The Value of Rigorous Statistical and Analytical Techniques in Test & Evaluation

Purpose of Operational Testing

Provide realistic and objective assessments of how systems improve mission accomplishment under realistic combat conditions

- **Evaluate systems in operational scenarios**
  - Employed by representative warfighters
  - Realistic threats
- **Provide objective information before a system is used in combat**
  - Inform warfighters on capabilities and limitations
  - Facts for acquisition executives prior to full rate production decision
- **Ensure testing is adequate to support defensible evaluation**
- **DOT&E's fundamental purpose has been codified in Title X and DoD 5000 for many years and has not changed**

Rigorous Analytical and Statistical Techniques (1)

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

- **Data Analysis and Evaluation**
  - Using statistical analysis methods to maximize information gained from test data
    - » Regression Analysis
    - » Hypothesis Testing
    - » Confidence Intervals
  - Incorporate all relevant information in analyses
  - Ensure conclusions are objective and robust

Rigorous Analytical and Statistical Techniques (2)

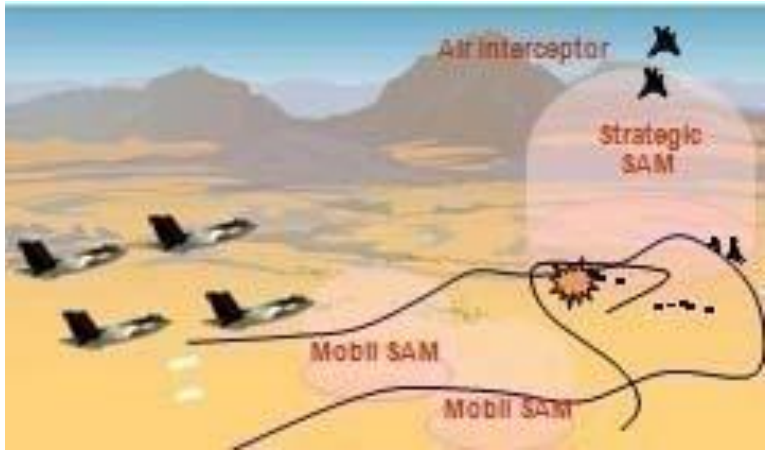
- **Best practices for survey data**
  - Empirically verified surveys
    - » NASA TLX
    - » System Usability Scale
  - Quantitative data about subjective experiences
  - Provide context for traditional measures

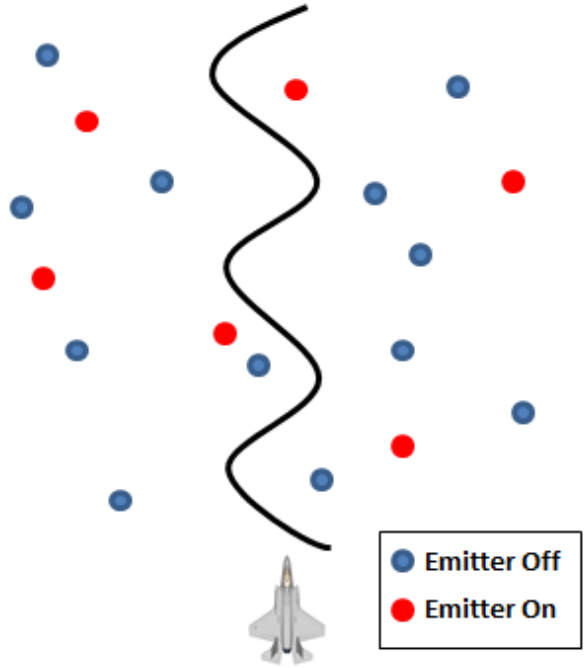
- **Modeling and simulation validation**
  - Consistent approach for validating model & simulation
  - Apply statistical techniques to quantify differences between M&S and live testing
  - Examine performance in difficult or impossible to test scenarios

Experimental Design for Efficient Testing
F-35 Joint Strike Fighter Mission Data File Scan Schedule Optimization

The F-35 is a tri-Service, multi-national, family of strike aircraft. It will replace the F-16 and A-10 in the Air Force, the AV-8B in the Marine Corps, and augment the F-18 in the Navy.

