Implementing Design and Analysis of Experiments in the U.S. Department of Defense Testing Community

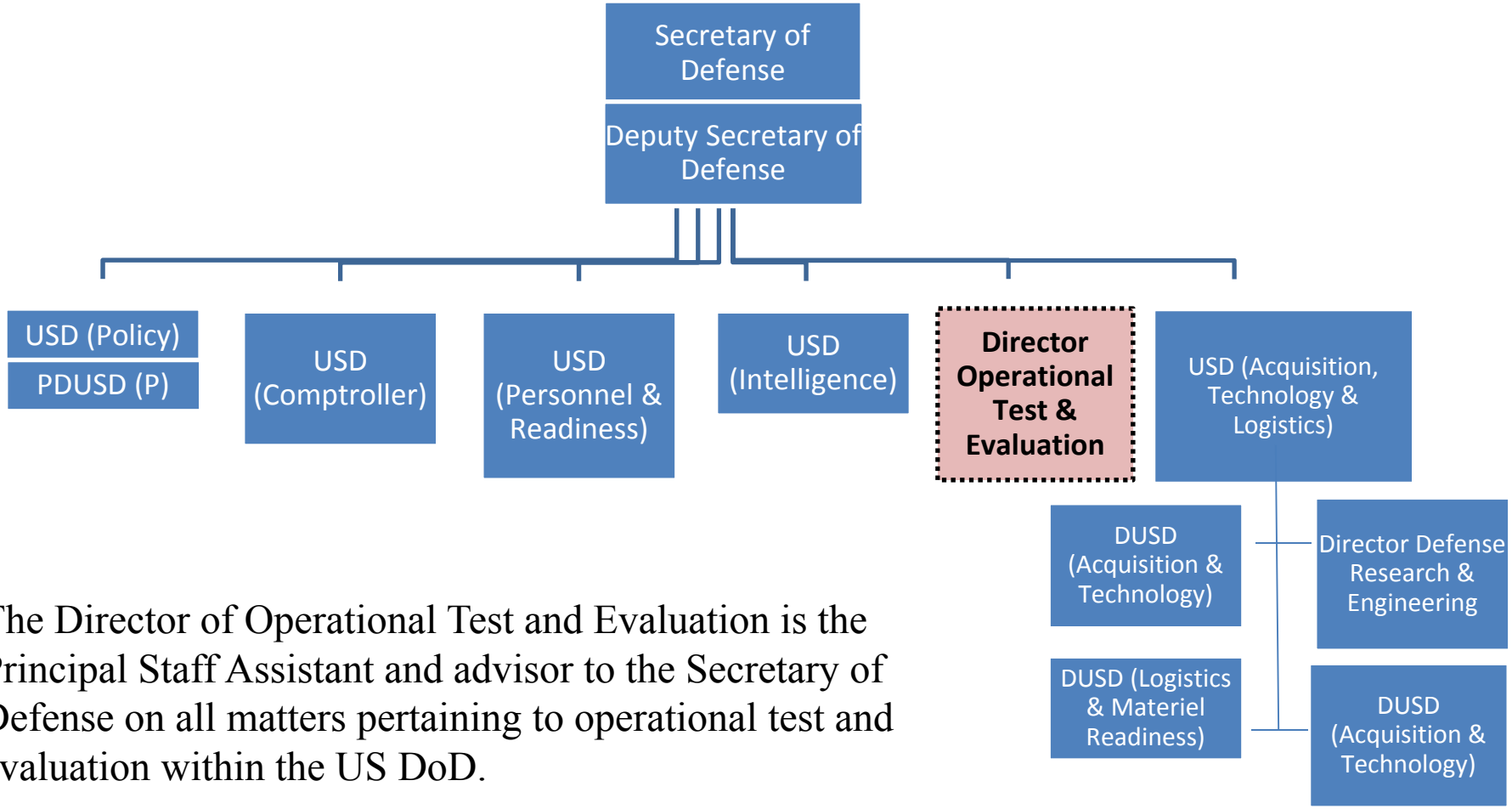
European Network for Business and Industrial Statistics
10 September 2012

Laura J. Freeman
Research Staff Member
lfreeman@ida.org

Outline

- **Overview of DoD Test and Evaluation**
  - Director, Operational Test & Evaluation
- **Design of Experiments in Test and Evaluation**
- **Statistical Analysis to inform Effectiveness and Suitability Decisions**
- **Challenges: Cultural, Educational, and Statistical**
- **Conclusions**

Office of the Secretary of Defense



The Director of Operational Test and Evaluation is the Principal Staff Assistant and advisor to the Secretary of Defense on all matters pertaining to operational test and evaluation within the US DoD.

DOT&E Background

- **DOT&E was created by Congress in 1983.**
- **Director is appointed by the President and confirmed by the Senate.**
- **Director's reports, by statute, go directly to the Secretary of Defense and Congress**
- **Responsible for all operational test and evaluation, and live fire test and evaluation within DoD.**
- **Provides independent oversight and reporting.**

Operational Test Mission

- **"Operational test and evaluation means --**
  - the field test, under realistic combat conditions, of any item of (or key component of) weapons, equipment, or munitions for use in combat by typical military users; and the evaluation of the results of such test." – 10 USC Section 139

- **Focus:**
  - Is the OT&E and/or LFT&E adequate?
  - Is the system operationally effective?
  - Is the system operationally suitable?
  - Is the system survivable and lethal?

- **Operational testing is about assessing the mission!**
  - Systems do not have missions, units equipped with systems have missions.
  - Effectiveness: can a unit equipped with the system under test accomplish the mission?
  - Suitability: can the system be used in the operational environment by the user to accomplish the mission?

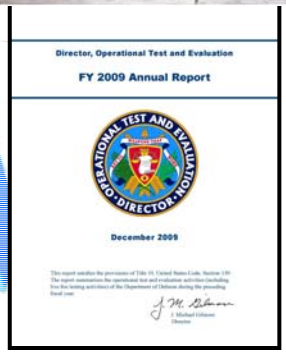
- **End-to-End mission oriented responses are essential for determining system effectiveness.**

- **Testing across the entire operational envelop is necessary for effectiveness and suitability.**

DOT&E Interactions

DOT&E Tools:

1. Test and Evaluation Master Plan approval
2. Test plan and Test Strategy approval
3. Beyond Low Rate Initial Production Reports
4. Early Fielding reports
5. Annual Report
6. Director's Memo, Testimony, Speeches
7. Close cooperation with Service Test Agencies



