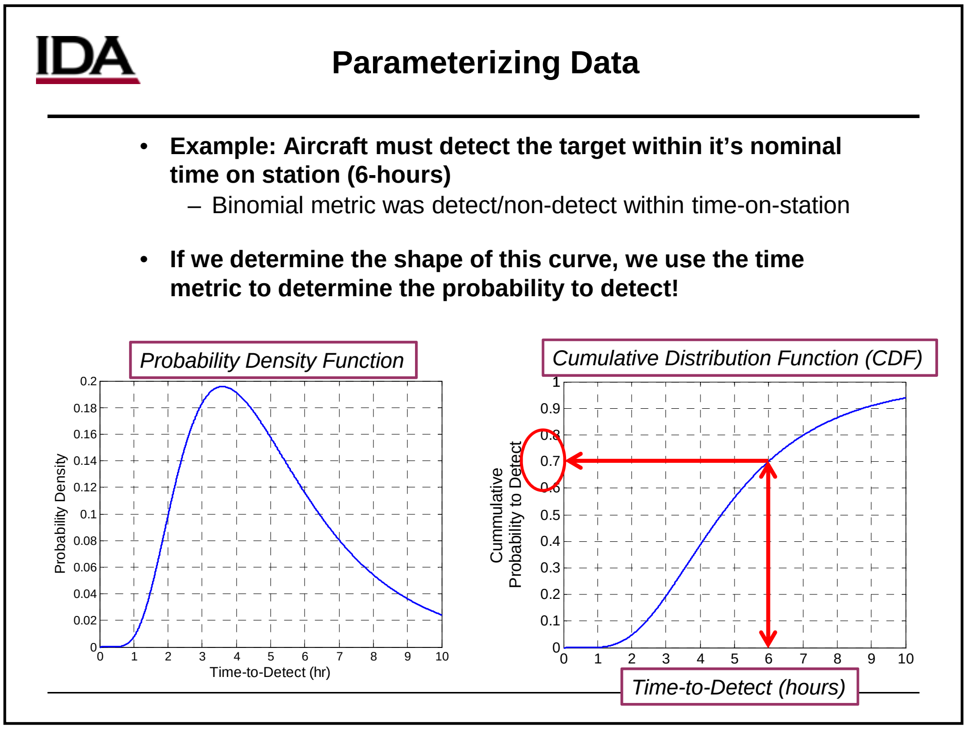
Parameterizing Data

- **Assume that the time data come from an underlying distribution, such as the log-normal distribution**
  – Other distributions may apply – *must consider carefully*, and check the assumption when data are analyzed (may have to pick a better parameterization)
- **That parameterization will enable us to link the time metric to the probability of detection metric.**

Probability Density Function

Cumulative Distribution Function (CDF)

Time-to-Detect (hours)



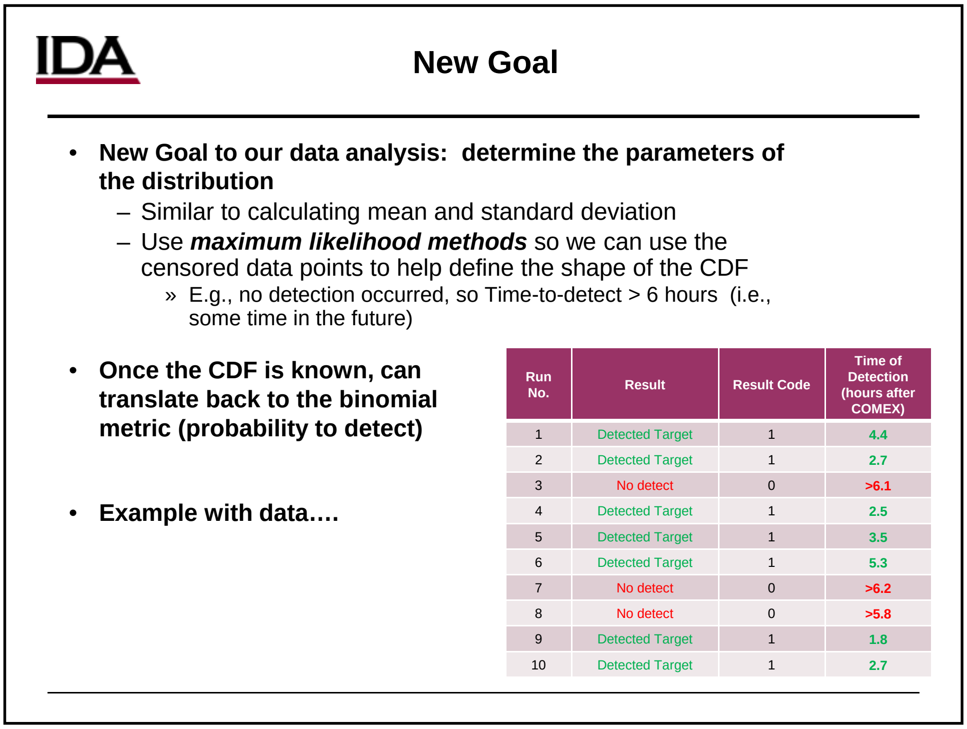
Parameterizing Data

- **Example: Aircraft must detect the target within it's nominal time on station (6-hours)**
  – Binomial metric was detect/non-detect within time-on-station
- **If we determine the shape of this curve, we use the time metric to determine the probability to detect!**

Probability Density Function

Cumulative Distribution Function (CDF)

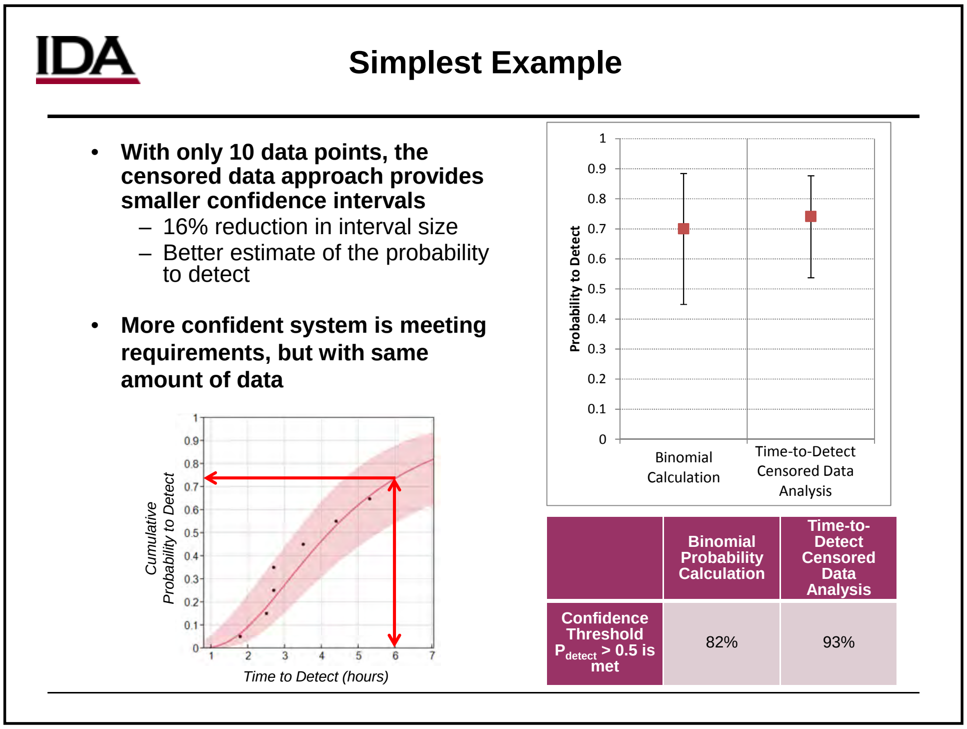
Time-to-Detect (hours)



New Goal

- **New Goal to our data analysis: determine the parameters of the distribution**
  - Similar to calculating mean and standard deviation
  - Use ***maximum likelihood methods*** so we can use the censored data points to help define the shape of the CDF
    - » E.g., no detection occurred, so Time-to-detect > 6 hours (i.e., some time in the future)
- **Once the CDF is known, can translate back to the binomial metric (probability to detect)**
- **Example with data….**

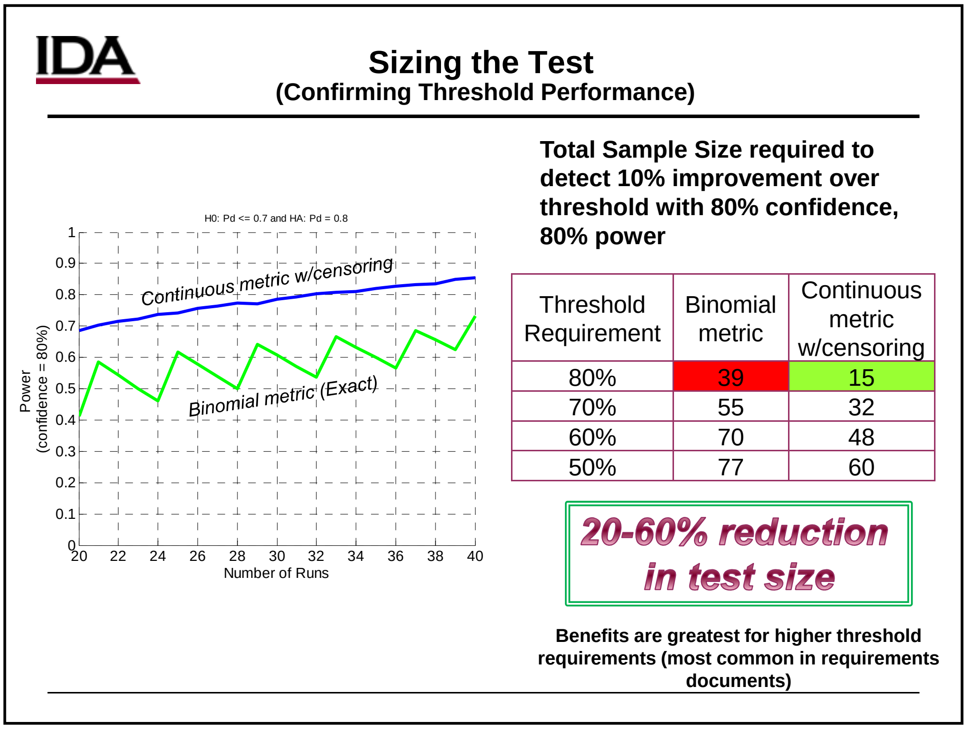
Run No.	Result	Result Code	Time of Detection (hours after COMEX)
1	Detected Target	1	4.4
2	Detected Target	1	2.7
3	No detect	0	>6.1
4	Detected Target	1	2.5
5	Detected Target	1	3.5
6	Detected Target	1	5.3
7	No detect	0	>6.2
8	No detect	0	>5.8
9	Detected Target	1	1.8
10	Detected Target	1	2.7



Simplest Example

- **With only 10 data points, the censored data approach provides smaller confidence intervals**
  - 16% reduction in interval size
  - Better estimate of the probability to detect
- **More confident system is meeting requirements, but with same amount of data**

	Binomial Probability Calculation	Time-to-Detect Censored Data Analysis
Confidence Threshold $P_{detect} > 0.5$ is met	82%	93%



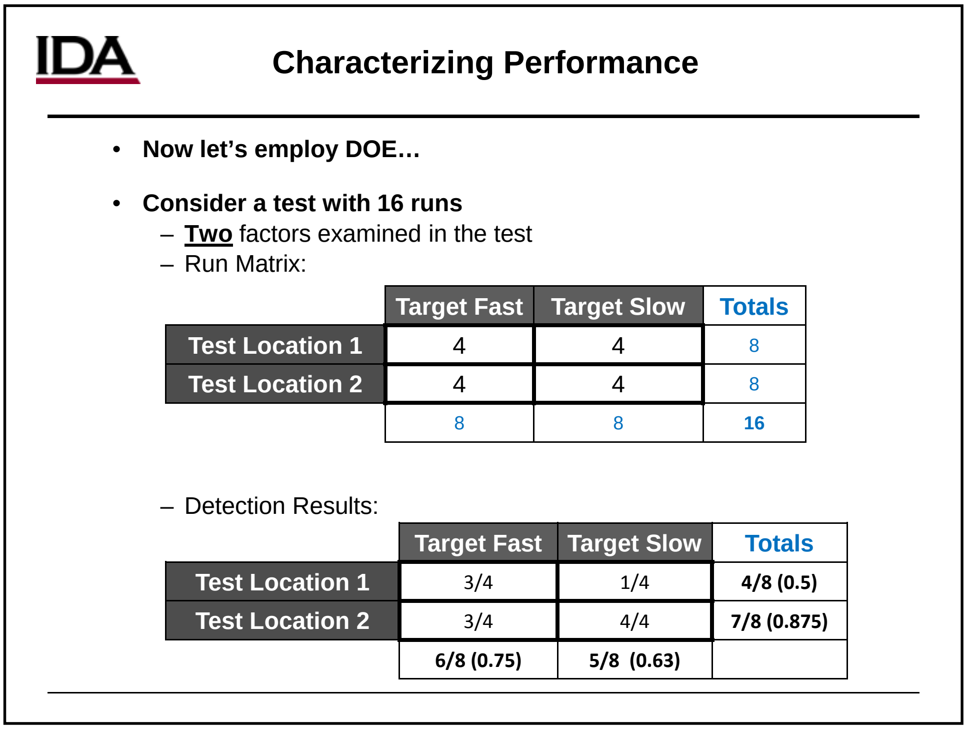
Sizing the Test
(Confirming Threshold Performance)

Total Sample Size required to detect 10% improvement over threshold with 80% confidence, 80% power

Threshold Requirement	Binomial metric	Continuous metric w/censoring
80%	39	15
70%	55	32
60%	70	48
50%	77	60

20-60% reduction in test size

Benefits are greatest for higher threshold requirements (most common in requirements documents)



Characterizing Performance

- **Now let's employ DOE…**
- **Consider a test with 16 runs**
  - <u>**Two**</u> factors examined in the test
  - Run Matrix:

	Target Fast	Target Slow	Totals
Test Location 1	4	4	8
Test Location 2	4	4	8
	8	8	**16**

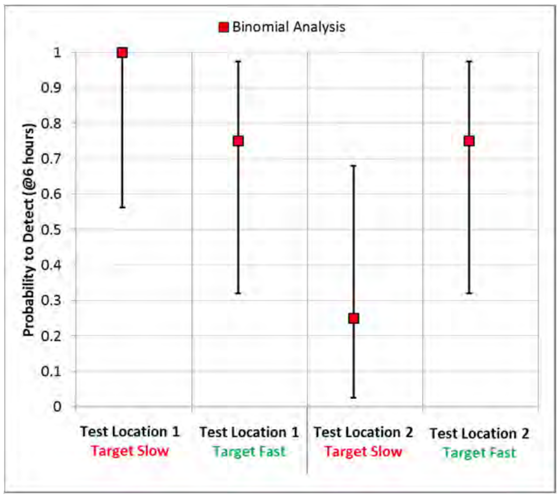
  - Detection Results:

	Target Fast	Target Slow	Totals
Test Location 1	3/4	1/4	**4/8 (0.5)**
Test Location 2	3/4	4/4	**7/8 (0.875)**
	6/8 (0.75)	**5/8 (0.63)**	

IDA

Attempt to Characterize Performance

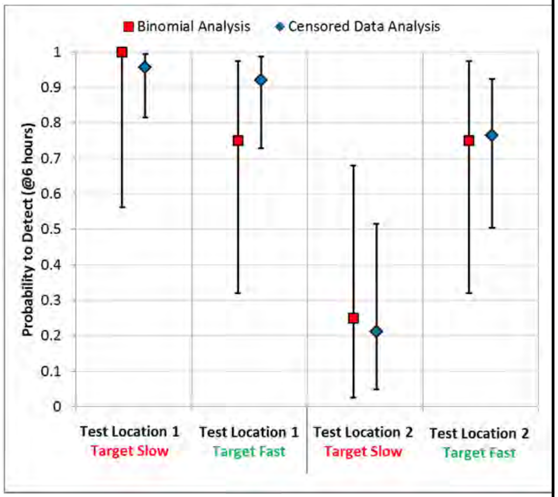
- **As expected, 4 runs in each condition is *insufficient* to characterize performance with a binomial metric**
- **Cannot tell which factor drives performance or which conditions will cause the system to meet/fail requirements**
- **Likely will only report a ‘roll-up’ of 11/16**
  - 90% confidence interval: [0.45, 0.87]



IDA

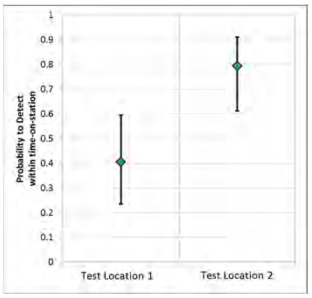
Characterizing Performance Better

- **Measure *time-to-detect* in lieu of binomial metric, employ censored data analysis…**
- **Significant reduction in confidence intervals!**
  - Now can tell significant differences in performance
    - » E.g., system is performing **poorly** in Location 2 against slow targets
  - We can confidently conclude performance is above threshold in three conditions
    - » Not possible with a “probability to detect” analysis!



Sizing Tests

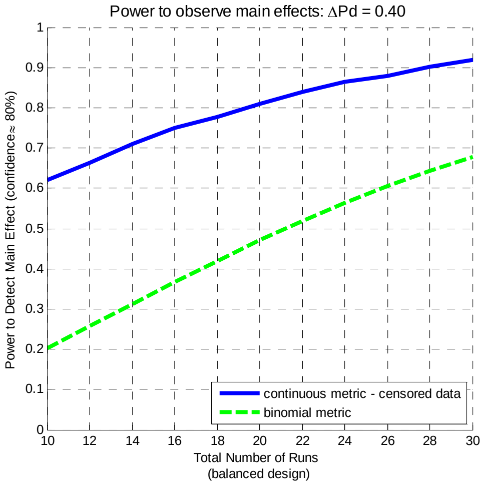
- **Why size a test based on ability to detect differences in P_{detect}?**
  - This is standard way to employ power calculations to detect factor effects in DOE methodology
  - We <u>are</u> interested in performance differences – this is how we *characterize performance* across the operational envelope
  - This is also how we ensure a level of precision occurs in our measurement of P_{detect} (size of the “error bars” will be determined)



If we size the test to detect this difference, then the confidence intervals on the results will be approx. this big

If the measured delta is different than assumed, still ensure a level of accuracy in the measurement

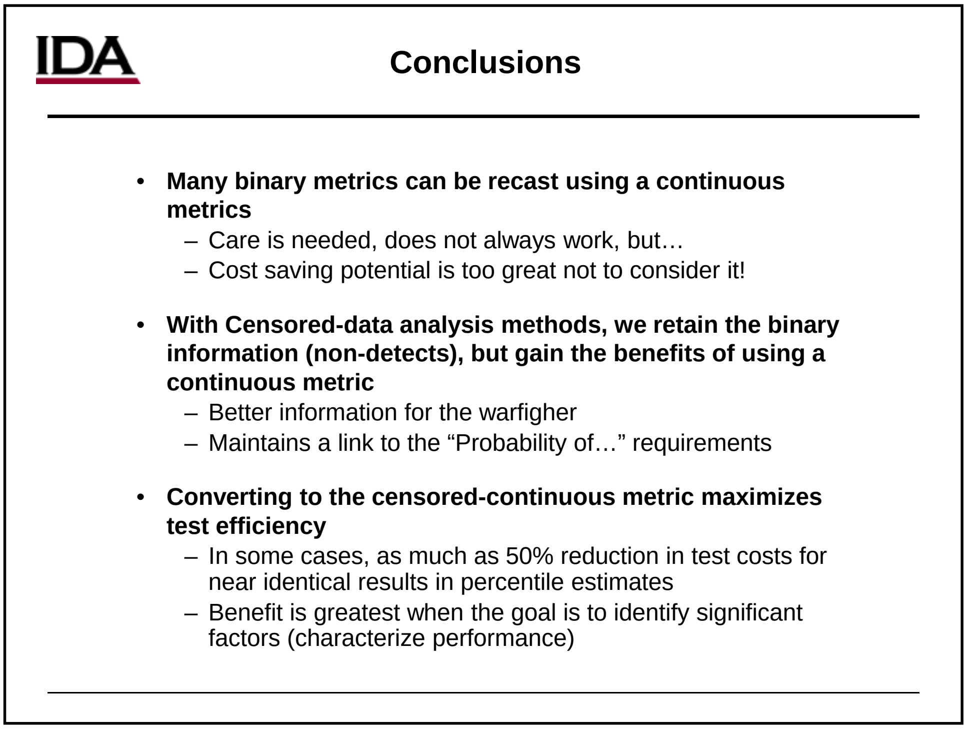
Sizing Tests



Total Sample Size required to detect Factor Effects with 90% confidence, 80% power

ΔP detectable	Binomial metric	Continuous metric w/censoring
40%	44	24
30%	74	38
20%	166	98

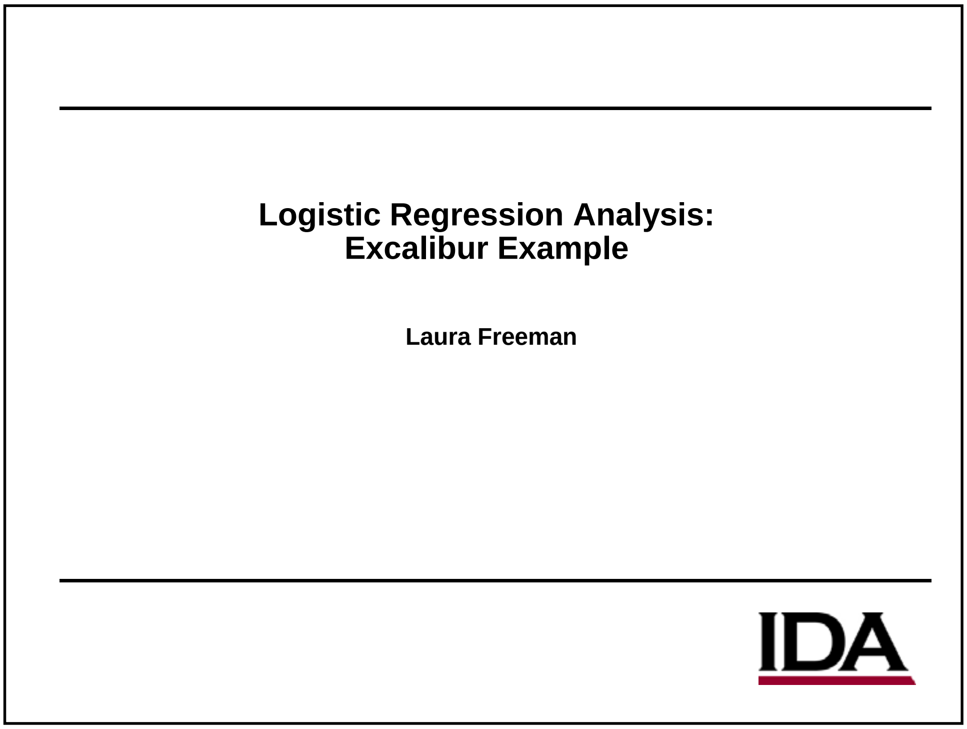
40-50% reduction in test size



Conclusions

- **Many binary metrics can be recast using a continuous metrics**
  - Care is needed, does not always work, but…
  - Cost saving potential is too great not to consider it!
- **With Censored-data analysis methods, we retain the binary information (non-detects), but gain the benefits of using a continuous metric**
  - Better information for the warfigher
  - Maintains a link to the “Probability of…” requirements
- **Converting to the censored-continuous metric maximizes test efficiency**
  - In some cases, as much as 50% reduction in test costs for near identical results in percentile estimates
  - Benefit is greatest when the goal is to identify significant factors (characterize performance)

Appendix 4-7. Excalibur Logistic Regression



Logistic Regression Analysis: Excalibur Example

Laura Freeman

Logistic Regression Example: Excalibur

- **Test objective (retrospective): Characterize Excalibur reliability as a function of potential causes of system failure**
- **Response variable: Hit/Miss**
  - **System requirement**
    - Probability of success = 80%
- **Large reliability dataset**
  - Spans several phases of DT, Integrated Testing (IT), and OT
  - 392 test points
- **Robust dataset**
  - Test conditions recorded
    - Temperature
    - Charge

Logistic Regression

- **Goal: Identify factors, interactions, and higher-order model terms important in explaining changes in probability success**
- **Appropriate analysis for pass/fail (binary) response variables**
- **Requires lots of data**
- **Model "log-odds" as a linear function of factors and their interactions**

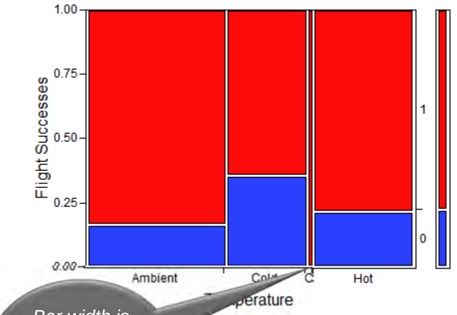
$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

  - If β is statistically different from zero then the model term is important
- **Probability model:** $$p = \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}$$
- **More sophisticated analysis than regression**
  - Statistical analysis packages make it accessible

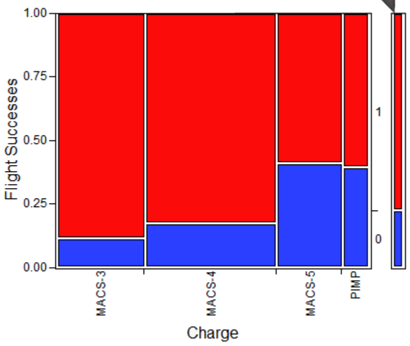
Graphical Data Analysis

- **Mosaic plots summarize the data quickly and easily**
- **Provides intuition on best analysis model**

Summary bar provides total number of success(red) compared to total number of failures (blue)