- **Goal**: determine robust settings of the USRL programmable scan schedule to:
  - Maximize Accuracy and Timeliness
  - Minimize Non-detections



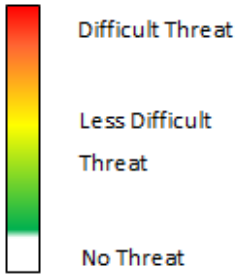


- **Complex operational space:**
  - Multiple emitter types
  - Multiple scenarios/geometries
  - Completely customizable scan schedule settings
- **DOE Solution**
  - **Robust parameter design** for the scan schedule settings crossed with **optimal designs** to span the complex emitter/scenario space
  - **Multivariate optimization process** to determine robust settings of the scan schedule in the laboratory

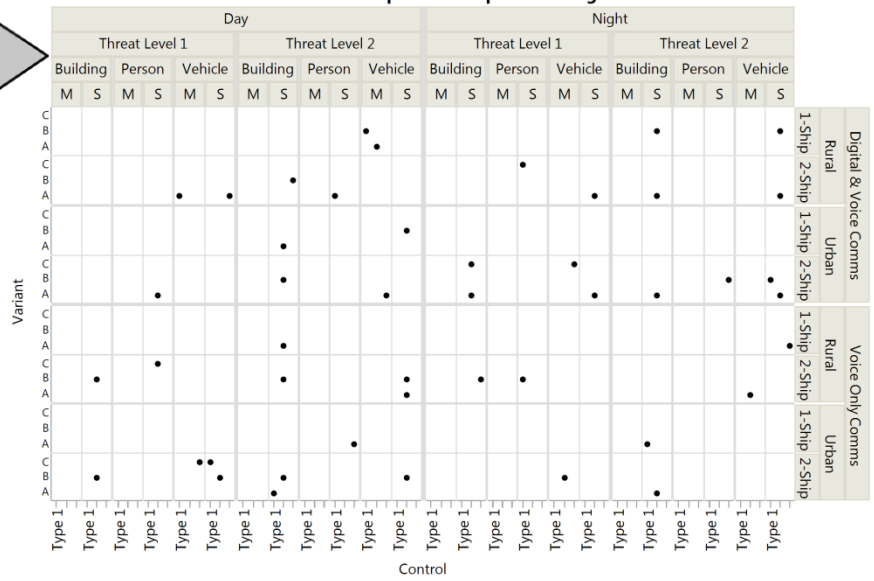
Experimental Design for Efficient Testing
F-35 Joint Strike Fighter Initial Operational Test & Evaluation (IOT&E) Planning

The F-35 is a tri-Service, multi-national, family of strike aircraft. It will replace the F-16 and A-10 in the Air Force, the AV-8B in the Marine Corps, and complement the F-18 in the Navy.

	Core Mission Areas				
	Air-Surface	OCA / Strike / SEAD	DCA	CAS	CSAR/ FAC[A]
JSF Variant	✓	✓	✓	✓	✓
Day/Night	✓	✓	✓	✓	✓
Air Threat					
Ground Threat					
Mission Specific Factors					



CAS Mission Operational Space Coverage



- **Multi-mission**
  - Air-to-Surface Attack, Offensive Counter Air, Suppression of Enemy Air Defenses, Defensive Counter Air, Close Air Support, Combat Search and Rescue, Forward Air Controller (Airborne).
- **DOE enabled defensible, <u>efficient</u> methodology for covering the operational space in each mission area**
  - Maximize problem identification opportunities
  - Understand performance differences
- **Provides a defensible methodology for comparison testing (F-35A compared to A-10 in the CAS Mission)**

Littoral Combat Ship

Quantifying the Precision of Estimates

Threshold Reliability

Probability = 0.80
(30-day mission no failures)

