INSTITUTE FOR DEFENSE ANALYSES

Statistical Design & Analysis Challenges in Defense Testing

Heather Wojton, Project Leader

Kelly Avery

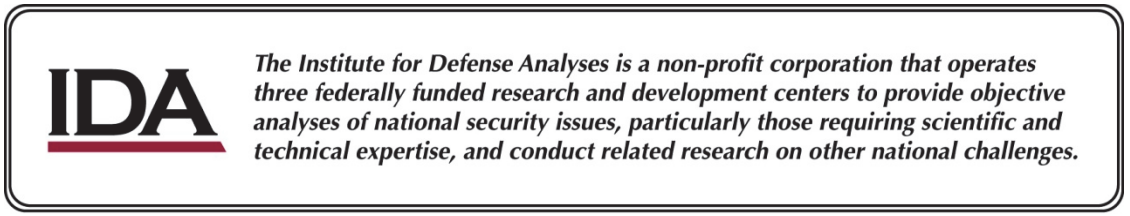
January 2019

Approved for public release.
Distribution is unlimited.

IDA Non-Standard Document
NS D-9225

Log: H 2018-000338

INSTITUTE FOR DEFENSE ANALYSES
4850 Mark Center Drive
Alexandria, Virginia 22311-1882



About This Publication
This work was conducted by the Institute for Defense Analyses (IDA) under contract HQ0034-14-D-0001, Task BD-9-2299(90),"Test Science Applications," for the Office of the Director, Operational Test and Evaluation. The views, opinions, and findings should not be construed as representing the official position of either the Department of Defense or the sponsoring organization.

Acknowledgments
The IDA Technical Review Committee was chaired by Mr. Robert R. Soule and consisted of Yevgeniya Pinelis and Matthew Avery from the Operational Evaluation Division.

For more information:
Heather Wojton, Project Leader
hwojton@ida.org • (703) 845-6811

Robert R. Soule, Director, Operational Evaluation Division
rsoule@ida.org • (703) 845-2482

INSTITUTE FOR DEFENSE ANALYSES

IDA Non-Standard Document NS D-9225

Statistical Design & Analysis Challenges in Defense Testing

Heather Wojton, Project Leader

Kelly Avery

Executive Summary

Before the DoD acquires any major new capability, that system must undergo realistic testing in its intended environment with military users. The complex, data-limited, highly variable nature of the test environment presents many unique statistical challenges. The set of conditions in which a system will operate is typically large, and important variables are often uncontrollable during test, making rigorous experimental design a challenge. Data sets obtained from tests are almost always messy. Issues such as lurking variables, small and unbalanced sample sizes, and ordinal responses necessitate creative and sometimes sophisticated data analysis approaches. This document examines some of these defense-testing situations in detail and discusses how statisticians in the test and evaluation community have approached associated design and analysis challenges.

Testing in the DoD

All systems undergo operational testing before fielding or full rate production. While contractor and developmental testing tends to be requirements-driven, operational testing focuses on mission success. The goal is to evaluate operational effectiveness and suitability in the context of a realistic environment with representative users.

Congress established the office of the Director, Operational Test and Evaluation (DOT&E) separate from the Services' operational testing agencies. This office writes reports on system performance to Congress and the Secretary of Defense, and sets policy and guidance for conducting operational testing. Statistics has been an area of emphasis in recent years as DOT&E strives to institutionalize test science and rigor in test and evaluation.

Statistical Challenges (and Some Solutions)

For large, multi-mission programs such as the F-35 Joint Strike Fighter, testers want information on system performance across multiple system variants, mission areas, and environmental conditions. However, each mission is extremely expensive and collecting data in every combination of conditions is impractical or impossible. A design of experiments approach can span a complex operational mission space efficiently and defensibly. In the case of the F-35 Joint Strike Fighter, a D-optimal design approach was used to maximize information gain in this resource-limited situation.

Designing a good test does not eliminate the need to think carefully about analysis. If not properly accounted for, uncontrolled or lurking variables can lead to seemingly

counterintuitive results. As an initial analysis of an unmanned aerial vehicle designed to image and track targets, analysts inspected probability of detection performance across each factor. These "roll-up" results make it impossible to differentiate the complex effects of multiple factors and can lead to incorrect conclusions, such as radar performance appearing worse for faster targets. A deeper dive revealed confounding between controlled and uncontrolled factors. Using mixed effects models, analysts could appropriately account for factor effects and characterize performance across this complex battlespace.

Models and simulations have increasingly become an essential element of operational test and evaluation. Since model runs are cheaper than live testing, analysts can typically collect multiple simulation data points for any given live data point, thus creating an unbalanced data set. This poses a challenge for validating a simulation capability, since most traditional statistical comparison techniques are designed for large, balanced samples. However, Monte Carlo power simulations with type 1 error corrections can tell analysts which "standard" statistical techniques work well for these non-standard situations.

Data collected from tests can look different than any standard textbook problem. For example, when testing an upgraded video link between an unmanned aircraft and its home station, analysts can record altitude and distance every minute, as well as video quality (on a three-point categorical scale). Thus, the outcome of interest is both continuously observed and ordinal, a combination not commonly encountered. In this particular case, since altitude, distance, and video quality remained constant for relatively large periods of time, the data could be significantly reduced without loss of information via change point detection and smoothing. Cumulative logistic mixed models regression could then be used to estimate performance and measure mission-to-mission variation.

Statistical Design & Analysis Challenges in Defense Testing

Dr. Kelly Avery

Institute for Defense Analyses

9th International Purdue Symposium on Statistics

Outline

- Testing in the DoD
- Statistical Challenges (and Some Solutions)
  - Designing tests in large and uncontrollable factor spaces
  - Dealing with lurking variables
  - Validating models using small & unbalanced sample sizes
  - Analyzing continuously observed ordinal data

Note: All data presented are either transformed or notional

Testing in the DoD

All military systems undergo operational testing before fielding or full rate production…



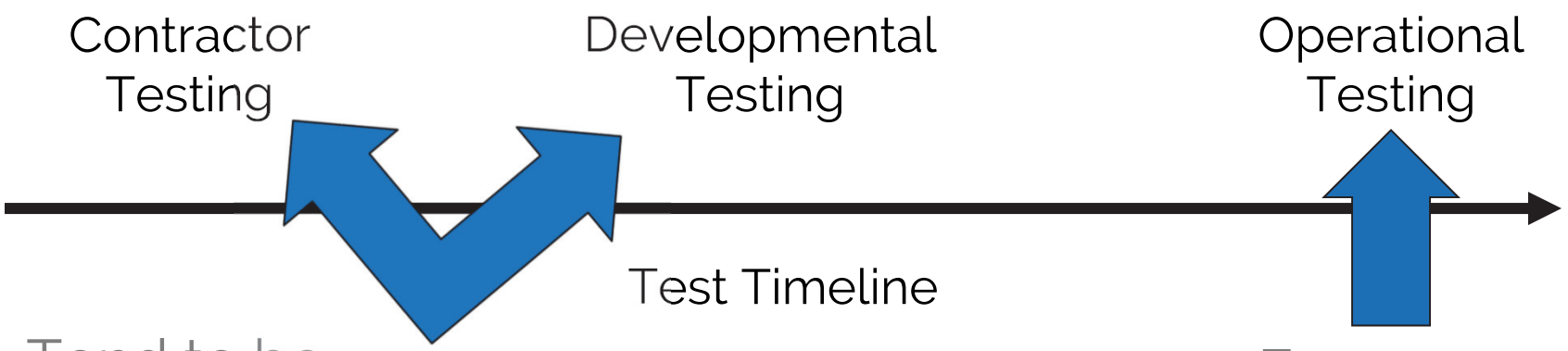





…Even the ones you don't normally think of

- Biometrics systems
- Personnel management systems
- Logistics and readiness systems
- Command & control systems
- Pilot trainers