Bar width is proportional to amount of data in each condition



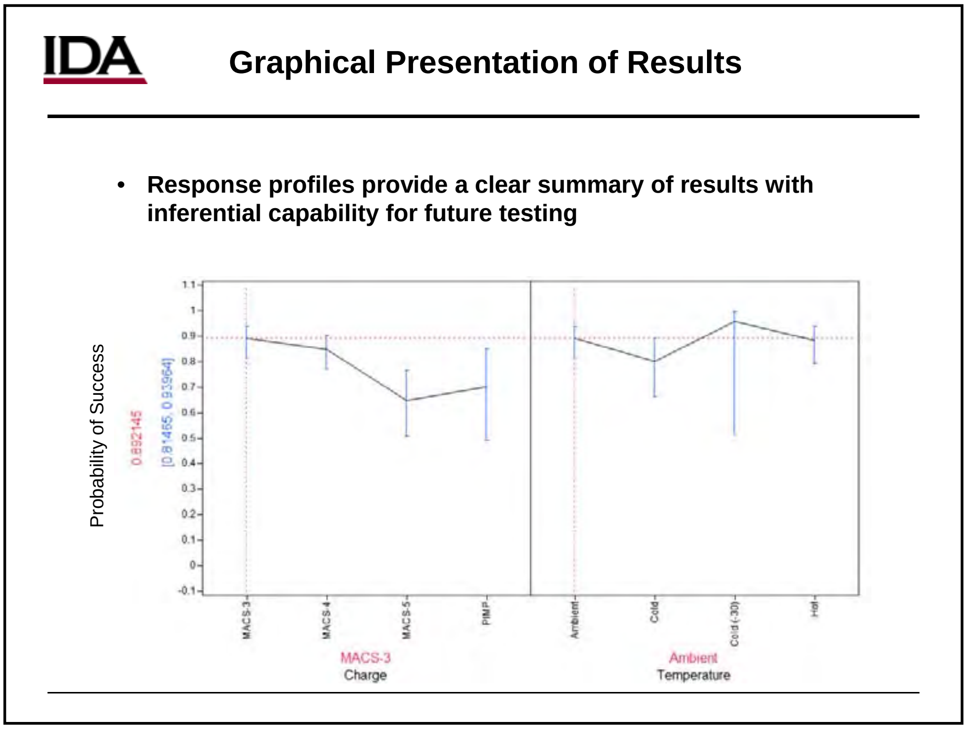
Model Analysis

- **Overall model analysis determines whether the factors significantly affect test outcomes**

Source of Variation	Levels	P-value
Charge	4	<0.0001
Temperature	4	0.0021

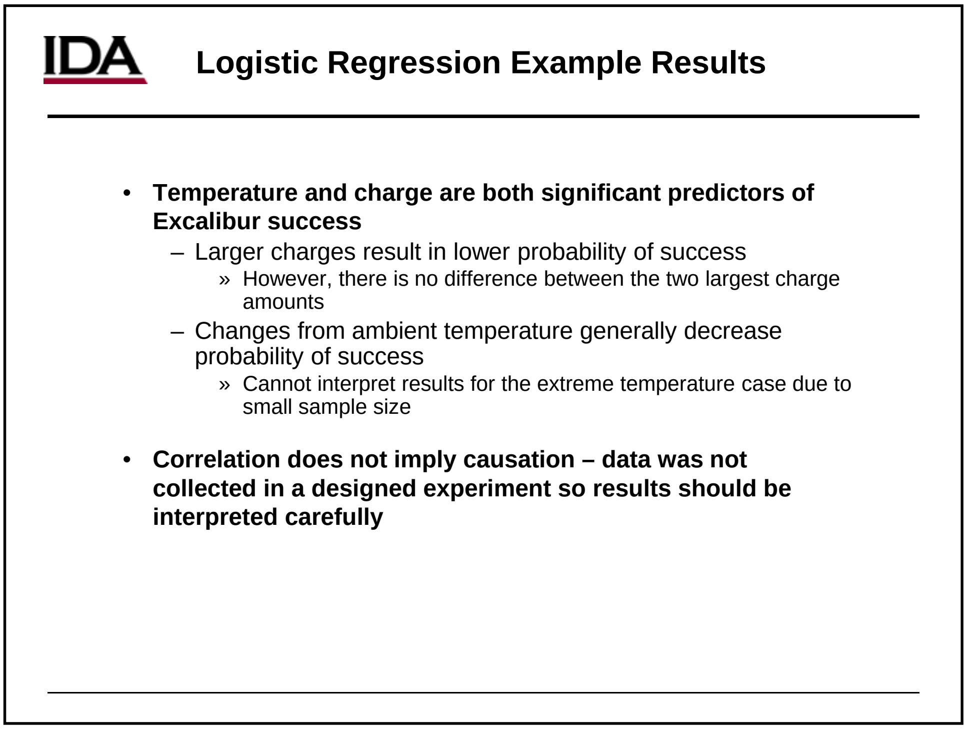
- **Customizable tests of contrast provide information on where differences exist**
  - For example: compare ambient temperature to cold temperature and hot temperature

Contrast	Estimate	Difference	P-value
Ambient versus Cold	83%-64%	19%	.02
Ambient versus Hot	83%-78%	5%	.30
Hot versus cold	78%-64%	14%	.02



Graphical Presentation of Results

- **Response profiles provide a clear summary of results with inferential capability for future testing**



Logistic Regression Example Results

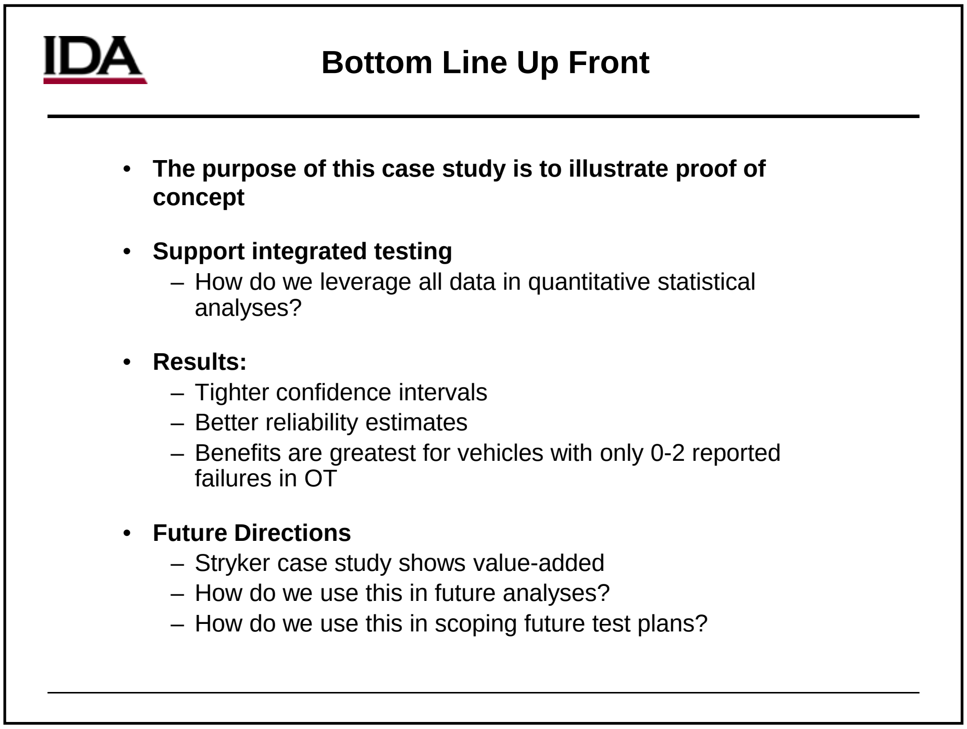
- **Temperature and charge are both significant predictors of Excalibur success**
  - Larger charges result in lower probability of success
    - However, there is no difference between the two largest charge amounts
  - Changes from ambient temperature generally decrease probability of success
    - Cannot interpret results for the extreme temperature case due to small sample size
- **Correlation does not imply causation – data was not collected in a designed experiment so results should be interpreted carefully**

Appendix 4-8.
Stryker Reliability Case Study

Statistical Models for Combining Information: Stryker Reliability Case Study

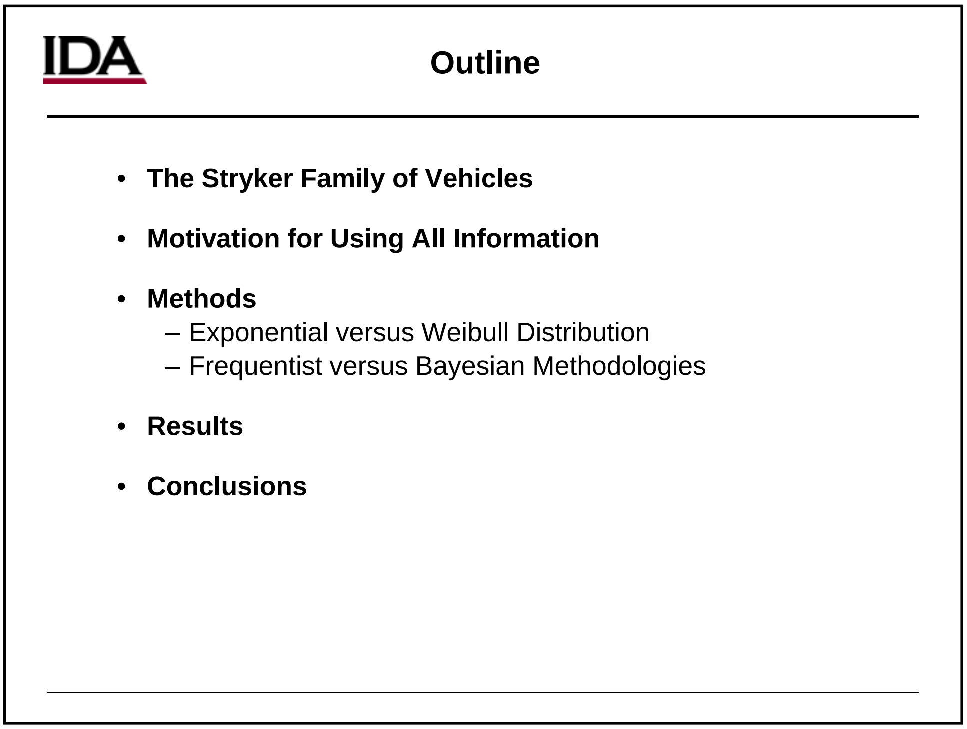
Rebecca Dickinson, Virginia Tech
Laura Freeman, IDA
Alyson Wilson, IDA
Bruce Simpson, IDA

IDA



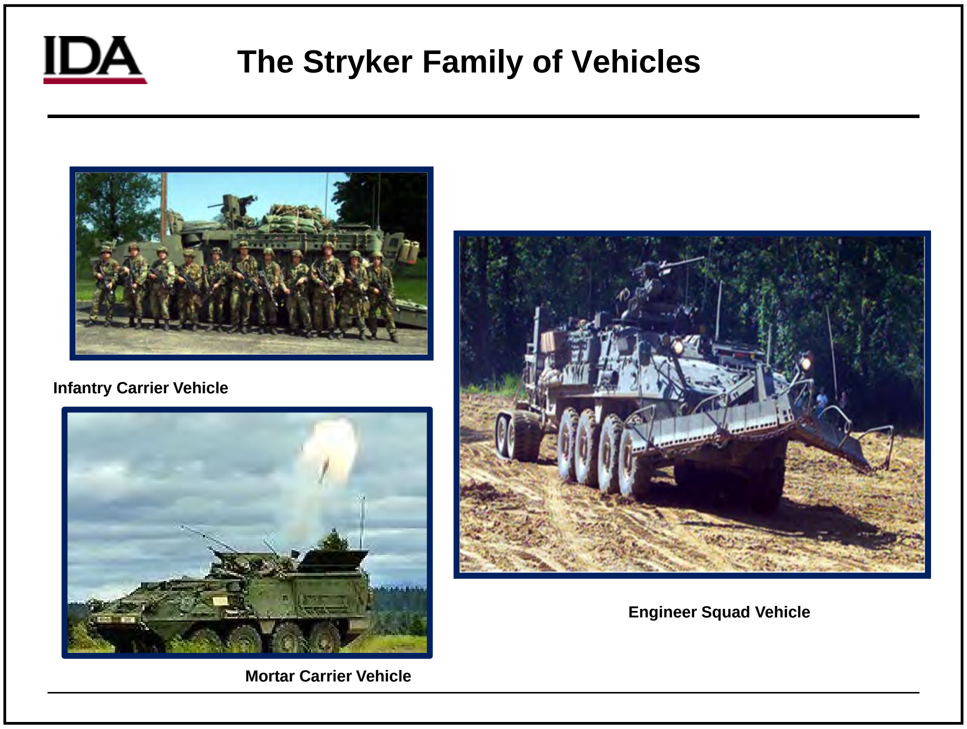
Bottom Line Up Front

- **The purpose of this case study is to illustrate proof of concept**
- **Support integrated testing**
  – How do we leverage all data in quantitative statistical analyses?
- **Results:**
  – Tighter confidence intervals
  – Better reliability estimates
  – Benefits are greatest for vehicles with only 0-2 reported failures in OT
- **Future Directions**
  – Stryker case study shows value-added
  – How do we use this in future analyses?
  – How do we use this in scoping future test plans?



Outline

- **The Stryker Family of Vehicles**
- **Motivation for Using All Information**
- **Methods**
  – Exponential versus Weibull Distribution
  – Frequentist versus Bayesian Methodologies
- **Results**
- **Conclusions**

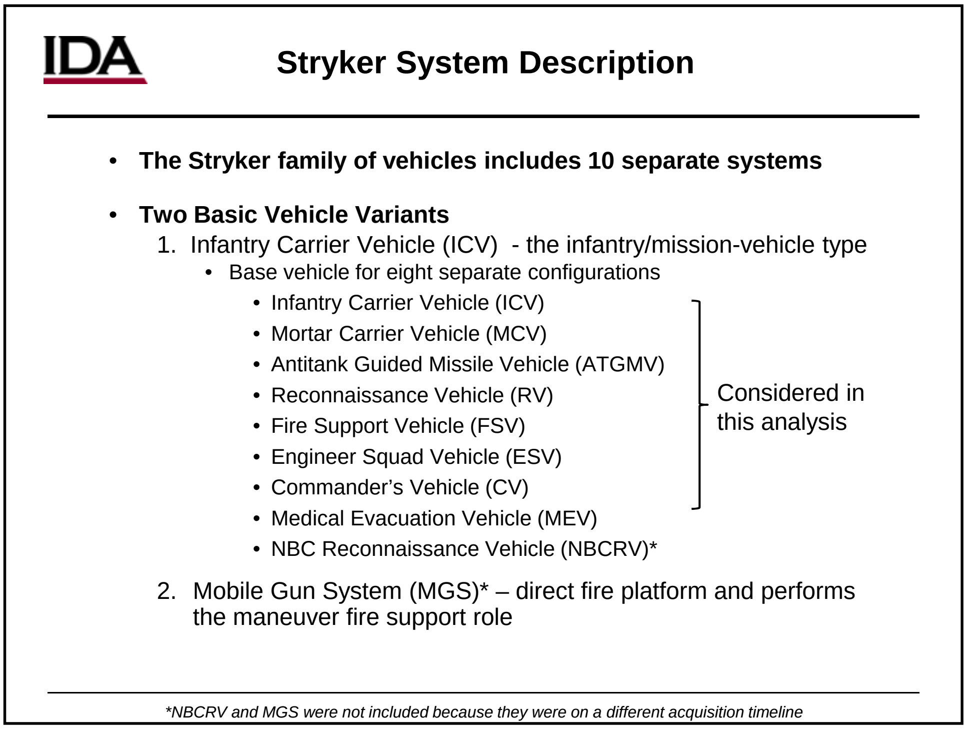


The Stryker Family of Vehicles

Infantry Carrier Vehicle

Engineer Squad Vehicle

Mortar Carrier Vehicle



Stryker System Description

- **The Stryker family of vehicles includes 10 separate systems**
- **Two Basic Vehicle Variants**
  1. Infantry Carrier Vehicle (ICV) - the infantry/mission-vehicle type
     - Base vehicle for eight separate configurations
       - Infantry Carrier Vehicle (ICV)
       - Mortar Carrier Vehicle (MCV)
       - Antitank Guided Missile Vehicle (ATGMV)
       - Reconnaissance Vehicle (RV)
       - Fire Support Vehicle (FSV)
       - Engineer Squad Vehicle (ESV)
       - Commander's Vehicle (CV)
       - Medical Evacuation Vehicle (MEV)
       - NBC Reconnaissance Vehicle (NBCRV)*
  2. Mobile Gun System (MGS)* – direct fire platform and performs the maneuver fire support role

Considered in this analysis

**NBCRV and MGS were not included because they were on a different acquisition timeline*

IDA

Stryker Mission Essential Functions

- **There are four essential functions**
  - Move
  - Shoot
  - Command and Control, Communications, Computers, Intelligence, Surveillance and Reconnaissance (C4ISR)
  - Survive
- **A failure is an event in which an item or part of an item does not perform as specified**
- **The Army failure definition scoring criteria (FDSC) categorizes the severity of failures**
  - **System Abort**
    - » The vehicle is unable to complete the mission
  - Essential Function Failure
  - Non-essential Function Failure
- **Reliability requirement:**
  - Mean miles between system aborts = 1,000 miles

IDA

Developmental and Operational Testing

Developmental Testing

- **Controlled Conditions**
- **Experienced Technicians operating the vehicles.**
  - They have done this for years and they know the courses really well
- **Courses**
  - Use courses that are designed to replicate the primary roads, secondary roads, and trail like conditions

DT And OT Are Different!
•Operators
•Environments
•Test Durations

Operational Testing

- **Operational Conditions**
- **An army unit comes in to do this testing**

- **Courses**
  - OT data set comes from testing that was done at Fort Knox
  - Most of the testing was done using secondary road type conditions
- **Limited amount of Time**
  - Due to operator availability and range availability
  - Operational testing may be too short to discover many reliability deficiencies

Motivation For Using All Information

- **What is the Current Practice?**
  - DOT&E in most cases uses only operational test data for reliability analyses
    - » Stryker Beyond Low Rate Initial Production (BLRIP) Report
    - » Benefit: ensures data is representative of operational test conditions
    - » Drawback: discards information from previous testing that provides information on system reliability
- **Why use all test data?**
  - Testing is expensive
  - Lose valuable information by not using all information
- **National Research Council Studies**
  - *Statistics, Testing and Defense Acquisition*, **1998**
    - » Emphasizes that all relevant information be examined for possible use in both the design and evaluation of operational tests …
    - » State-of-the-art statistical methods for combining information should be used, when appropriate, to make tests and their associated evaluations as cost-efficient as possible
  - *Improved Operational Testing and Evaluation*, **2006**
    - » Focuses specifically on methods of combining information for the Stryker family of vehicles

Reliability Analysis

- **Reliability is an essential component of the assessment of operational suitability**
- **Examples of reliability data:**
  - » Miles driven until failure, hours of use until a failure, number of on-off cycles until a failure
- **Commonly used distributions in reliability:**

Exponential Distribution

- Historically used in DoD reliability assessment
- Simple model: only one parameter to estimate

$$f(t_i) = \frac{1}{\lambda} e^{-\left(\frac{t_i}{\lambda}\right)}$$

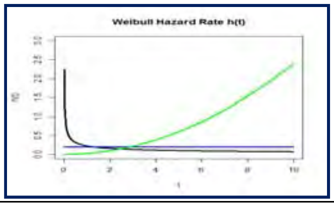
- Easy to interpret: under this parameterization, λ is the mean time between failures

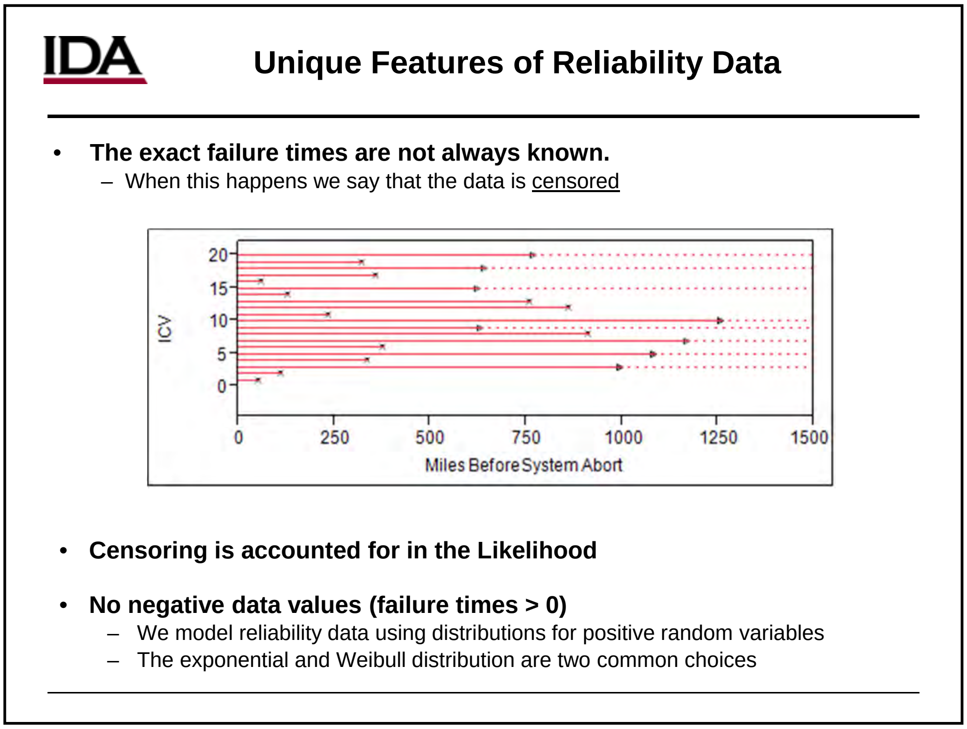
Weibull Distribution

- Flexible distribution: two parameters

$$f(t_i) = \frac{\beta}{\eta}\left(\frac{t_i}{\eta}\right)^{\beta-1} e^{-\left(\frac{t_i}{\eta}\right)^{\beta}}$$

- Can describe multiple failure mechanisms

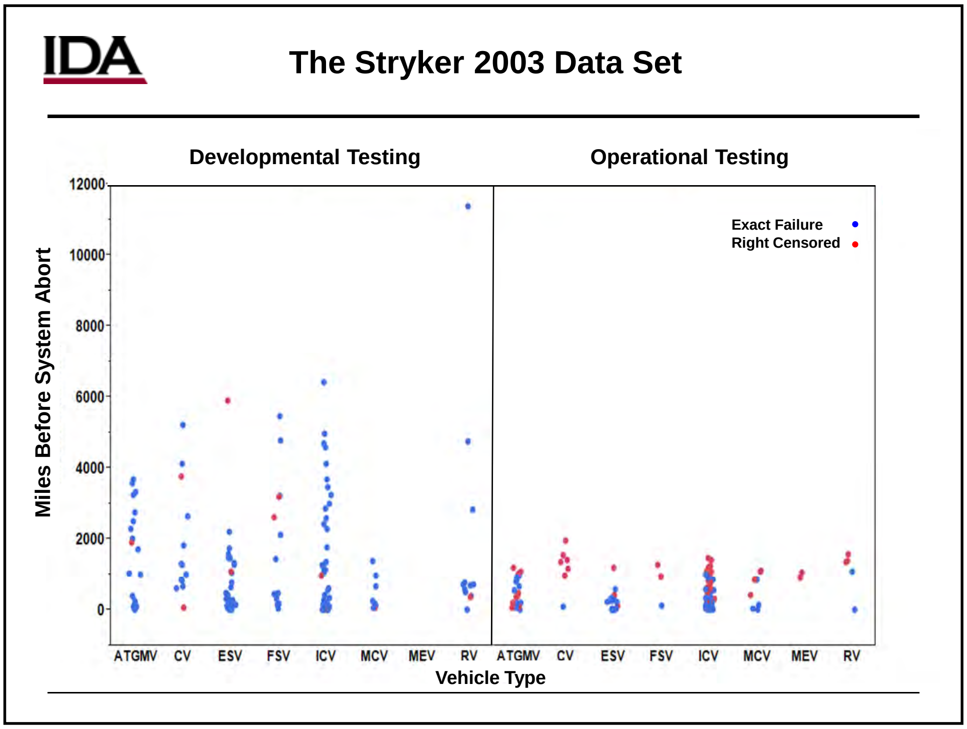




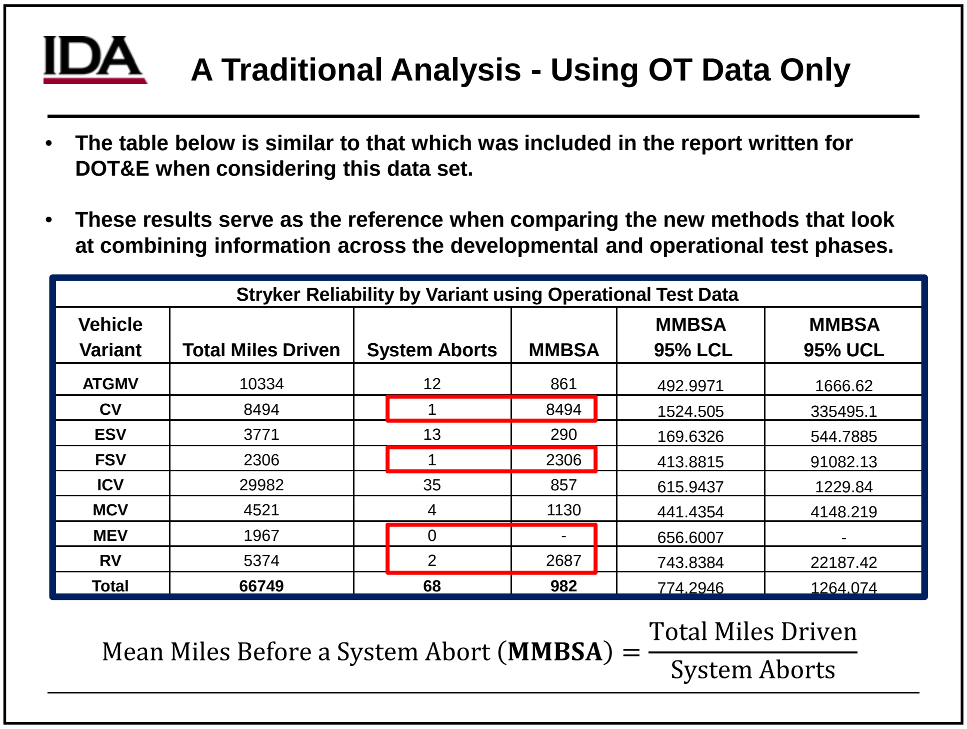
Unique Features of Reliability Data

- **The exact failure times are not always known.**
  - When this happens we say that the data is censored

- **Censoring is accounted for in the Likelihood**
- **No negative data values (failure times > 0)**
  - We model reliability data using distributions for positive random variables
  - The exponential and Weibull distribution are two common choices



The Stryker 2003 Data Set



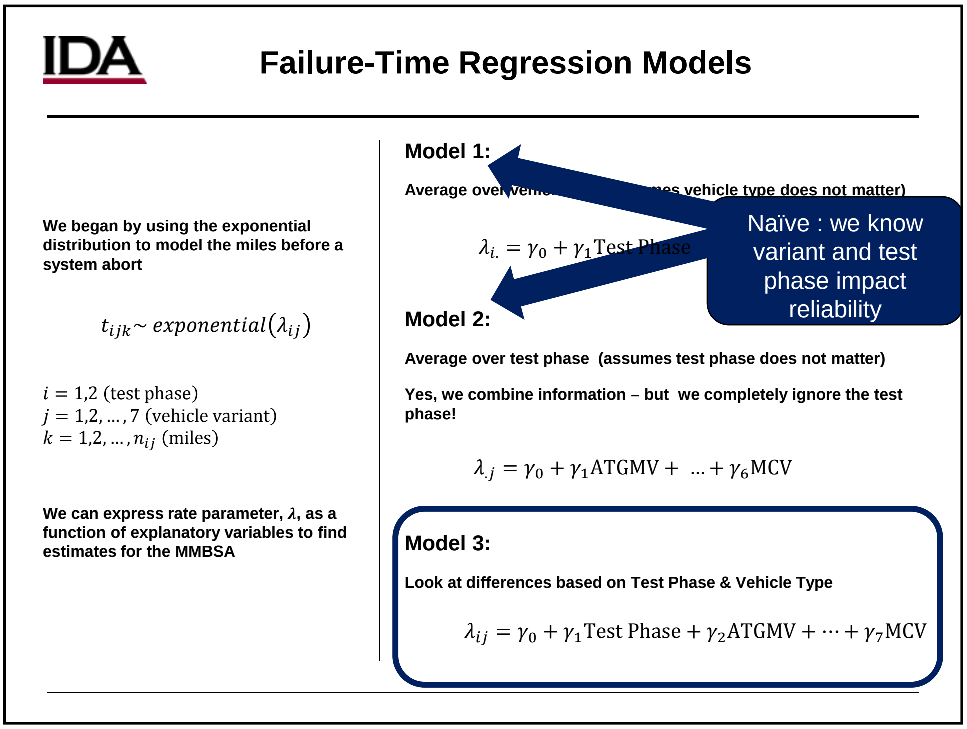
A Traditional Analysis - Using OT Data Only

- The table below is similar to that which was included in the report written for DOT&E when considering this data set.
- These results serve as the reference when comparing the new methods that look at combining information across the developmental and operational test phases.