Core Mission Systems

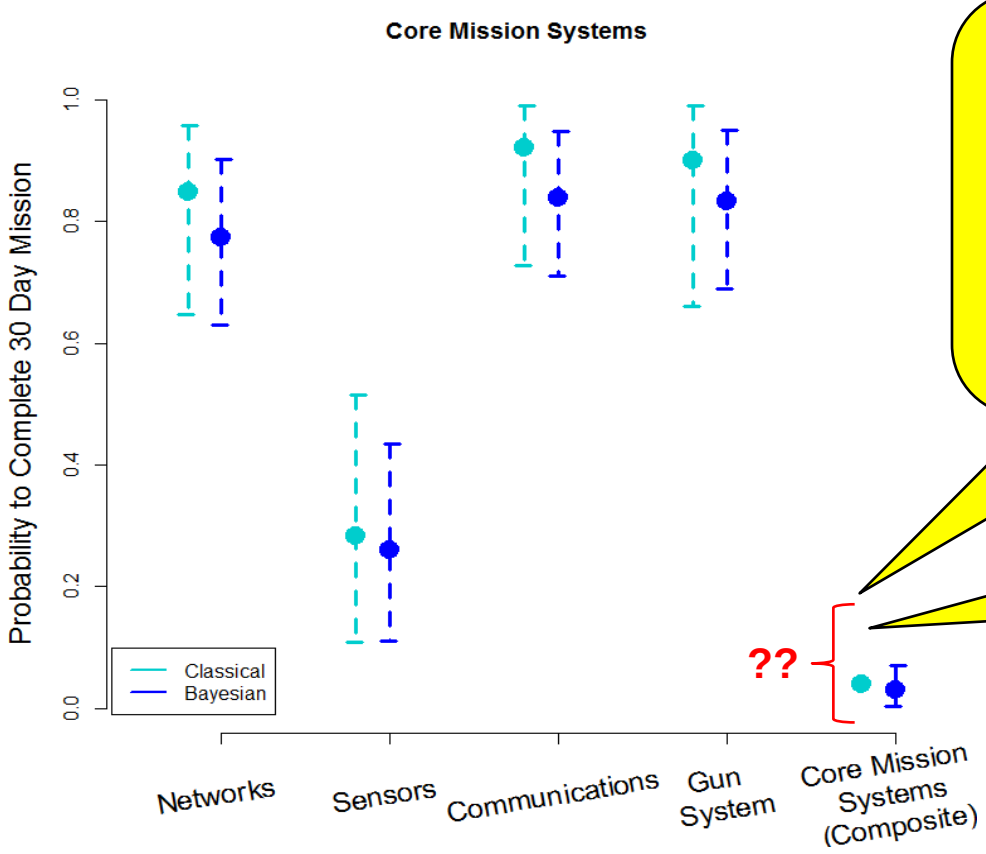
- *Computing Environment (Networks)*
- *Sensors*
- *Communications Systems*
- *Gun System*

Continuous Use

On Demand

Littoral Combat Ship

Quantifying the Precision of Estimates



No universally acceptable classical approach for combining data, estimating uncertainty

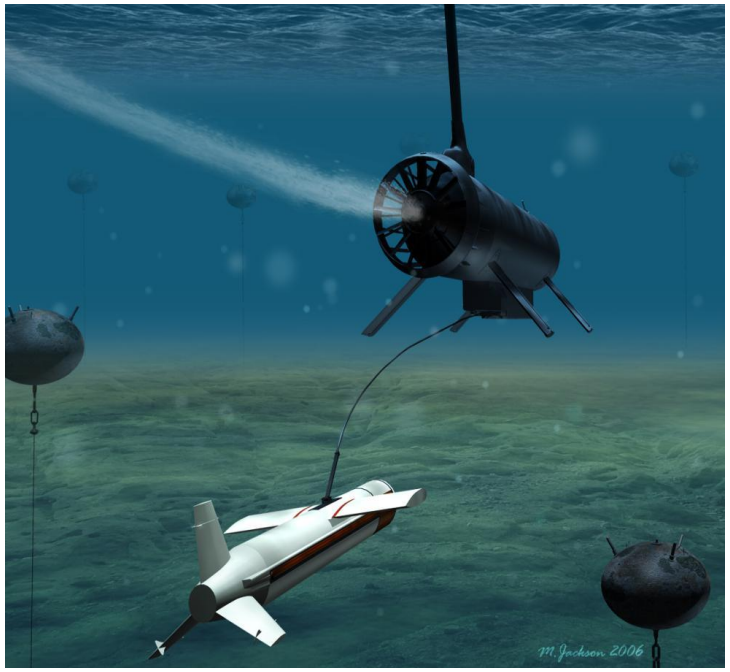
Impossible to estimate with classical methods if <u>zero</u> failures observed

Modern analytical techniques allow us to provide point estimates and uncertainty quantification for complex cases.