DoD Test Paradigm



Tend to be requirements driven

Focuses on mission success

Requirements documents are often missing important mission considerations

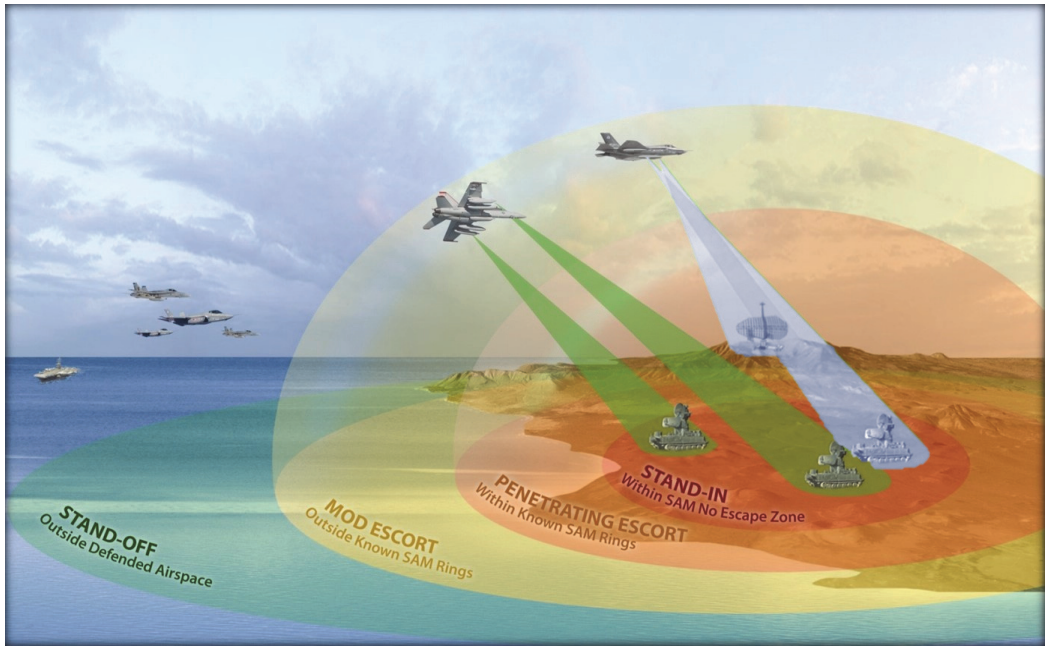
Goal of operational test: evaluate operational effectiveness and suitability

Operational Environment

Representative Users

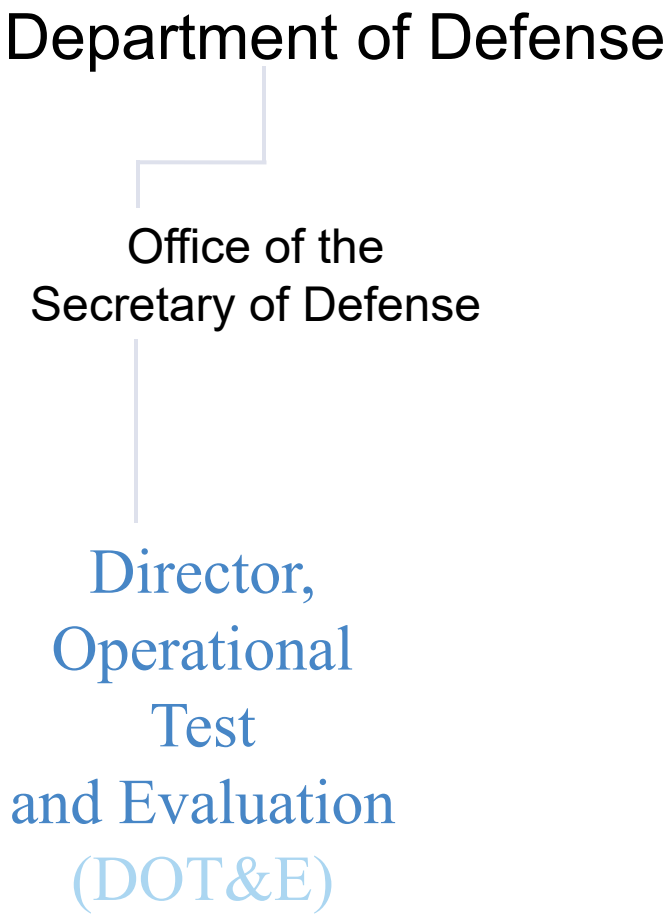
"Real" Threats

Conducting Missions

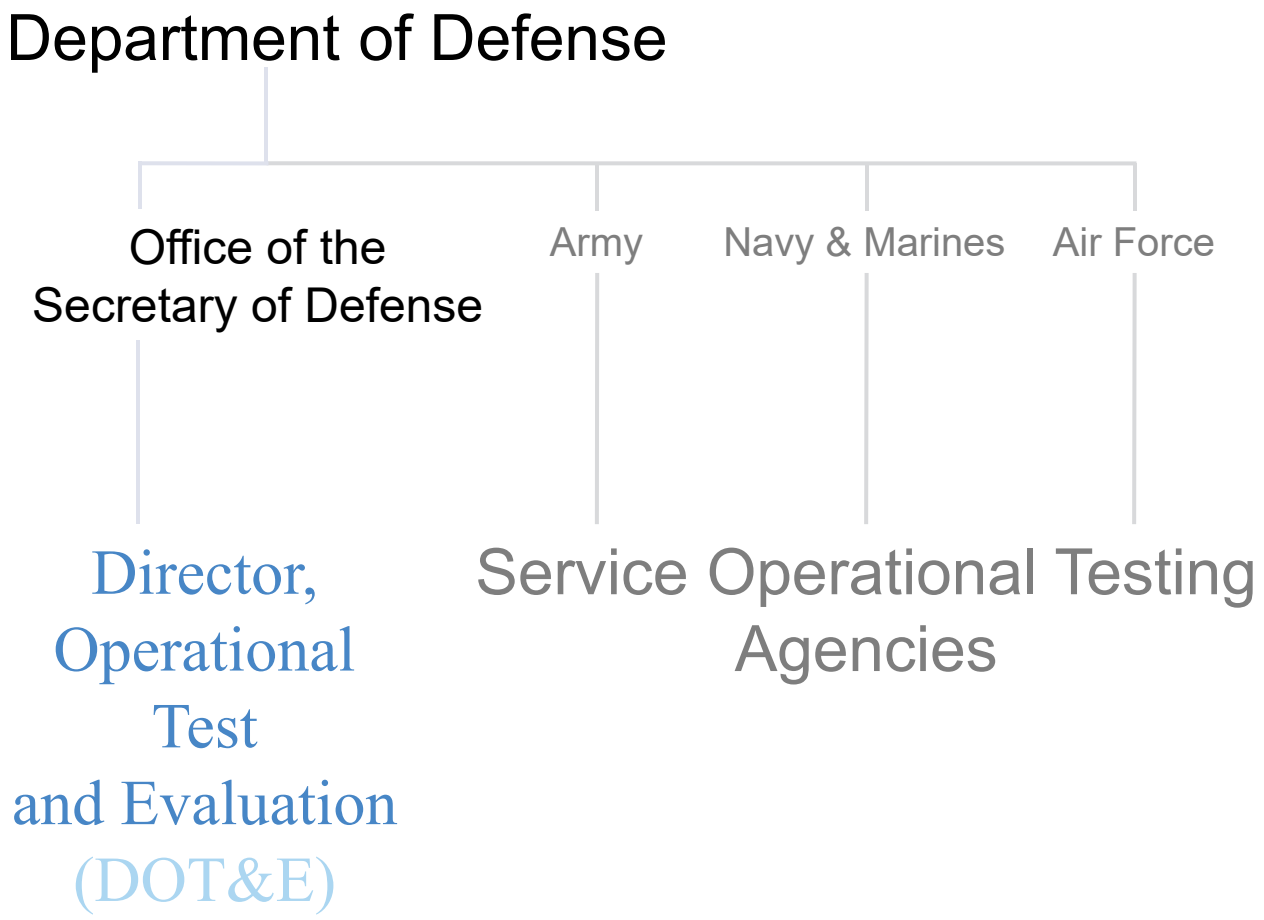


We have to think carefully about the range of realistic combat scenarios a system may face