Stryker Reliability by Variant using Operational Test Data

Vehicle Variant	Total Miles Driven	System Aborts	MMBSA	MMBSA 95% LCL	MMBSA 95% UCL
ATGMV	10334	12	861	492.9971	1666.62
CV	8494	1	8494	1524.505	335495.1
ESV	3771	13	290	169.6326	544.7885
FSV	2306	1	2306	413.8815	91082.13
ICV	29982	35	857	615.9437	1229.84
MCV	4521	4	1130	441.4354	4148.219
MEV	1967	0	-	656.6007	-
RV	5374	2	2687	743.8384	22187.42
Total	**66749**	**68**	**982**	774.2946	1264.074

$$\text{Mean Miles Before a System Abort } (\textbf{MMBSA}) = \frac{\text{Total Miles Driven}}{\text{System Aborts}}$$



Failure-Time Regression Models

We began by using the exponential distribution to model the miles before a system abort

$$t_{ijk} \sim exponential(\lambda_{ij})$$

$i = 1,2$ (test phase)
$j = 1,2, \dots, 7$ (vehicle variant)
$k = 1,2, \dots, n_{ij}$ (miles)

We can express rate parameter, λ, as a function of explanatory variables to find estimates for the MMBSA

Model 1:

Average over vehic [illegible] mes vehicle type does not matter)

$$\lambda_{i.} = \gamma_0 + \gamma_1 \text{Test Phase}$$

Naïve : we know variant and test phase impact reliability

Model 2:

Average over test phase (assumes test phase does not matter)

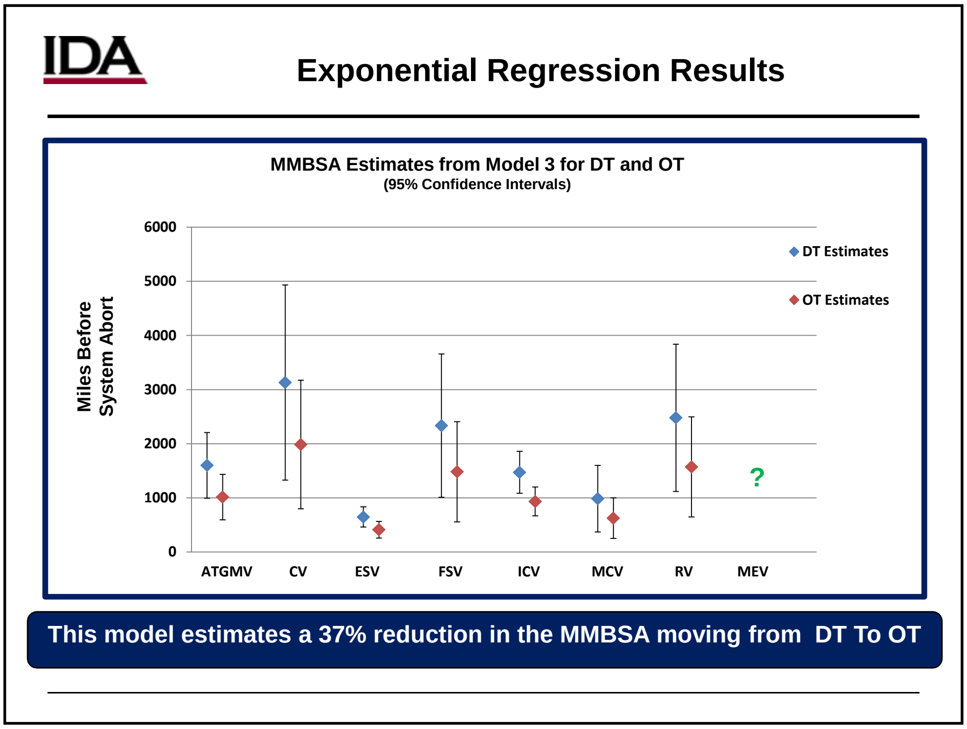
Yes, we combine information – but we completely ignore the test phase!

$$\lambda_{.j} = \gamma_0 + \gamma_1 \text{ATGMV} + \dots + \gamma_6 \text{MCV}$$

Model 3:

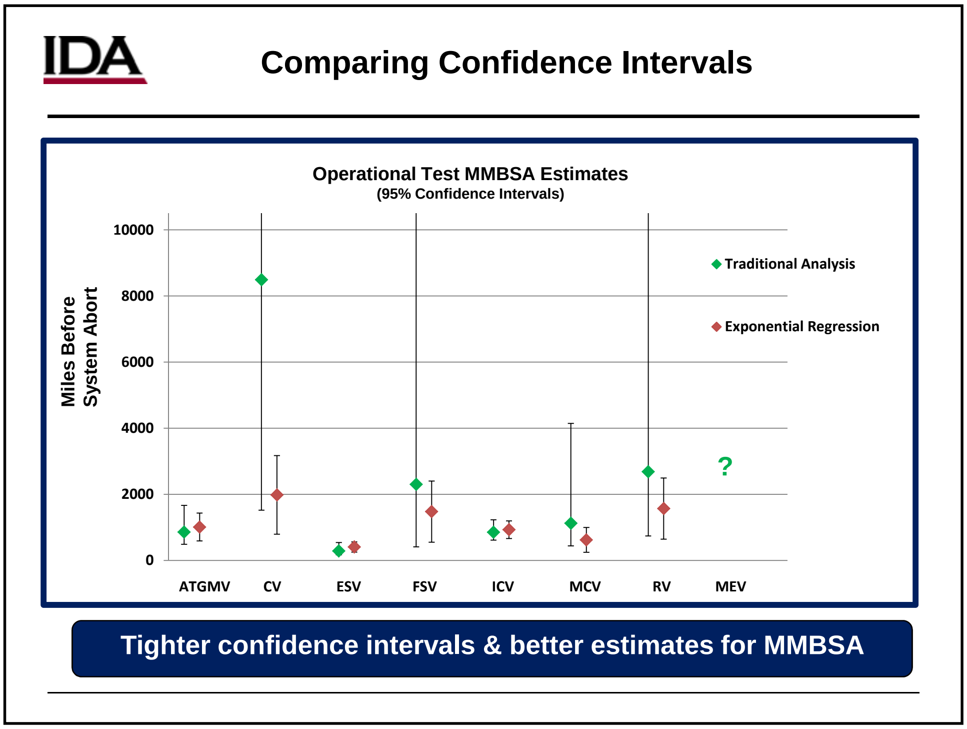
Look at differences based on Test Phase & Vehicle Type

$$\lambda_{ij} = \gamma_0 + \gamma_1 \text{Test Phase} + \gamma_2 \text{ATGMV} + \cdots + \gamma_7 \text{MCV}$$



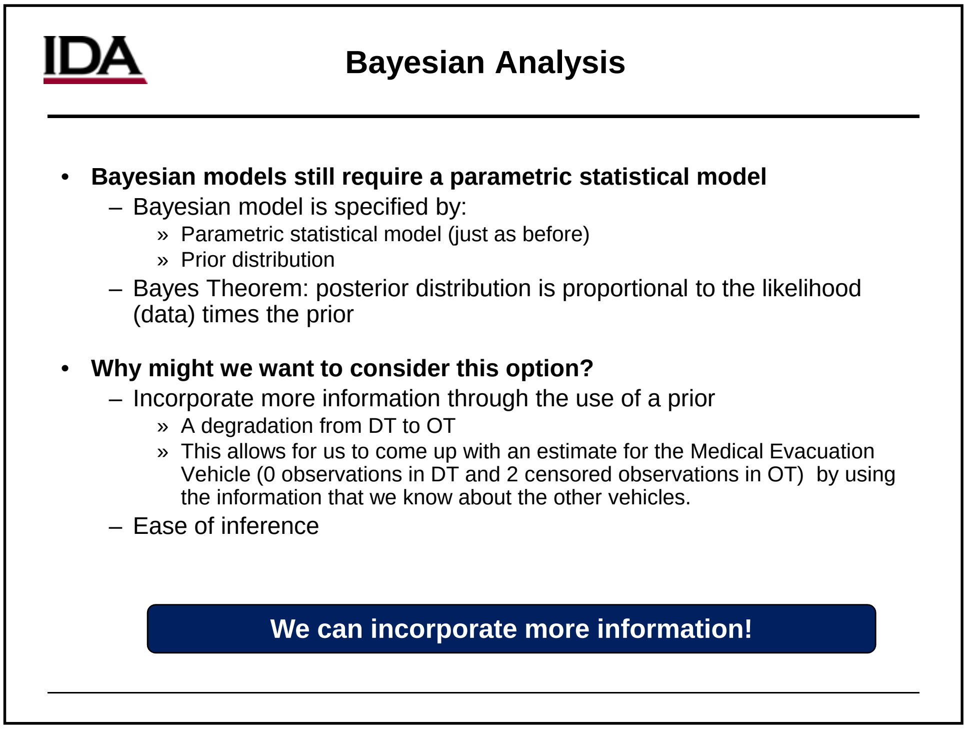
Exponential Regression Results

This model estimates a 37% reduction in the MMBSA moving from DT To OT



Comparing Confidence Intervals

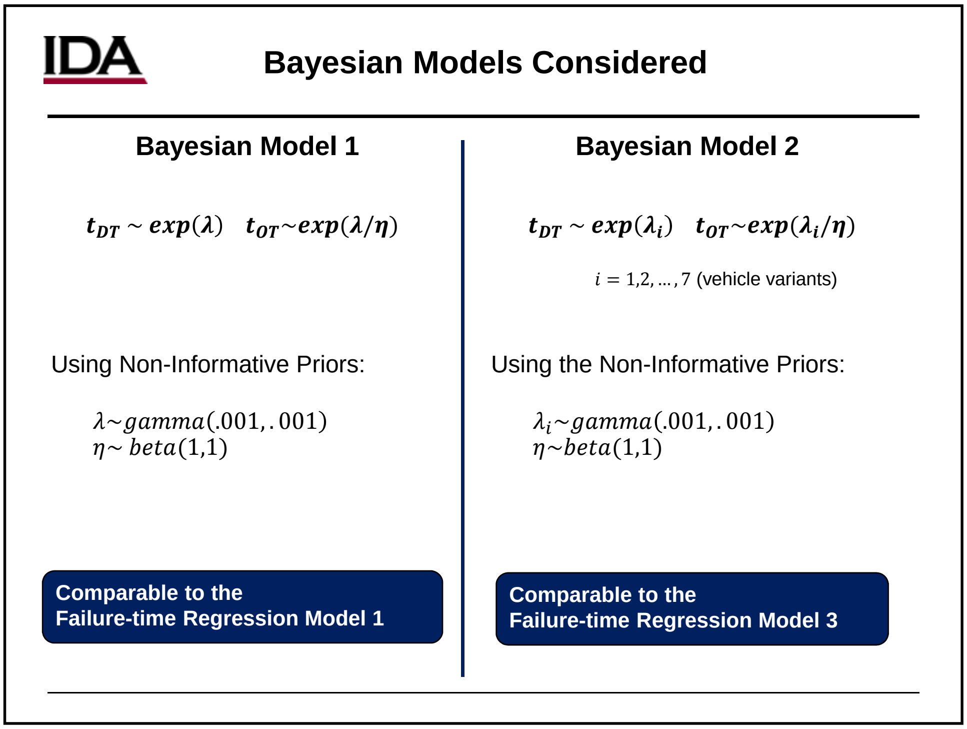
Tighter confidence intervals & better estimates for MMBSA



Bayesian Analysis

- **Bayesian models still require a parametric statistical model**
  - Bayesian model is specified by:
    - Parametric statistical model (just as before)
    - Prior distribution
  - Bayes Theorem: posterior distribution is proportional to the likelihood (data) times the prior
- **Why might we want to consider this option?**
  - Incorporate more information through the use of a prior
    - A degradation from DT to OT
    - This allows for us to come up with an estimate for the Medical Evacuation Vehicle (0 observations in DT and 2 censored observations in OT) by using the information that we know about the other vehicles.
  - Ease of inference

We can incorporate more information!



Bayesian Models Considered

Bayesian Model 1

$\boldsymbol{t_{DT} \sim exp(\lambda)} \quad \boldsymbol{t_{OT} \sim exp(\lambda/\eta)}$

Using Non-Informative Priors:

$\lambda \sim gamma(.001, .001)$
$\eta \sim beta(1,1)$

Comparable to the Failure-time Regression Model 1

Bayesian Model 2

$\boldsymbol{t_{DT} \sim exp(\lambda_i)} \quad \boldsymbol{t_{OT} \sim exp(\lambda_i/\eta)}$

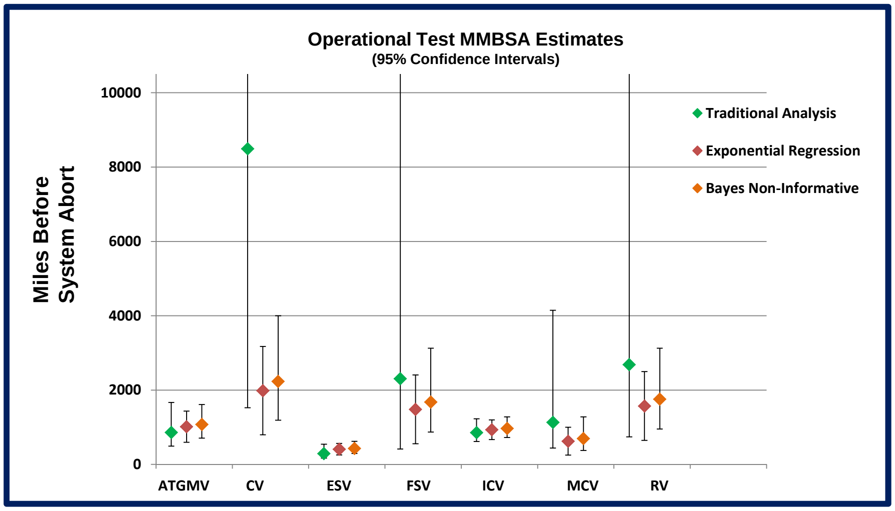
$i = 1,2, \ldots, 7$ (vehicle variants)

Using the Non-Informative Priors:

$\lambda_i \sim gamma(.001, .001)$
$\eta \sim beta(1,1)$

Comparable to the Failure-time Regression Model 3

Comparing Intervals



Point and interval estimates for MMBSA are nearly identical

Incorporating More Information

- **Informative Priors**
  - Based on subject matter expertise
    - » Data is already included in model
- **Hierarchical Models**
  - Assumes the parameters are related, the data tells us how closely related
  - Hierarchical models for the Stryker case study allow us to estimate MEV reliability based on other data

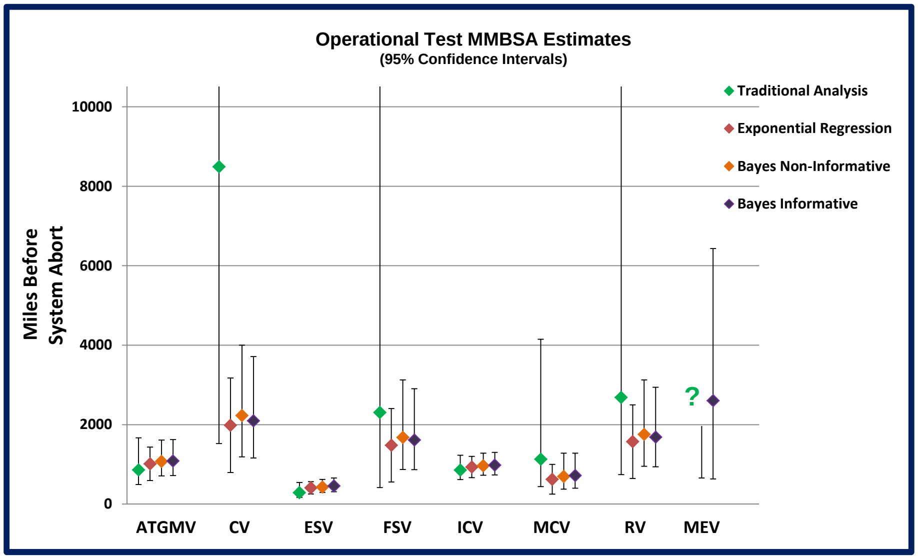
A Model That Allows Us To Estimate MEV Reliability

$$t_{DT} \sim exp(\lambda_i) \quad t_{OT} \sim exp(\lambda_i/\eta)$$

$i = 1,2,\ldots,8$ (vehicle variants including MEV)

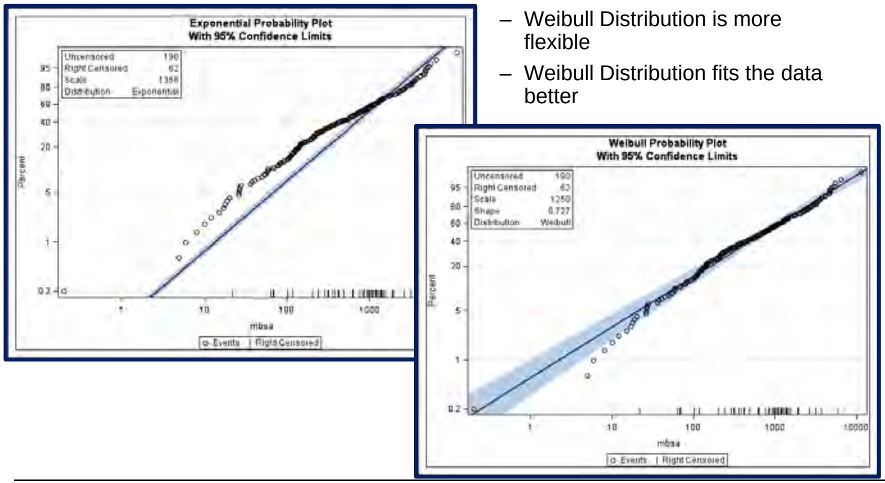
$$\lambda_i \sim gamma(a,b)$$
$$\eta \sim beta(1,1)$$
$$a \sim gamma(.001,.001)$$
$$b \sim gamma(.001,.001)$$

Incorporating More Information

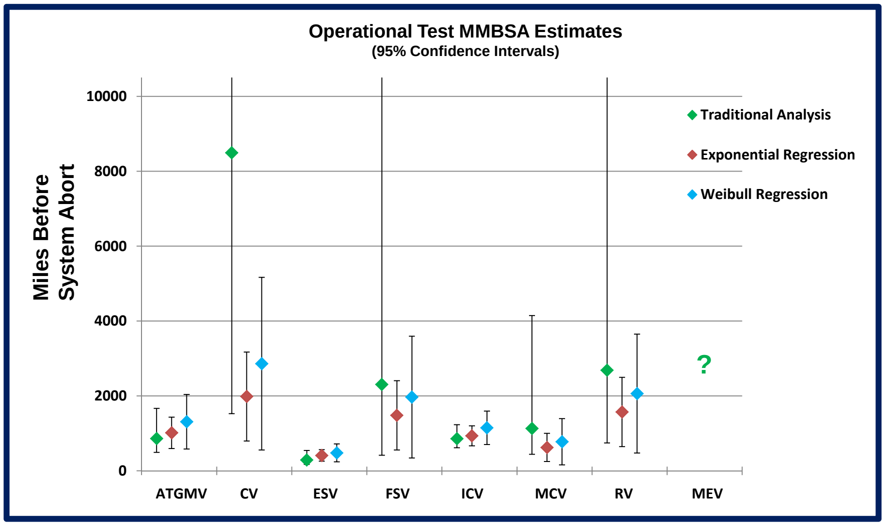


Reality Check

- **Is the exponential distribution appropriate?**
  - Weibull Distribution is more flexible
  - Weibull Distribution fits the data better

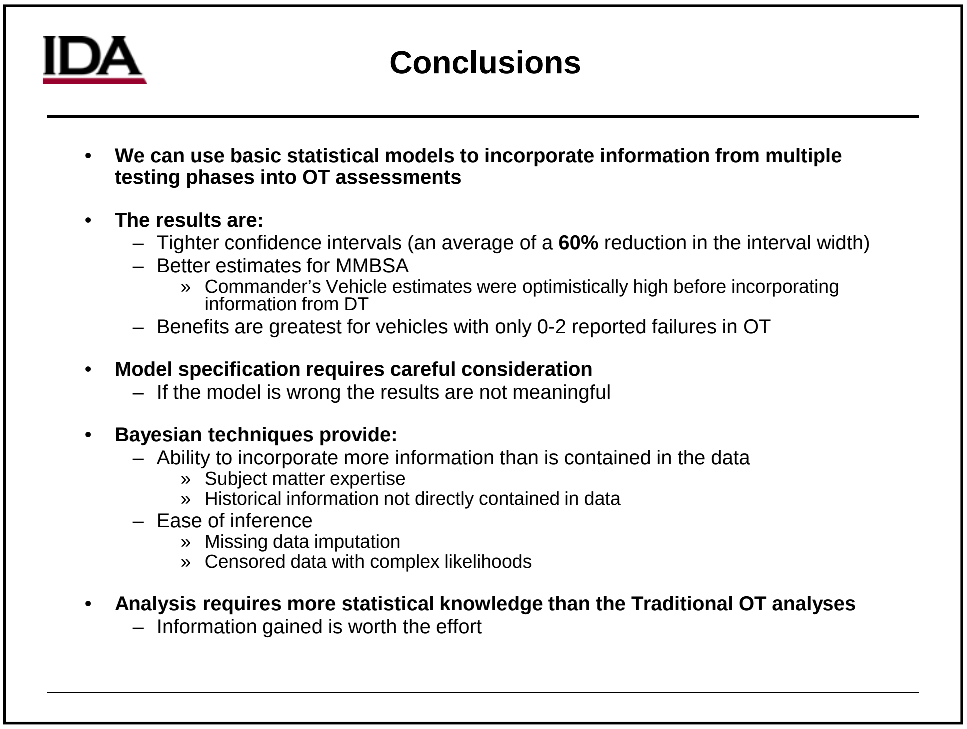


Comparing Exponential and Weibull Results



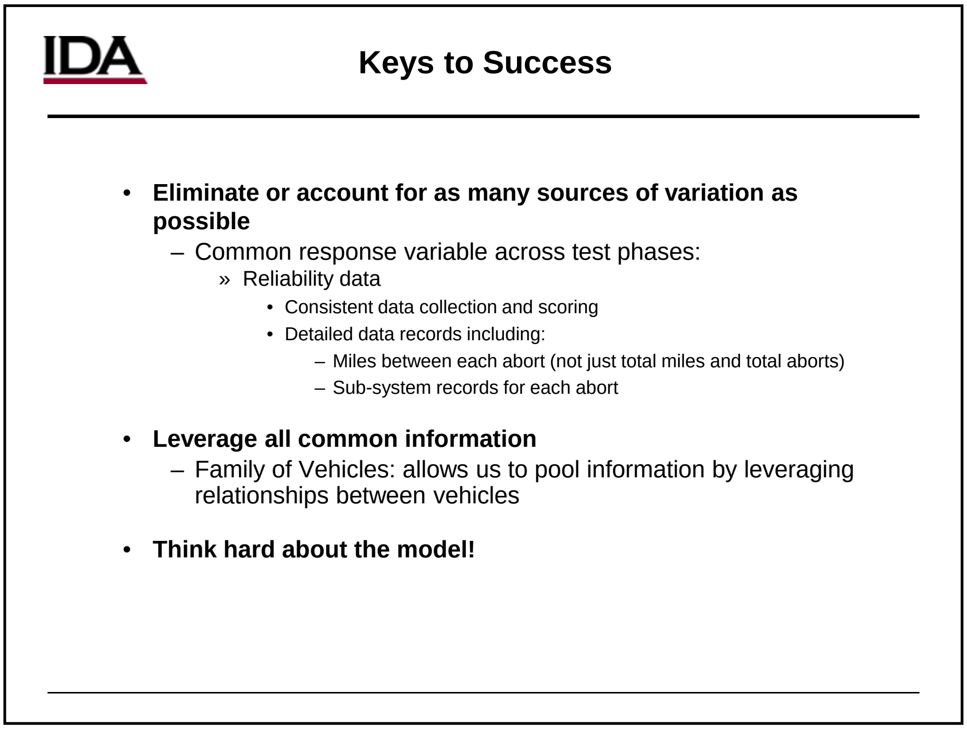
Statistical Challenges

- **Likelihood based inferences**
  - Cannot always be done in standard statistical software
  - Multivariate delta method
- **Censored data**
- **Might need to write your own code.**
  - Software packages don’t always provide enough flexibility
- **No data set is ever perfect**
  - Missing data
    - Multiple imputation
    - Bayesian imputation