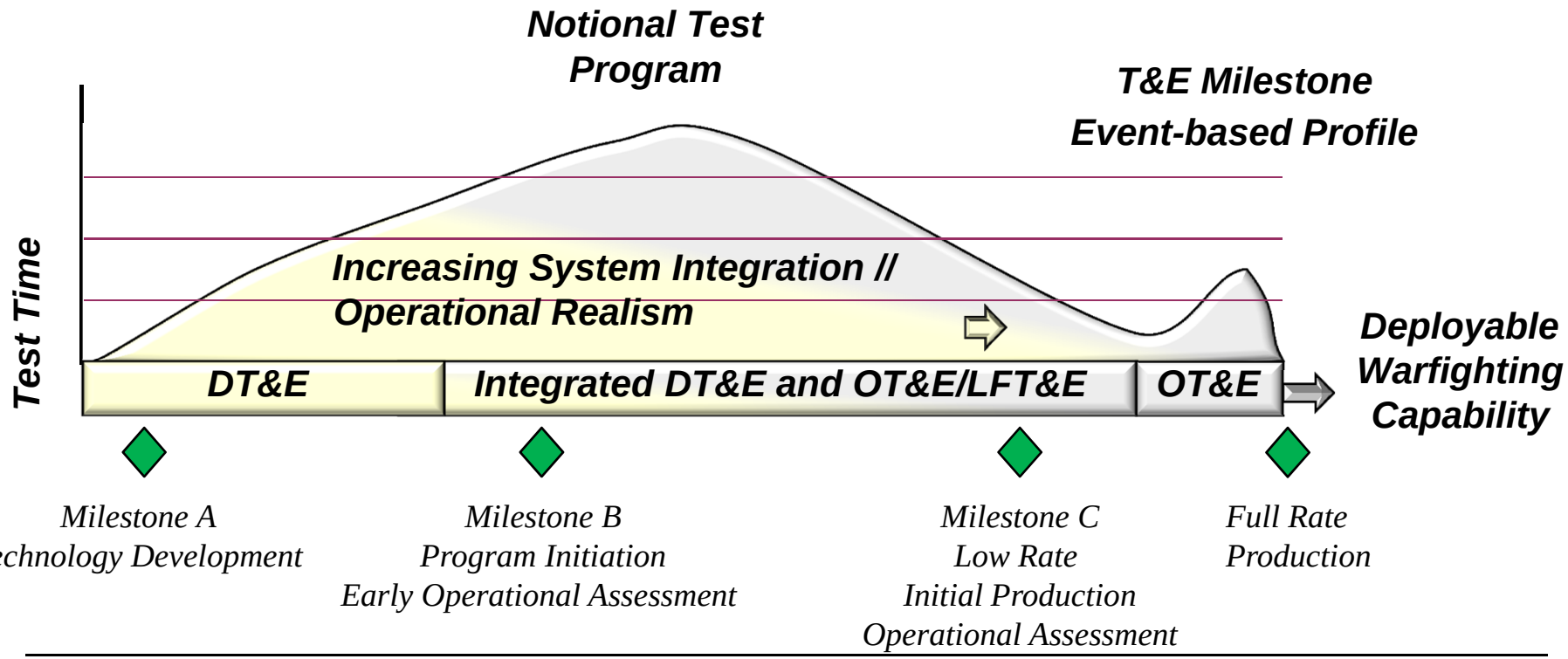
Guidance

BLRIPs, annual reports, testimony, etc.

Guidance and consultation

Acquisition Timeline

- **Operational Testing supports full rate production decision**
- **Report on programs, before full-rate production decision:**
  - Test adequacy, Operational Effectiveness, Suitability, Survivability and Lethality

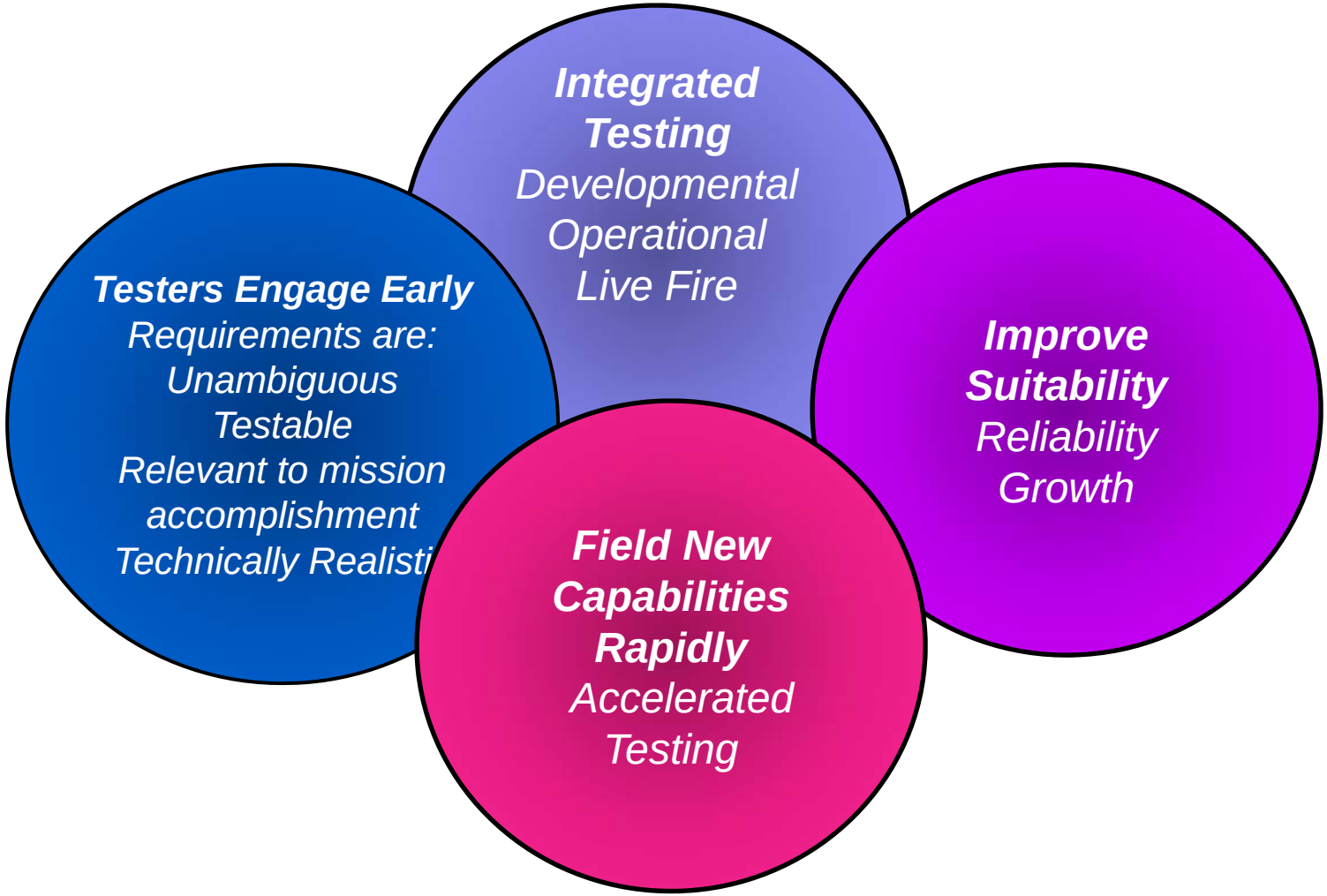


National Research Council Study (1998)
Panel on Statistical Methods for T&E of Defense Systems

- **Conclusions**
  - Major advances can be realized by applying selected industrial principles and practices in restructuring the paradigm for operational testing…
  - The current practice of statistics in defense testing design and evaluation does not take full advantage of the benefits available from the use of state-of-the-art statistical methodology.
- **Recommendations**
  - All estimates of the performance of a system from operational test should be accompanied by statements of uncertainty through use of confidence intervals…
  - The service test agencies should examine the applicability of state-of-the-art experimental design techniques and principles and, as appropriate, make greater use of them in the design of operational tests.
  - Operational test agencies should promote more critical attention to the specification of statistical models of equipment reliability, availability, and maintainability and to supporting the underlying assumptions…

The majority of the recommendations have not been implemented 13 years later

DOT&E Initiatives



A Brief and Selective History of DOE in T&E

- **National Research Council Study (1998)**
  - "The current practice of statistics in defense testing design and evaluation does not take full advantage of the benefits available from the use of state-of-the-art statistical methodology."
  - "The service test agencies should examine the applicability of state-of-the-art experimental design techniques and principles…"

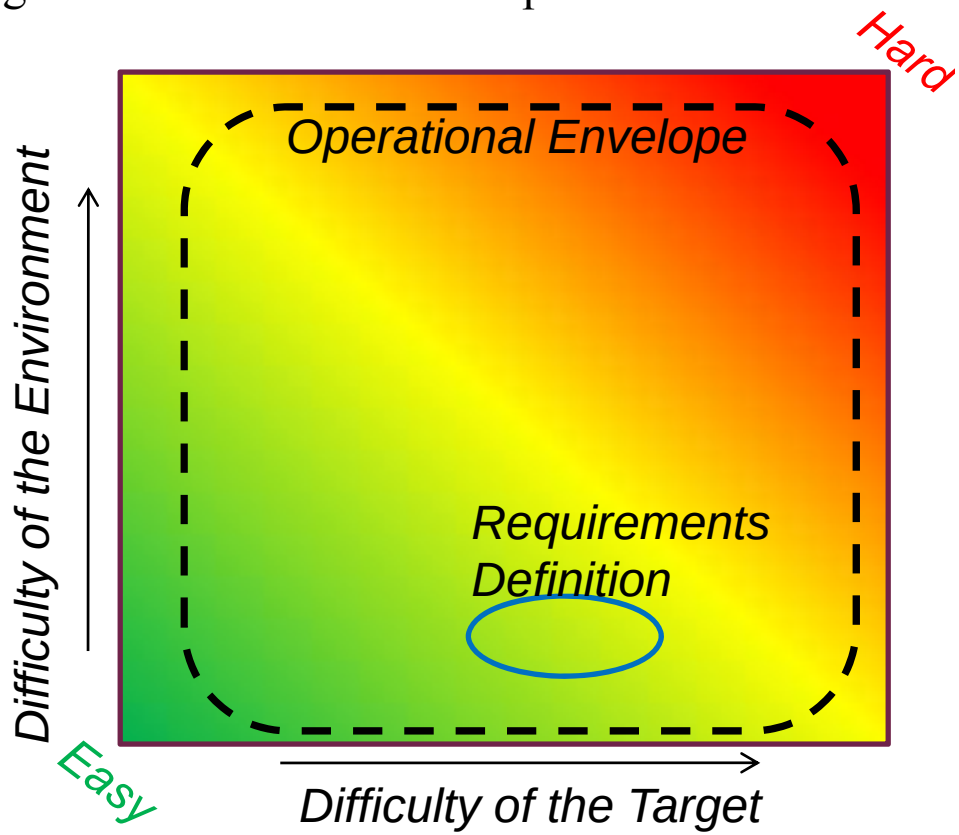
- **Operational Test Agency Memorandum of Agreement (2009)**
  - "This group endorses the use of DOE as a discipline to improve the planning, execution, analysis, and reporting of integrated testing."