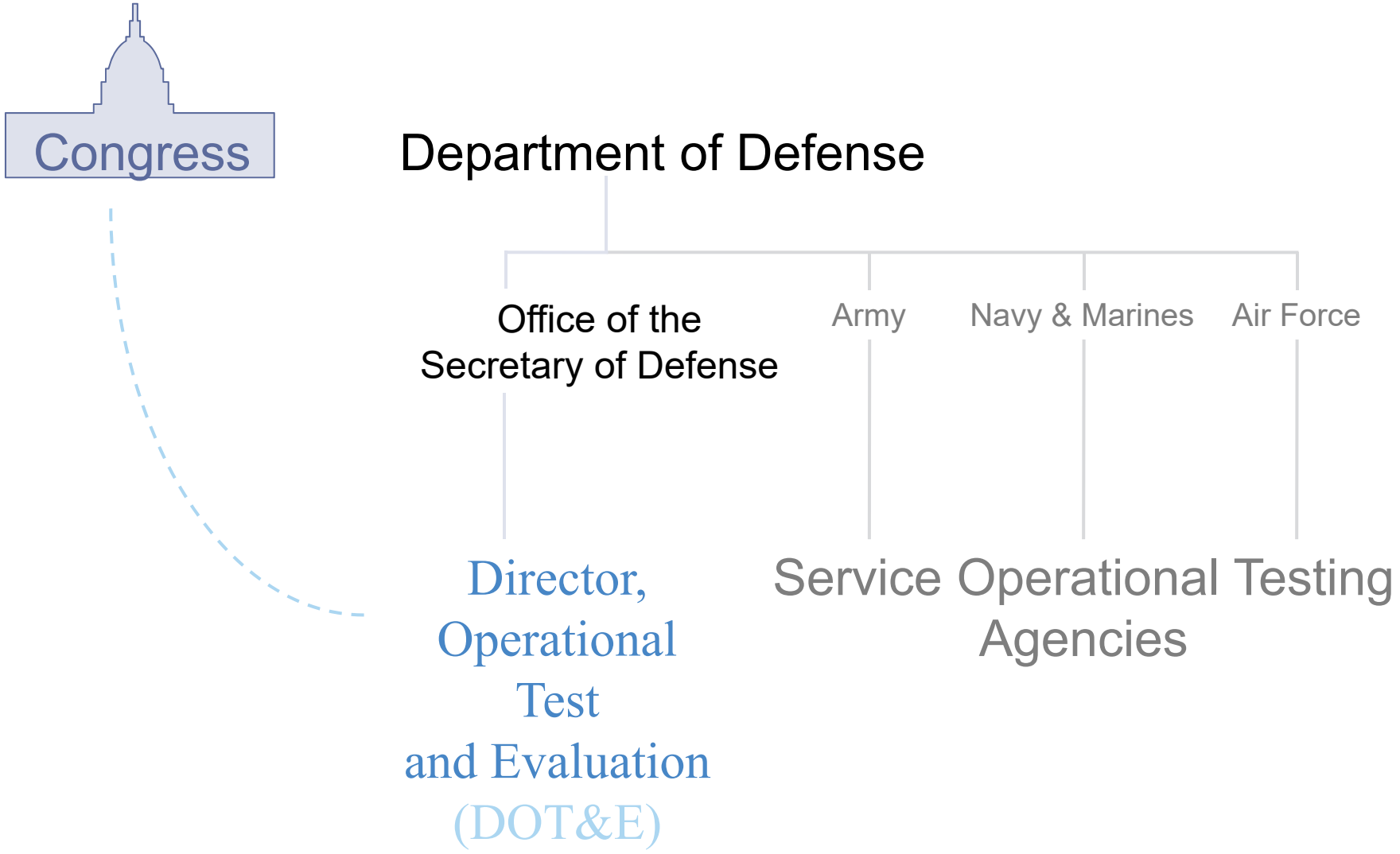
Congress established DOT&E separate from the Services' operational testing agencies



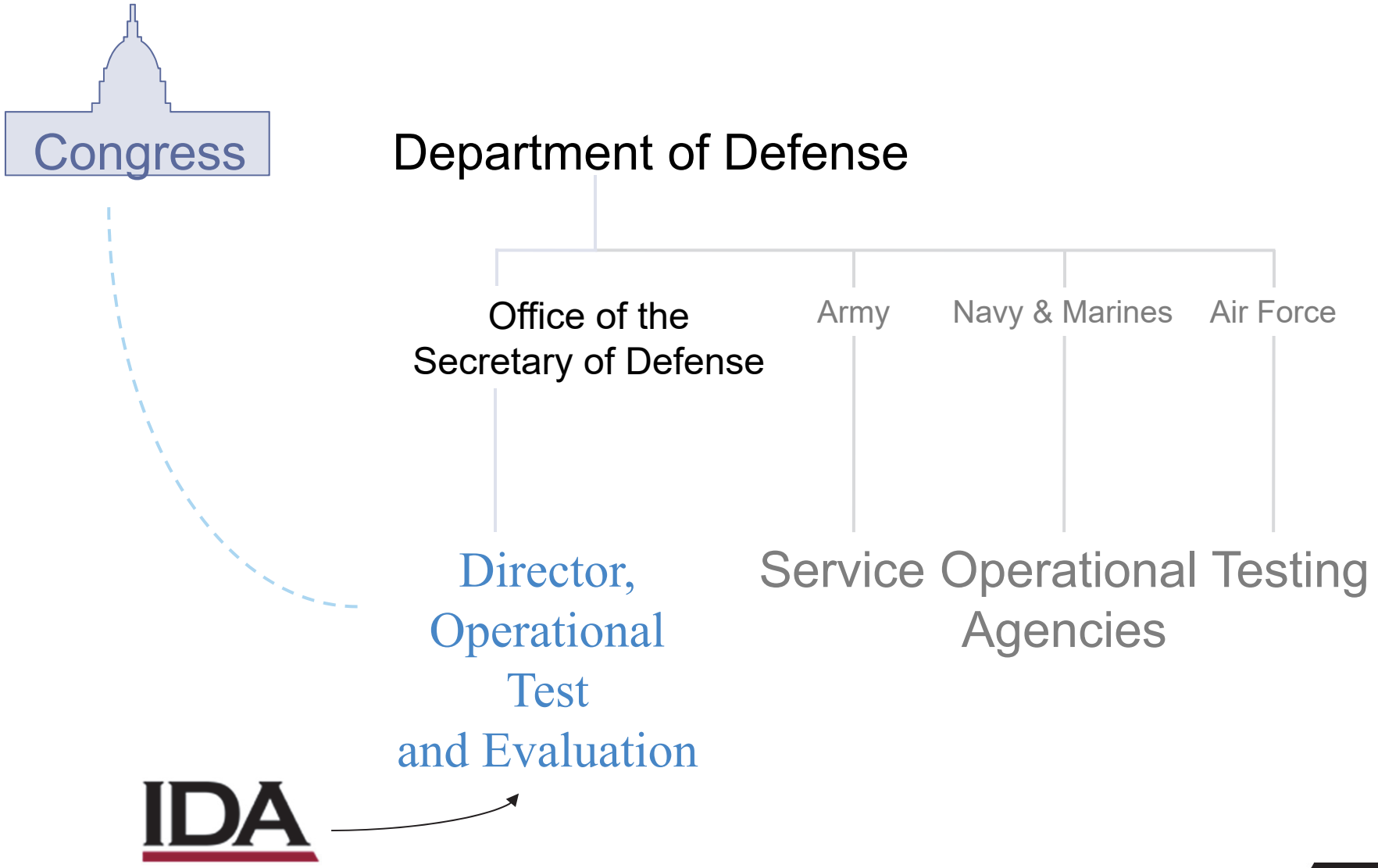
Congress established DOT&E separate from the Services' operational testing agencies



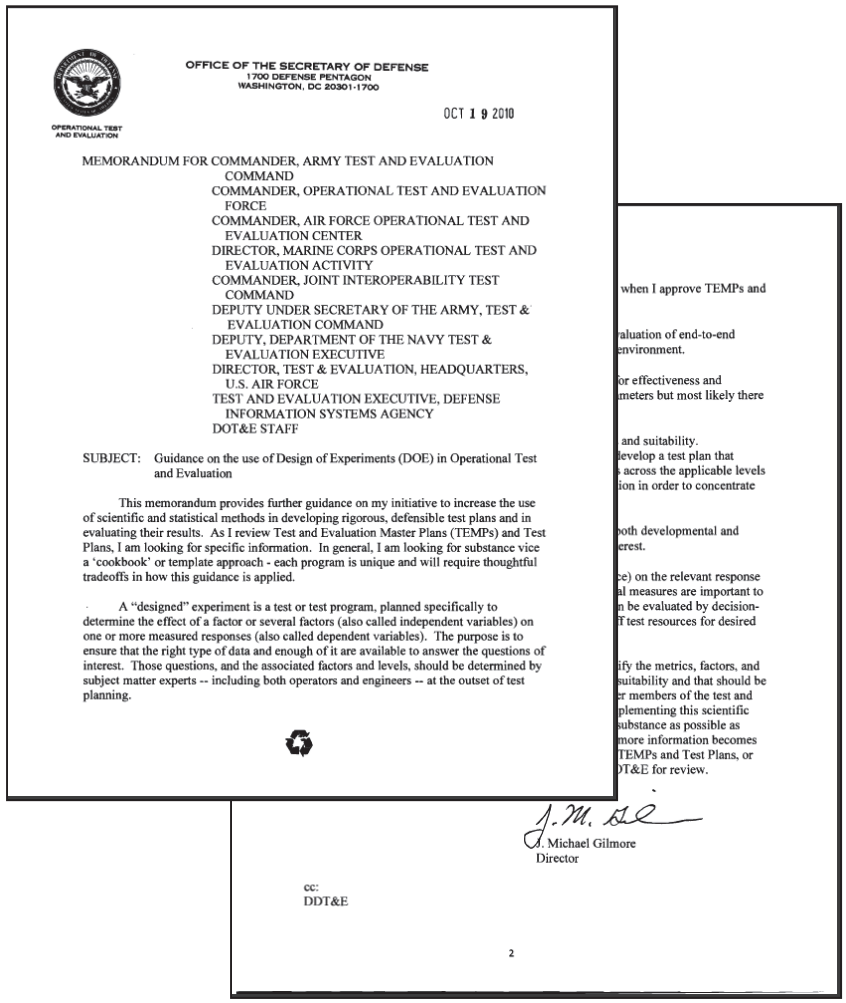
Congress established DOT&E separate from the Services' operational testing agencies