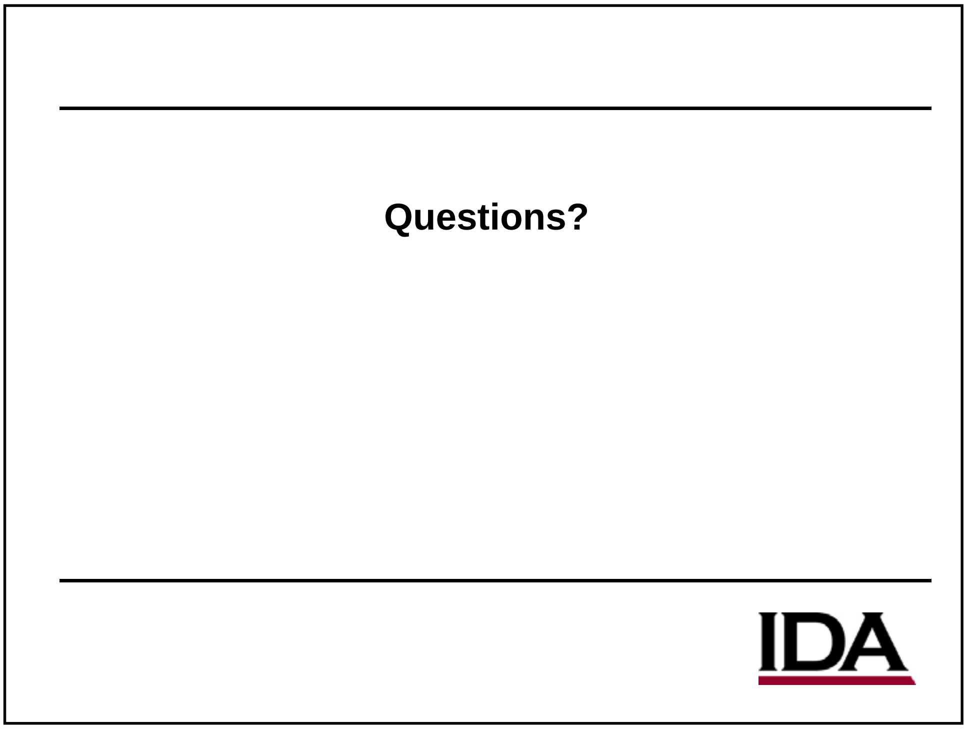
Conclusions

- **We can use basic statistical models to incorporate information from multiple testing phases into OT assessments**
- **The results are:**
  - Tighter confidence intervals (an average of a **60%** reduction in the interval width)
  - Better estimates for MMBSA
    - » Commander's Vehicle estimates were optimistically high before incorporating information from DT
  - Benefits are greatest for vehicles with only 0-2 reported failures in OT
- **Model specification requires careful consideration**
  - If the model is wrong the results are not meaningful
- **Bayesian techniques provide:**
  - Ability to incorporate more information than is contained in the data
    - » Subject matter expertise
    - » Historical information not directly contained in data
  - Ease of inference
    - » Missing data imputation
    - » Censored data with complex likelihoods
- **Analysis requires more statistical knowledge than the Traditional OT analyses**
  - Information gained is worth the effort

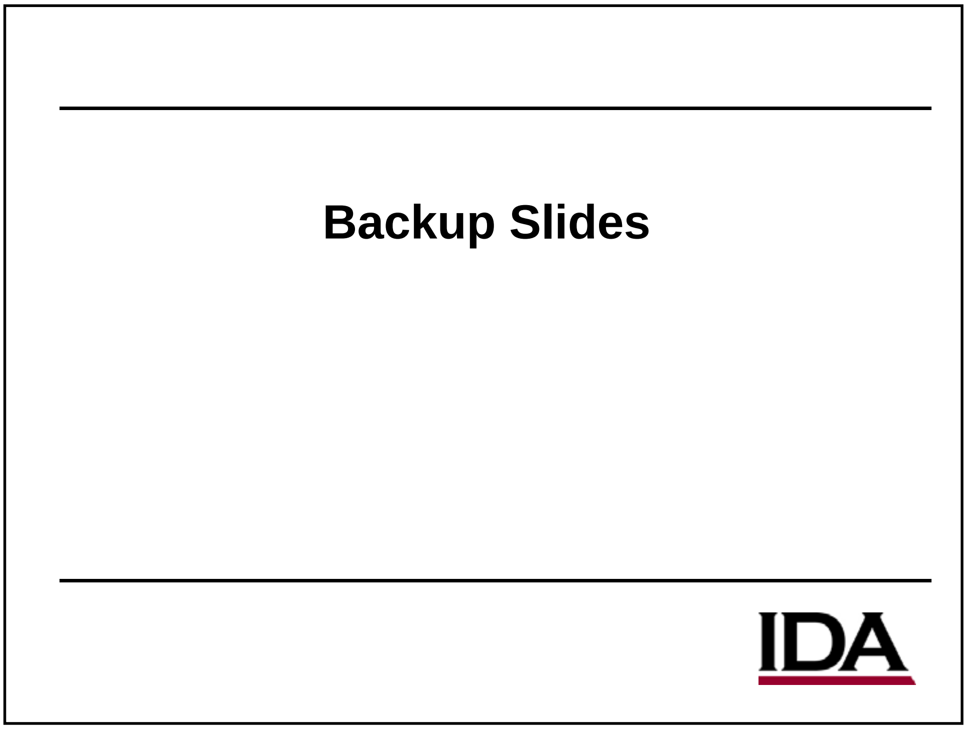


Keys to Success

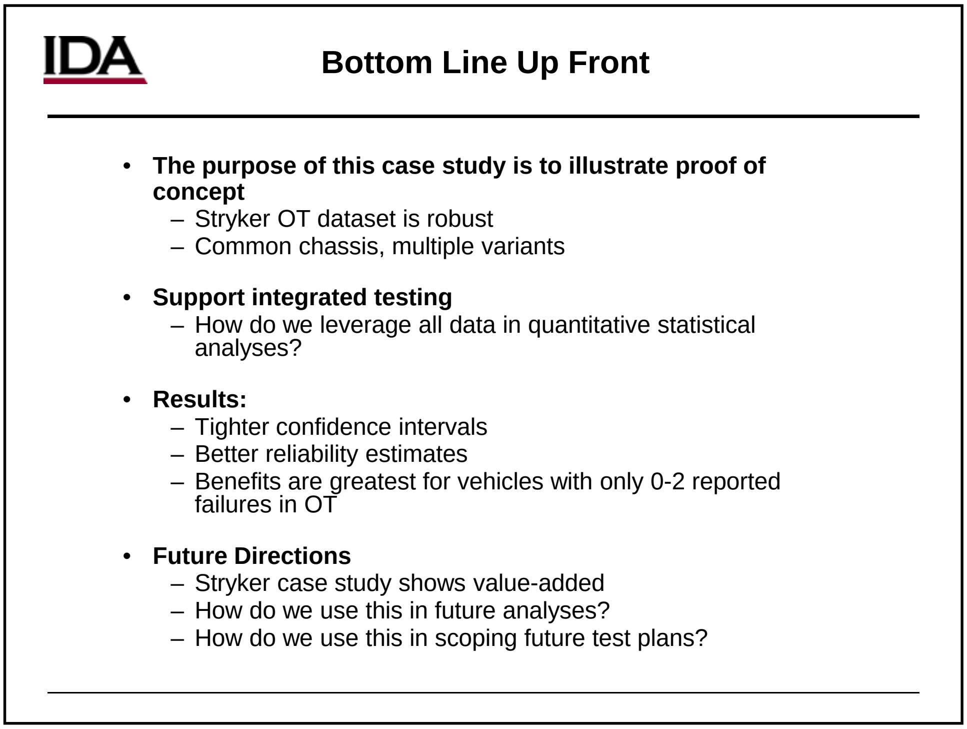
- **Eliminate or account for as many sources of variation as possible**
  - Common response variable across test phases:
    - » Reliability data
      - Consistent data collection and scoring
      - Detailed data records including:
        - Miles between each abort (not just total miles and total aborts)
        - Sub-system records for each abort
- **Leverage all common information**
  - Family of Vehicles: allows us to pool information by leveraging relationships between vehicles
- **Think hard about the model!**



Questions?

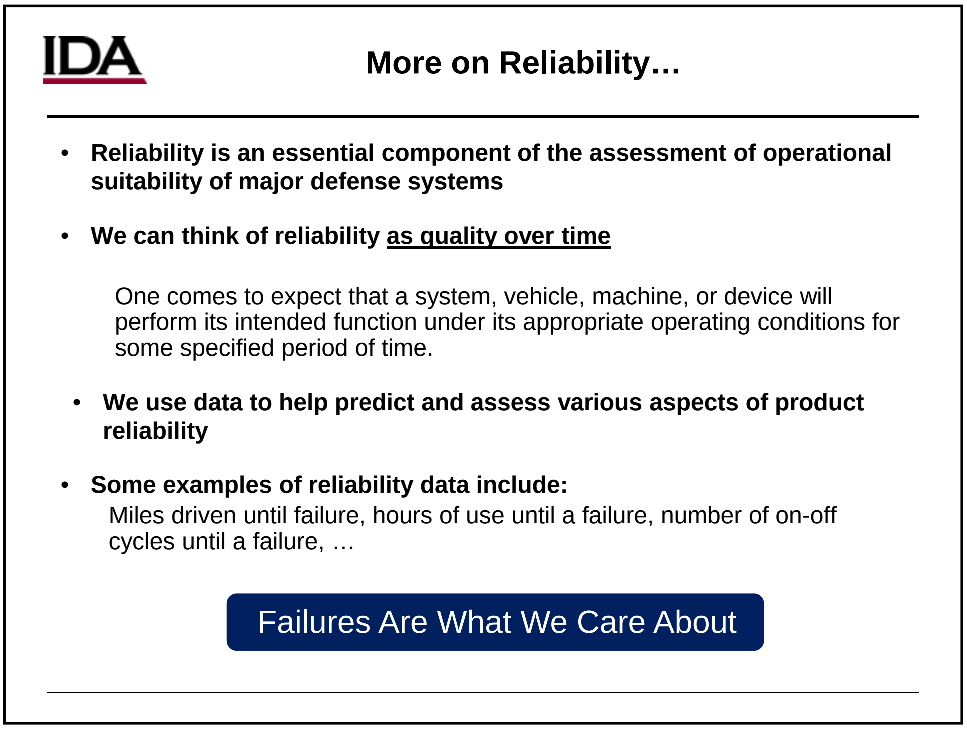


Backup Slides



Bottom Line Up Front

- **The purpose of this case study is to illustrate proof of concept**
  - Stryker OT dataset is robust
  - Common chassis, multiple variants
- **Support integrated testing**
  - How do we leverage all data in quantitative statistical analyses?
- **Results:**
  - Tighter confidence intervals
  - Better reliability estimates
  - Benefits are greatest for vehicles with only 0-2 reported failures in OT
- **Future Directions**
  - Stryker case study shows value-added
  - How do we use this in future analyses?
  - How do we use this in scoping future test plans?



More on Reliability…

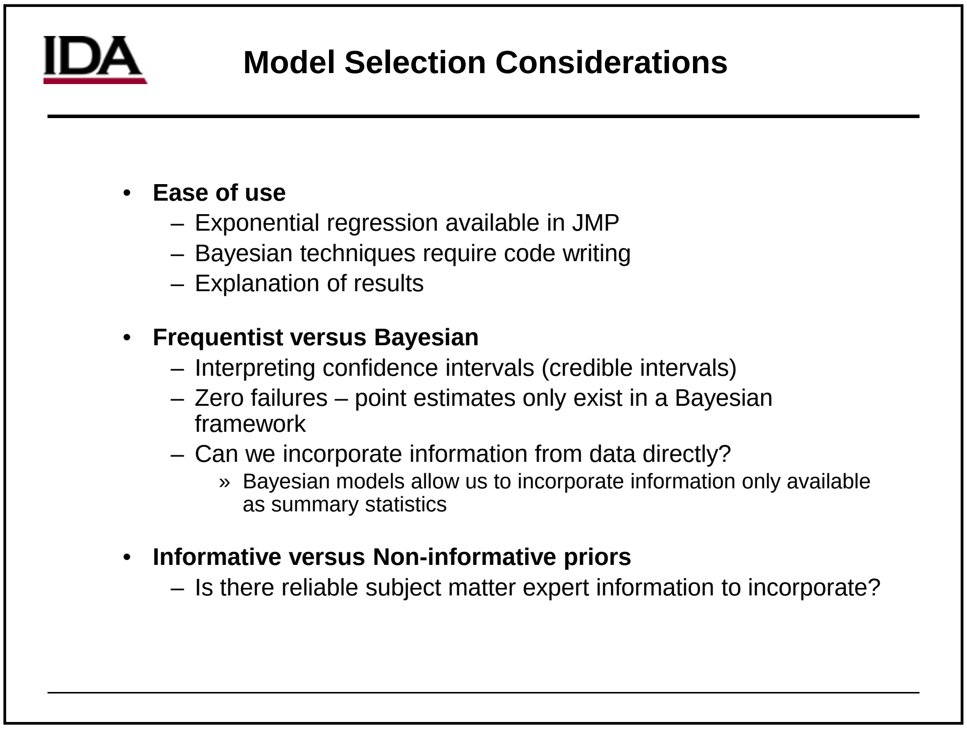
- **Reliability is an essential component of the assessment of operational suitability of major defense systems**
- **We can think of reliability <u>as quality over time</u>**

  One comes to expect that a system, vehicle, machine, or device will perform its intended function under its appropriate operating conditions for some specified period of time.

- **We use data to help predict and assess various aspects of product reliability**
- **Some examples of reliability data include:**

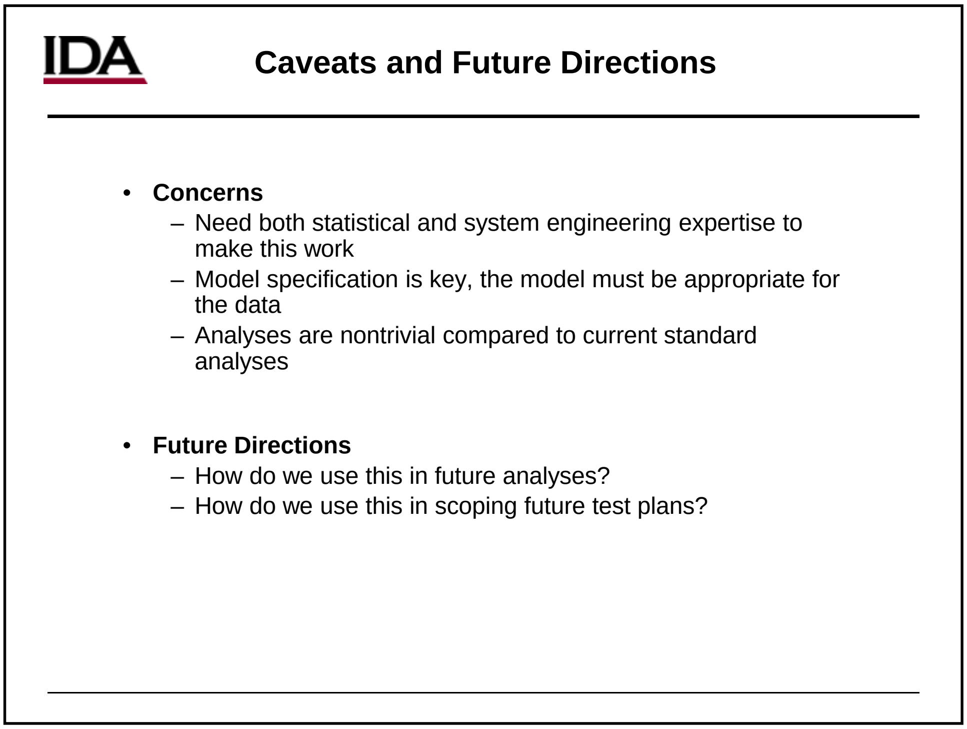
  Miles driven until failure, hours of use until a failure, number of on-off cycles until a failure, …

Failures Are What We Care About



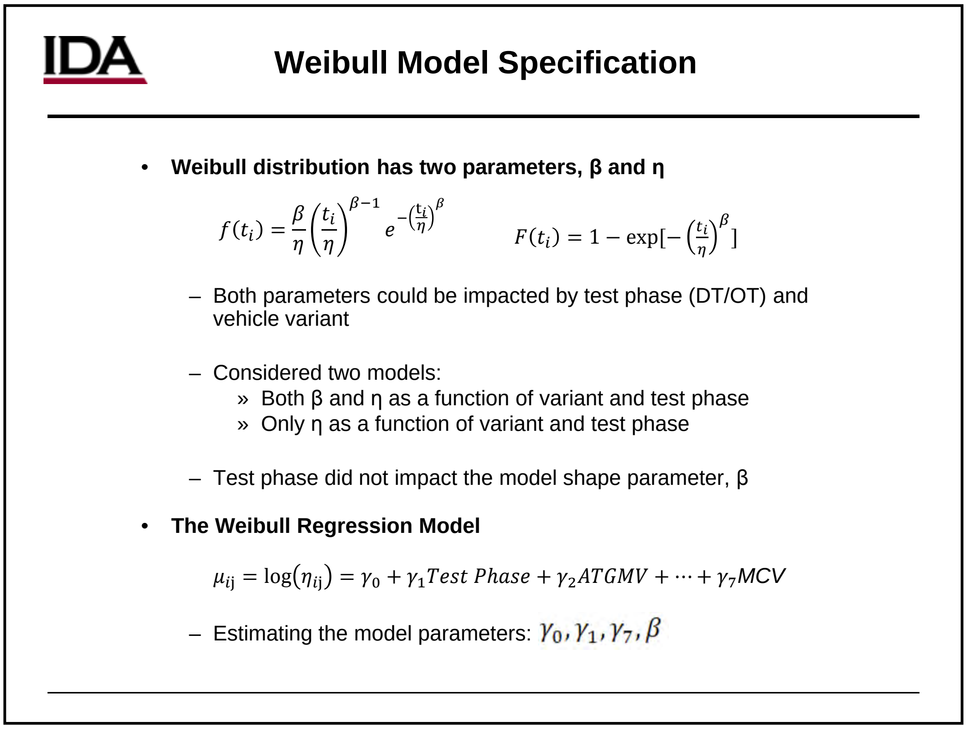
Model Selection Considerations

- **Ease of use**
  - Exponential regression available in JMP
  - Bayesian techniques require code writing
  - Explanation of results
- **Frequentist versus Bayesian**
  - Interpreting confidence intervals (credible intervals)
  - Zero failures – point estimates only exist in a Bayesian framework
  - Can we incorporate information from data directly?
    - Bayesian models allow us to incorporate information only available as summary statistics
- **Informative versus Non-informative priors**
  - Is there reliable subject matter expert information to incorporate?



Caveats and Future Directions

- **Concerns**
  - Need both statistical and system engineering expertise to make this work
  - Model specification is key, the model must be appropriate for the data
  - Analyses are nontrivial compared to current standard analyses
- **Future Directions**
  - How do we use this in future analyses?
  - How do we use this in scoping future test plans?



Weibull Model Specification

- **Weibull distribution has two parameters, β and η**

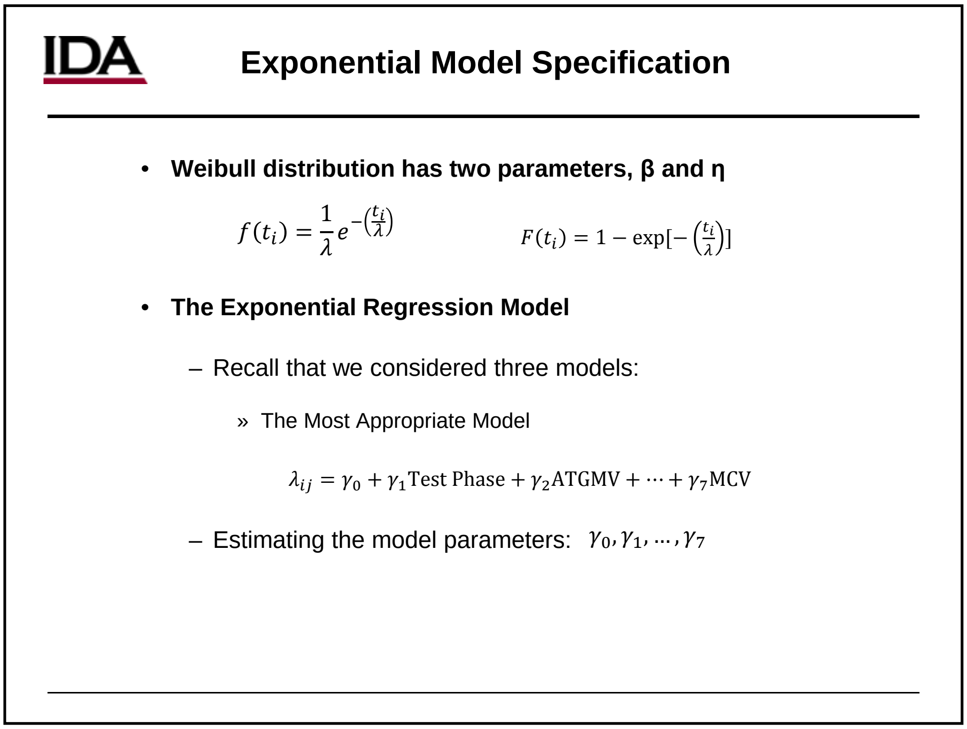
$$f(t_i) = \frac{\beta}{\eta}\left(\frac{t_i}{\eta}\right)^{\beta-1} e^{-\left(\frac{t_i}{\eta}\right)^{\beta}} \qquad F(t_i) = 1 - \exp[-\left(\frac{t_i}{\eta}\right)^{\beta}]$$

  - Both parameters could be impacted by test phase (DT/OT) and vehicle variant
  - Considered two models:
    - » Both β and η as a function of variant and test phase
    - » Only η as a function of variant and test phase
  - Test phase did not impact the model shape parameter, β

- **The Weibull Regression Model**

$$\mu_{ij} = \log(\eta_{ij}) = \gamma_0 + \gamma_1 Test\ Phase + \gamma_2 ATGMV + \cdots + \gamma_7 MCV$$

  - Estimating the model parameters: $\gamma_0, \gamma_1, \gamma_7, \beta$



Exponential Model Specification

- **Weibull distribution has two parameters, β and η**

$$f(t_i) = \frac{1}{\lambda} e^{-\left(\frac{t_i}{\lambda}\right)} \qquad F(t_i) = 1 - \exp[-\left(\frac{t_i}{\lambda}\right)]$$

- **The Exponential Regression Model**
  - Recall that we considered three models:
    - » The Most Appropriate Model

$$\lambda_{ij} = \gamma_0 + \gamma_1 \text{Test Phase} + \gamma_2 \text{ATGMV} + \cdots + \gamma_7 \text{MCV}$$

  - Estimating the model parameters: $\gamma_0, \gamma_1, \cdots, \gamma_7$

Failure-Time Regression Models: Censoring and MLE

- **We need to estimate the regression model parameters!**
  - We do this using Maximum Likelihood Estimation (MLE)
    - » The estimates for the model parameters are the values that maximize the likelihood function

- **Total Likelihood for right censored data**
  - Product of the likelihood contributions:

$$L(\boldsymbol{\Theta}|\mathrm{t}_1,\ldots,\mathrm{t}_\mathrm{n}) = C\prod_{i=1}^{n}\ \underbrace{[f(t_i)]^{\delta_i}}_{\uparrow\ \text{Exact Failure}}\ \underbrace{[1-F(t_i)]^{1-\delta_i}}_{\nwarrow\ \text{Right Censored Contribution}},$$

  - Where:

$$\delta_i = \begin{cases} 1 \text{ exact failure} \\ 0 \text{ right censored} \end{cases}$$

$\boldsymbol{\Theta}$ is a vector of model parameters

$f(t_i)$ is the pdf for the distribution under consideration

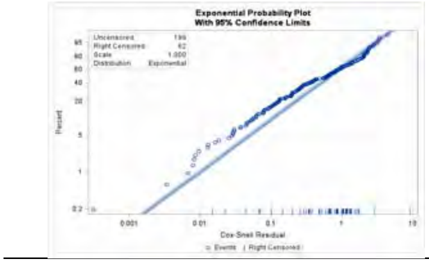
$F(t_i)$ is the cdf for the distribution under consideration

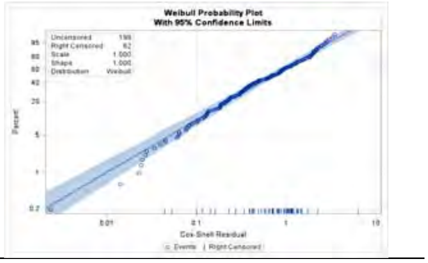
Assessing The Model Adequacy Of Failure Time Regression Models

- **Model Comparisons**
  - Weibull is the best distribution to use based on the model comparison AIC and BIC values.

- **A Whole Model Test**
  - Exponential Regression: p < .0001
  - Weibull Regression: p< .0001

- **Probability Plots of Residuals for Exponential and Weibull Regression**





MCMC Routine

- **The Steps below are outlined under the assumption that the data follows a Weibull distribution (easy to modify for exponential distribution)**

- **Calculate the Log-Posterior:**

$$= logL(\gamma_0, \dots \gamma_7, \gamma_8, \beta | \mathrm{t}_1, \dots, \mathrm{t}_\mathrm{n}) + \sum_{i=1}^{8} \pi(\gamma_i) + \pi(\beta) + \pi(\eta)$$

- **Algorithm**

Step 0:
Initialize starting value for $\gamma_1, \gamma_2, \gamma_3, \dots \gamma_7, \beta, \eta, t_{missing}$

Step 1:
Propose γ_1 -> accept or reject using log-posterior (using current values of other parameters).
Propose γ_2 -> accept or reject using log-posterior (updated γ_1 value and current values of other parameters).
“ “ … for other parameters ($\gamma_3, \dots \gamma_7, \beta, \eta$)

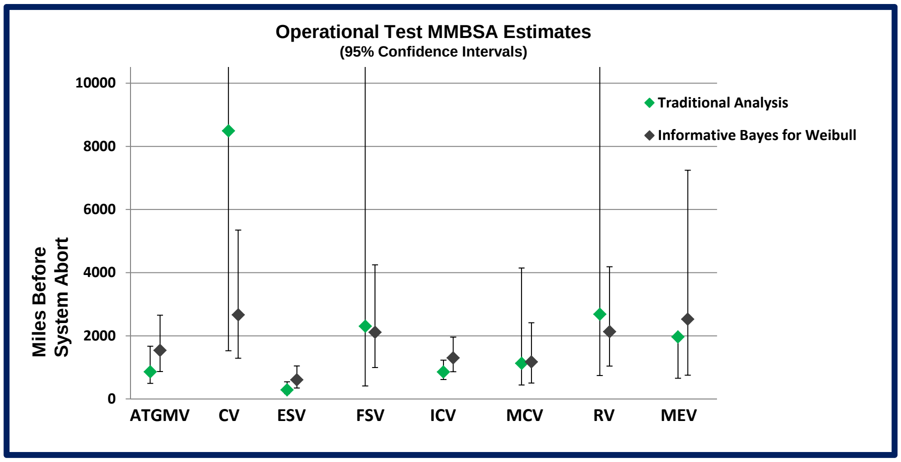
Step 2:
Update missing data and adjust the other failure times accordingly. In this step we can sample using the fact that:

$$t_{missing} | \, \gamma_{phase,variant} \, , \beta \sim Weibull \, (\gamma_{phase,variant} \, , \beta)$$

Step 3:
Repeat Steps 1 and 2 a total of N times.