- **DOT&E Initiatives (2009)**
  - "The DT&E and OT&E offices are working with the OTAs and Developmental Test Centers to **apply DOE across** the **whole development and operational test cycle** for a program."
  - "Whenever possible, our evaluation of performance must include a rigorous **assessment of** the **confidence level of the test**, the **power of the test** and some measure of how well the **test spans the operational envelope** of the system."

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

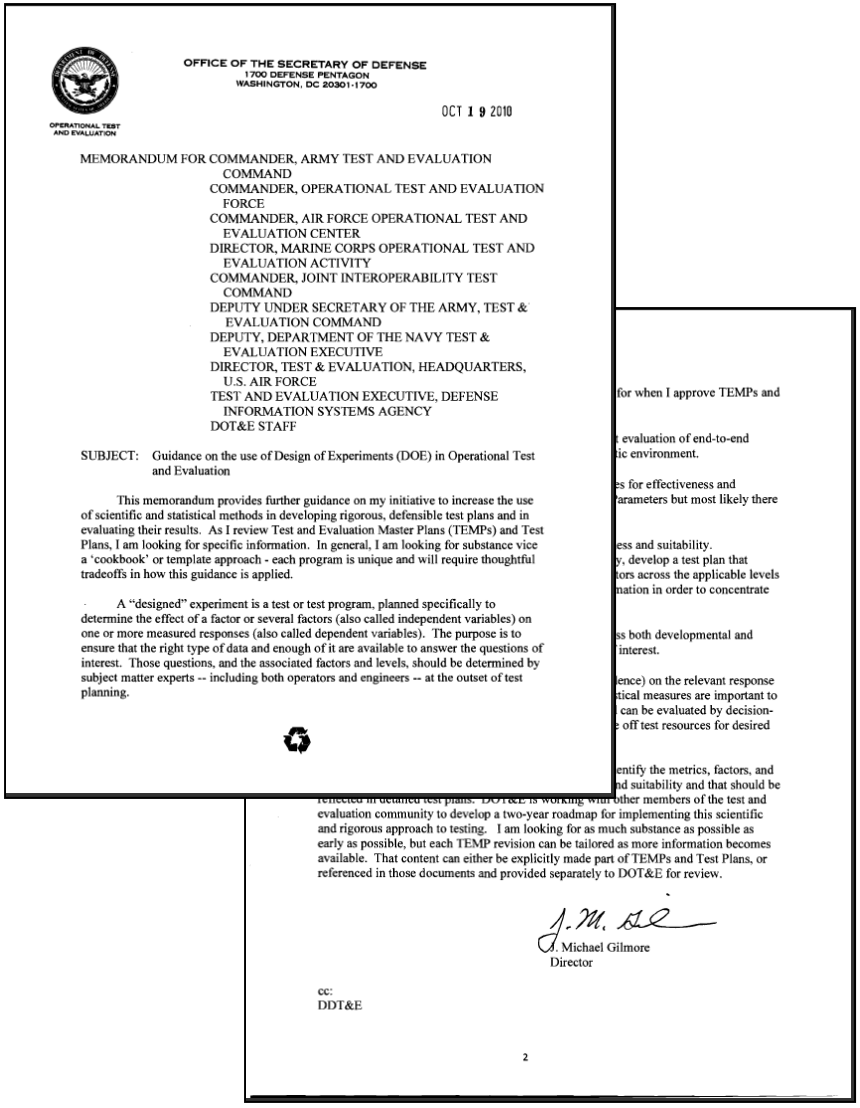
- **DOE Provides:**
  - a scientific, structured, objective way to span the operational envelope
  - the most powerful allocation of test resources for a given number of tests.
  - an approach to integrated test.
  - a structured analysis for summarizing test results



DOE changes "I think" to "I know"

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 19 2010

MEMORANDUM FOR COMMANDER, ARMY TEST AND EVALUATION COMMAND
COMMANDER, OPERATIONAL TEST AND EVALUATION FORCE
COMMANDER, AIR FORCE OPERATIONAL TEST AND EVALUATION CENTER
DIRECTOR, MARINE CORPS OPERATIONAL TEST AND EVALUATION ACTIVITY
COMMANDER, JOINT INTEROPERABILITY TEST COMMAND
DEPUTY UNDER SECRETARY OF THE ARMY, TEST & EVALUATION COMMAND
DEPUTY, DEPARTMENT OF THE NAVY TEST & EVALUATION EXECUTIVE
DIRECTOR, TEST & EVALUATION, HEADQUARTERS, U.S. AIR FORCE
TEST AND EVALUATION EXECUTIVE, DEFENSE INFORMATION SYSTEMS AGENCY
DOT&E STAFF

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

This memorandum provides further guidance on my initiative to increase the use of scientific and statistical methods in developing rigorous, defensible test plans and in evaluating their results. As I review Test and Evaluation Master Plans (TEMPs) and Test Plans, I am looking for specific information. In general, I am looking for substance vice a 'cookbook' or template approach - each program is unique and will require thoughtful tradeoffs in how this guidance is applied.

A "designed" experiment is a test or test program, planned specifically to determine the effect of a factor or several factors (also called independent variables) on one or more measured responses (also called dependent variables). The purpose is to ensure that the right type of data and enough of it are available to answer the questions of interest. Those questions, and the associated factors and levels, should be determined by subject matter experts -- including both operators and engineers -- at the outset of test planning.

for when I approve TEMPs and

evaluation of end-to-end
ic environment.

s for effectiveness and
arameters but most likely there

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

ss both developmental and
interest.

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

entify the metrics, factors, and
nd suitability and that should be
reflected in detailed test plans. DOT&E is working with other members of the test and evaluation community to develop a two-year roadmap for implementing this scientific and rigorous approach to testing. I am looking for as much substance as possible as early as possible, but each TEMP revision can be tailored as more information becomes available. That content can either be explicitly made part of TEMPs and Test Plans, or referenced in those documents and provided separately to DOT&E for review.

