Statistical Thinking in DoD Test & Evaluation: F-35 Case Study

Dr. Laura Freeman
Improving Operational Testing: A case study from my past 8 years
Goal of Operational Test: Evaluate Operational Effectiveness, Suitability, and Survivability

Operational Environment

Representative Users

“Real” Threats

Conducting Missions
DoD Test Paradigm In Terms of Your New Corolla

Contractor Testing

Developmental Testing

Operational Testing

Test Timeline

Tend to be requirements driven
Requirements documents are often missing important mission considerations.
DoD Test Paradigm In Terms of Your New Corolla

Contractor Testing

Developmental Testing

Operational Testing

Test Timeline
DoD Test Paradigm In Terms of Your New Corolla

Contractor Testing

Developmental Testing

Operational Testing

Test Timeline
DoD Test Paradigm In Terms of Your New Corolla

Contractor Testing

Developmental Testing

Operational Testing

Test Timeline
Congress established DOT&E separate from the Services’ operational testing agencies

Department of Defense

Office of the Secretary of Defense

Director, Operational Test and Evaluation

Army
Navy & Marines
Air Force

Service Operational Testing Agencies
The goal of the experiment. This should reflect evaluation of end-to-end mission effectiveness in an operationally realistic environment.

Quantitative mission-oriented response variables for effectiveness and suitability. (These could be Key Performance Parameters but most likely there will be others.)

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

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

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

1. Establish a sense of urgency
2. Form a powerful coalition
3. Create a vision
4. Communicate the vision
5. Empower others to act
6. Create short term wins
7. Consolidate improvements and produce more change
8. Institutionalize new approaches
Project Campions

Rigor and Objectivity in T&E: An Update
J. Michael Gilmore, Ph.D.
Director, Operational Test and Evaluation,
Office of the Secretary of Defense, Washington, D.C.

The Director of Operational Test and Evaluation (OT&E) began four Test and Evaluation (T&E) initiatives after his confirmation by Congress in fall 2009. Underlying his four initiatives were the need for rigorous and objective T&E. Since his original initiatives the Director has advocated for the use of statistically designed experiments as a methodology for increasing the rigor of test planning resulting in efficient tests yielding statistically defensible results. Additionally, he continues to emphasize the need for reliable systems and reliability growth plans and accordingly defensible reliability growth models in T&E.

I began my term as the Director of Operational Test & Evaluation (DOT&E) with four initiatives to increase scientific rigor in T&E. I published those initiatives in the June 2010, ITEA Journal, and I am happy to use this opportunity to provide an update. During the past year, I have seen several success stories as well as areas for improvement. I would like to commend ITEA for the theme of this journal, “The Rigor of the Scientific Method.” And I appreciate the many articles others have authored on applying rigorous and objective scientific approaches to their specific test challenges.

J. Michael Gilmore, Ph.D.

The DOE associate with the test results. Finally, DOE provides the tester with methods for developing and analyzing sequences of tests. Before testing, DOE enables decision makers to clearly see the tradeoffs between test resources and risk. During testing, DOE enables testers to use early results to strengthen and refine subsequent tests. After testing, DOE gives decision makers a framework for understanding and weighing the importance of the results.

In October 2010, I outlined the specific elements of DOE that I am looking for when I review TEMPs and test plans.
Rigor and Objectivity in T&E: An Update

J. Michael Gilmore, Ph.D.
Director, Operational Test and Evaluation,
Office of the Secretary of Defense, Washington, D.C.

The Director of Operational Test and Evaluation (OT&E) began four Test and Evaluation (T&E) initiatives after his confirmation by Congress in fall 2009. Underlying his four initiatives were the need for rigorous and objective T&E. Since his original initiatives, the Director has advocated for the use of statistically designed experiments as a methodology for increasing the rigor of test planning resulting in efficient tests yielding statistically defensible results. Additionally, he continues to emphasize the need for reliable systems and reliability growth plans and accordingly defensible reliability growth models in T&E.

I began my term as the Director of Operational Test & Evaluation (DOT&E) with four initiatives to increase scientific rigor in T&E. I published those initiatives in the June 2010, ITEA Journal, and I am happy to use this opportunity to provide an update. During the past year, I have seen several success stories as well as areas for improvement. I would like to commend ITEA for the theme of this journal, “The Rigor of the Scientific Method.” And I appreciate the many articles others have authored on applying rigorous and objective scientific approaches to their specific test challenges.

Finally, DOE provides the tester with methods for developing and analyzing sequences of tests. Before testing, DOE enables decision makers to clearly see the tradeoffs between test resources and risk. During testing, DOE enables testers to use early results to strengthen and refine subsequent tests. After testing, DOE gives decision makers a framework for understanding and weighing the importance of the results.