Results: Incorporating More Information

We can use Bayesian methods for t ~ Weibull too!



Summarizing Confidence Intervals

Reduction in Intervals
(compared to Traditional Analysis)

Vehicle	Under the Assumption t ~ Exponential
ATGMV	0.25
CV	0.99
ESV	0.13
FSV	0.98
ICV	0.10
MCV	0.77
RV	0.91
MEV	
Column Average	0.59

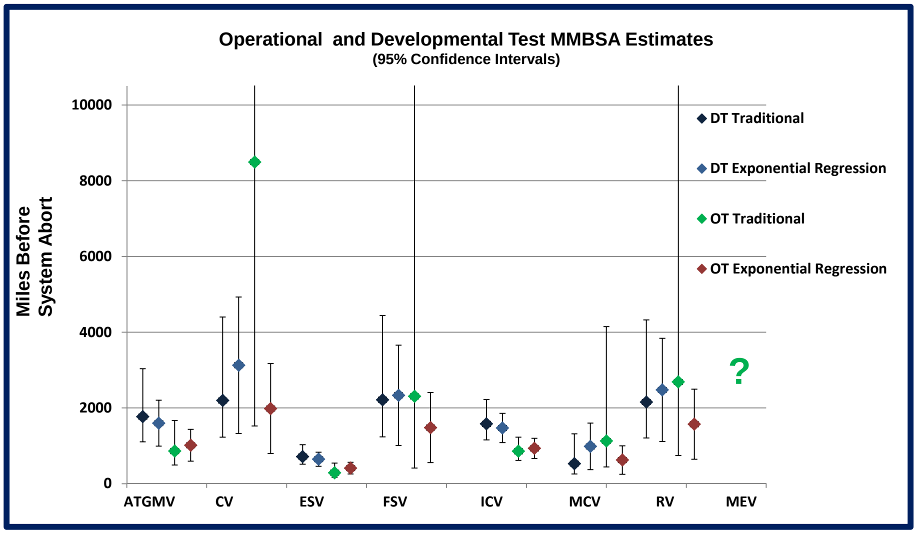
A Traditional Analysis - Using DT Data Only

Stryker Reliability by Variant using Developmental Test Data

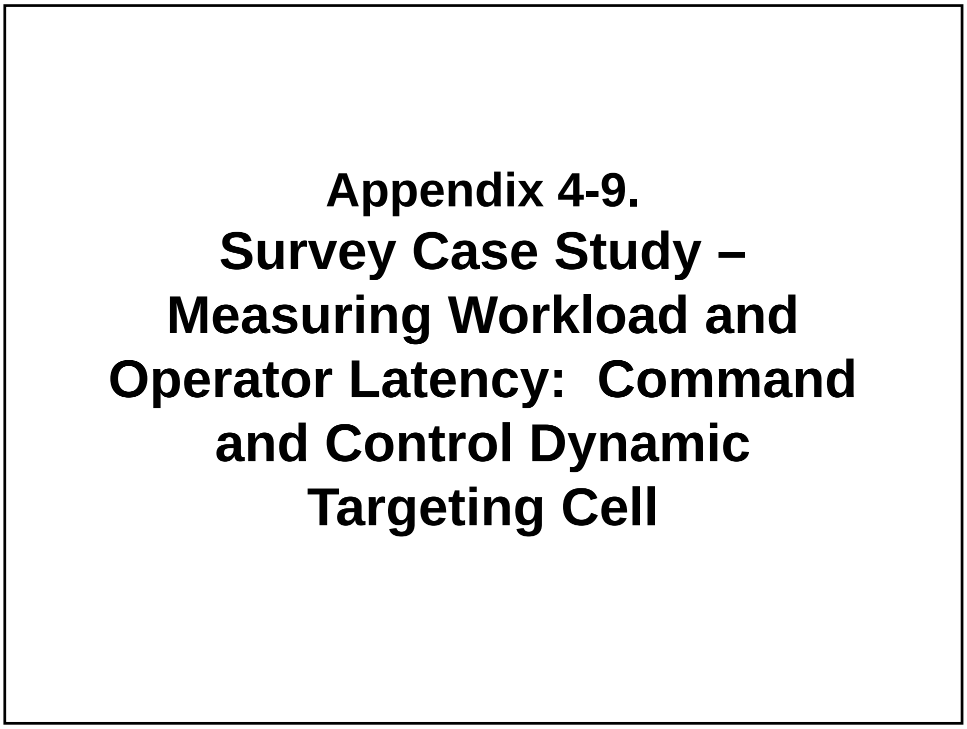
Vehicle Variant	Total Miles Driven	System Aborts	MMBSA	MMBSA 95% LCL	MMBSA 95% UCL
ATGMV	30086	17	1770	1105	3038
CV	24160	11	2197	1228	4400
ESV	25095	35	717	516	1029
FSV	24385	11	2217	1239	4441
ICV	61623	39	1580	1156	2222
MCV	3702	7	529	257	1315
MEV	-	-	-	-	-
RV	23742	11	2158	1206	4324
Total	**192793**	**131**	1472	1240	1760

$$\text{Mean Miles Before a System Abort } (\mathbf{MMBSA}) = \frac{\text{Total Miles Driven}}{\text{System Aborts}}$$

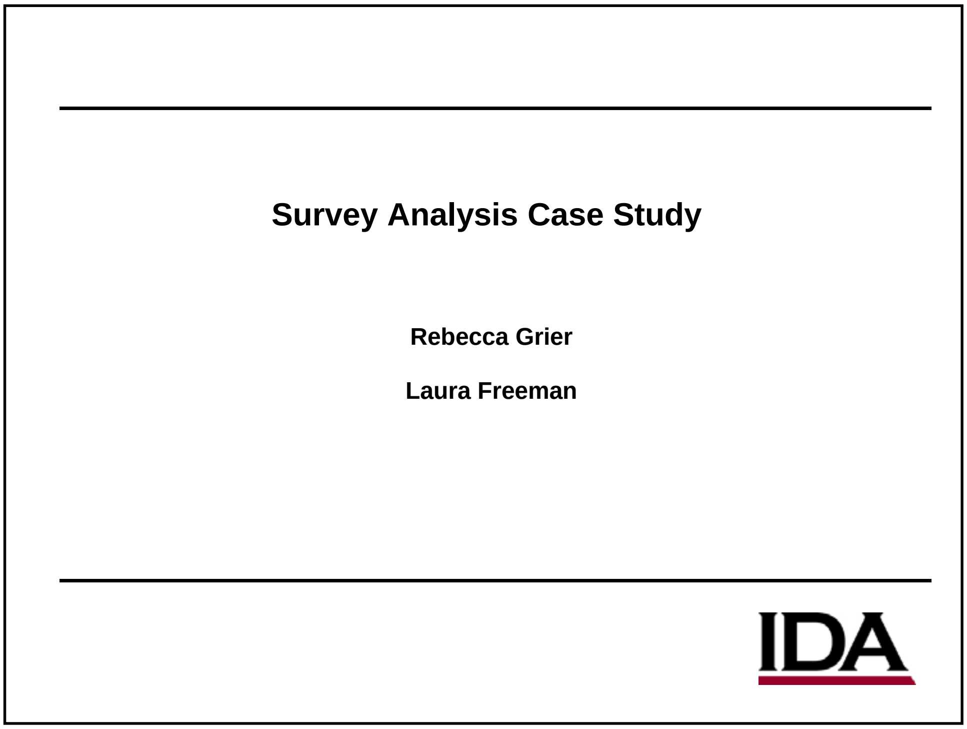
Comparing Traditional Results For DT And OT To Exponential Regression Results



This page intentionally left blank.



Appendix 4-9. Survey Case Study – Measuring Workload and Operator Latency: Command and Control Dynamic Targeting Cell

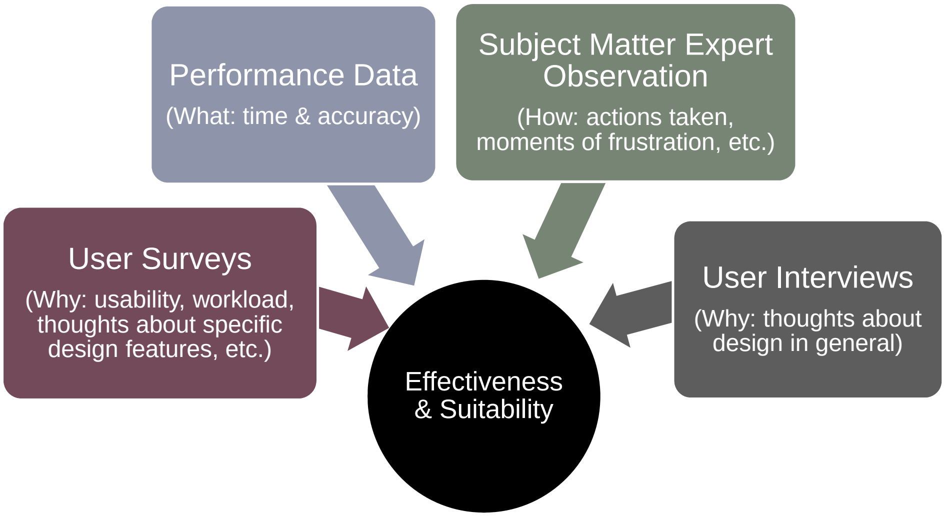


Survey Analysis Case Study

Rebecca Grier

Laura Freeman

Surveys & Interviews: Important Parts of OT&E



Performance Data

(What: time & accuracy)

Subject Matter Expert Observation

(How: actions taken, moments of frustration, etc.)

User Surveys

(Why: usability, workload, thoughts about specific design features, etc.)

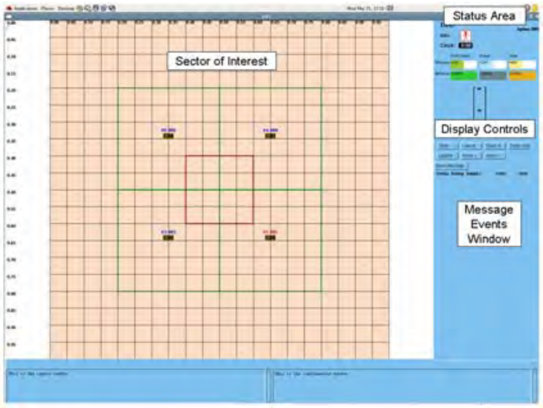
User Interviews

(Why: thoughts about design in general)

Effectiveness & Suitability

Command and Control Dynamic Targeting Cell

- **Study evaluated the latency and workload for participants monitoring a no fly zone for targets**
- **Messages were delivered either via text or audio**
- **Responses**
  - When Target in No Fly Zone – Send Asset
  - Send Text Message – Saying Asset Deployed
  - NASA TLX Workload
- **Air defense scenario hosted on the DDD simulator**



DDD display window for simulation

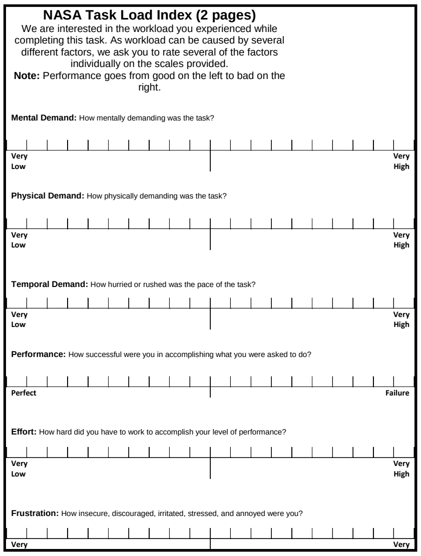
NASA TLX Administration & Scoring

- **Administer Immediately After Task Completed**
- **Considerations:**
  - Only ~90min Assessed
  - Electronic Version Available
- **2 parts**
  - Workload Experience: 0-100 for 6 Types of Workload Contributors
  - Weights: Degree 6 Types Contributed Most to Experience of Workload
- **Formula:**

[MD(MDw) + PD(PDw) + TD (TDw) + P(Pw) + E(Ew) + F(Fw)]/ 15 = workload

	MD	PD	TD	P	E	F	MDw	PDw	TDw	Pw	Ew	Fw	WKLD
A	100	50	0	50	75	0	3	5	1	2	4	0	63.33
B	20	0	50	15	35	20	3	5	1	2	4	0	19.67
C	20	0	50	15	35	20	2	0	5	1	4	3	33.67

NASA TLX Survey



NASA Task Load Index (2 pages)

We are interested in the workload you experienced while completing this task. As workload can be caused by several different factors, we ask you to rate several of the factors individually on the scales provided.

Note: Performance goes from good on the left to bad on the right.

Mental Demand: How mentally demanding was the task?

Very Low — Very High

Physical Demand: How physically demanding was the task?

Very Low — Very High

Temporal Demand: How hurried or rushed was the pace of the task?

Very Low — Very High

Performance: How successful were you in accomplishing what you were asked to do?

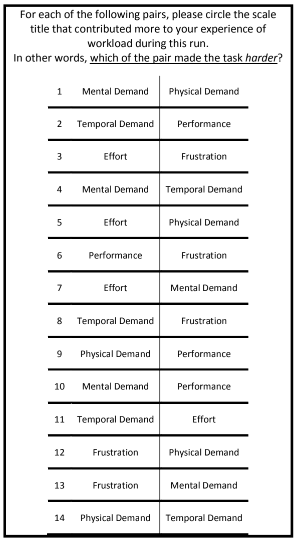
Perfect — Failure

Effort: How hard did you have to work to accomplish your level of performance?

Very Low — Very High

Frustration: How insecure, discouraged, irritated, stressed, and annoyed were you?

Very — Very

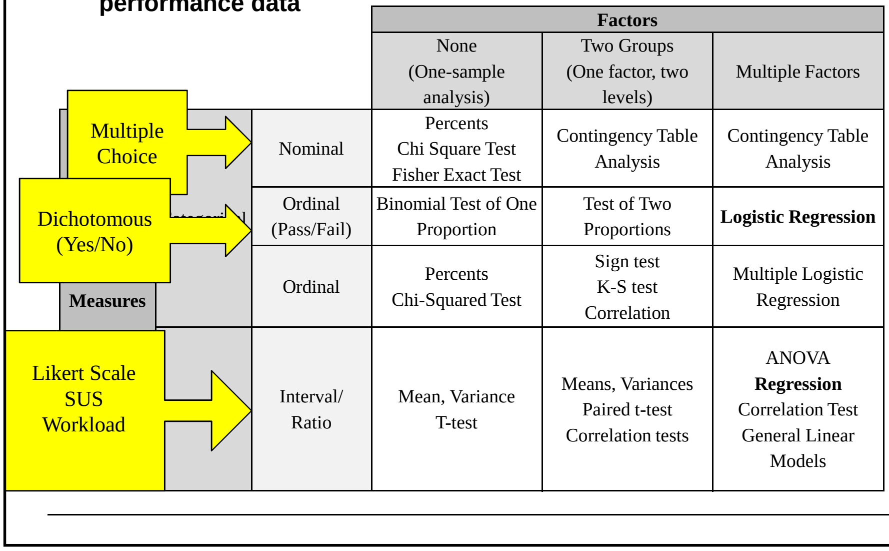


For each of the following pairs, please circle the scale title that contributed more to your experience of workload during this run.
In other words, which of the pair made the task *harder*?

1	Mental Demand	Physical Demand
2	Temporal Demand	Performance
3	Effort	Frustration
4	Mental Demand	Temporal Demand
5	Effort	Physical Demand
6	Performance	Frustration
7	Effort	Mental Demand
8	Temporal Demand	Frustration
9	Physical Demand	Performance
10	Mental Demand	Performance
11	Temporal Demand	Effort
12	Frustration	Physical Demand
13	Frustration	Mental Demand
14	Physical Demand	Temporal Demand

Statistical Analysis Methods for Survey Data

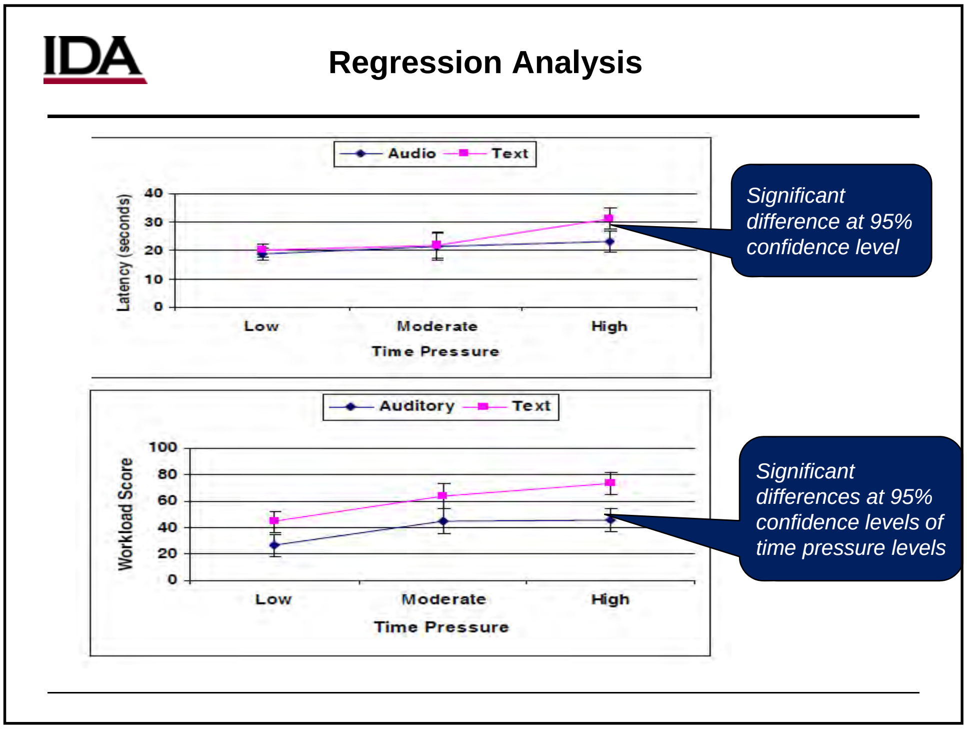
- **Survey data is analyzed using the same statistical models as performance data**



Measures	Factors: None (One-sample analysis)	Two Groups (One factor, two levels)	Multiple Factors
Nominal	Percents Chi Square Test Fisher Exact Test	Contingency Table Analysis	Contingency Table Analysis
Ordinal (Pass/Fail)	Binomial Test of One Proportion	Test of Two Proportions	**Logistic Regression**
Ordinal	Percents Chi-Squared Test	Sign test K-S test Correlation	Multiple Logistic Regression
Interval/ Ratio	Mean, Variance T-test	Means, Variances Paired t-test Correlation tests	ANOVA **Regression** Correlation Test General Linear Models

Experimental Design

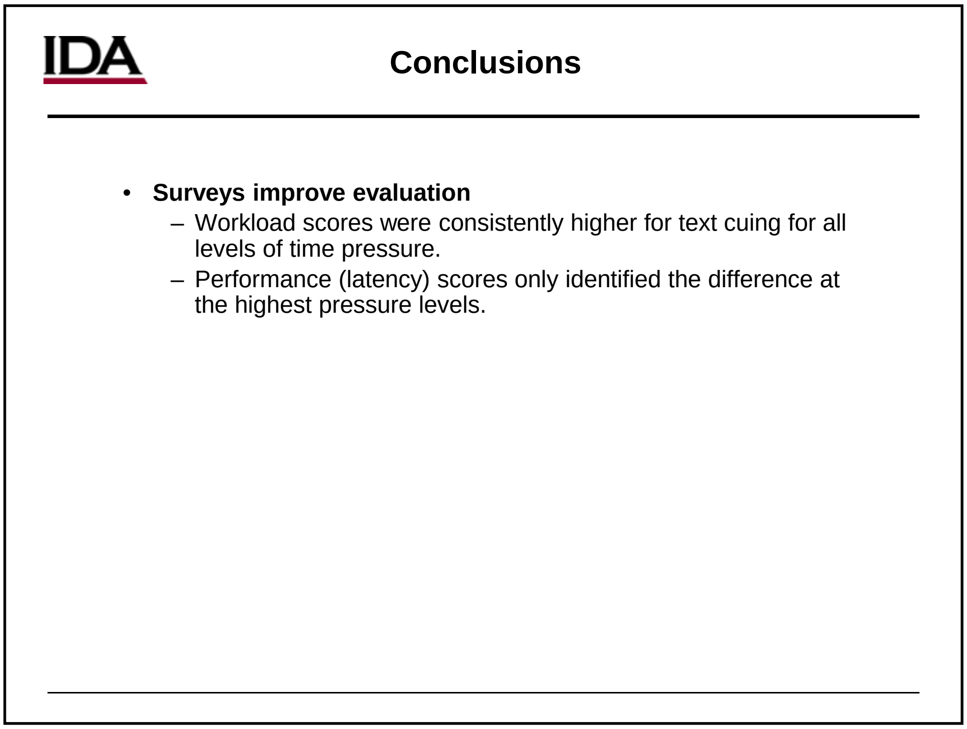
- **12 Participants**
- **Two Factors**
  - Message modality (between subjects):
    - Text (intramodality)
    - Audio (Intermodality)
  - Time pressure (targets per unit time; within subjects):
    - Low – 7 targets in 15min
    - Moderate – 15 targets in 15min
    - High – 30 targets in 15min
- **Response variables:**
  - Latency
  - Workload (TLX)



Regression Analysis

Significant difference at 95% confidence level

Significant differences at 95% confidence levels of time pressure levels



Conclusions

- **Surveys improve evaluation**
  - Workload scores were consistently higher for text cuing for all levels of time pressure.
  - Performance (latency) scores only identified the difference at the highest pressure levels.

This page intentionally left blank.

Appendix 5
White Papers

5-1. Case Studies for the Use of DOE in Developmental Testing

5-2. Mine Susceptibility Comparison Study

5-3. Fuel Leakage Comparative Analysis

This page intentionally left blank.

Appendix 5-1
Case Studies for the Use of DOE in Developmental Testing

Summary

The Director, Operational Test and Evaluation (DOT&E) has advocated the more rigorous application of scientific experimental design in test and evaluation, which includes the application of Design of Experiments (DOE). In this regard, DOT&E policy is not intended to be prescriptive. The director's T&E initiatives letter of 24 November 2009 notes that DOE is "One important means to achieve integrated test…." DOT&E policy recognizes the limitations of DOE and the applicability of other scientific and statistical techniques.

To understand the applicability of DOE to operational test and evaluation (OT&E), we previously conducted a retrospective analysis of OT&E and concluded that DOE was being underutilized. That analysis determined that structured test and evaluation was generally used and that in some test programs DOE techniques had been applied. However, there were many instances where DOE and other statistical techniques could have been applied and improved the test program, but had not.

To supplement our previous analysis of OT&E reports, you asked for preliminary information concerning the use of DOE in test and evaluation activities of a developmental nature. This memorandum examines cases where DOE has been applied, considers why DOE was used, and examines the benefits that the practitioners sought. We note that the cases that we examined are almost exclusively in industry and non-defense government agencies. We have not examined the used of DOE in developmental test and evaluation (DT&E) of Department of Defense (DoD) systems to any significant extent, and we have not studied the distinctions between the use of DOE by defense contractors and government agencies involved in DoD DT&E. We are not aware of retrospective analyses of the potential use of DOE in DoD DT&E similar to the ones we performed for OT&E. Analyses of such cases might provide additional insights on DOE applicability to DoD DT&E.

This memorandum concludes that DOE is applicable to DT&E in many instances, there is long history of its use in industry, and it is considered a "best practice" in industry. We understand the Director, Developmental Test and Evaluation (DDT&E) is developing policy for the application of scientific test and evaluation design (STED) methods to DT&E events. The information in this memorandum may assist you in your discussions on these issues with your DDT&E counterparts.

Background

In order to coherently discuss the use of DOE in DT&E, we must begin by defining an experiment. An experiment is a test event or a series of test events in which purposeful changes are made to the input variables and factors of a process or system so that we can observe and identify the reasons for change in the output response.[1]

DOE is the scientific process of planning the experiment so that appropriate data will be collected, resulting in statistically valid, objective conclusions. The process for applying scientific experimental design to test and evaluation can be divided into the following steps[2]:

(1) Identify the questions to be answered, also known as the objectives.

(2) Identify the quantitative metrics, also known in the statistical world as response variables, in support of those questions.

(3) Identify the factors that affect the response variables. Factors are broad categories of test conditions that affect the outcome of the test. In developmental testing, factors might include system configuration, temperature, and pressure.

(4) Identify the levels for each factor. For example, a factor such as temperature might have levels such as high temperature and low temperature. The levels represent various subcategories between which analysts and engineers expect system performance to vary significantly. When performance is expected to vary linearly, two levels are used. Nonlinear performance typically results in three or more levels.

(5) Identify applicable DOE techniques. Examples of DOE techniques include factorial designs, response surface methodology, and combinatorial designs. The applicable DOE technique depends on the question, the metrics, the types of factors (numeric or categorical), and available test resources.

(6) Identify which combinations of factors and levels will be addressed in each test period (i.e., coverage of the envelope). In statistical terms, this is often referred to as blocking.

(7) Identify relevant statistical measures of the test (e.g., confidence, power, effect size).