J. Michael Gilmore
Director

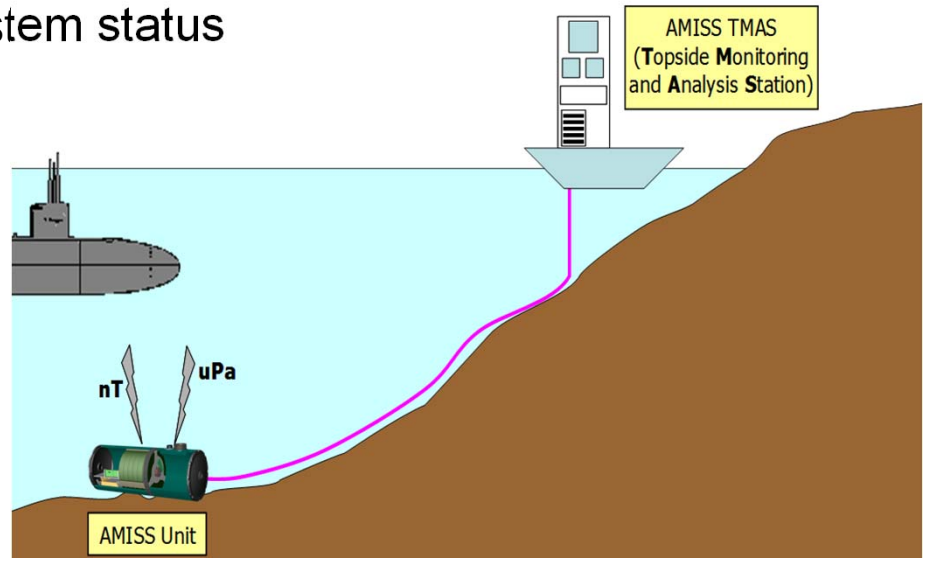
cc:
DDT&E

2

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

Example: Adequate Test Plans for Mine Susceptibility

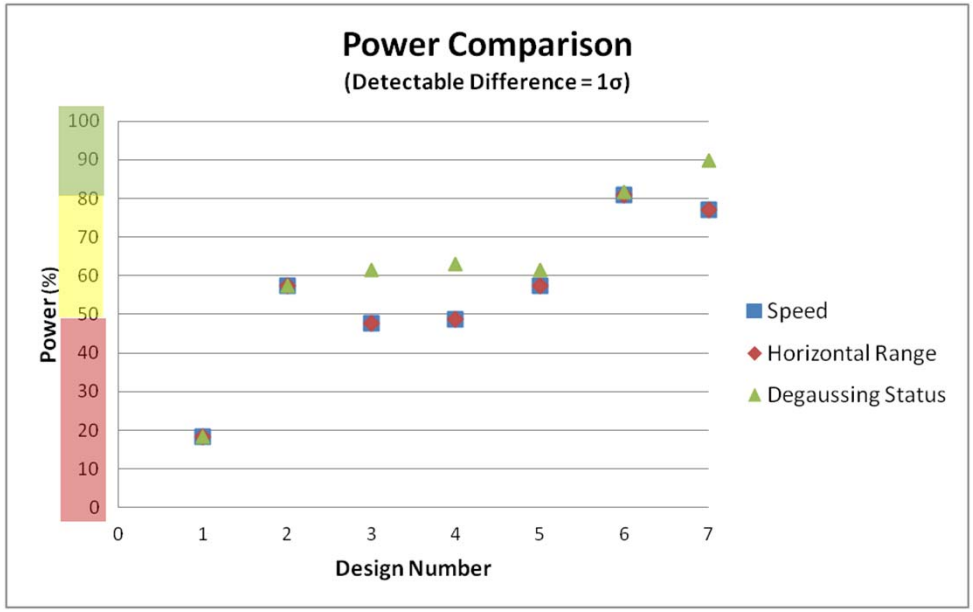
- **Goal:**
  - Develop an adequate test to assess the susceptibility of a cargo ship against a variety of mine types using the Advanced Mine Simulation System (AMISS).
- **Responses:**
  - Magnetic signature, acoustic signature, pressure
- **Factors:**
  - Speed, range, degaussing system status



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**

	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



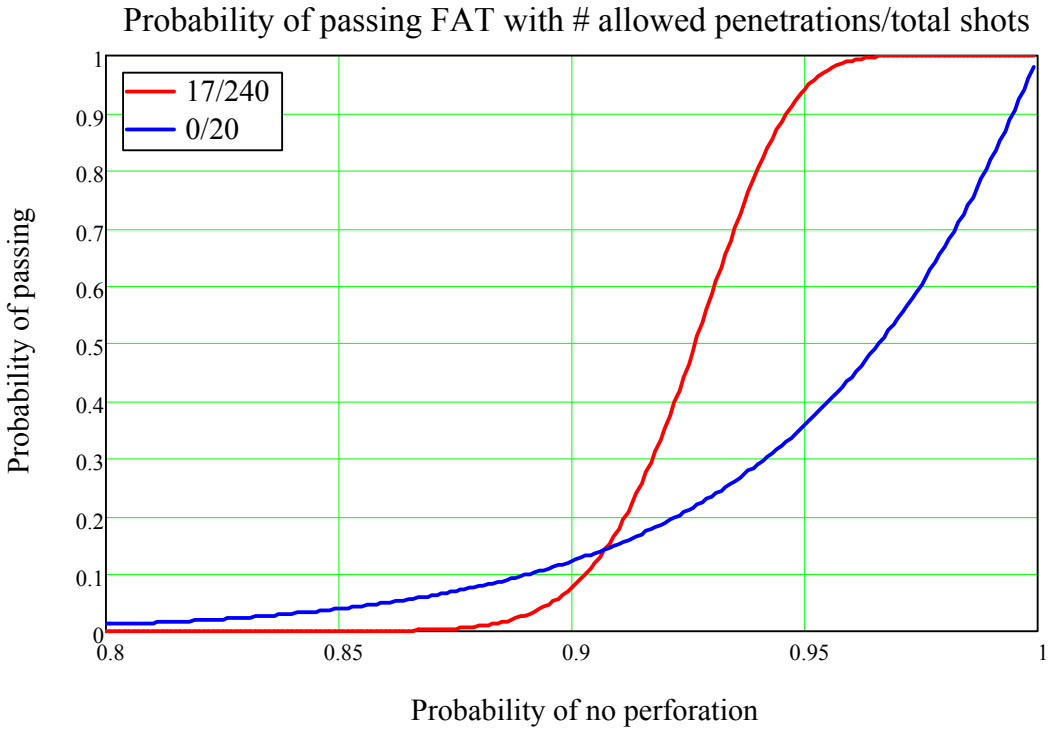
Example: Combat Helmets
First Article Test (FAT) and Lot Acceptance Testing (LAT)

- **Develop adequate test protocols**
  - First article testing (FAT)
  - Lot acceptance testing (LAT)

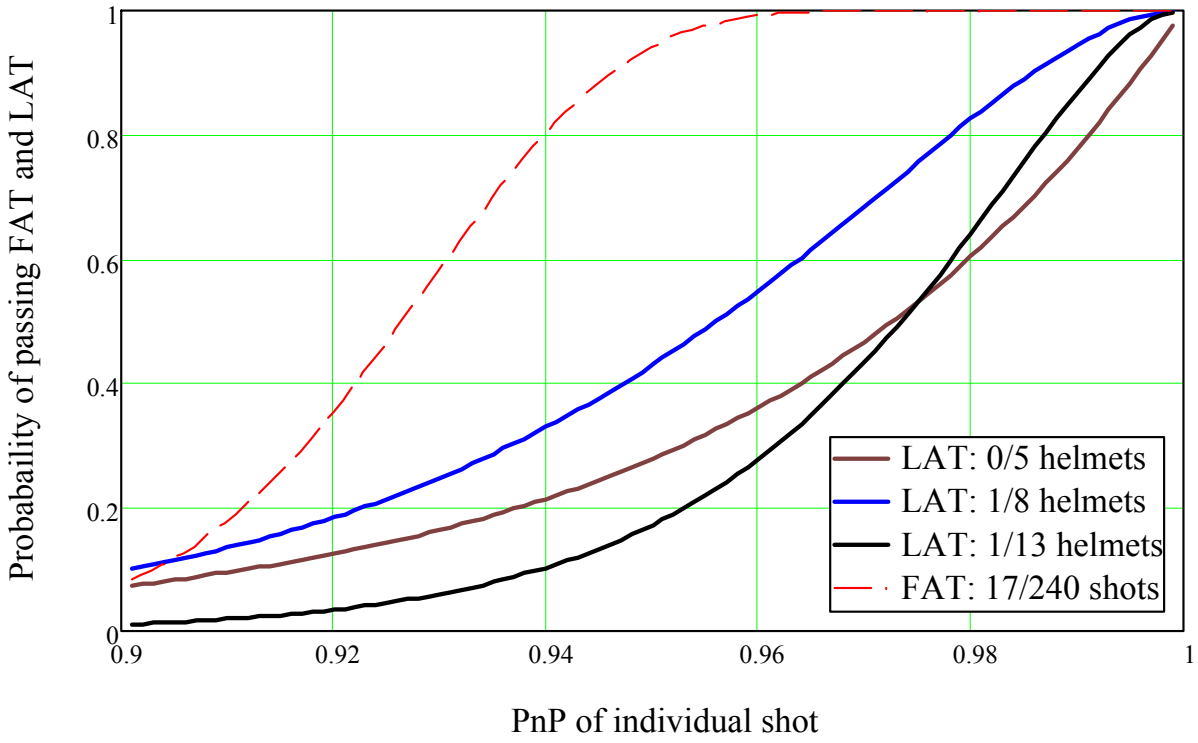
- **Design of Experiments and Operating Characteristic Curves**