Congress established DOT&E separate from the Services' operational testing agencies



DOT&E sets policy and guidance for conducting operational testing, and statistics has been a point of emphasis in recent years



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.

when I approve TEMPs and

aluation of end-to-end environment.

or effectiveness and meters but most likely there

and suitability. develop a test plan that across the applicable levels ion in order to concentrate

both developmental and erest.

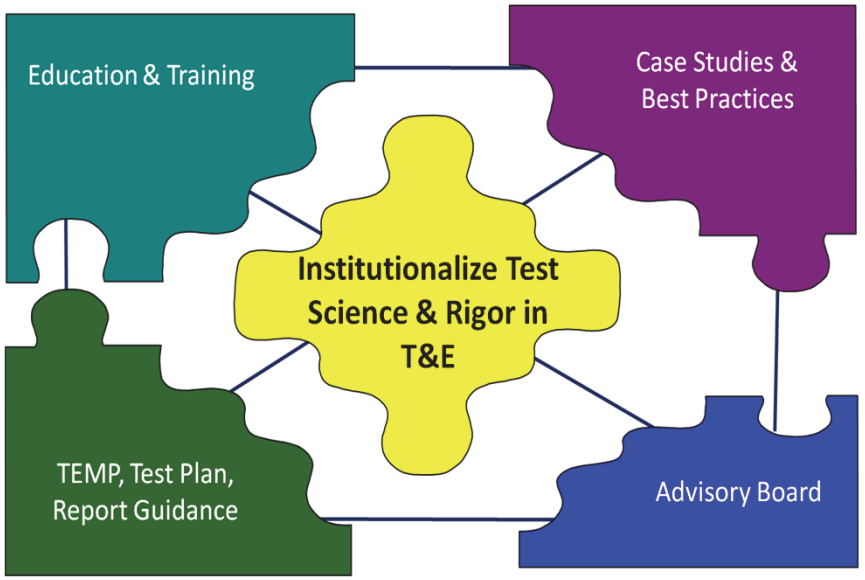
e) on the relevant response al measures are important to n be evaluated by decision- F test resources for desired

ify the metrics, factors, and suitability and that should be r members of the test and plementing this scientific substance as possible as more information becomes TEMPs and Test Plans, or DT&E for review.

J. Michael Gilmore
Director

cc:
DDT&E

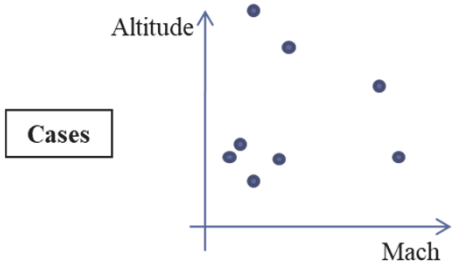
2

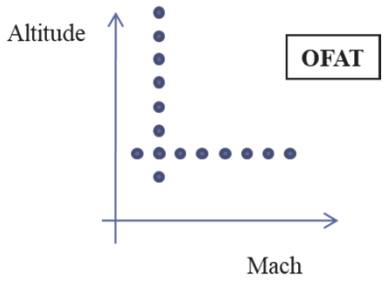


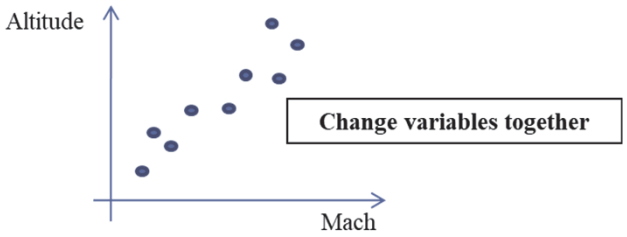
Design of Experiments (DOE) is not just for baking cookies:

Designing tests in large and uncontrollable factor spaces

There are several statistical shortcomings associated with common ways of designing operational tests





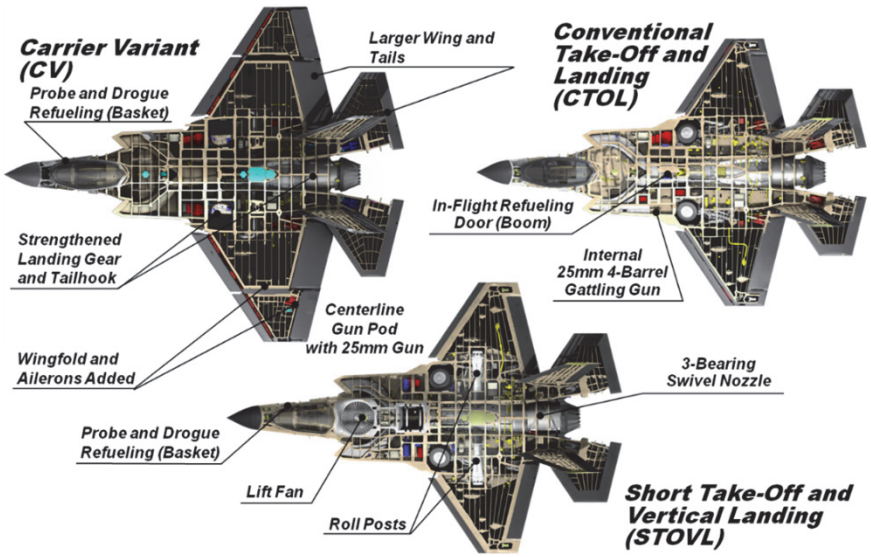


- Case-based
  - Little predictive ability; loss of ability to determine cause and effect
  - Limited to the specific conditions selected – might miss important performance shortfalls
  - Often poor statistical precision (demos)
- One factor at a time
  - Often is overkill, unnecessarily large test sizes
  - Interactions between conditions often not examined
- Observational studies
  - Confounding data
  - Loss of ability to determine cause & effect

DOE can mitigate these problems but implementation is challenging!

The F-35 Joint Strike Fighter (JSF) is tri-Service, multi-national family of strike aircraft

- Goal of test is to evaluate effectiveness of JSF on nine core mission areas
- Three different variants – test must provide adequate information for each
- Need to understand the effect of numerous other environmental/mission conditions on performance
  - These depend on core mission area, but could include physical environment (urban/rural), time of day, target type/movement, air threat, communications type, formation size, and many others



We want information across all variants, core areas, and conditions…but every JSF mission is extremely expensive!