Many of the steps outlined above are part of the longstanding practices of the test and evaluation community. What the emphasis on DOE brings is a shift in those practices to apply scientific experimental design principles. In the retrospective analysis, we noted that most operational testing employed a structured approach to testing due to the fact that many of the steps described above were already being employed, particularly steps 1 through 4. That analysis also noted, however, that in many areas a more rigorous application of DOE principles would have improved test and evaluation. Specifically, it was noted that step 4, while generally considered, could have been conducted in a more rigorous and systematic fashion. Additionally, if steps 5 through 7 had been implemented, they would have identified holes in the testing where performance was not examined and would have provided an assessment of the uncertainties in the measurements and conclusions.

[1] Definition adapted from: Montgomery, Doug, *Design and Analysis of Experiments, 6th Edition*, 2005, John Wiley & Sons, Inc.

[2] These steps directly map to steps 1 through 4 in Montgomery's Text (see note 2), page 14, Table 1.1.

Objectives of DOE

DOE is a rich scientific methodology, containing many tools. The specific tool that is employed depends on the question to be answered (step 1). The question, or in other words the objective of the test, can vary significantly from one developmental test to the next. And the questions and objectives can change as the system under test matures. The choice of DOE technique (step 5) should reflect the objective. Table 1 below lists several common objectives and the corresponding designs one might select to satisfy the corresponding objective. This list is intended to show the breadth of tools that are included in DOE, but is far from exhaustive.

Table 1. Test Objectives and Corresponding DOE Designs

Test Objective	DOE Design Method	Examples in this Memorandum
Product design and development	Super-Saturated Designs, Factorial and Fractional Factorial Designs	Trade Studies and Engineering Analyses
Process optimization	Response Surface Designs, Optimal Designs	Trade Studies and Engineering Analyses
Test for problems	Combinatorial Designs, Orthogonal Arrays, Space Filling Designs	Software Testing Integration and Interoperability Testing
Evaluation of material properties	Accelerated Life Tests, Mixture Designs	Accelerated Life Tests
Screen for important factors	Factorial and Fractional Factorial Designs	Characterizing Performance
Characterize a system or process over an envelope	Factorial and Fractional Factorial Designs, Response Surface Designs, Optimal Designs	Characterizing Performance
Develop robust processes (i.e., affected minimally by input conditions)	Taguchi Arrays, Orthogonal Arrays, Response Surface Designs	Not covered in this memorandum

In addition, to the examples in Table 1, DOE is applicable to various certifications. As an example, MIL-STD-1763 describes the process to demonstrate compatibility between an aircraft and specific stores for use on that aircraft. The process involves numerous steps, including structural analysis, flutter analysis, fit tests, and separation tests. Many of these steps are amenable to experimental design. For example, wind tunnel tests are an important step in the certification process, and as will be discussed below, DOE offers substantial benefits when applied to wind tunnel testing. Similarly, CJCSI 6212.01 describes the process for developing, coordinating, reviewing, and approving Interoperability and Supportability (I&S) needs for Information Technology (IT) systems. Part of the process is demonstrating IT standards conformance, and as discussed below, DOE is applicable to examining compliance with communication protocols and interfaces.

In the discussion below, we examine a variety of DT&E papers. We provide examples of using DOE to meet the test objectives, given in Table 1, associated with various systems. The systems considered are not always military systems; however, the examples illustrate types of testing that are applicable to military systems. The goal was to identify how DOE and other scientific experimental design principles have been employed in DT&E. The goal was **not** to provide a comprehensive examination of DOE in DT&E. Because DOE and DT&E are both broad subjects, such an endeavor would be impossible.[3] Instead, the goal was to sample the use of DOE in DT&E to illustrate its applicability. The cases include DOE applied to trade studies and engineering analyses, software and hardware testing, integration and interoperability testing, accelerated life testing, and characterizing performance.

Trade Studies and Engineering Analyses

Trade studies and engineering analyses are a common task early in the development of a new system; Rhew and Parker[4] have described the application of DOE techniques[5] to such analyses. In their example, a trade study and engineering analysis was conducted for the Launch Abort System (LAS) for NASA's manned launch system, Ares I. The LAS is a rocket tower and shroud mounted on the crew vehicle; it is used to separate the crew vehicle from the Ares rocket in the event of an emergency. In assessing various LAS designs, NASA wanted to identify which factors (e.g., tower length, tower diameter, nose shape) affected drag the most. The study used parametric Computational Fluid Dynamics (CFD) models to rank the factors based on their contributions to aerodynamic drag over the vehicle's ascent trajectory. Ultimately, the CFD results fed into wind tunnel analyses.

A DOE approach was used to ensure that important interactions between factors were understood, to examine non-linear behavior, and to limit the scope of the analysis. A traditional analytic approach would have required an examination of all possible combinations of factors and levels, changing one factor at a time.[6] Such an approach would have required an analysis of at least 1,556 LAS configurations to study seven factors, and it would have ignored important interactions between the factors. Under a DOE approach, however, only 84 configurations were required to study the same seven factors. In addition, the DOE approach allowed critical interactions between factors to be examined, and it allowed an analysis of non-linear performance. Rhew and Parker noted that the DOE approach represented a starting point for experimental activities that would eventually explore the entire design space.

Holcomb, Montgomery, and Carlyle,[7] in another study, employ the use of DOE[8] in the development of a turbine engine. They note that during product development there is usually a

[3] The authors also recognize that this is by no means the first attempt to conduct such overview. The literature is filled with such studies.

[4] Rhew and Parker, *A Parametric Geometry Computational Fluid Dynamics (CFD) Study Utilizing Design of Experiments (DOE).*

[5] In their paper, they used fractional factorial designs with center points.

[6] This type of analysis is known as One Factor At a Time analysis.

[7] Holcomb, Montgomery, and Carlyle, *The Use of Supersaturated Experiments in Turbine Engine Development.* Quality Engineering, 2007.

[8] In their paper, they use a supersaturated design.

significant time constraint. DOE offers a useful method of examining many design factors with only a few tests. Once the factors influencing the design's performance are identified, the designer can rapidly make meaningful design decisions.

The goal of the study was to identify factors that affect the performance of a turbine engine. Engineers identified 27 potential factors, including heat transfer coefficients, shaping of specific components, and loads. DOE allowed for the investigation of the 27 factors with between 12 and 20 tests, depending on the DOE selected. However, by using such a small number of tests, there was a high risk of mistakenly concluding that a factor was not significant (the design had low power). DOE allows this risk to be quantified.

Software Testing

Many military systems employ complex software (and hardware) that is developed in an evolutionary manner, with functionality being developed incrementally and tested in each iteration. The number of combinations of input data, operator actions, etc., can be huge. As a result, testing can be overwhelming.

Burr and Young have described the application of DOE[9] to software testing.[10] Others have described similar applications to software and hardware suites.[11] In the Burr and Young example, they examined testing of an email system. Traditional testing would have required 27 trillion test cases. They note that under traditional approaches, test cases take too long to create, too long to automate, too long to run, too long to verify, and for new software and hardware builds there is no easy way to know which test cases need to be re-run for regression testing.

Burr and Young describe the DOE approach as a "best practice" for industry, and by applying DOE in their problem, they were able to reduce the number of test cases from 27 trillion to 100. Within the smaller number of test cases, they were able to cover 97 percent of the branches (conditional statements) within the software and 93 percent of the testable code. In contrast, they note that typical software testing covers only 40 to 60 percent of the code.

In a similar study, National Institute for Standards and Technology (NIST) researchers, Kuhn and Reilly,[12] use DOE techniques in software testing. They employ a DOE approach[13] that allows for a large number of input conditions to be covered in a small number of runs. They examined two open source projects: the Mozilla web browser and the Apache web server. Both projects have large sets of code, large user bases, and extensive databases of reported bugs. Kuhn and Reilly conclude that 89 (Apache) to 95 percent (Mozilla) of reported bugs could be found using only three small DOE designs, and 100 percent of reported bugs could be found using six small DOE designs. The advantage of using DOE in this case was that Kuhn and

[9] These papers describe the application of combinatorial DOE designs.

[10] Burr and Young, *Combinatorial Test Techniques: Table-based Automation, Test Generation and Code Coverage*, Software Engineering Analysis Lab, Nortel.

[11] Hartman, *Software and hardware Testing Using Combinatorial Covering Suites,* IBM Haifa Research Laboratory.

[12] Kuhn and Reilly, *An Investigation of the Applicability of Design of Experiments to Software Testing.* NASA IEEE Software Engineering Workshop, 2002.

[13] They employ combinatorial designs as their DOE approach.

Reilly were able to find the majority of reported bugs quickly. These techniques are applicable early in software development when code segments are being tested.

Integration Testing

During DT&E, it is common to conduct integration tests to examine whether systems have properly implemented communication protocols, interfaces, and other requirements; Burroughs, Jain, and Erickson described the application of DOE[14] to such testing.[15] In their example, Burroughs, Jain, and Erickson examined testing of telecommunication switches using Integrated Services Digital Network (ISDN) protocols. ISDN is a set of communications standards for the simultaneous digital transmission of voice, video, data, and other network services over telephone circuits.

The problem encountered in this case is common: the number of possible combinations of message types, message originator, interface configurations, etc., is large. Traditional testing approaches do not provided sufficient breadth of coverage.

Burroughs, Jain, and Erickson noted that DOE allowed integration testing to be conducted that provided "much broader coverage of the test space without leaving any systematic holes." Testing is easily implemented in automated test systems, and the "improved quality of testing leads to faster detection of non-conformances, and a higher quality of products in a shorter development interval."

Interoperability Testing

Also common to DT&E is interoperability testing; Brownlie, Prowse, and Phadke describe a DOE approach for such testing.[16] Their problem was to examine interoperability of a new email software release within an environment that included multiple operating systems, hardware configurations, and client and server software. Testing examined interoperability at the functional level (e.g., copy function).

Brownlie, Prowse, and Phadke noted that testing takes up a significant portion of development resources and that a DOE approach improved testing. They concluded that DOE-based testing was completed in less staff time, provided systematic testing of the product functionality, higher confidence in coverage of the requirements, and discovered more faults (in their case, 22 percent more faults).

Accelerated Aging

Accelerated aging is a common procedure during DT&E. In a presentation to the DOT&E Science Advisor (February 2009), NIST described the use of DOE in accelerated aging programs to determine the lifetime of compact disks (CD). The testing was conducted in

[14] In their examples, they use orthogonal arrays.

[15] Burroughs, Jain and Erickson, *Improved Quality of Protocol Testing Through Techniques of Experimental Deign*, IEEE, 1994.

[16] Brownlie, Prowse and Phadke, *Robust Testing of AT&E PMX/StarMAIL using OATS*, June 1992, AT&T Technical Journal.

cooperation with the Library of Congress to examine archiving of data. It was known that high temperatures and humidity could degrade CDs. The objective of the testing was to estimate the lifetime of commercially available CDs.

NIST has described the testing in publications.[17] The DOE approach taken allowed a specific life expectancy model to be applied in a systematic way. A sample of 100 CDs was divided into six groups. Each group was exposed to one of six levels of stress (higher temperature and humidity). After each period of exposure, each CD was tested to evaluate any degradation in performance. Statistical analysis of the data allowed the team to estimate the life expectancy of the CDs. It also allowed them to estimate how stresses from temperature and humidity mighty reduce life expectancy.

Characterizing Performance

During development, it is common that requirements must be verified by characterizing the performance of the system or subsystem; DOE is applicable to these tests. As an example, the Joint Chemical Agent Detector had a test requirement to characterize its ability to detect chemical agents as a function of agent concentration, atmospheric water vapor content, and temperature. The goal was to determine the mathematical equations that related these quantities to probability of detection, time to detect, and other relevant metrics.

The testing was conducted in the laboratory under developmental test conditions and employed a DOE approach.[18] In this case, DOE was selected in order to provide what is known as a response surface model (the mathematical relationship between factors mentioned above). This approach has been used throughout the program's history as the system has been developed. It has provided test results with high statistical confidence.

In another example, Landman, Simpson, Mariani, Ortiz, and Britcher[19] use DOE techniques to characterize the aerodynamic behavior of the X-31 Enhanced Fighter Maneuverability program. The aerodynamic behavior of an aircraft is characterized through aerodynamic equations. Traditionally, one factor at a time experiments have been used to vary the factors in the wind tunnel. For aerodynamic analysis, this often requires more than 1,000 test points.

Such testing can require weeks of wind tunnel time and is complicated by instrument drift over the lengthy test periods. Instrument drift leads to biases in the results. Landman *et al* use DOE[20] in this example to characterize the aircraft's aerodynamic performance as a function of altitude and aerodynamic control inputs in only 104 test points. The dramatic reduction in the number of test points reduces instrument drift concerns. Additionally, based on the response surface models, the DOE allowed for predictions accurate to within one percent of the true value. It also allowed for the characterization of experimental error through an analysis of variance.

[17] *NIST/Library of Congress (LoC) Optical Disc Longevity Testing Procedure,* NIST Special Publication 500-263.

[18] JCAD employed a D-Optimal test design.

[19] Landman, Simpson, Mariani, Ortiz, and Britcher, *A High Performance Aircraft Wind Tunnel Test using Response Surface Methodologies.* U.S. Air Force T&E Days, 2005.

[20] They employ a Response Surface Design the Face Centered Cube (FCC) DOE technique.

Finally, the DOE revealed unexpected interactions. The interactions would have been impossible to detect using traditional experiments.

Characterizing Performance across DT&E and OT&E

In addition to characterizing performance solely in developmental testing or in operational testing, it might be important to characterize performance and ensure coverage of the envelope across DT&E and OT&E.

Hutto and Kowalski[21] use a DOE approach to ensure adequate testing of the MAU-209 B guidance kit across developmental and operational testing. The guidance kit straps on to the MK-82 and MK-84 bomb, which turns it into a laser-guided bomb. Several factors were identified as affecting the performance of the guidance kit. A factorial design was used to ensure that all important combinations of factors and levels were covered between DT&E and OT&E with adequate confidence and power. The DOE provides important understanding of where the DT&E could be improved. It also provides information on where DT&E and OT&E testing can be synergistic.

Careful Planning

In this memorandum, we provided examples of successful implementation of DOE techniques to DT&E. These case studies omit one of the most important aspects of DOE. The DOE process requires critical thought in the planning stages of potential factors and levels using the expertise of engineers and scientists. This process can prevent gaps in testing by initiating the thought process on causal factors and environmental factors that might affect the outcome of the test. The worst unknown is the unknown-unknown. The DOE process, properly executed, helps to reduce the risk of unknown-unknowns.

Conclusion

It is clear that DOE is applicable to many areas of DT&E and that it has a wide range of benefits – systematic coverage of the envelope, improved quality of testing with faster detection of problems, a higher probability of detecting faults, potential cost and time efficiencies, and the ability to quantify the risks inherent to any test program.

The case studies presented in this memorandum represent only a fraction of the publically available literature on DOE in DT&E. Nonetheless, they represent cases that span the range of developmental test and evaluation activities, from early engineering analyses, through incremental development of software and hardware, to final verification of system requirements. The application of DOE to DT&E in the open literature is dominated by examples from industry. Only limited information is available on the application of DOE to DT&E of military systems.

[21] Hutto, Drenth, Kowalski, and Sparkman, *Design of Experiments: Meeting the Central Challenge of Flight Test,* Page 16-27.

Appendix 5-2
Mine Susceptibility Comparison Study

Summary

Design of Experiments (DOE) is a methodology for planning and analyzing tests. In this memorandum, we compare multiple design methodologies for the mine susceptibility test of the *Lewis and Clark* Class (T-AKE-1) Dry Cargo/Ammunition Ship using the Advanced Mine Simulation System (AMISS). The comparison study determines the trade space between the number of test conditions (factors) examined, the sample size (test cost), and the associated test risk. A two-part comparison study first compares seven different statistically optimum designs to determine the trade-off between sample size and statistical power, which is a measure of test risk. The result from this comparison study shows that designs between 20 and 28 test points are adequate to fully characterize the performance of T-AKE-1 against AMISS as a function of range, ship speed, and whether or not the degaussing system is turned on.

A second comparison study examines the impact of adding and removing additional factors form the design on statistical power. From this study, one can see that there is only a minimal impact of adding or removing factors from the design in terms of statistical power.

Overview

The goal of this mine susceptibility trade study is to evaluate potential test designs for a mine susceptibility test and determine the trade space between the number of factors, the sample size (test cost), and the associated test risk. This is accomplished by a two-step comparison study. In the first comparison study we compare different test designs of varying size for a fixed number of factors and investigate trade-offs in test risk as a function of design type (sample size). In the second comparison study, for a fixed number of samples (16 and 36) we investigate the trade-off between risk and the number of factors included in the test design.

The goal of the test is to characterize the detonation distance for a variety of mine types for a surface ship. The factors that may influence the range at which the mine detonation occurs are:

- Speed of the surface ship
- Horizontal range of the ship to the simulated mine
- Degaussing status of the ship
- Machine line-up (correlated with speed)
- Ship’s direction (north/south approach versus east/west approach)

The first three factors (speed, range, degaussing status) are the most important factors to investigate. Therefore, these three factors will be used to determine the base designs for the first comparison study. Since, machine line-up is correlated with the speed of the ship it will be treated as a recordable factor and not considered in any of the test designs. In the second comparison study the impact of adding and removing a factor from the base design are considered. Table 1 shows the factors considered in the second comparison study.

Table 1. Factors Considered in Factor Trade-off Study

Number of Factors	Factors Considered
2	Range, degaussing status
3	Speed, range, degaussing status
4	Speed, range, degaussing status, ship's direction

To compare the design we will use two metrics, the first is the number of model terms that are estimable based on the design type. Consider the following generic statistical model:

$$y_i = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i=1}^{k} \sum_{j \neq i}^{k} \beta_{ij} x_{ij} + \sum_{i=1}^{k} \beta_{ii} x_i^2$$

where k is the number of factors considered in the design. The first summation $\sum_{i=1}^{k} \beta_i x_i$ provides the "main effect" of the factor on the outcome. In our case, these terms provide the estimated mean shift in response (detonation distance) for the factors. The second summation, $\sum_{i=1}^{k} \sum_{j \neq i}^{k} \beta_{ij} x_{ij}$, provides the interaction effects, which provide information on how factors work synergistically to impact the detonation distance. The final summation, $\sum_{i=1}^{k} \beta_{ii} x_i^2$, provides the quadratic effects, which account for non-linear relationships between the continuous factors (range, speed) and the test outcome. The ability to estimate more model terms provides increase flexibility in the analysis and therefore is desirable. We could continue to expand upon the model he to higher order terms (three-way interactions, cubic terms) However, from the principle of sparsity of effects we know that typically second order models are adequate to characterize the response (think Taylor series). For the three factors considered in the first comparison study, an ideal number of model terms is eight (three main effects, three two-way interactions, and two squared terms).

The second metric considered is the power for estimating model terms. For a designed experiment, the power calculations tell us about our ability to detect an effect of a factor as different from zero. This is one estimate of test risk. Power is the probability that given β_i has a non-zero effect on the detonation range that we will be able to conclude that based on our testing. This is a key element for determining an adequate test. The remainder of this document is laid out as follows:

- Overview of common statistical designs that are viable candidates for the mine susceptibility test.
- Comparison study of sample size/design type versus test risk
- Comparison study of number of factors versus test risk
- Recommendations

Potential Test Designs

Table 2Table below provides seven common statistical designs for the three primary factors considered in this comparison study (speed, range, and degaussing status). These designs have been shown by the statistical literature to be the best designs available for three factor tests.

Table 2. Designs Evaluated in Comparison Study

	Design Type	Number of Runs	Estimable Model Terms	Design Properties
1	Full Factorial (2-level)	8	6	Smallest possible design to investigate 3 factors and their interactions. Very low power for detecting factor effects.
2	Full Factorial (2-level) replicated	16	7	Increased power over non-replicated 2-level factorial design. Adds the ability to estimate a three-way interaction over the un-replicated design.
3	General Factorial (3x3x2), also referred to as a Face Centered Cube (CCD) Design	18	9	Three-level designs for the continuous factors allow for the estimation of squared model terms.
4	Central Composite Design (w/ 1 center point)	18	9	Five – level design produces a rotatable design that balances variance and increases power.
5	Central Composite Design (replicated center point)	20	9	Center point replication allows for an estimate pure error (variability between runs under the same conditions) in addition to all other design benefits.
6	Central composite Design with replicated factorial points (Large CCD)	28	9	Large design has great power and the ability to estimate all desired model terms.
7	Replicated General Factorial	36	9	Large design with good power but not as optimum as the Large CCD.

Notice in Table, that the smallest two designs support a smaller model than the other designs.

Figures 1 – 3 provide a pictorial view of what these designs look like.

In Figure 1, one can see the layout of the design for the 2-level full factorial design. The scales are in coded units, one the actual ranges of interest for both horizontal range and airspeed are determined the scales can be adjusted to match the low and high values. The purple boxes with the number “2” next to them indicate that two runs will be executed at this point, one with the degaussing system turned on, the other with the degaussing system turned off.

The second design simply replicates the full-factorial design illustrated in Figure 1 such that there are 4 points run at each design point (2 with degaussing, 2 without degaussing).

Figure 2 shows the layout for Design 3, the general factorial design. Notice that the design adds “axial points” colored in green, and a “center point” colored in brown to the full factorial layout. These points allow for the estimation of the additional desired model terms.

Figure 3 illustrates Design 4, the CCD. Notice that this design pulls the green axial points out to make a spherical design region. This balances the information across the design space, resulting in lower variance for each for estimating each of the model terms.

Design 5 simply replicates the brown center point of Figure 3.

Design 6, the Large CCD, replicates the purple factorial points and the brown center point from Figure 3 resulting in 28 total runs.

Design 7, the replicated General Factorial Design, replicates all of the design points in Figure 2.

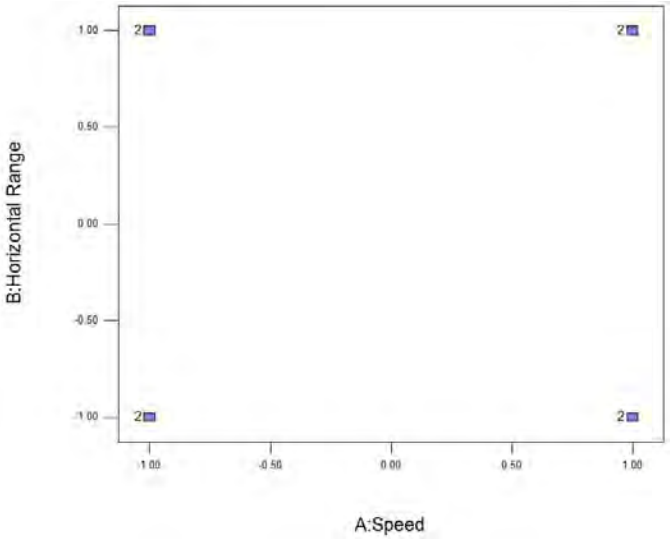


Figure 1. Full Factorial Design (2-level)

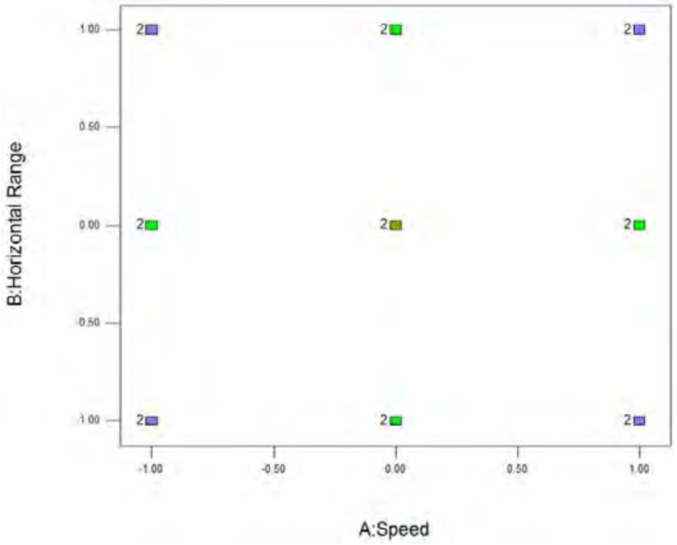


Figure 2. General Factorial Design (3x3x2)

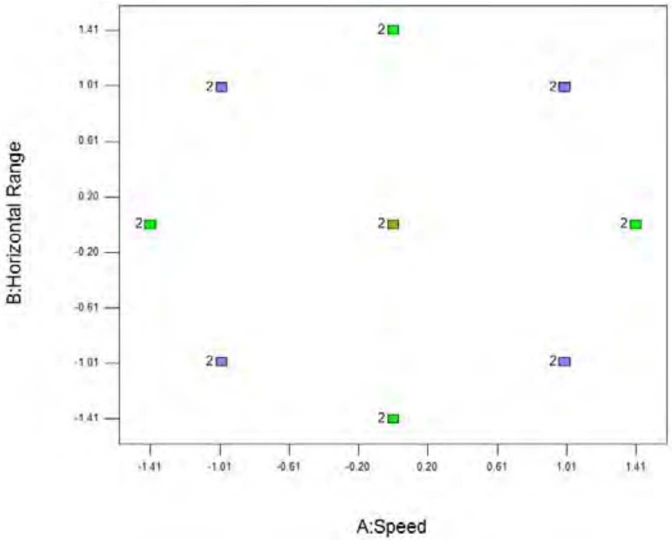


Figure 3. Central Composite Design

Design/Sample Size Comparison

Figures 4 and 5 examine the trade space between the design type, and therefore sample size, and power. Typically, power levels above 80 percent are considered favorable for adequately covering the design space. A test with 80 percent power means that if a factor, for example degaussing status, has an effect on the test outcome, we will have an 80 percent probability of being able to conclude that based on the data collected in the test. The detectable difference of one standard deviation (σ) tells us about the magnitude of the difference in the test outcome that we will be able to detect. Figures 4 and 5 show the power levels for the main effects factors in each of the designs for a detect able difference of one standard deviation and two standard deviations respectively. The power results for the two-way interactions are similar in magnitude due to the inherent balanced of the all the designs.

Notice only the smallest design (Design 1) provides extremely low power, meaning that this test is high risk for failing to detect the impact of the degaussing system (or any other factor). Figures 4 and 5 show that if one is interested in effects on the order of twice the standard deviations any of the Design 2 – 7 will be adequate. However, if one is interested in effects on the order of the one standard deviation, the larger designs (Design 6 with 28 runs and Design 7 with 36 runs) are recommended.