Size	Ambient		Hot		Cold		Seawater	
	Shot 1	Shot 2	Shot 1	Shot 2	Shot 1	Shot 2	Shot 1	Shot 2
Small	Cr	B	L	F	Cr	B	L	F
	L	F	Cr	B	R	F	Cr	B
	Cr	B	R	F	Cr	B	R	F
Medium	R	F	Cr	B	R	F	Cr	B
	Cr	B	R	F	Cr	B	L	F
	L	F	Cr	B	L	F	Cr	B
Large	Cr	B	R	F	Cr	B	R	F
	R	F	Cr	B	L	F	Cr	B
	Cr	B	L	F	Cr	B	L	F
X-Large	L	F	Cr	B	L	F	Cr	B
	Cr	B	L	F	Cr	B	R	F
	R	F	Cr	B	R	F	Cr	B

F: Front B: Back Cr: Crown



Comparing FAT and LAT For Combat Helmets



Probability of passing FAT is based on PnP of individual shots: max 17 out of 240 is allowed; probability of passing a LAT is based on a number of failed helmets, which is computed [based on the PnP] and used for calculating the probability of passing three different LATS

Note that the LATs fail a substantial amount of helmets that passed FAT.

Example: Chemical Agent Detector

- **Goal: Determine the probability of detection within one minute**
  - Threshold is least 85% 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)**

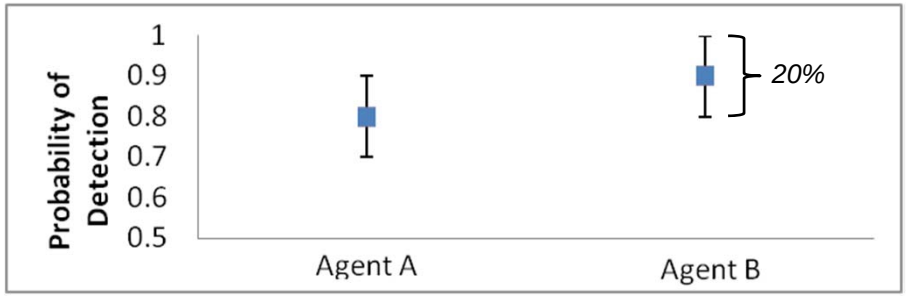
DOE Matrix

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?

Adequate Test Resources

- **Goal: Determine an adequate sample size to determine a 10% change in probability of detection across all factor levels (across the operational envelope)?**



- **Steps**
  - Determine detectable difference for binary response (10%)
  - Calculate sample size for binary response variable
  - Determine the appropriate continuous response (detection time)
  - Determine equivalent effect size of interest using percentiles of appropriate continuous response distribution (e.g. lognormal)
  - Calculate sample size for continuous response variable & compare
- **Results**
  - Detectable difference = 10%
  - 90% Confidence Level, 80% Power
    - » 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 costs!

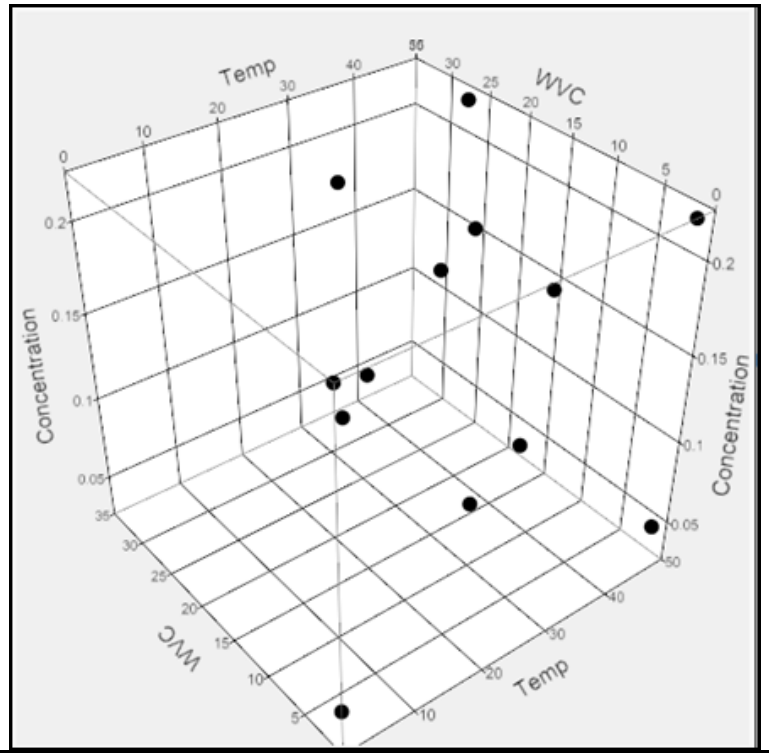
This example results in a 65% reduction in test cost!

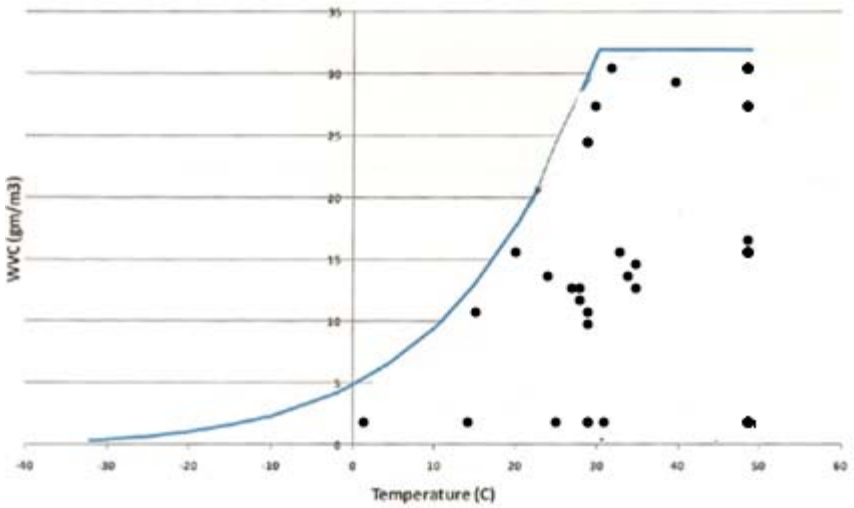
Benefits of Statistical Analyses for DoD T&E

- **Historical Analyses have focused on one-sample "*roll-up*" methods or selectively binning of data and calculating *averages***
- **Regression, ANOVA, Response Surface Models, and General Linear Models techniques allow the data to determine which factors are significant as opposed to subject matter experts**
  - Provides object analysis methodologies with inferential abilities
  - Provides standardized methodologies to approaching analysis
- **Quick and easy in commercial statistical software**

Example Analysis: Chemical Agent Detector

- **Design from Joint Chemical Agent Detector**
  - Employed an optimal design methodology
  - Responses times are hypothetical
  - What is the implication in test analysis?