A design of experiments approach can span a complex operational mission space efficiently and defensibly

- D-optimal design techniques are ideal to cover a large factor space and retain the ability to estimate all desired model terms
- Treating the F-35 variant as a factor allows us to leverage relevant performance and mission-level data across variants
- Building the DOE requires subject matter expertise on which higher order interactions are expected to occur

Optimal designs are computer-generated based on variance-oriented criteria and are highly flexible

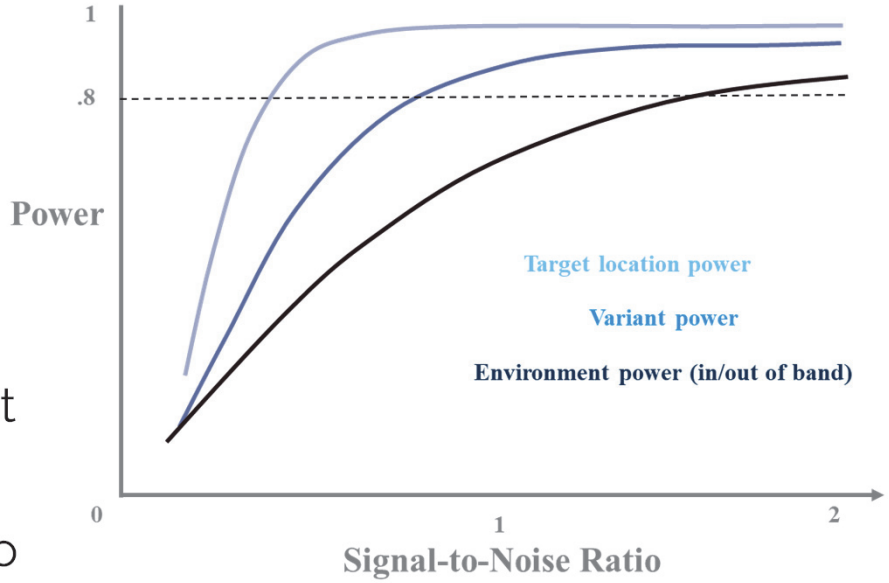
The D-Optimal objective is to maximize the determinant of the moment matrix M over all designs, hence minimizing the variances of the regression coefficients

where p is the number of model parameters

DOE maximizes information gain in a resource-limited world

The operational space for the F-35 operational test is vast! But taking a DOE approach:

- Spans across missions, variants, and threats efficiently
- Enables testers to make informed decisions about covering the space
- Provides a defensible rationale for test adequacy
- Reduces sorties required compared to legacy platforms



Strategic DOE enabled testers to adequately cover nine core mission areas, each with multiple factors in a combined total of just 110 trials

Results aren't always as expected:

Dealing with lurking variables

DOE solves many testing challenges but is not a panacea

Data analysis is rarely as simple as summaries or models based on the factors you built into the design

A test was conducted to evaluate an unmanned aerial vehicle (UAV) designed to image and track targets

Primary question of interest: Does the UAV have a high probability of detecting moving ground targets?

Controlled	
Target Size	Ground Target Category (GTC) 1, GTC 2, GTC 3
Range to Target	Near, Medium, Far, Very Far
Radar Mode	Dedicated GMTI, Concurrent GMTI and SAR Imagery
Terrain	Mountain, Flat, Littoral
Surface Cover	Barren, Foliage, Urban, Snow
Recorded	
Squint Angle	Low, High
Grazing Angle	Low, High
Target Speed	Slow, Medium, Fast

A D-optimal design was executed across several factors

Several other uncontrollable variables were recorded

An intuitive first step might be to inspect performance across the levels of each factor…

Single-Factor Probability of Detection Roll-Up*

Target Size	*GTC 1*	*GTC 2*	*GTC3*	
	0.6	0.7	0.8	
Range To Target	*Near*	*Medium*	*Far*	*Very Far*
	0.7	0.8	0.5	0.3
Radar Mode	*Dedicated*	*Concurrent*		
	0.8	0.6		
Terrain	*Mountain*	*Flat*	*Littoral*	
	0.8	0.6	0.7	
Surface Cover	*Barren*	*Foliage*	*Urban*	*Snow*
	0.8	0.7	0.4	0.6
Target Speed	*Slow*	*Medium*	*Fast*	
	0.4	0.9	0.2	

Results are counterintuitive for certain factors

Radar performance worse at Near ranges than Medium ranges

Radar appears to detect targets with higher probability in mountainous terrain than in flat terrain, but mountainous terrain should have more line-of-sight blockage.

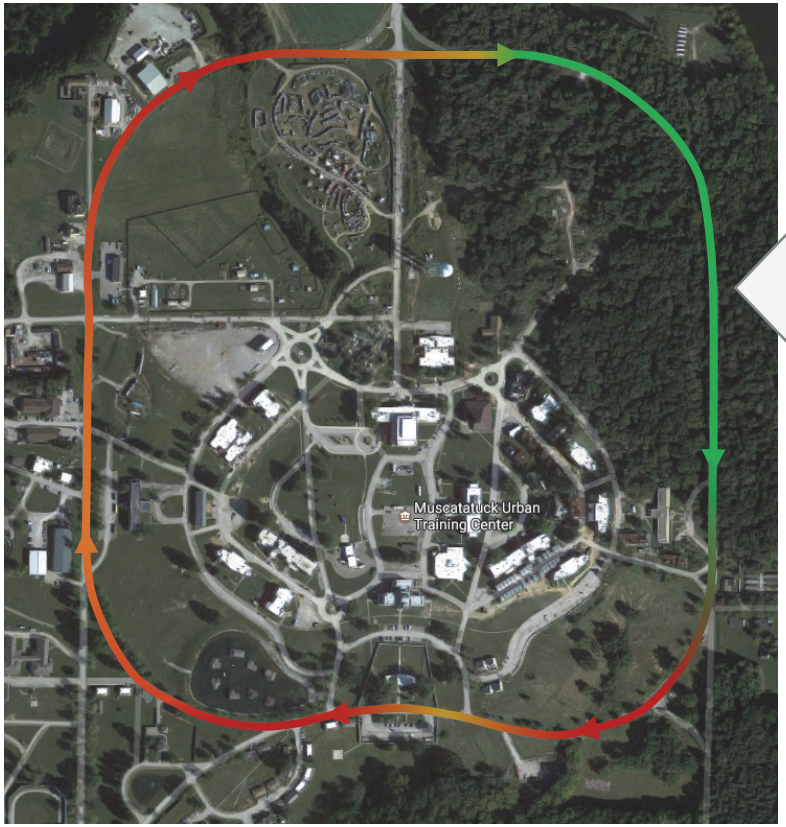
Radar performance appears worse for faster targets, which defies expectations

**Data are notional*

…but roll-ups can make it impossible to differentiate the complex effects of multiple factors and can lead to incorrect conclusions

Digging deeper, we see a much more complicated situation than originally expected

- Surface cover and terrain varied substantially at each range
- Applying the same surface cover or terrain level to all runs at a range did not appropriately control for the variance!
- This led to seemingly counterintuitive conclusions



Range restrictions and roads induced a correlation between target speed and position. In this case, the segment where the targets drove fastest was obscured by dense forest.

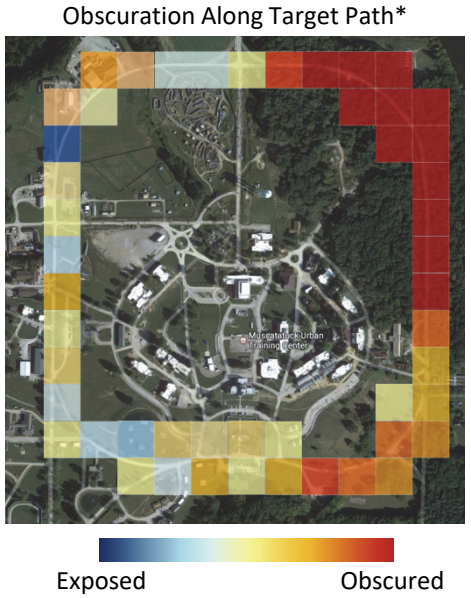
Target Speed*

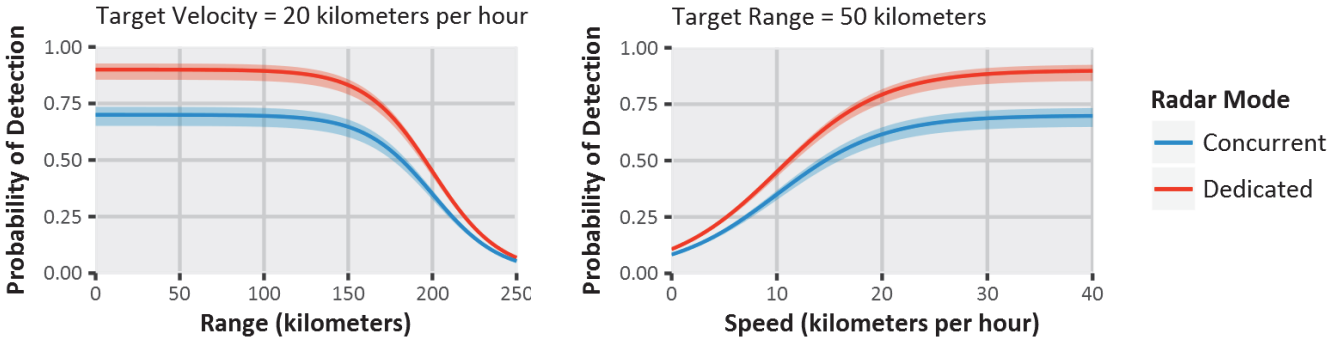
Slowest — Fastest

Mixed effects models can characterize performance across this complex battlespace