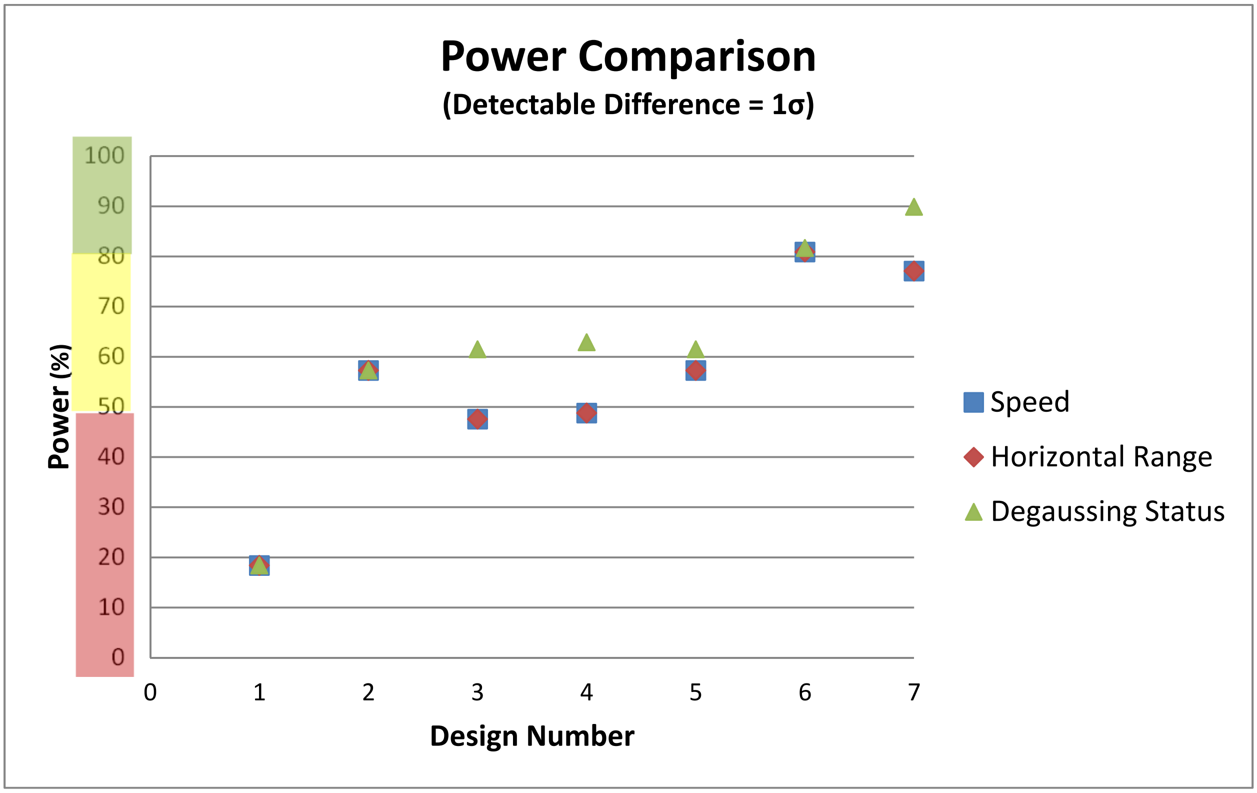


Figure 4. Power Comparison for the Model Main effects at the 90% Confidence Level

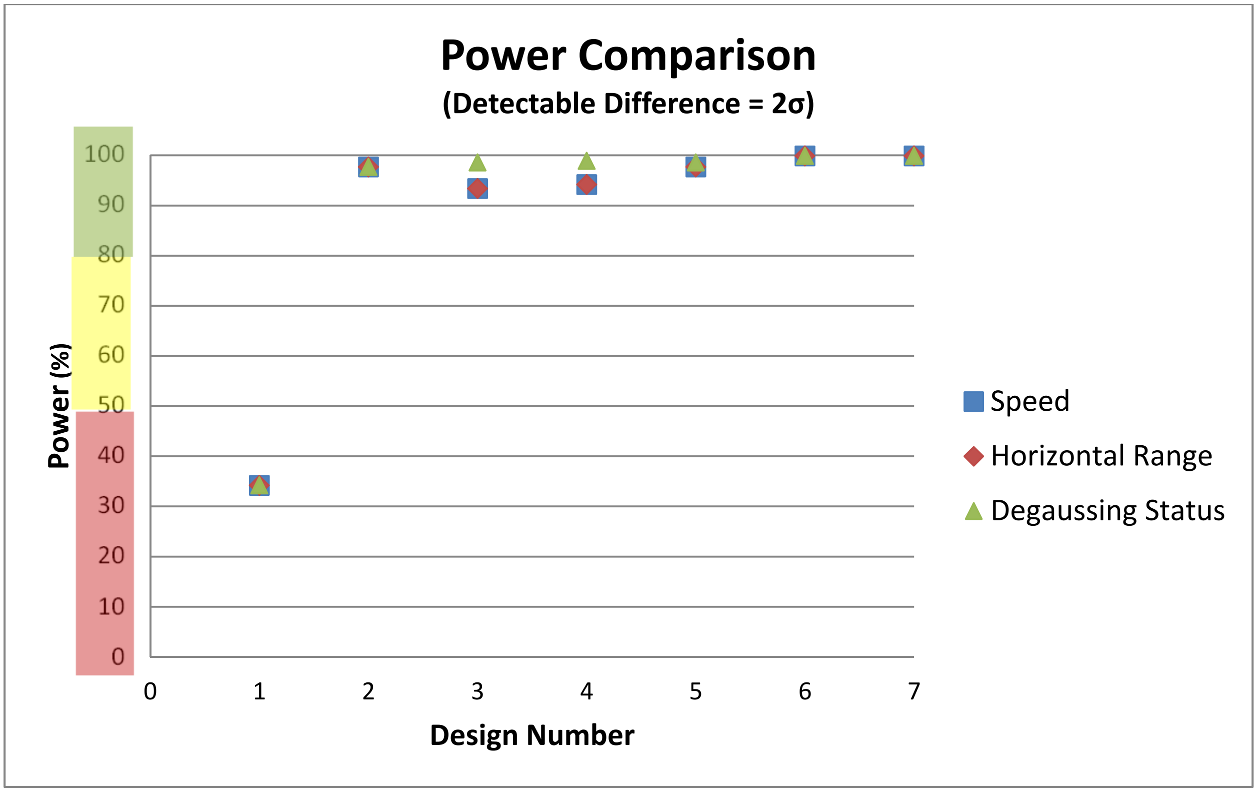


Figure 5. Power Comparison for the Model Main effects at the 90% Confidence Level

Number of Factors Comparison

The second comparison examines the trade space between the number of factors and power. Figure 6 shows the power for testing main effects as a function of the number of factors considered in the design. Notice that there is a decline in power, as expected, when the number of factors is increased. However, the decrease in power is minimal compared to the risk of not having any information on that factors impact on the outcome of the test if it is not considered at all. For a constant test size, the power for each factor main effects only decreases by on average 5.75 percent when increasing the number of factors from two to four factors.

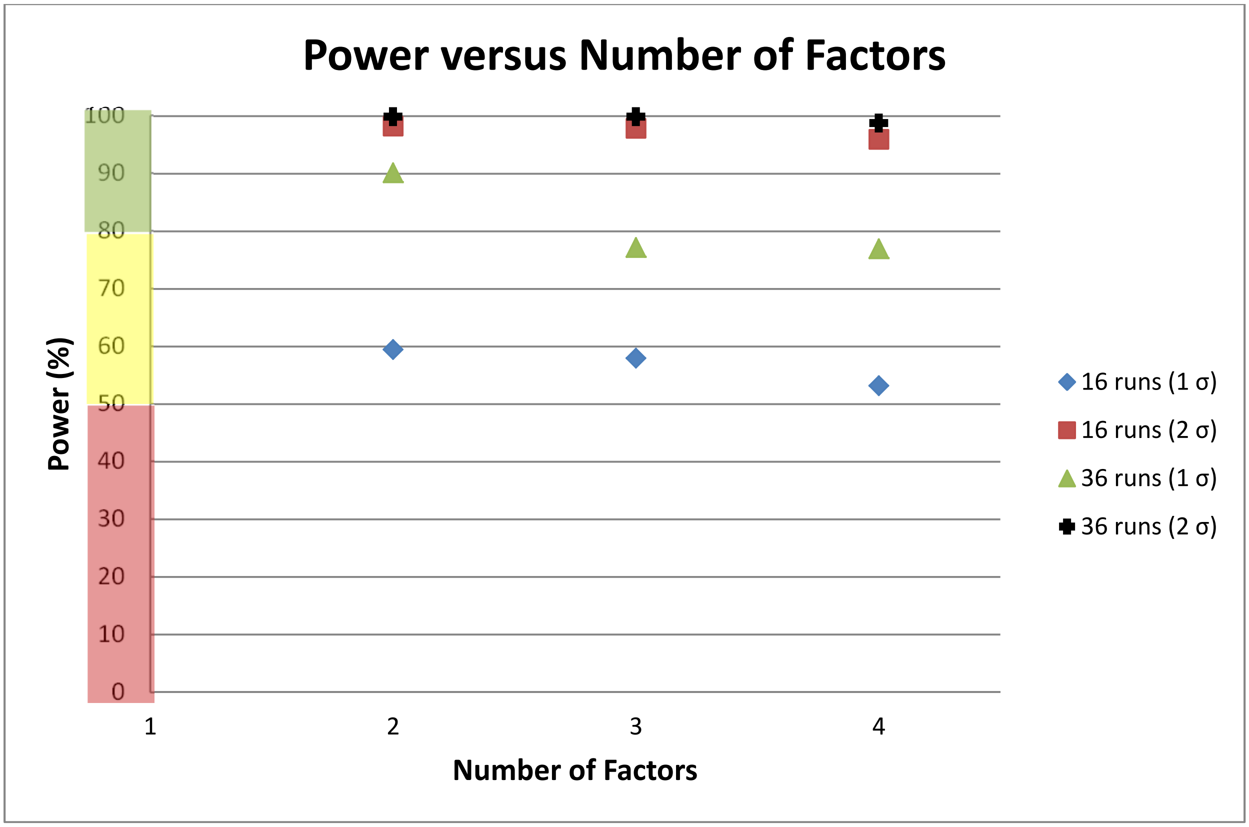


Figure 6. Power for the Model Main effects for 2, 3, and 4 factors at the 90% Confidence Level

Recommendations

Design 5 and Design 6 both provide excellent coverage of the factors that impact the outcome of the mine detection simulation test. It would be prudent to plan for Design 6 to provide more discriminatory ability between the factors levels and their effect on the outcome of the test. Additionally, these designs provide five levels of the horizontal range, which allows for flexibility in the test setup. One of the unknowns going into testing is the exact values of the horizontal ranges needed to ensure useful data is collected. Five levels allows for maximum flexibility in moving between different levels as data is collected throughout the test to determine the most appropriate sets of ranges for the ship from AMISS. However, if achieving five levels of the speed and horizontal range is not possible, then Design 3 is another competitive test design option.

Another point of interest is the building block nature of all of the test designs. In fact, design 1 is actually a subset design of all the other designs. A good test execution strategy might be to execute the subset of Design 6 that aligns with Design 1 first. A preliminary data analysis of the eight runs can be done to determine the relative impact of each of the factors on the test outcome. Adjustments based on the outcome of the initial analysis can be made to maximize the benefits of the remaining test points. Potential adjustments include, adding/removing an additional factor, reducing the required number of test points, and rescaling the levels of either the range and/or speed factors.

This page intentionally left blank.

Appendix 5-3
Fuel Leakage Comparison Analysis

Summary

The Naval Air Systems Command conducted live fire testing to determine the impact of fuel type on the self-sealing properties of aircraft fuel bladders. The objective of the test was to collect data to determine if switching fuel types, from traditional petroleum based fuels with high aromatic contents to a bio-fuels negatively impacts self-sealing. Four fuels were considered in the experiment, JP-5 (20.5 percent aromatics), JP-8 (11.5 percent aromatics), hydrotreated renewable jet fuel (HRJ-5) (0 percent aromatics), and a 50/50 blend of JP-5 and HRJ-5 (9 percent aromatics). The four fuel types were placed in similar test setups consisting of a metal test cubes with fuel bladder panel/backing board facing the gun. The panels were impacted by fully tumbled 7.62-millimeter (mm) round and the leakage of fuel was measured for 6 minutes.

Prior to the completion of the analysis described in this memorandum, two separate analyses were performed on the data collected by the Navy Live Fire and NAVAIR. The two analyses focused on comparing only a subset of the fuels tested (i.e. each vendor was treated as an independent subset) and resulted in difference conclusions about the impact of the biofuel on self-sealing properties of fuel cubes. IDA conducted a third analysis described in this memorandum to independently determine if the use of biofuels impacts the self-sealing ability of fuel cubes.

The analysis that follows uses linear mixed modeling to determine if the fuel type impacts the leakage rate for the data under consideration. We conclude that there is no statistical difference between three of the four fuel types: JP-8, HRJ-5, and the 50/50 blend. JP-5 fuel results in a statistically significant reduction in the fuel leakage over the six-minute test period from the JP-8 fuel, but there is no statistical difference between JP-5 and HRJ-5 of the 50/50 blend over the six-minute test period. Additionally, the analysis shows that all of the fuel types exhibit some degree of self-sealing within approximately two minutes.

Overview

The Navy recently conducted live fire testing at the Naval Air Warfare Center Weapons Division (NAWCWD), China Lake, Weapons Survivability Laboratory (WSL) in support of the support of the Navy’s Alternative Fuels program. The testing was conducted to help clarify the potential vulnerabilities associated with the use of biofuels in military aircraft. The objective of the live fire testing was to provide data regarding the relative self-sealing performance of fielded military aircraft fuel bladder materials when used in conjunction with biofuels with reduced aromatic content. Fuel bladder materials came from four different vendors: Meggitt, GKN, METS, and AmFuel, but comparing the self-sealing capabilities of the different vendors was not a goal of the testing. The alternative fuel used in testing was a hydrotreated renewable jet fuel (HRJ-5) designed to meet the JP-5 specification. The hydrocarbons present in this fuel are nearly identical to petroleum fuels, but lack the aromatic compounds found in petroleum. Table 1 summarizes the fuels used in this live fire test.

Table 1. Fuels and Corresponding Aromatic Contents

Fuel	Aromatic Content
JP-5	20.5%
JP-8	11.5%
Neat HRJ-5 (Neat)	0%
50/50 Blend of HRJ-5 and JP-5	9%

After the completion of testing, two separate analyses were conducted on the raw data. The first analysis used statistical t-tests to determine if the mean leakage rates were different at each time step in the data collection for between JP-8 and the 50/50 blend. The analysis focused on these two fuel types because they provided the closest match in aromatic content. The first analysis concluded that the data did not support the conclusion that there was a difference in performance between the two fuel types. The second analysis used a linear extrapolation of the aromatic content of the traditional fuels (JP-5 and JP-8) to match the 50/50 Blend. The second analysis concludes that there is a significant difference between a hypothetical traditional petroleum based fuel at 9% aromatic content and the biofuel 50/50 blend fuel with 9 percent aromatic content. These two analyses focused on comparing only a subset of the fuels tested (i.e. each vendor was treated as an independent subset) resulting in difference conclusions about the impact of the biofuel on self-sealing properties of fuel cubes.

In this memorandum, IDA provides a third analysis that incorporates all the data in a statistically rigorous manor. We account for problems with normality that were observed in the first analysis. We conclude that there is no statistical difference in the leakage amounts for JP-8, Neat, and the 50/50 Blend. Therefore, since these fuel types span 3 different aromatic contents levels it does not appear that for these lower levels of aromatic content that there is a difference between the petroleum based JP-8 fuel and the biofuel blend or the pure biofuel. Additionally, we find that JP-5 has significantly lower leakage rates than JP-8. The reason for this difference is unknown based on the test results.

Data Description

Fuel was placed in a metal test cube with fuel bladder panel/backing board facing the gun. Panels were impacted by a fully tumbled 7.62-millimeter (mm) round and observed for 6 minutes. Amount of fuel leakage was recorded at regular intervals. Raw data from each fuel type is shown in Figure 1.

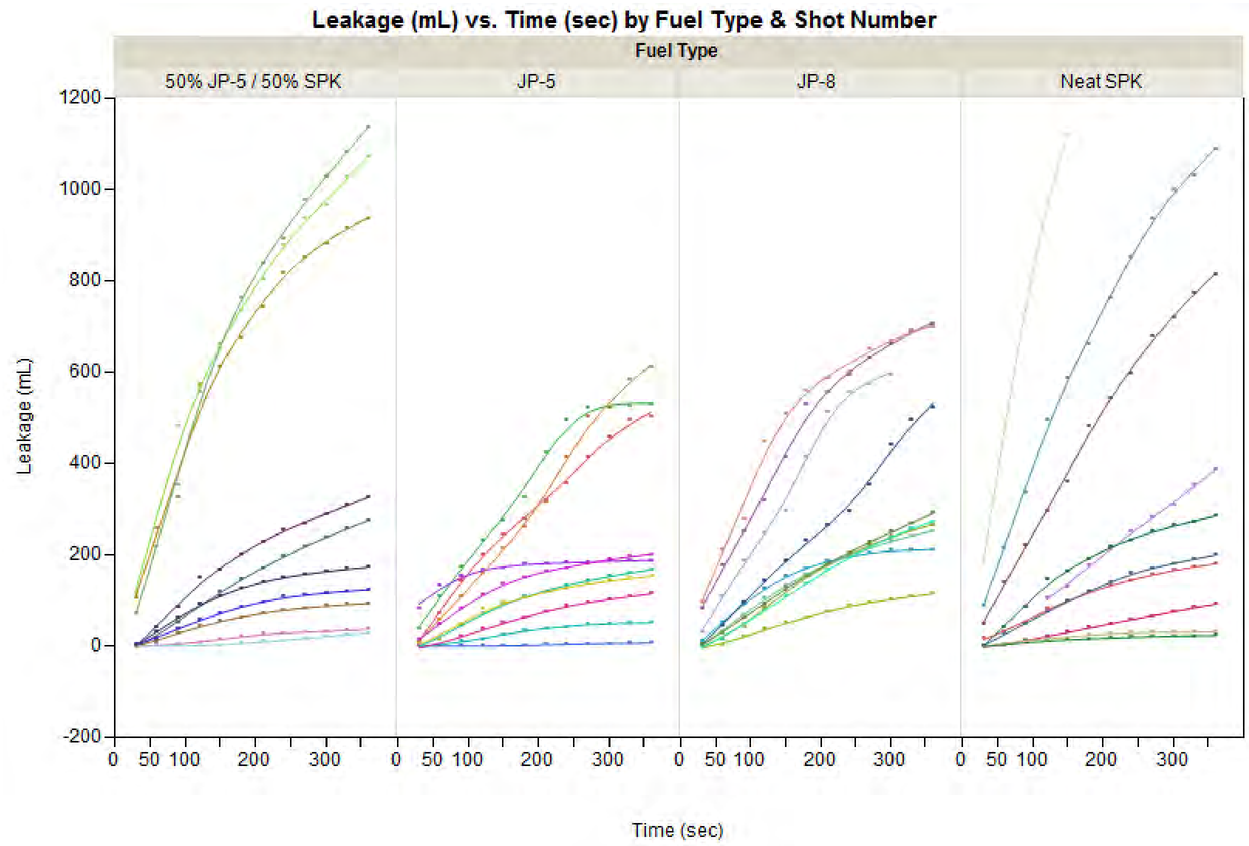


Figure 1. Leakage Amounts for Six Minutes by Fuel Type

Data Analysis

The data are correlated between the time increments because the total leakage at a time point always includes previous leakage. In this analysis we calculate the leakage amount within a given 30 second time bin to use as the primary response variable for two reasons: (1) to remove some of the correlation in the data; (2) it provides an easier understanding of the fuel leakage rate relative to the current time. Figure 2 provides the fuel leakage amounts within a given time increment. Figure 2 clearly shows that for most of the trials, there is some sealing effect within all the data. The leakage amounts tend to increase for a short period of time and then appear to level-off or decrease after that initial window.

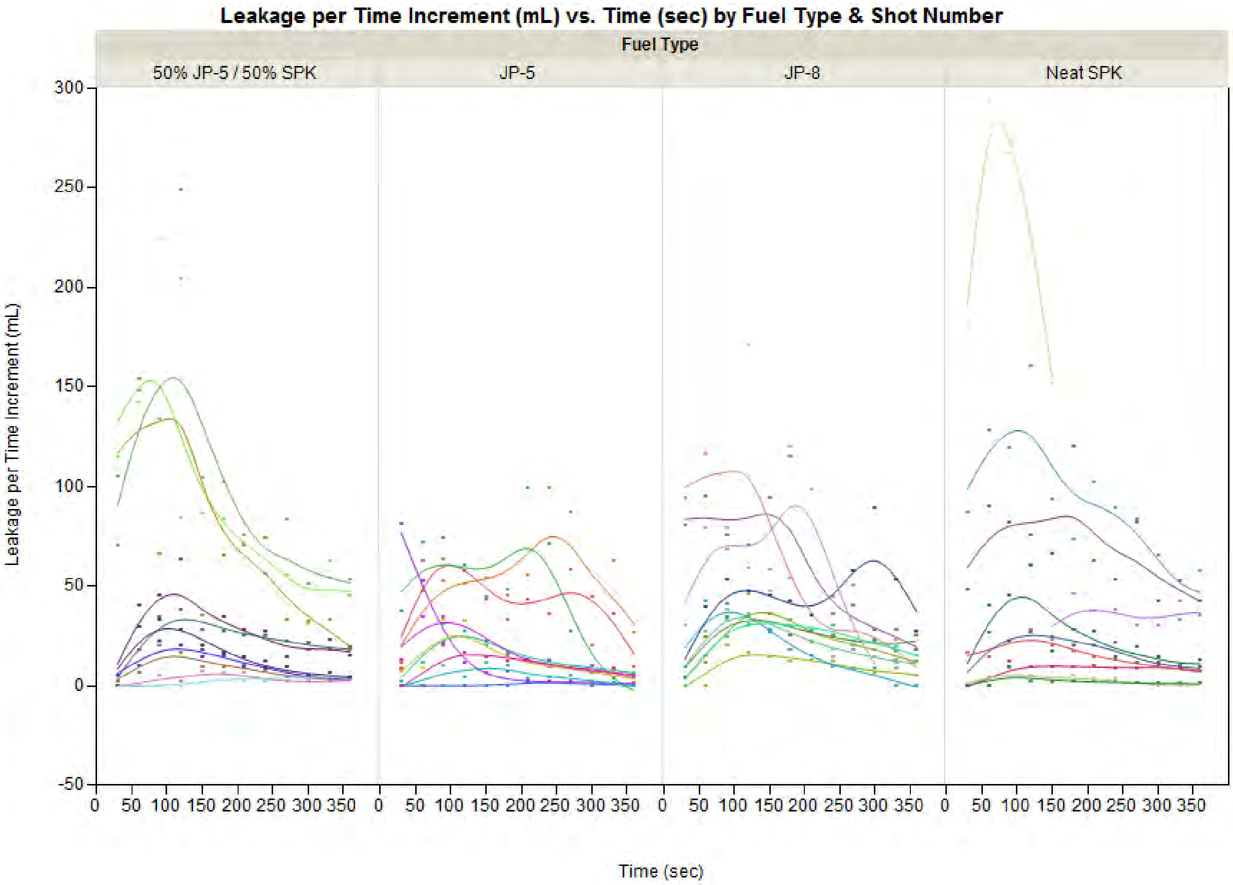


Figure 2. Leakage Amounts within a Time Increment by Fuel Type

Figure 3 shows the average leakage rates by fuel type per 30-second time increment. The leakage rate was calculated by subtracting the total leakage amount from the previous time period from the new leakage amount to get the leakage total for each 30-second time increment. This was done to reduce the amount of correlation between each time bin to improve the power of the statistical analysis. These leakage rates were plotted against time (the raw data points are dots in Figure 2), and then used cubic splines to fit a smooth trend line to the data for each Fuel Type. In Figure 3, one can see that all four fuel types follow a similar leakage pattern. Initially, we seen an increasing trend in the leakage rates, however, and after around 100-120 seconds all of the fuel types show some degree of sealing and leakage amounts begin to decrease.

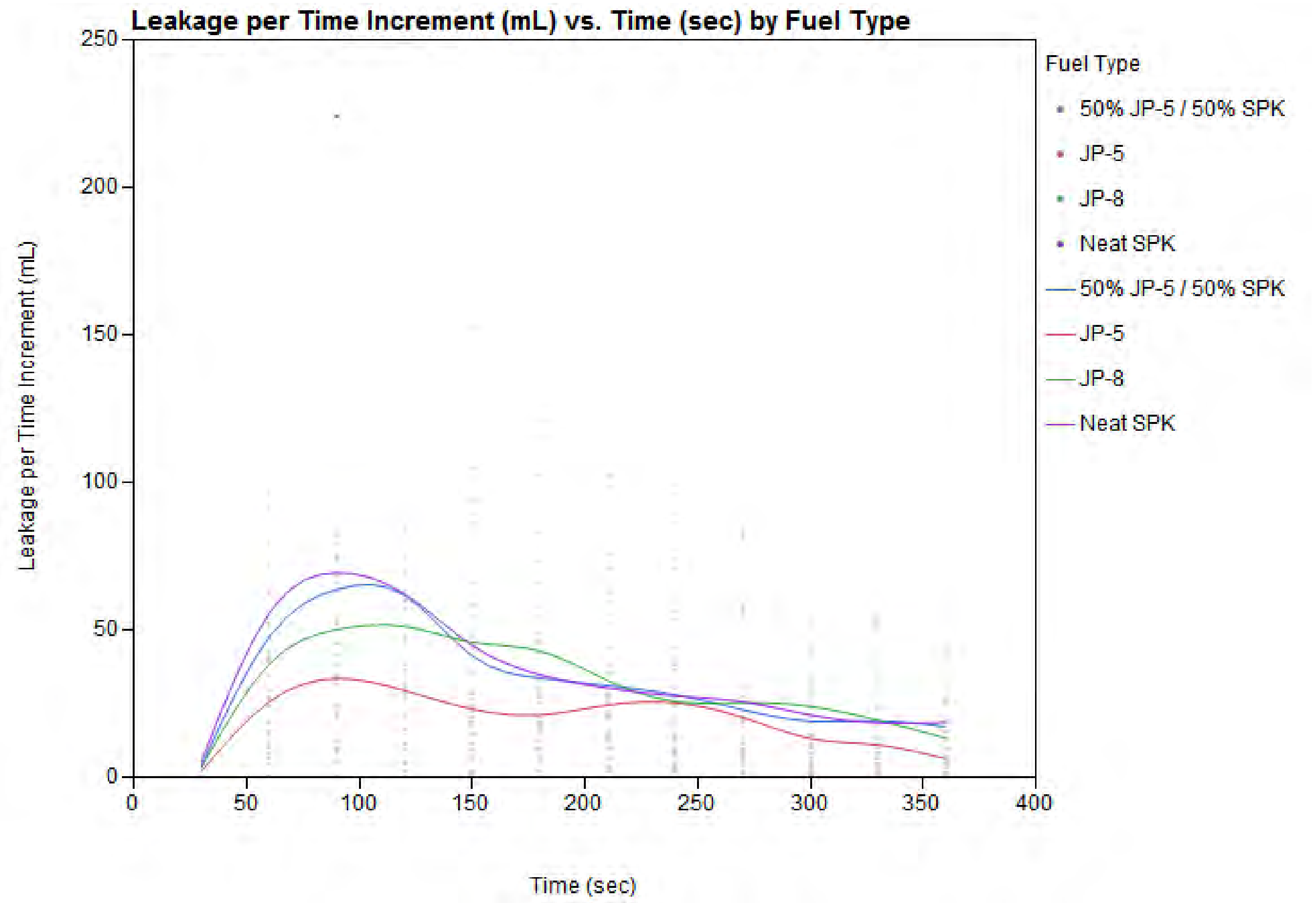


Figure 3. Smoothed Average Leakage Amount by Fuel Type

Statistical Data Analysis

To determine if the fuel type (and its corresponding aromatic content) significantly impacts the self-sealing properties, we use a linear mixed model. The mixed model allows for random effects that account for correlations in the dataset. Additionally, because the leakage rate is not normally-distributed we must transform the data. Figure 4 below shows the distribution of the leakage amounts per 30-second time interval for the raw data. Clearly, these data are not normal; they are highly right-skewed. Figure 5 shows the distribution of the data after a log transformation. It is reasonable to assume the normal distribution for this data because it is has a single peak and is close to symmetric.