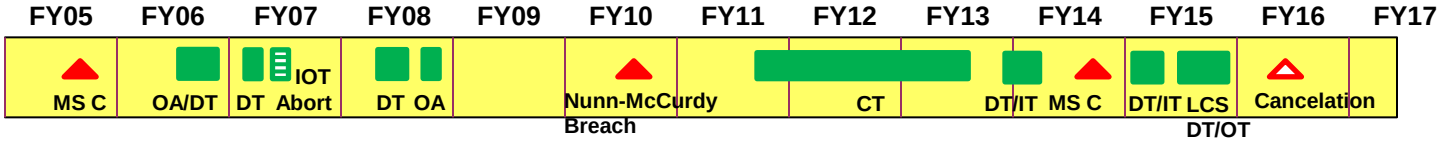
Remote Minehunting System

Quantifying the Precision of Estimates

- **Remote semi-submersible vehicle and towed sonar set to detect, localize and identify mines; key component of Littoral Combat Ship Mine Countermeasures Mission Package**

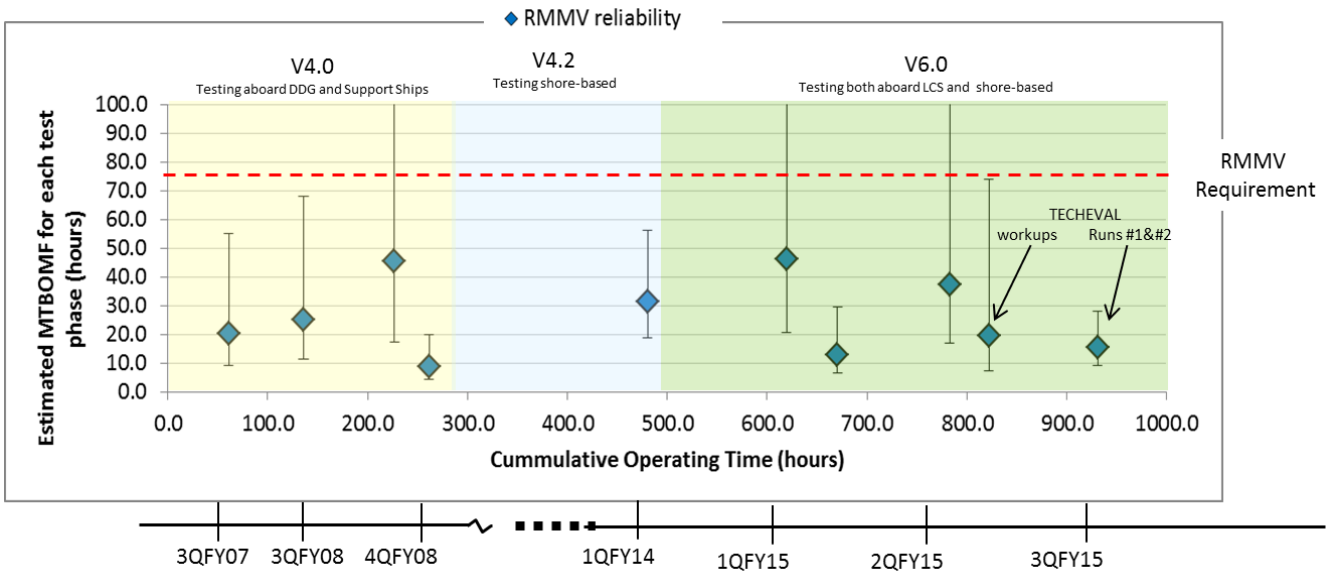


- **Long history of poor reliability plagued program development**
- **Reliability growth program began in FY11**

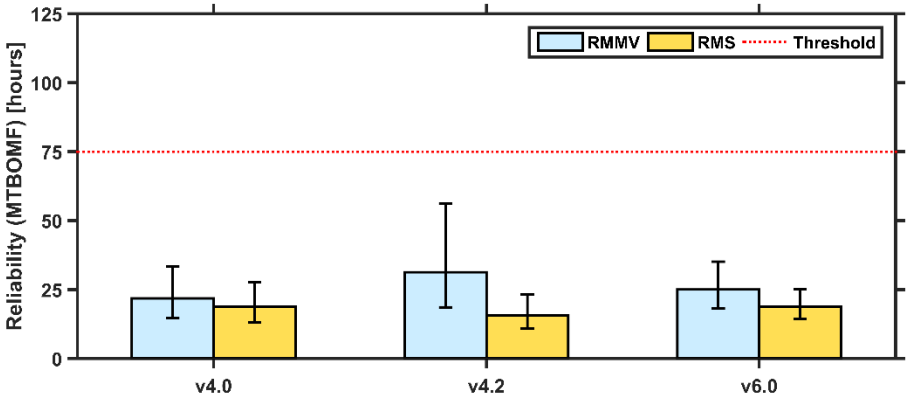


Remote Minehunting System

Quantifying the Precision of Estimates



- **No evidence of growth in reliability over time**
- **Formal statistical growth models used to confirm quantitatively**
- **Growth parameter = 0.02** (-0.38, 0.30)