Random effects allow an estimation of line-of-sight blockage without having to define the surface cover and terrain at each point on the range

After accounting for the effects of terrain, target range and velocity have the expected behavior (i.e., reduced probability of detection at farther target range and higher target speed)*





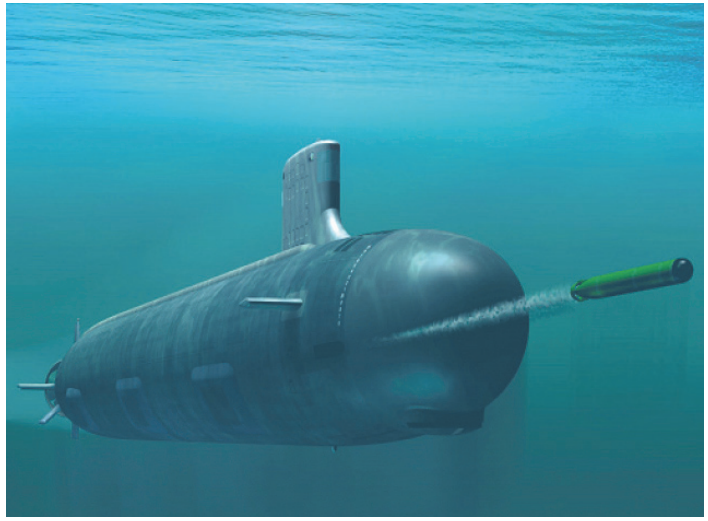
*Data are notional

Assumptions are for the birds:

Validating models using small & unbalanced sample sizes

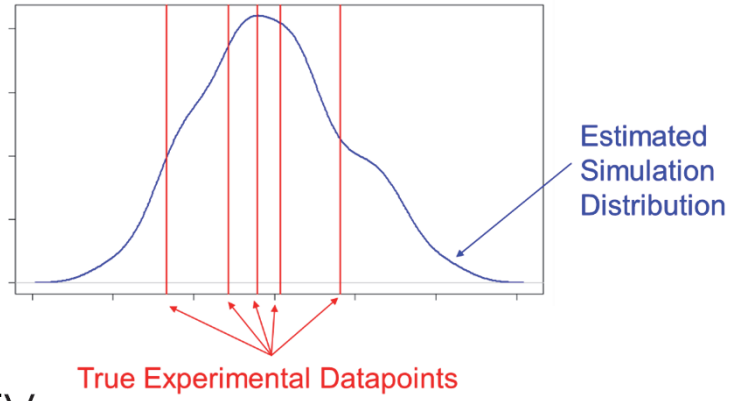
Models and simulations have increasingly become an essential element of operational test and evaluation

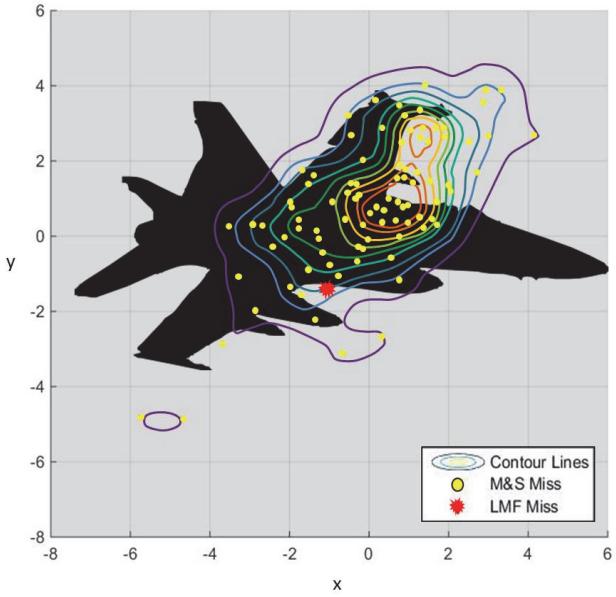
- Supplement or augment live test data when experiments are cost and/or safety prohibitive
- Examine threats incapable of being reproduced for testing
- Characterize rare events or threats
- Allow for end-to-end mission evaluation
- Inform experimental design decisions

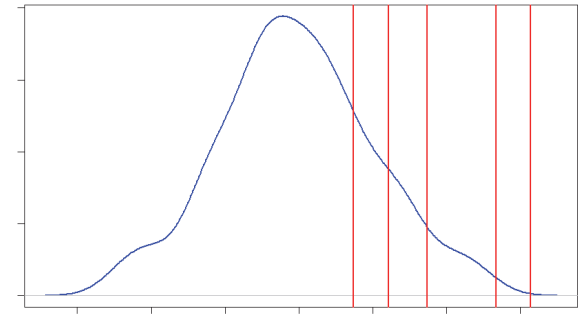


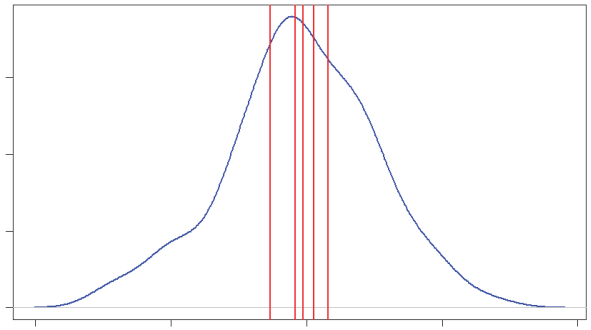
We can typically collect multiple simulation data points for any given live data point

- The goal is to determine if the simulation data and the live experimental data agree and to understand the associated uncertainty
- And if they don't agree, can we identify the specific conditions where they disagree?



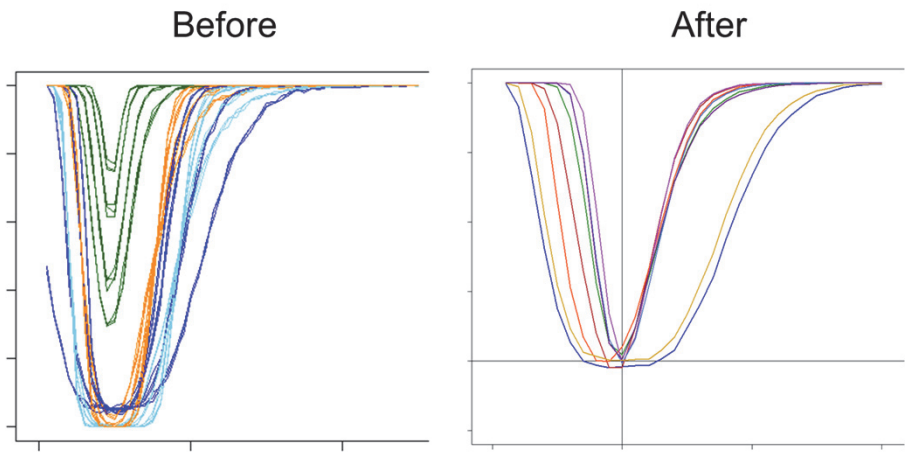






Monte Carlo power simulations can tell us which "standard" statistical techniques work well for these non-standard situations

Type 1 error had to be empirically corrected to support meaningful comparisons



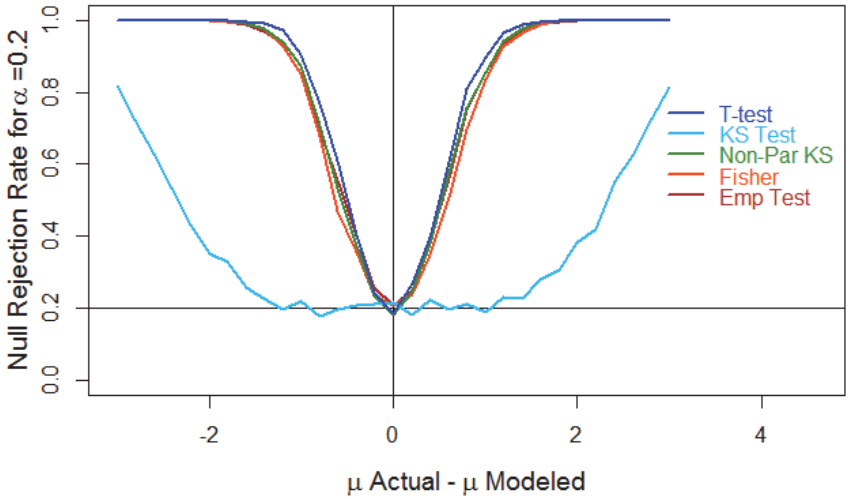
Distribution	Structure Of Factors	Small Sample Sizes	Moderate Samples Sizes	Large Sample Sizes
Skewed	Univariate			
	Distributed Level Effects			
	Designed Experiment			
Symmetric	Univariate			
	Distributed Level Effects			
	Designed Experiment			
Binary	Univariate			
	Distributed Level Effects			
	Designed Experiment			