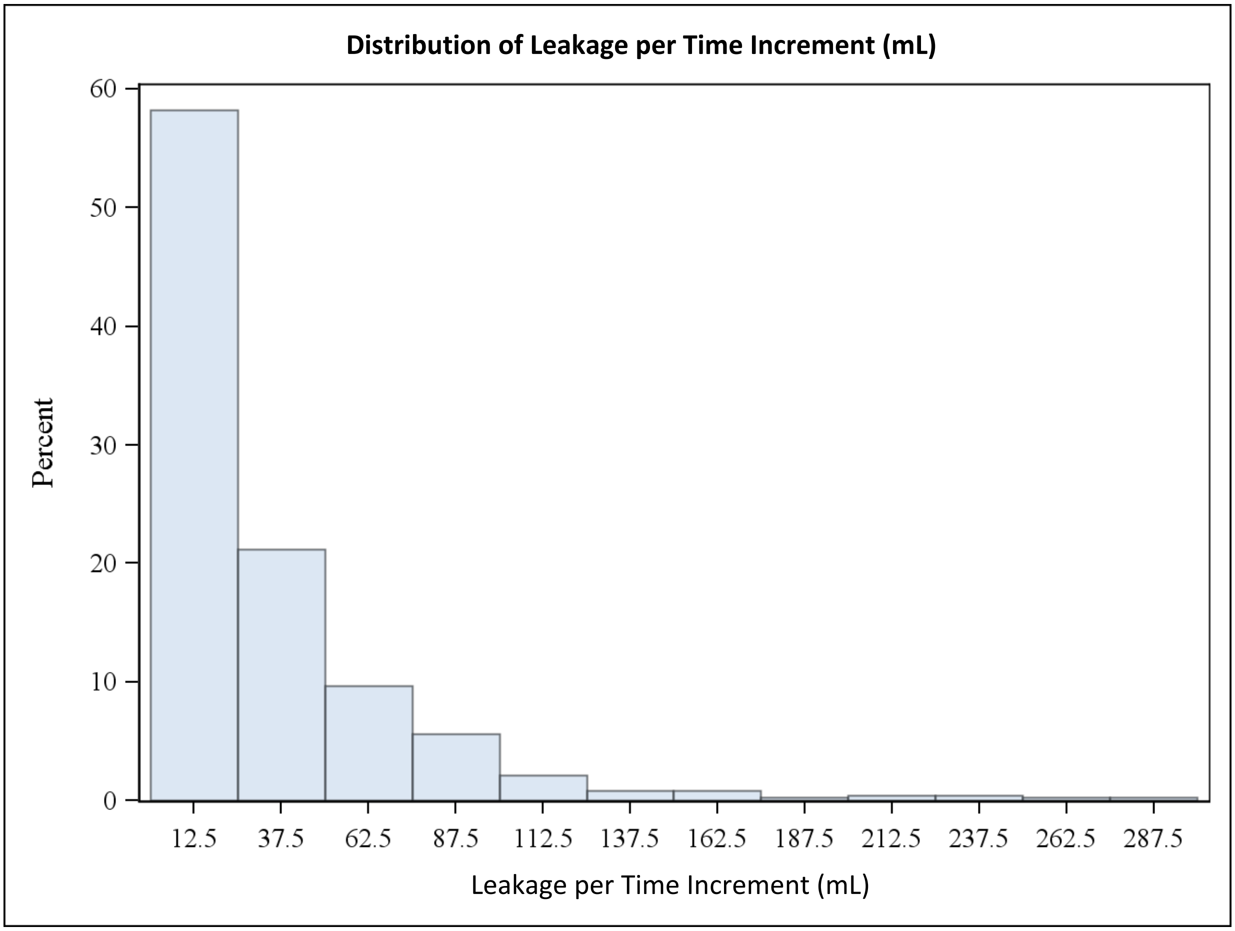


Figure 4. Histogram of Leakage Amounts

The linear mixed model used in the analysis also allows for the inclusion of additional factors that may influence fuel leakage amounts. We model the log leakage amounts as a function of time period, fuel type, and velocity. Additionally, to determine if the leakage amounts vary by fuel type as a function of time (i.e. sealing occurs faster for one fuel type than another) we include the interaction term between fuel type and time period.

Table 2 below shows the least squares estimates of the mean log leakage amounts by fuel type and time. Recall, that all of these values have been transformed to be on the log scale so to get the actual mean leakage amounts one needs to exponentiate the values in Table 1. Figure 6 plots the actual least squares estimates of leakage rates (not log transformed) by time increment and fuel type.

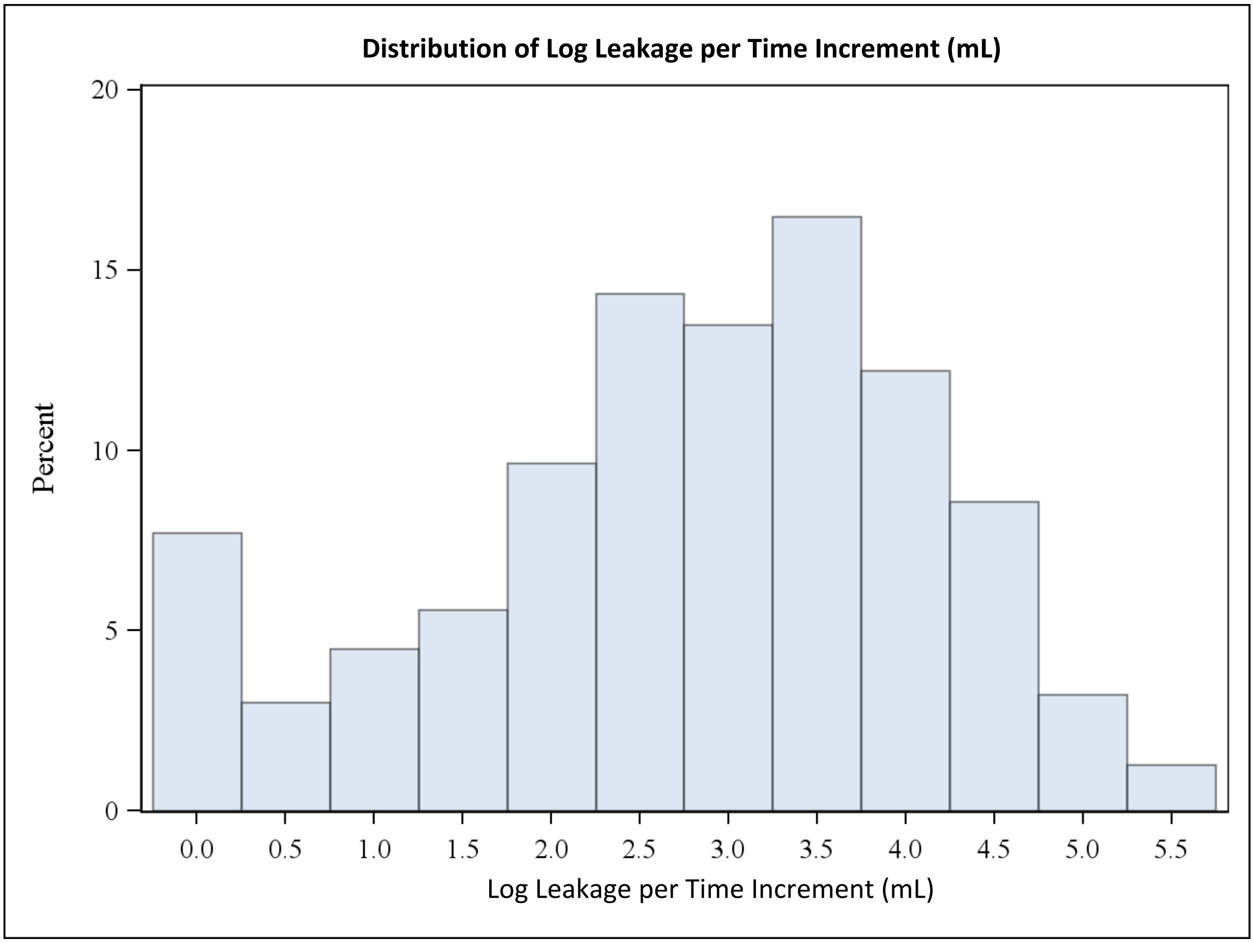


Figure 5. Histogram of Log Transformed Leakage Amounts

Table 2. Least Square Estimates of Mean Leakage Rate per Fuel Type and Time Increment

Fuel Type	Time (sec) (Standard Error)											
	30	60	90	120	150	180	210	240	270	300	330	360
50/50	1.66 (.415)	2.95 (.415)	3.35 (.415)	3.35 (.415)	2.99 (.415)	3.20 (.415)	3.00 (.415)	2.84 (.415)	2.65 (.415)	2.47 (.415)	2.55 (.415)	2.40 (.415)
JP-5	1.83 (.418)	2.43 (.418)	3.13 (.418)	2.92 (.418)	2.74 (.418)	2.61 (.418)	2.63 (.418)	2.60 (.418)	2.44 (.418)	1.82 (.418)	1.94 (.418)	1.42 (.418)
JP-8	2.11 (.416)	3.31 (.416)	3.78 (.416)	3.83 (.416)	3.65 (.416)	3.65 (.416)	3.33 (.416)	3.14 (.416)	3.15 (.416)	3.01 (.416)	2.72 (.416)	2.42 (.416)
Neat	1.89 (.421)	3.00 (.421)	3.62 (.421)	3.49 (.421)	3.02 (.421)	3.30 (.421)	3.18 (.421)	3.05 (.421)	2.89 (.421)	2.70 (.421)	2.57 (.421)	2.68 (.421)

The highlighted cells in Table 2 indicate that there was a significant difference between that cell and another cell within the same time step. Table 3 below summarizes all of the significant difference between the cells. Notice, all of the pair-wise significant differences contain JP-5. Therefore, this analysis shows that JP-5 does exhibit different leakage amounts from the other fuels indicated. Additionally, there is no statistically distinguishable difference between the 50/50 blend, JP-8, and Neat SPK.

Table 3. Significant Pair-wise Differences between Fuel Types at a given Time Increment

Fuel Type 1	Fuel Type 2	Time (sec)	Estimated Difference	Standard Error	t Value	p-value
JP-5	JP-8	60	-0.8822	0.5944	-1.48	0.1386*
JP-5	JP-8	120	-0.9118	0.5944	-1.53	0.1259*
JP-5	JP-8	150	-0.9034	0.5944	-1.52	0.1294*
JP-5	JP-8	180	-1.0379	0.5944	-1.75	0.0816**
JP-5	JP-8	300	-1.1906	0.5944	-2	0.0459***
JP-5	Neat	300	-0.8766	0.5926	-1.48	0.1399*
JP-5	JP-8	330	-0.7802	0.5987	-1.3	0.1933*
JP-5	Neat	360	-1.2589	0.5926	-2.12	0.0343***
JP-5	JP-8	360	-1.0008	0.5987	-1.67	0.0955*
JP-5	50%	360	-0.9784	0.5909	1.66	0.0986*

[a] Significant at the 80% Confidence Level

[b] Significant at the 90% Confidence Level

[c] Significant at the 95% Confidence Level

Table 4 provides an overall summary of the differences between fuels if we look at the differences averaged over all of the time points. Overall, the only significant difference between fuel types across all time points is JP-5 results in significantly lower leakage amounts than JP-8.

Table 4. Overall Differences between Fuel Types

Fuel Type 1	Fuel Type 2	Estimated Difference	Standard Error	t Value	p-value
50%	JP-5	0.4091	0.5212	0.78	0.4331
50%	JP-8	-0.3900	0.5156	-0.76	0.4498
50%	Neat	-0.1642	0.5163	-0.32	0.7507
JP-5	JP-8	-0.7991	0.5253	-1.52	0.1290*
JP-5	Neat	-0.5732	0.5185	-1.11	0.2696
JP-8	Neat	0.2259	0.5180	0.44	0.6631

* Significant at the 80% Confidence Level

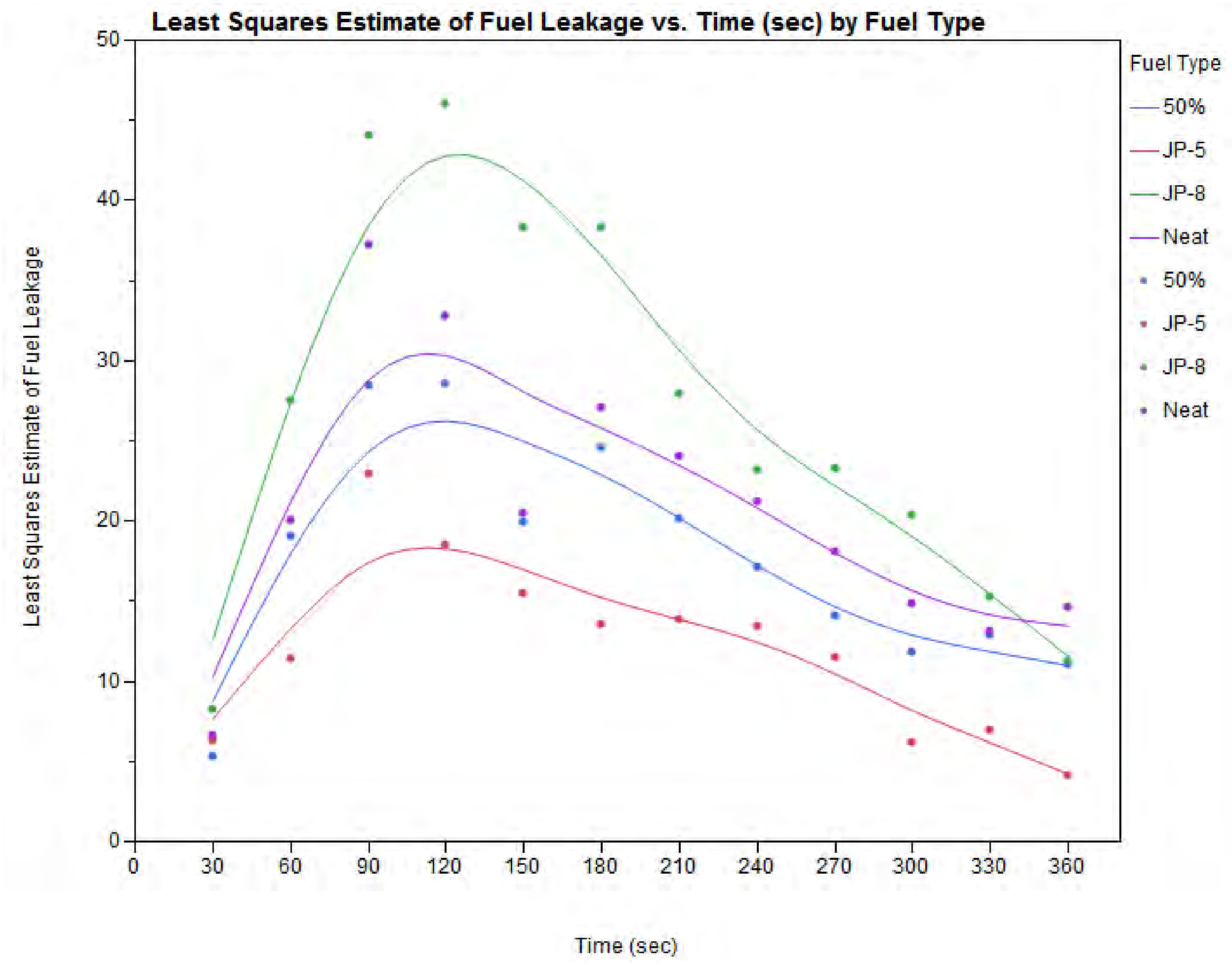
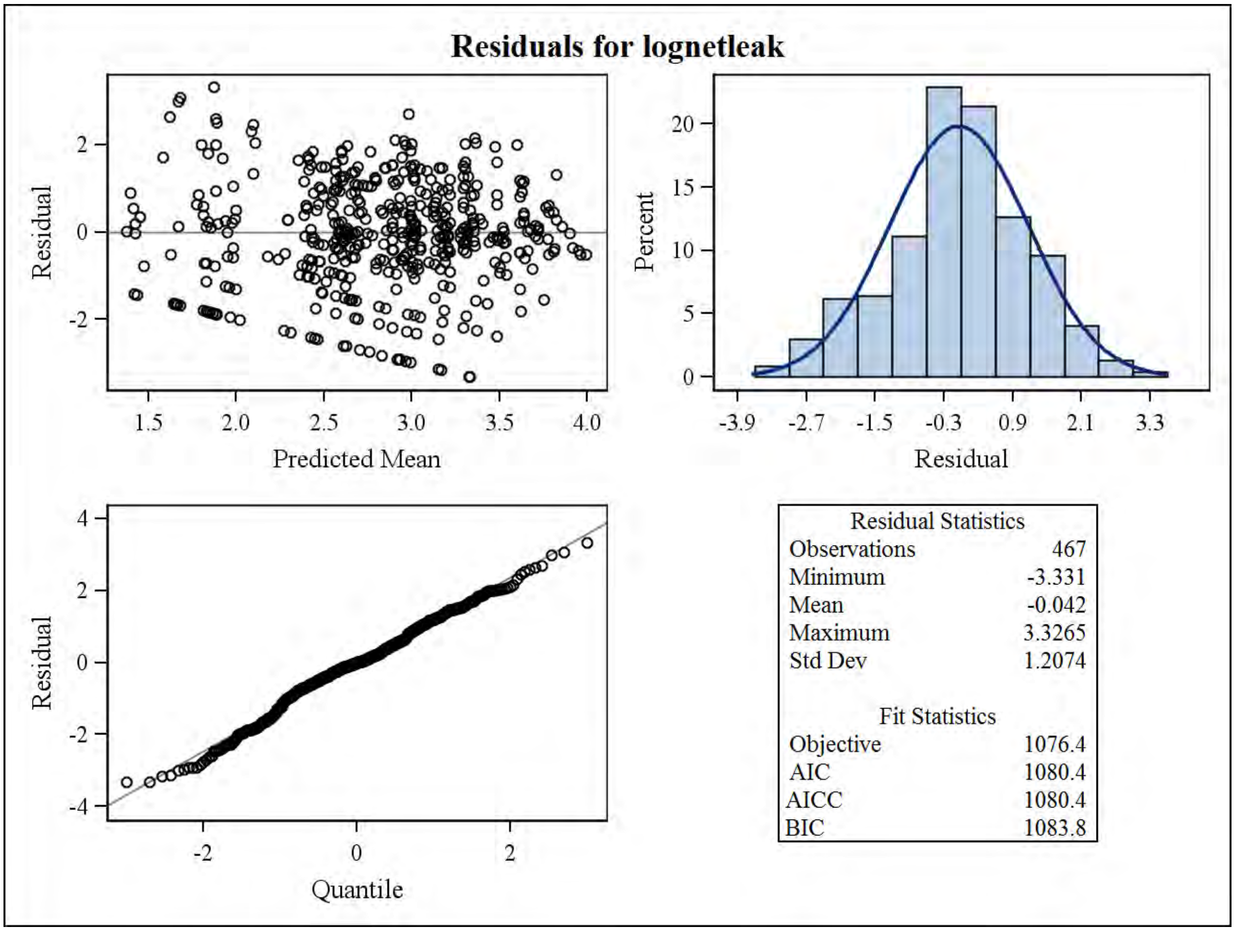


Figure 6. Least Square Estimates for Fuel Leakage amounts by Fuel Type

Additionally in Figure 6, it is interesting to note that the maximum leakage amount for all fuel types occurs at either 90 or 120 seconds, indicating that sealing is occurring after about 2 minutes across all fuel types. Additionally, it is important to notice that the amount of fuel leaking from the cube does not appear to be a function directly of aromatic content. JP-8, which has the second highest aromatic levels (11.5 percent), has the highest amount of fuel leaked in this experiment. This graph illustrates that the linear extrapolation method used in the second pervious analysis was not valid, at least for the given data set.

Figure 7 provides an analysis of the model assumptions by checking the distribution of the residuals from the model. The linear mixed model assumes normality and that the variance between observations can be properly accounted for by random effects. The residual scatter plot below shows that there are no trends in the residuals as a function of the mean predicted value. The histogram and the residual versus quantile plots show that the residuals follow an approximately normal distribution. Therefore, the assumptions have been met to use this model for statistical inference.



Residual Statistics	
Observations	467
Minimum	-3.331
Mean	-0.042
Maximum	3.3265
Std Dev	1.2074
Fit Statistics	
Objective	1076.4
AIC	1080.4
AICC	1080.4
BIC	1083.8

Figure 7. Residual Plots for Linear Mixed Model

Conclusions

The analysis provided in this document supports the conclusions of the first analysis conducted using standard t-tests. There is no statistically significant difference between JP-8 and the 50/50 biofuel blend. Additionally, it expands on that analysis to show that there are no statistical differences in leakage amounts between JP-8, the 50/50 Blend, and the Neat Fuel. JP-5 is statistically different from JP-8 across the six minute observation period, but there is insufficient evidence to conclude it is different from the 50/50 blend or the Neat biofuel overall.

A factor that could not be considered in this analysis is the degree of damage that occurred in each live fire shot. The amount of damage, as indicated by the previously conducted analyses, is causing more variability in the fuel leakage amounts than the fuel type. Figure 8 below illustrates this point by plotting the leakage amount as a function of the classified leak type. In the presence of such a highly variable factor, to detect differences in the fuels ability to seal leak types one would need a much larger experiment. However, there may be no operationally meaningful reason to conduct such an experiment because the impact of the fuel type on leakage sealing from the current analysis appear to be small.

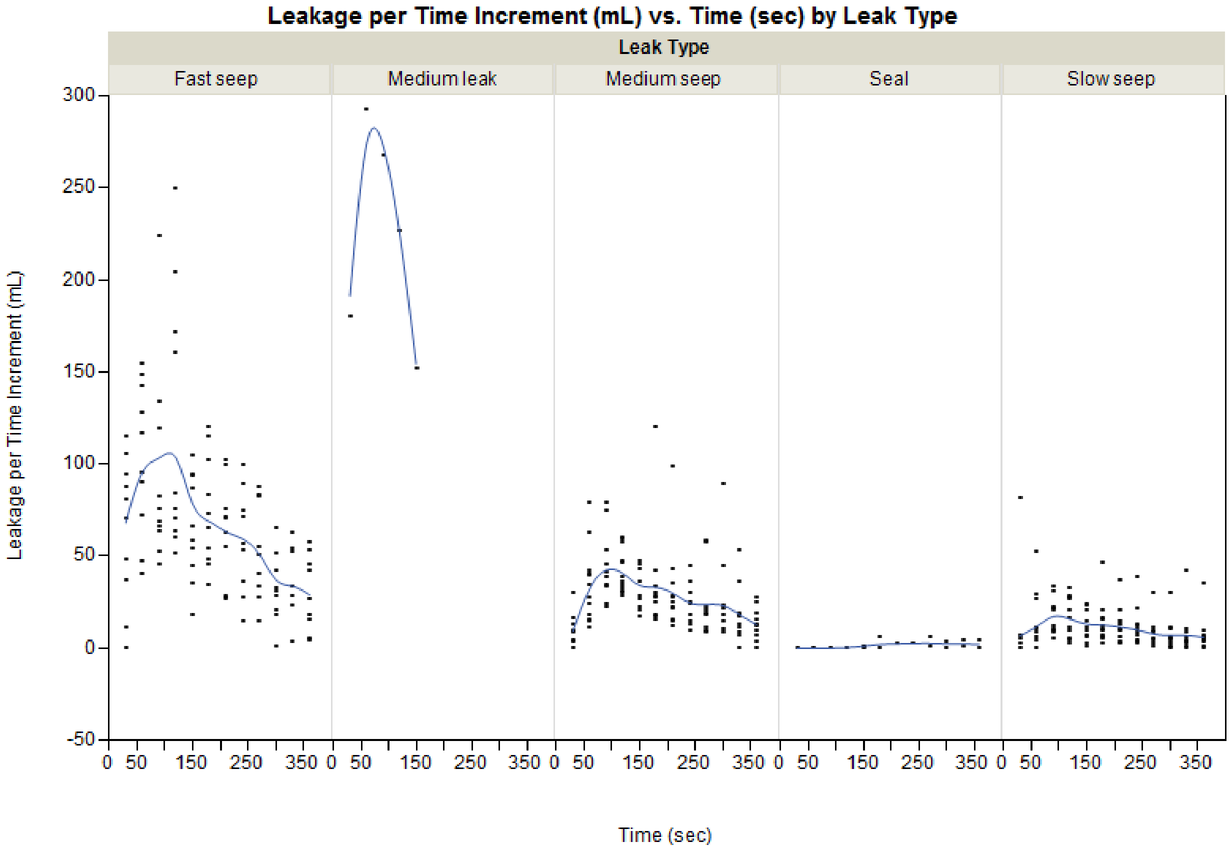


Figure 8. Fuel Leakage by Leak Type