Quantification of uncertainty essential for drawing conclusions about system performance

Systematic approach for Model & Simulation Validation

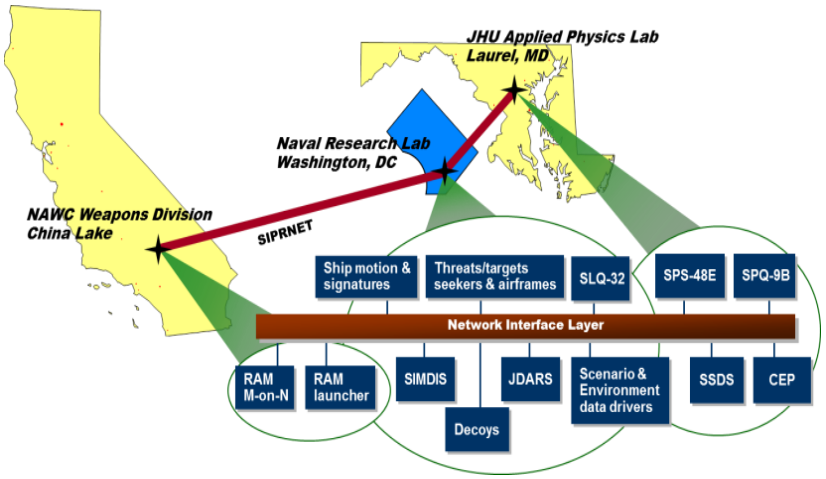
Do the Model & Simulation reasonably represent the real world?

Probability of Raid Annihilation (PRA) Testbed Validation

Analysts use the Navy's PRA Testbed to extend the results of live testing to threats not available on test ranges and other environmental conditions that may affect ship performance in defeating cruise missile attacks

In 1987, two Iraqi Exocets hit USS Stark, increasing the focus on ship self-defense

Cost and safety restrictions limit the number of live test events



PRA Testbed is complex federation of models that models the environment, threats, radars, missiles, and other systems

How do testers use <u>four</u> live test events to validate that the PRA Testbed provides meaningful results?

Systematic approach for Model & Simulation Validation

Does the Model & Simulation reasonably represent the real world?

Probability of Raid Annihilation (PRA) Testbed Validation

- **As part of validation, develop a statistical regression model of intermediate metrics:**

$$\textbf{Initial Detection Range}^{*} = \beta_0 + \beta_1 TestType + \beta_2 TestThreat + \beta_3 (TestType * TestThreat) + \epsilon$$

  - » *Test Type*: Are the data from a PRA testbed run or a live shot?
  - » *Test Threat*: Which cruise missile threat was presented?

Differences expected

Statistically distinguishable?

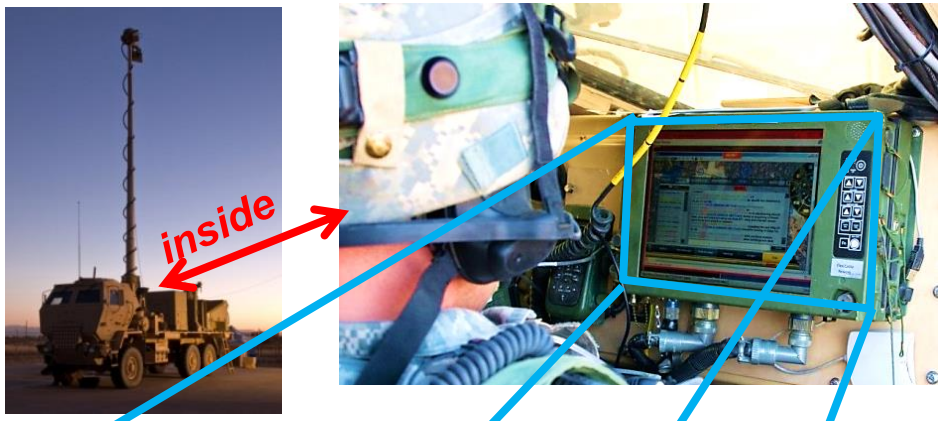
- **If more live data were available, other statistical techniques would be more appropriate**