Simulations conducted across multiple data distributions, sizes,* and structures

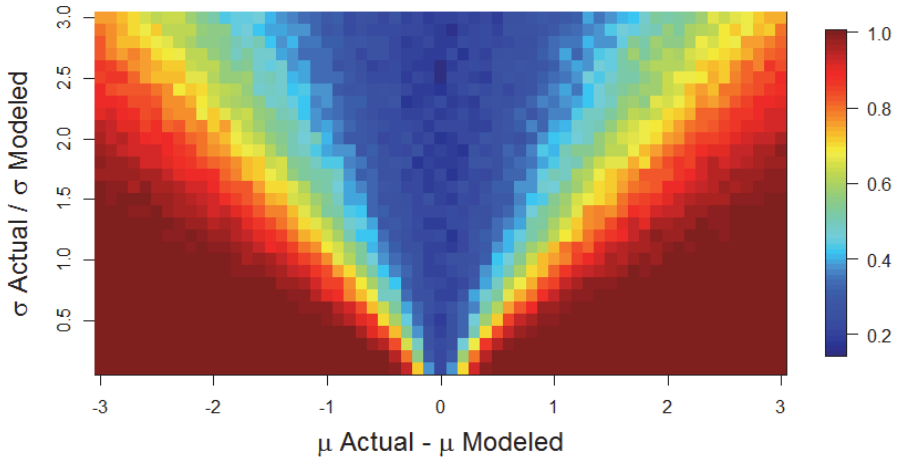
* A small sample size is two to four samples for continuous data or 20 for binary data, a medium size is six to 10 samples for continuous data or 50 for binary data, and a large size is 11 to 20 samples for continuous data or 100 for binary data.

Results show that simple techniques are still effective, even with extremely small & unbalanced samples

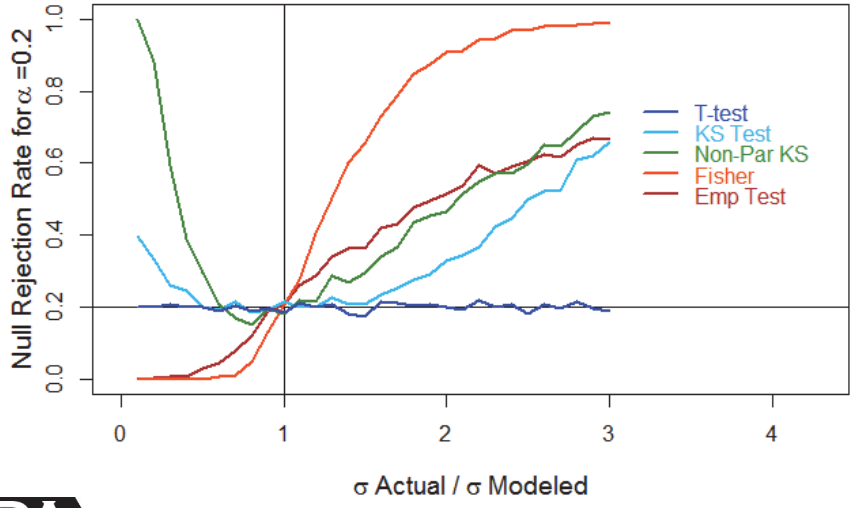
Changes in Mean



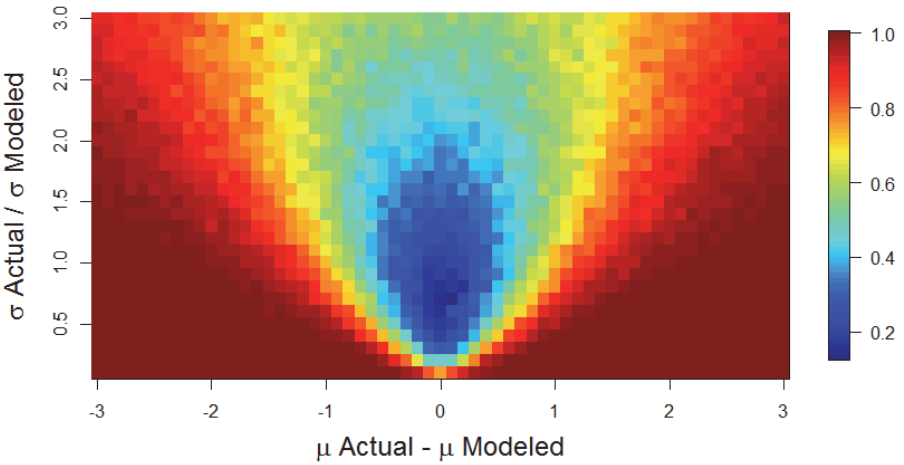
Changes in Mean & Variance – t-test



Changes in Variance



Changes in Mean & Variance – KS Test



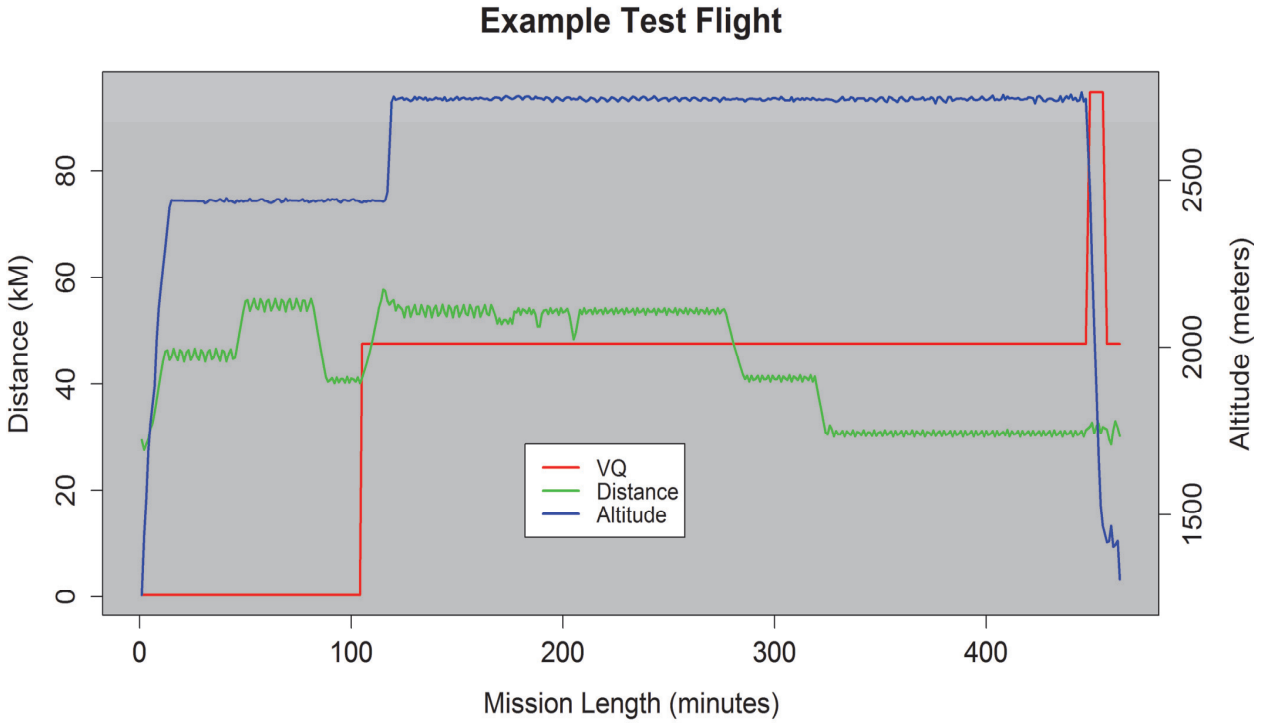
When textbooks don’t have the solution:

Analyzing continuously observed ordinal data

Consider testing an upgraded video link between an unmanned aircraft and its home station