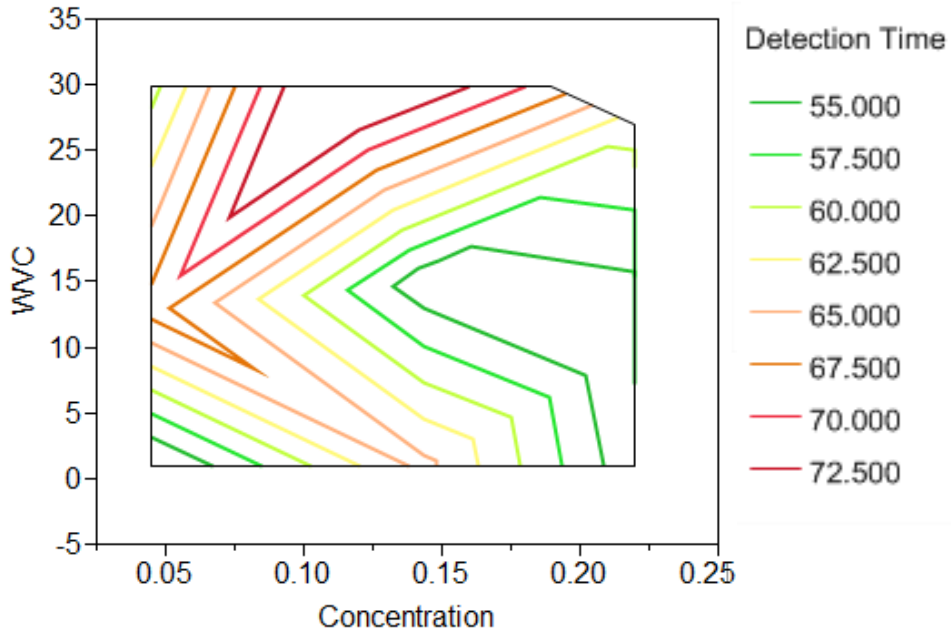
Chemical Agent Detector Results

(notional analysis – not based on actual data)

- **Data determine significant factors:**

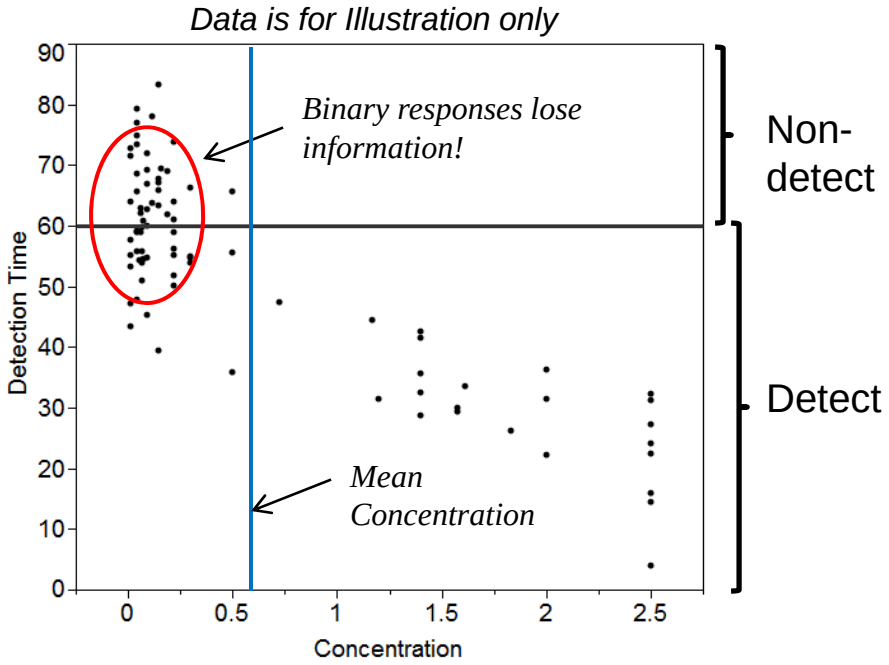
Factor	Model Coefficient Estimate	Standard Error	F-Ratio	P-value
Temperature	-7.07	1.30	29.7	<0.001
Water Vapor Content	5.13	1.06	23.6	<0.001
Agent Concentration	5.13	2.01	96.5	<0.001
Agent Type	N/A	N/A	4.34	<0.001

- **Allows for understanding of performance across the operational envelope.**
- **Note: All results are for Illustration only**



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 **400% increase** in confidence interval width

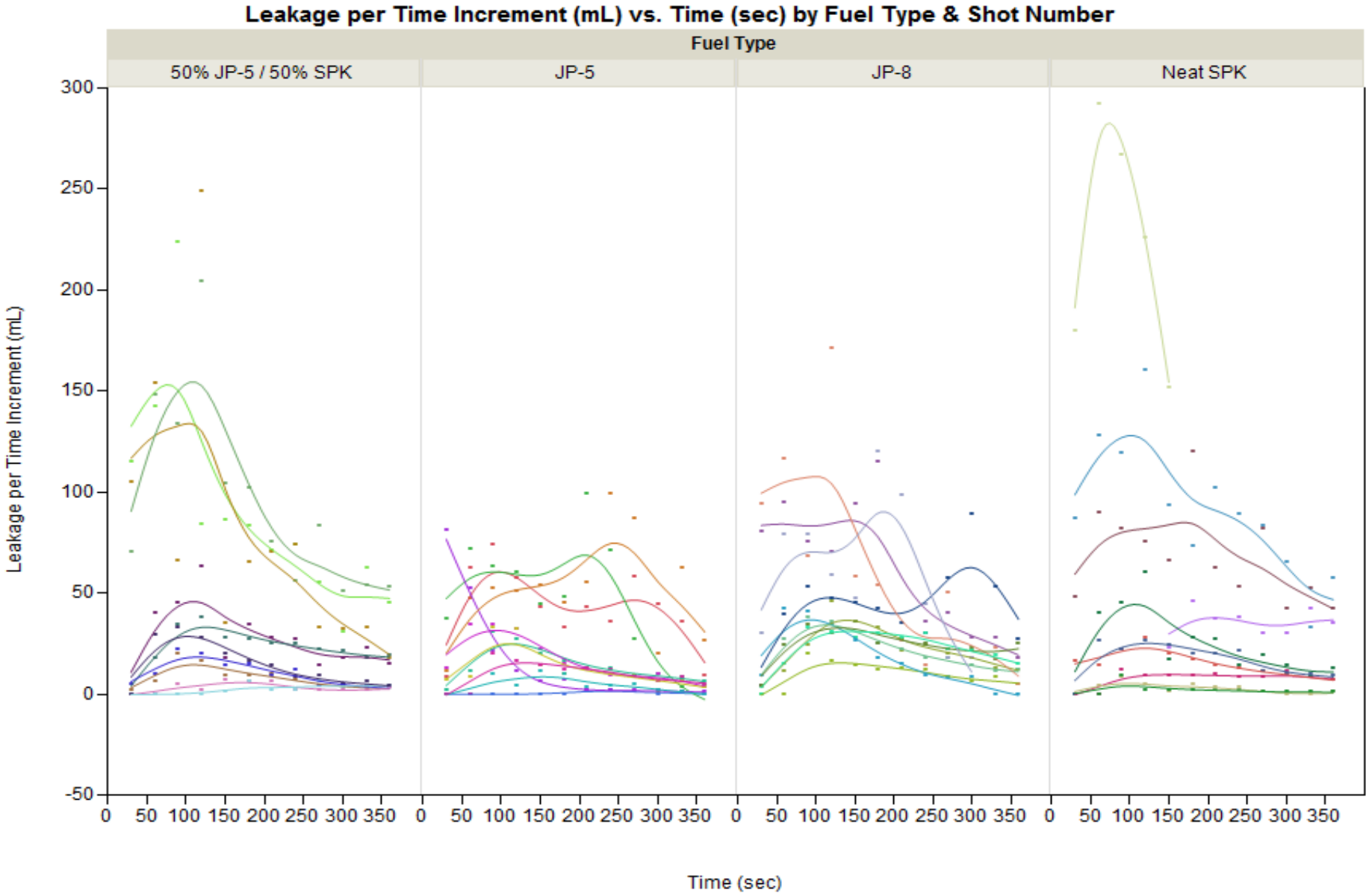


Response	Probability of Detection within 60 seconds at mean	Lower 90% Confidence Bound	Upper 90% Confidence Bound	Confidence Interval Width
Binary (Detect: Yes/No)	83.5%	60.5%	94.4%	33.9%
Continuous (Time)	91.0%	86.3%	94.5%	8.2%