In October 2010, I outlined the specific elements of DOE that I am looking for when I review TEMPs and test plans.
Strategic Plan

- Education & Training
- Case Studies & Best Practices
- TEMP, Test Plan, Report Guidance
- Advisory Board

Institutionalize Test Science & Rigor in T&E
Design of Experiments for Test Planning
F-35 Case Study
The F-35 Program is Complex even by DoD Standards
And Required to Accomplish Many Diverse Missions

<table>
<thead>
<tr>
<th>Mission Areas</th>
<th>Air Threat</th>
<th>Ground Threat</th>
</tr>
</thead>
<tbody>
<tr>
<td>Air-Surface</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Strike</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Destruction/Suppression of Enemy Air</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Defenses</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Defensive counter air</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Offensive counter air</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Close air support</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Search and rescue</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Problem Identification

How do you evaluate the F-35’s ability to accomplish a diverse set of operational missions with limited test resources?
Characterization across operational envelope - Strike, Offensive Counter Air, and Destruction/Suppression Enemy Air Defense
Characterization across operational envelope - Strike, Offensive Counter Air, and Destruction/Suppression Enemy Air Defense
Characterization across operational envelope - Strike, Offensive Counter Air, and Destruction/Suppression Enemy Air Defense
Characterization across operational envelope - Strike, Offensive Counter Air, and Destruction/Suppression Enemy Air Defense
Characterization across operational envelope - Strike, Offensive Counter Air, and Destruction/Suppression Enemy Air Defense
Characterization across operational envelope - Strike, Offensive Counter Air, and Destruction/Suppression Enemy Air Defense
Characterization across operational envelope - Strike, Offensive Counter Air, and Destruction/Suppression Enemy Air Defense
Characterization across operational envelope - Response Variables

Lots of measures to capture:

Mission outcomes

Air to Air Performance

Air to Surface Performance

System sensor capabilities

Targeting Accuracy Striker
Striker First Track Range
Striker First Hostile Declaration Range
Striker First Shot Range
Red Air First Detection Range
Red Air First Shot Range
Striker SAM Track Time
Proportion of Valid Weapon Releases to Number of Valid Weapon Releases Required to Meet Mission Tasking
Proportion of Assigned Air to Surface Targets Removed
Proportion of Striker Kill Removed
Striker to Red Air Exchange Ratio
Geolocation Find Time
Fix Time
DEAD Time
Targeting Accuracy Escort
Escort SAM Track Time
Proportion of Assigned SAM Elements Removed
Proportion of Assigned SAM Elements Engaged Exchange Ratio
Closest Red Air Range to Strike Package
Blue Striker Encroachment Range
Escort First Track Range
Escort First Hostile Declaration Range
Escort First Shot Range
Red Air First Detection Range
Red Air First Shot Range
Proportion of Escort Blue Strikers that reach their Weapons Release Point
Proportion of Protected Aircraft (Strikers) Not Kill Removed
Proportion of Escort F-35 Kill Removed
Escort to Red Fighter Exchange Ratio
Experimental designs determine test adequacy

24 Run, D-Optimal 2nd Order Design

Disallowed Combinations
Two mission designs, executed in a 5\textsuperscript{th} generation scenario
Power calculations provided justification for number of trials

**Diagram:**

- **Power**: Y-axis
- **Signal-to-Noise Ratio**: X-axis
- Lines:
  - Target location power
  - Variant power
  - Environment power (in/out of band)

The graph illustrates how power changes with different signal-to-noise ratios, highlighting the significance of power calculations in determining the number of trials required.
We took a scientific approach to all operational testing

### Mission Areas

- **Air-Surface**
- **Strike**
- **Destruction/Suppression of Enemy Air Defenses**
- **Defensive counter air**
- **Offensive counter air**
- **Close air support**
- **Search and rescue**

### Air Threat

- **Carrier variant**
- **Short takeoff/vertical landing**
- **Conventional**

### Ground Threat

- **Carrier variant**
- **Short takeoff/vertical landing**
- **Conventional**
Impact so far

Congressional review of Close Air Support Testing
Still to come

Test Execution and Analysis

Execution Considerations
• Challenges with aircraft availability
• Confounding variables

Analysis Considerations
• Demand for quick answers
• Big Data, Little Information
Statistical Engineering Shortcomings

Initial focus was on tools

Processes are still highly dependent on individuals involved

Adherence to statistical rules

Leadership changes & final solution not fully deployed

Failing to see the big picture
We continue to increase the statistical defensibility of DoD Test and Evaluation
Needed a larger focus for statistical engineering efforts

“Shift Left” to reduce late discovery by emphasizing mission context throughout acquisition

“Synchronized” Test Continuum

EXPERIMENTAL TESTING (Formative Evaluation)

Research, Experimentation & Prototyping

- Mission context analysis of system designs
- Goal: Identify design elements that will drive mission critical capability

CAPABILITIES FOCUSED SYSTEM ENGINEERING

Test

SE

CAPABILITIES FOCUSED EARLY TESTING

- Targeted testing to components and subsystems that enable mission critical capabilities
- Goal: Inform evolution of component and system design

Test

SE

INTEGRATED TESTING

- Efficient system level testing incorporating mission and user context
- Goal: Inform evaluations of system capabilities to support users accomplishing missions across the operational envelope

Test

SE

Operational Testing Confirmatory

Integrates Credible Data into Operational Evaluations
Thank you!
Innovation Adoption

I consistently meet brilliant, creative, entrepreneurial people in DoD with novel and implementable ideas, but they are fighting against entrenched processes and regulations that—in some cases—haven’t been modified in decades. Incentives are often misplaced. Decision-making seems surprisingly diffuse for an organization known for its hierarchical structure and decisive leaders. Some of these intrapreneurs find workarounds to inflexible systems or receive temporary shelter under a like-minded commander; far more do not. Even the most senior leaders described responsibilities being so intricately nested across the organization that a sense of true ownership proved elusive to them. Early on, I reached a fundamental conclusion that has been borne out over time: DoD does not have an innovation problem; it has an innovation adoption problem.

Dr. Eric Schmidt,
Testimony to House Armed Services Committee
April 17, 2018
Laura’s conjecture

Statistician’s are uniquely equipped to lead & implement change, especially in data-centric fields!