Even when limited live data are available, statistical techniques can be informative

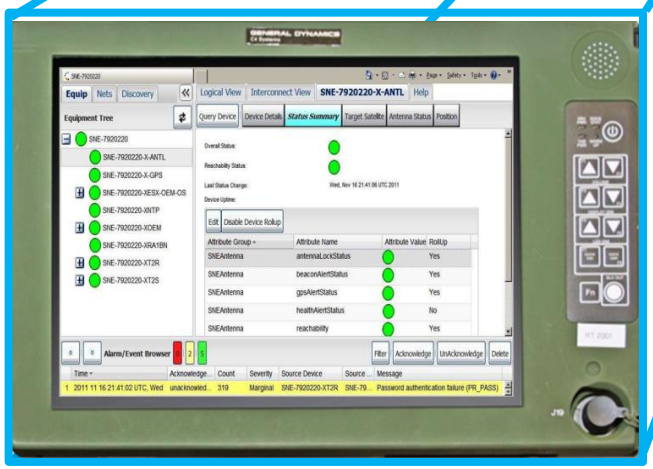
**Initial Detection range is just one of the many continuous metrics the will be used for validating the PRA Testbed*

Warfighter Information Network – Tactical (WIN-T)

WIN –T is a battlefield mobile Internet-like system providing Army Commanders with voice and data communications supporting command and control



- **Previous tests had focused on performance goals such as voice over internet protocol (VoIP) performance.**
- **Goal for the 2nd Follow-On Test & Evaluation (November 2014): Satisfy AT&L concerns on the usability, reliability and cyber security of the WIN-T system.**
  - Has the Army improved the start-up, shutdown and troubleshooting procedures?



OLD DISPLAY (Complex)

NEW DISPLAY (Simplified)

Warfighter Information Network – Tactical (WIN-T)

Old Survey Questions:

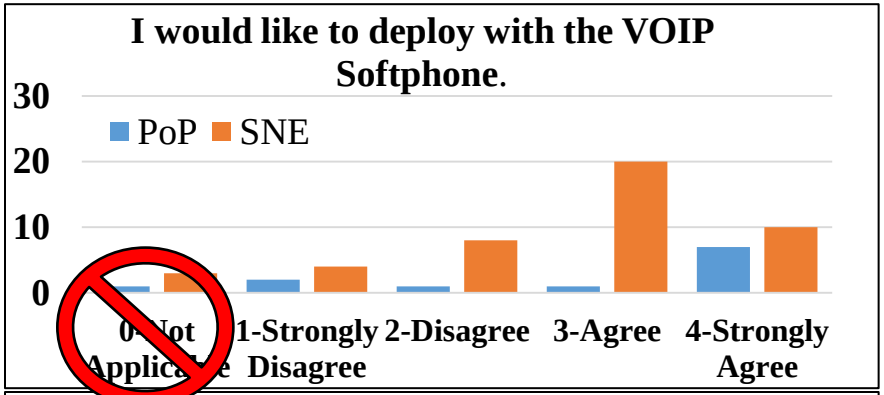
- Complex, long, branched questions
- Open ended responses
- Does not focus on subjective experience
- No data to report.

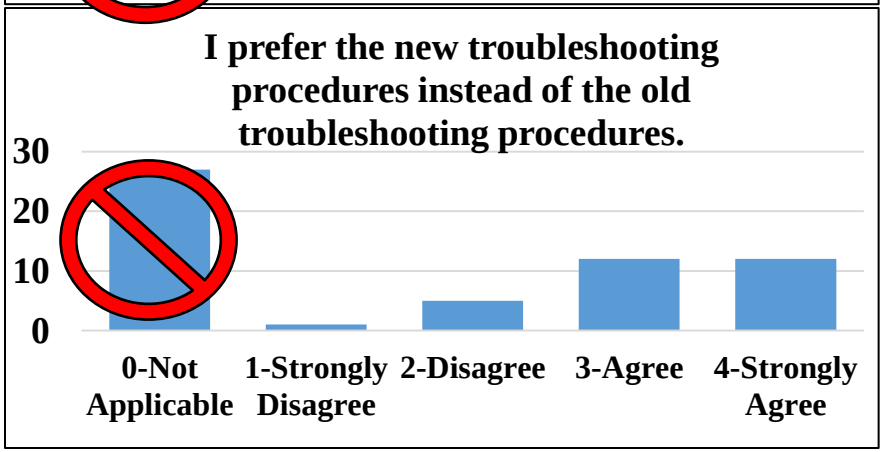
New Survey Questions:

- Simple questions
- Likert scale allows for statistical analysis

Example (poor) survey question:

"During movement were you able to communicate using the PoP? If there were times when you could not communicate, do you know why? What was the terrain? Approximately how fast were you going? [… question continues…]"





Working on future improvements:

- Use System Usability Survey (SUS)
  - Provides empirical comparison between WIN-T increments and to other systems
- Tailor surveys to group to remove "not applicable"

Reliability Growth Projection

Reliability Growth Potential Analysis

Two-seat Utility JLTVs (with JLTV trailer)

Four-seat Close Combat Weapons Carrier

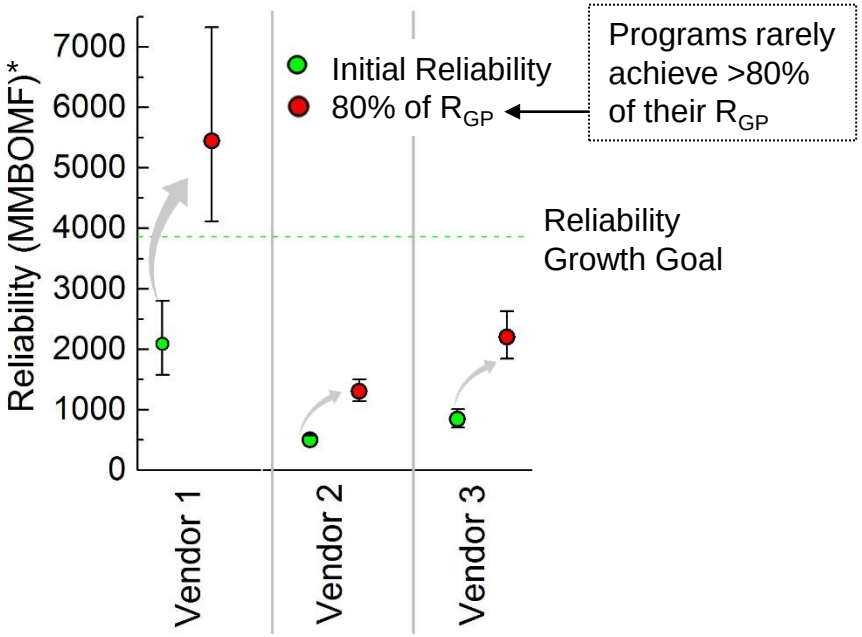
Reliability Growth Potential is one way to assess the feasibility of being able to reach reliability goals.

Reliability Growth Potential

$$R_{GP} = \frac{Initial\ Reliability}{1 - (FEF * MS)}$$

Based on JLTV Reliability Growth Plan:

- Fix Effectiveness Factor (FEF) = 0.73
- Management Strategy (MS) = 0.95



* The data shown here is notional since the actual data is source-selection sensitive information.

Why Use Rigorous Statistical Methods?