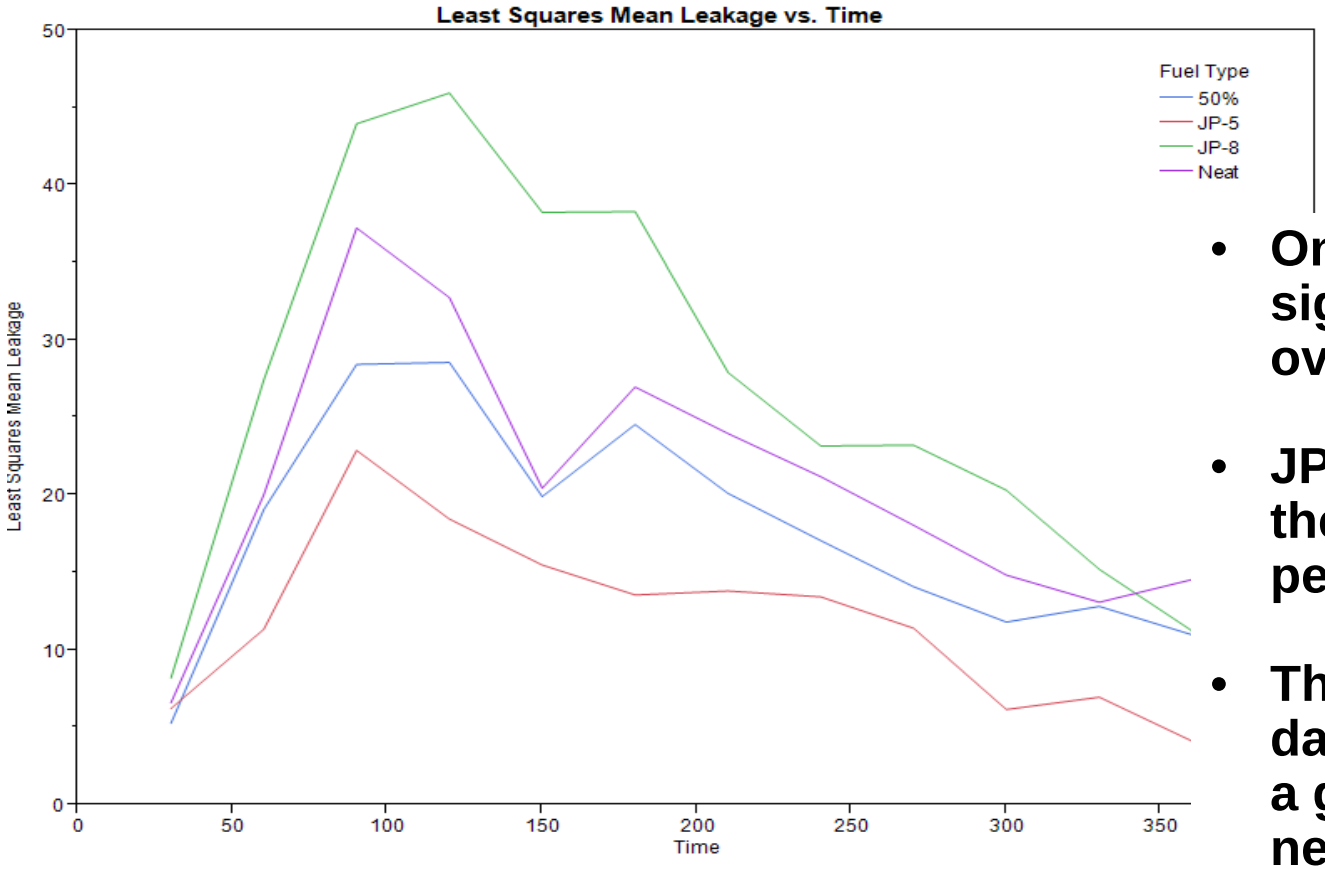
Example Analysis: Seal-Sealing Ability of Fuel Bladder Materials with Bio-Fuels

- **Bio-fuel alternatives are being incorporated to conventional fuel types.**
- **<u>Conjecture:</u> the low aromatic content of bio-fuels will reduce self-sealing properties of existing fuel bladders**
  - Navy conducted Live Fire testing to determine the effect of varying fuel types on the self-sealing abilities of operational fuel tanks
  - Shots fired at fuel tanks with multiple fuel types
  - Leakage amounts were measured over time
- **Different organizations used different analyses to reach different conclusions:**
  - Multiple t-tests at each time point resulted in no significant difference between fuel types and
  - Extrapolation over aromatic content followed by t-test concluded that there was a degradation in self-sealing due for bio-fuels
  - Difference due to failure to model the data correctly and erroneous extrapolations
- **Advanced analysis: time series model accounting for velocity of the projectile as a covariate**

Fuel Leakage Data by Time Increment



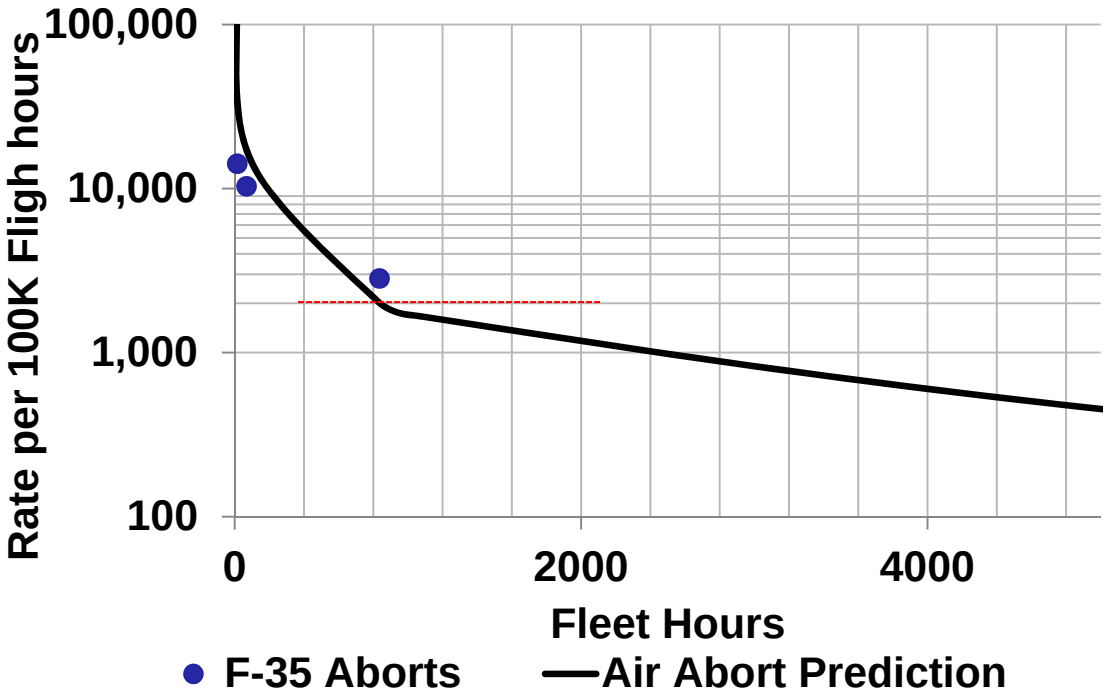
Average fuel leakage



- **Only one fuel type was significantly different overall (JP-5)**
- **JP-8, 50/50 Blend, and the 100% Bio-fuel performed similarly**
- **The variability due to damage was so large that a good deal more data is necessary to discriminate between fuel types**
  - Not an operationally relevant difference between fuel types

Joint Strike Fighter Aborts Analysis

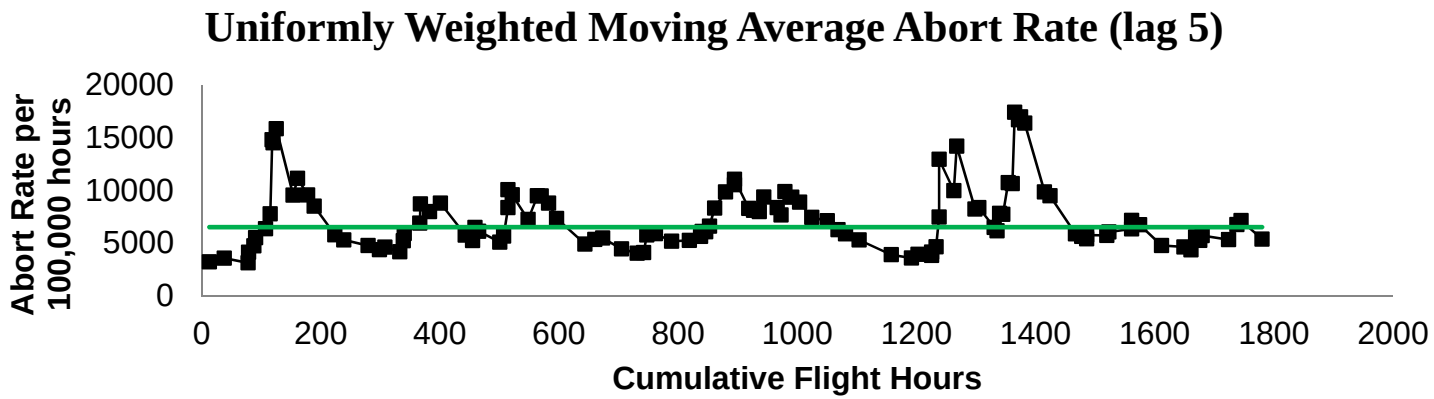
- **<u>Conjecture</u>: the abort rate of JSF is changing over time**
- **Leveraged control charts and reliability growth models to investigate the conjecture**
- **Conclusions: the abort rate for JSF to date is constant. There is no evidence of significant increases or decreases in the abort rate.**