- For each of multiple missions:
  - Altitude and distance from the control station are recorded automatically every minute
  - Video quality is human rated on a three-point scale (no video, intermittent video, continuous video)

- Each minute can be treated as a data point, but is this truly representative?
- Overfitting and varying autocorrelation structures are potential problems in this analysis

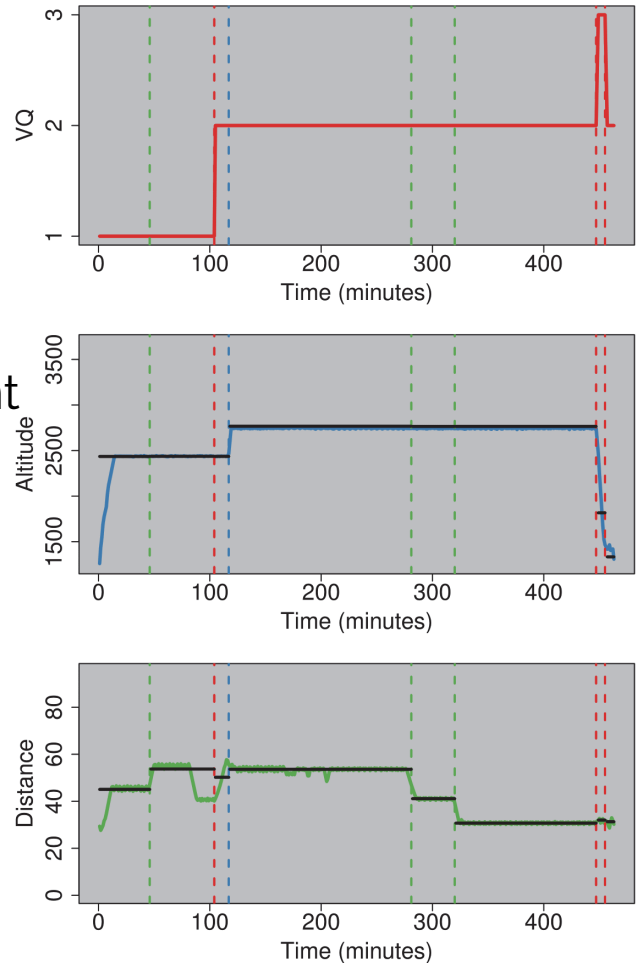


Data can be significantly reduced (without loss of information) via changepoint detection and smoothing

Although we collect hundreds of data points, the number of effective observations is much lower

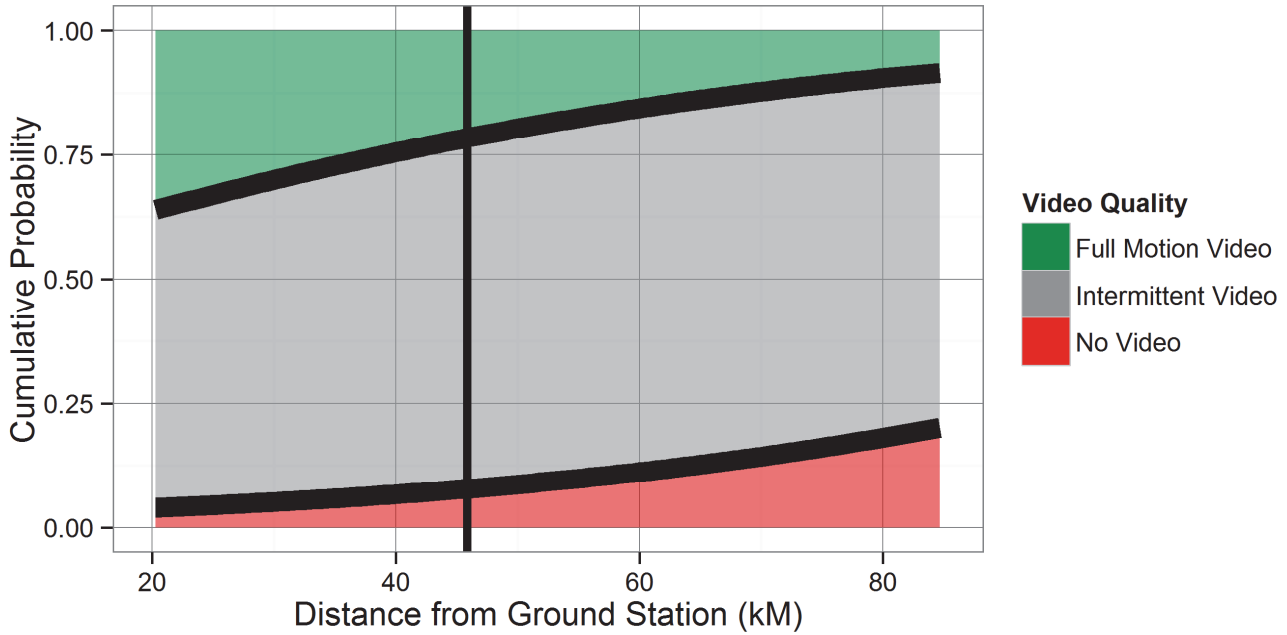
Automatic knot placement identifies regions of little change in covariates

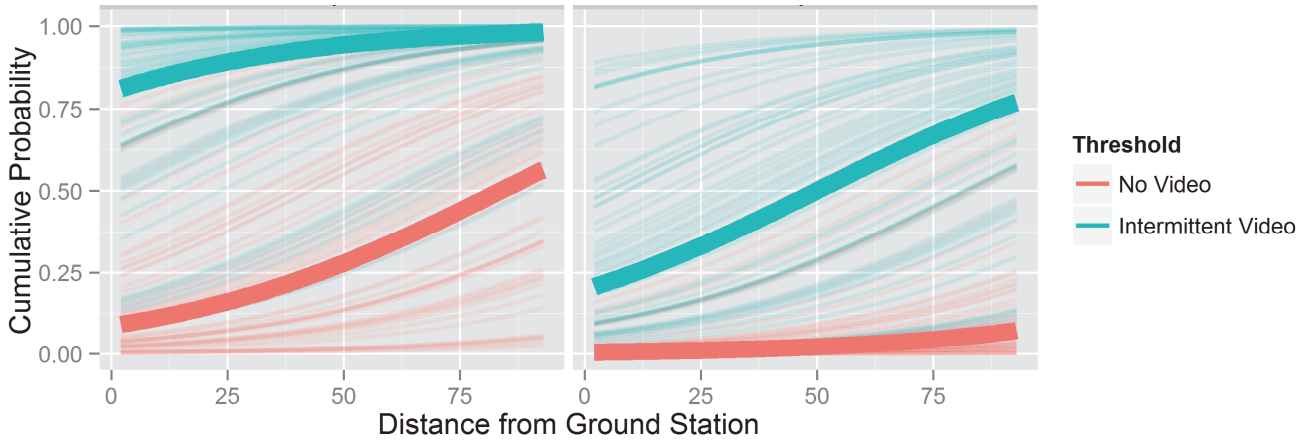
Use median to estimate the value of each factor for each interval between two changepoints



Distance (kM)	Altitude (m)	Video Quality	Observations	Mission	Day of Exercise
45.1	2437	1	47	7	4
53.7	2438	1	58	7	4
50.2	2437	2	13	7	4
53.6	2743	2	163	7	4
41.1	2742	2	39	7	4
30.7	2742	2	126	7	4
32	1864	3	7	7	4
31.3	1417	2	6	7	4

Cumulative logistic mixed model regression can estimate performance and measure mission-to-mission variation





Conclusions

We do rigorous design…

…and careful analysis!

Sometimes we adapt existing techniques…

…and sometimes we develop our own methods!

Defense testing presents some challenging, but not intractable, statistical problems…

…all in the context of real world systems with the goal of supporting our warfighters

Acknowledgements and References

Special thanks to Dr. Colin Anderson, Dr. Laura Freeman, and Dr. Matthew Avery for letting me showcase some of their work ☺

Avery, M., Orndorff, M., Robinson, T., & Freeman, L. (2016). Regularization for Continuously Observed Ordinal Response Variables with Piecewise-constant Functional Covariates. *Quality and Reliability Engineering International*, 32(6), 2033-2042.

Avery, K., Freeman, L. (2018, January). *Statistical Techniques for Modeling and Simulation Validation*. Paper presented at the 2018 Winter Simulation Innovation Workshop, Orlando, FL.

My contact info:
Dr. Kelly Avery
kavery@ida.org
(703) 845-2265