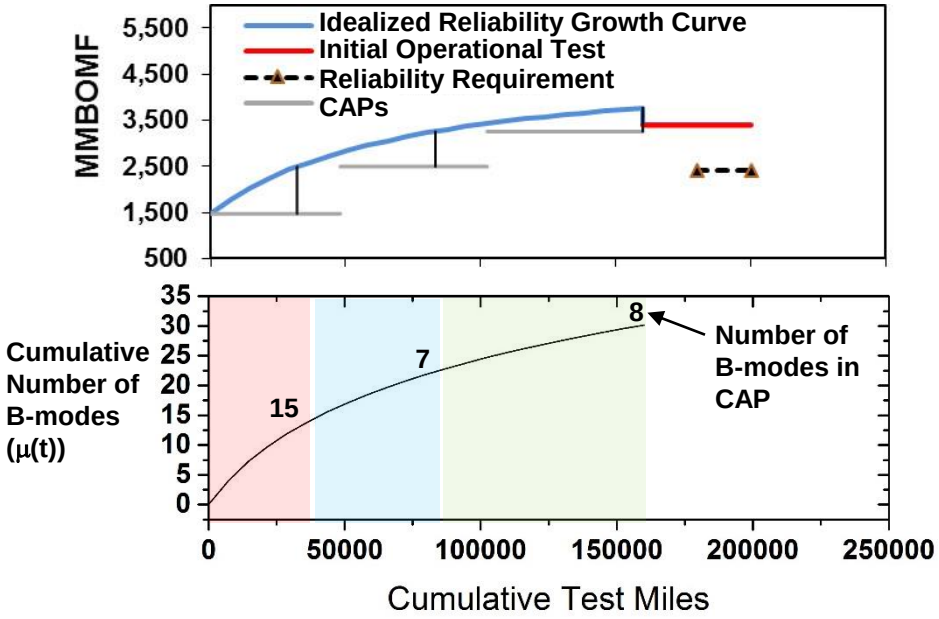
- **The answer seems obvious to me**
- **They Provide a Defensible Basis for Test and Evaluation**
  - Best practices used across many disciplines
  - Defensible test content
  - Defensible, informative evaluations

Reliability Growth Planning

Analysis of B-mode Discovery Rate

Four-seat Heavy Guns Carriers

JLTV EMD Reliability Growth Planning Curve



Reliability growth test duration should be long enough to support discovery and correction of failure modes.

CAP – Corrective Action Period
MMBOMF – Mean Miles Between Operational Mission Failures

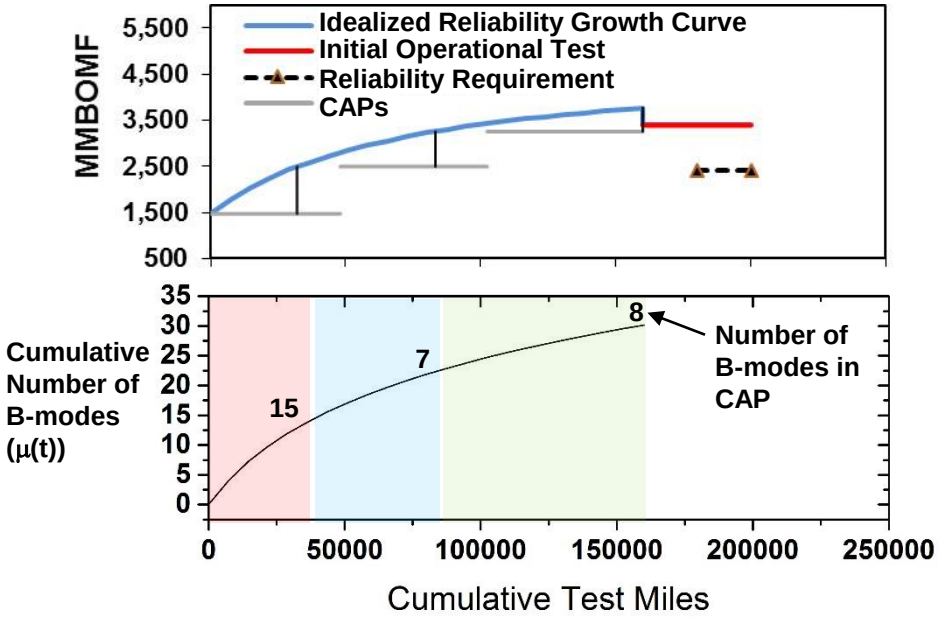
EMD – Engineering and Manufacturing Development

Reliability Growth Planning

Analysis of B-mode Discovery Rate

Four-seat Heavy Guns Carriers

JLTV EMD Reliability Growth Planning Curve



Reliability growth test duration should be long enough to support discovery and correction of failure modes.

CAP – Corrective Action Period
MMBOMF – Mean Miles Between Operational Mission Failures

EMD – Engineering and Manufacturing Development

Use of Rigorous Statistical Techniques Within DOD Acquisition

- **DOT&E Guidance Memos**
  - Experimental Design
  - Statistical Techniques
  - Surveys
  - Modeling & Simulation
- **TEMP Guidebook 3.0**
  - Updated with recent examples of successful implementation of this guidance
- **Service Test Agencies**
  - Increased staffing with technical backgrounds
  - Incorporation of DOE principals in test planning
- **Case Studies**