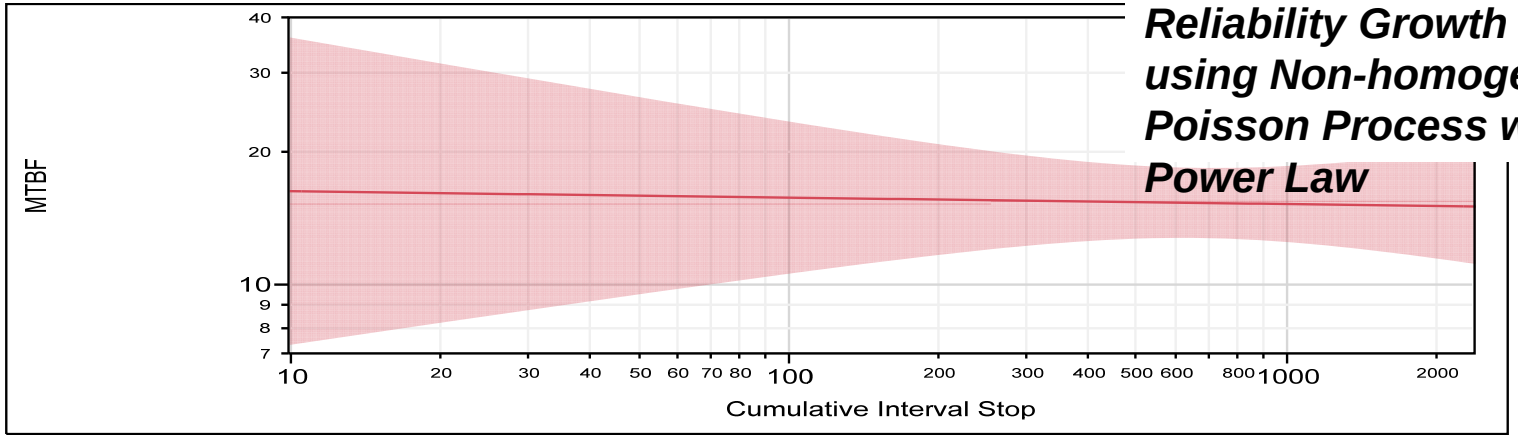


Original analysis suggested cumulative air abort rate is related to mishap rate

But this type of analysis converges to a mean and is difficult to see changes

IDA Joint Strike Fighter – Reliability Growth Analysis



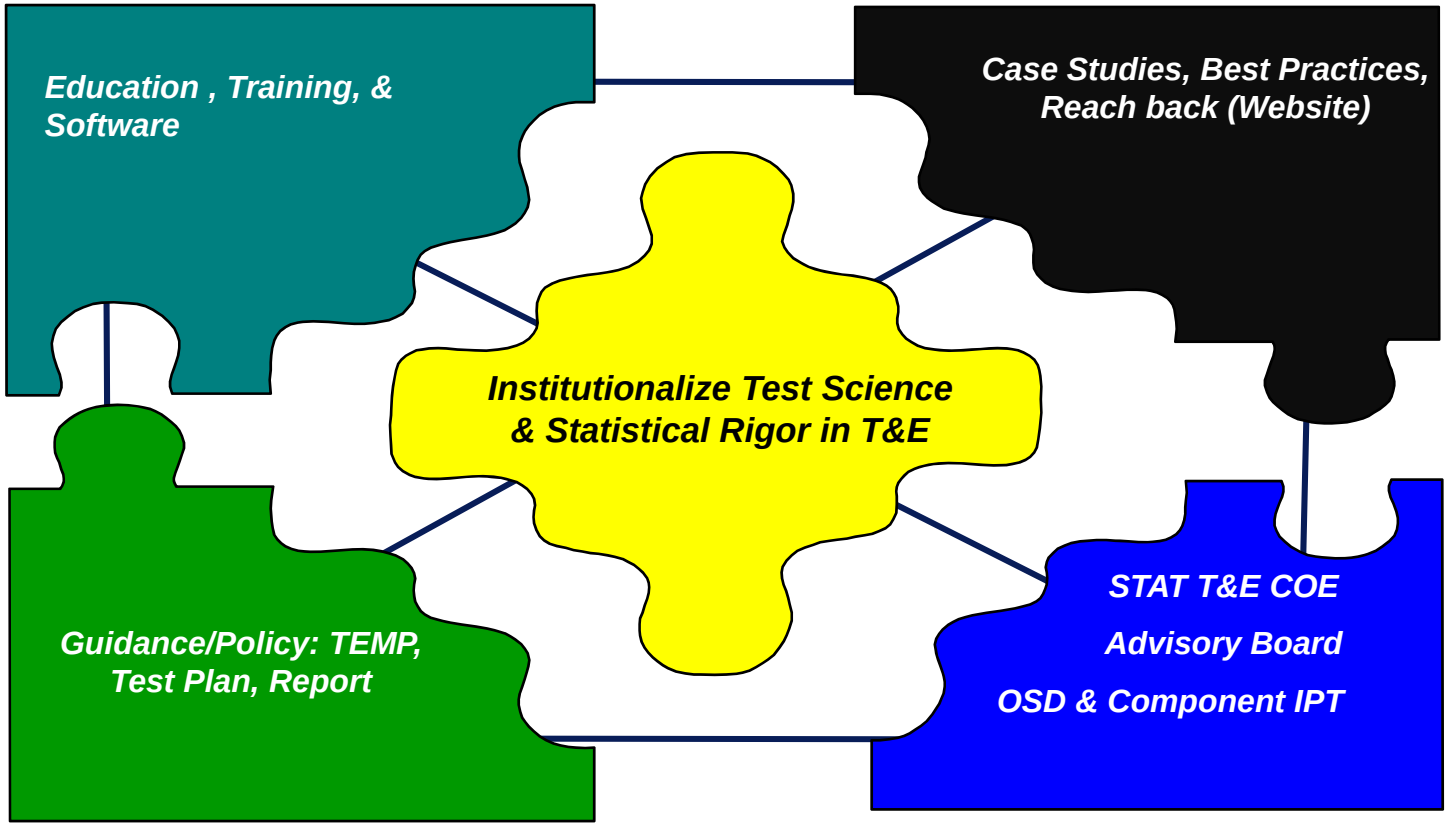


Reliability Growth Model using Non-homogeneous Poisson Process with Power Law

Aircraft	Parameter	Estimate	Standard Error	Lower 95% CI	Upper 95% CI
All	Λ	0.058467	0.04157435	0.0145095	0.2356007
	β	1.014490	0.09419305	0.8456989	1.2169702

Implementation Plan

- **Supported by DASD(DT&E) and DOT&E**



Current Efforts to Institutionalize Statistical Rigor in T&E

- **DOT&E and DDT&E have formed steering committees**
  - Investigate current workforce capabilities, education and new hire needs.
  - Develop toolbox of statistical design and analysis techniques appropriate for T&E
- **Development of Scientific Test and Analysis Techniques (STAT) Center of Excellence to provide support to programs**
- **Research Consortium**
  - Navel Post Graduate School, Air Force Institute for Technology, Arizona State University, Virginia Tech
  - Research areas:
    - » Case studies applying experimental design in T&E.
    - » Experimental Design methods that account for T&E challenges.
    - » Improved reliability analysis.
- **Current Training and Education Opportunities**
  - Air Force sponsored short courses on DOE
  - Army sponsored short courses on reliability
  - AFIT T&E Certificate Program
- **Review of current policy & guidance**
  - DOD 5000
  - Defense Acquisition Guidebook

Conclusions

- **Since October 2012, DOT&E and the Services have made significant improvements in increasing the statistical rigor of test designs.**
- **Analyses of these tests using statistical methods are just getting underway.**
- **Education is the number one challenge!**
  - Test Science Implementation Plan

Backup Material