Increasing Statistical Rigor in Operational Test and Evaluation

Dr. Michael Gilmore

Director, Operational Test & Evaluation

Keynote Address to the NASA Statistical Engineering Symposium

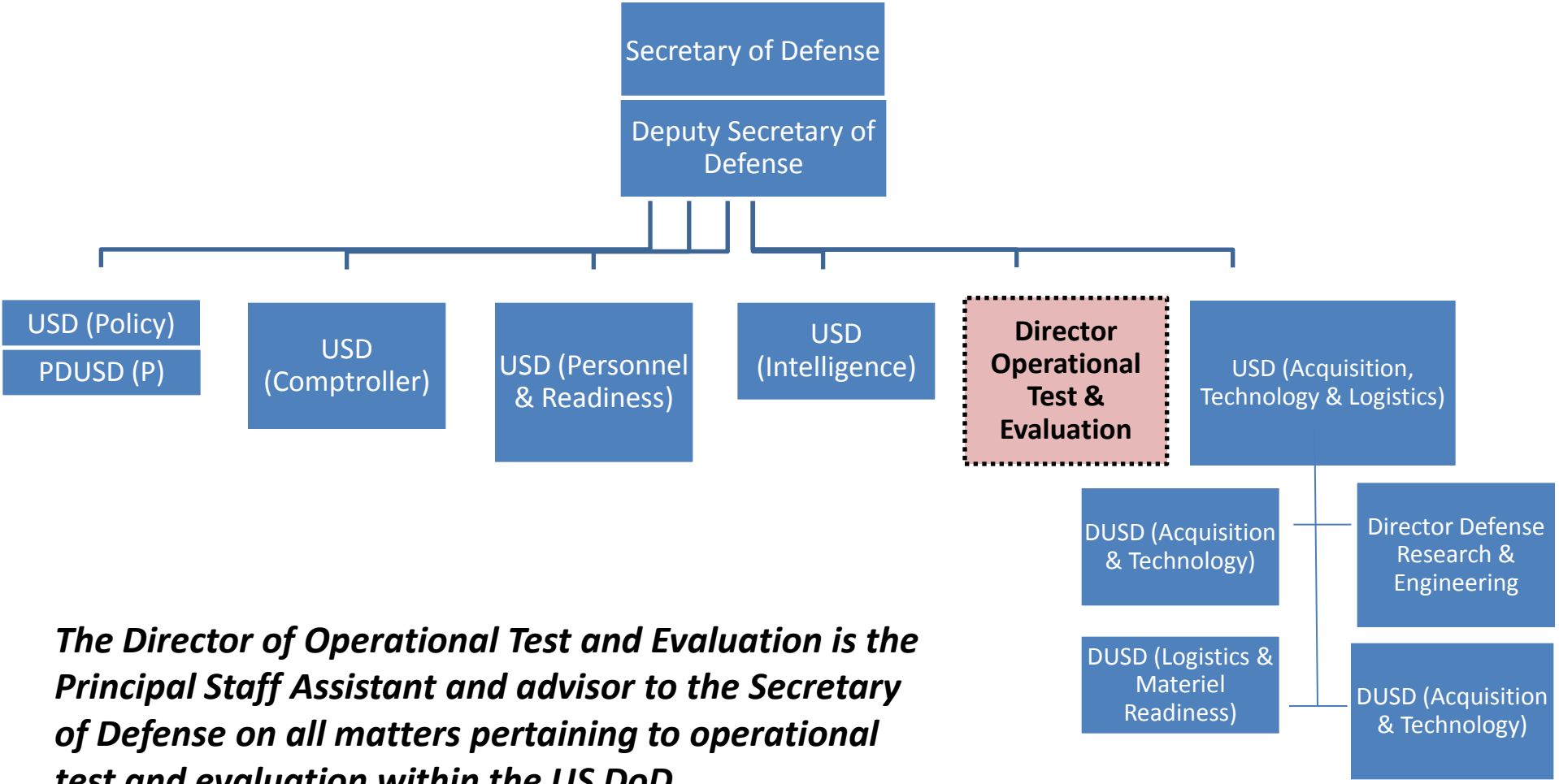
Outline

- DOT&E Background
- National Research Council Study
  - Statistics, Testing, and Defense Acquisition (1998)
- Current DOT&E Initiatives
- Increasing Statistical Rigor in Test & Evaluation
  - Design of Experiments
  - Reliability Growth

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.

Office of the Secretary of Defense

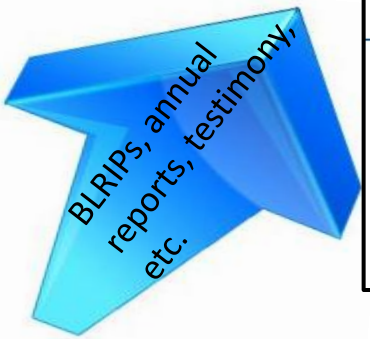


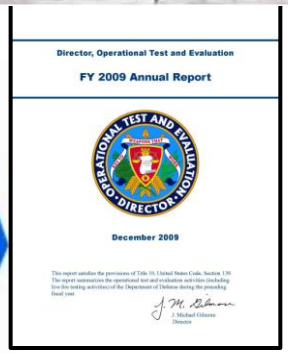
The Director of Operational Test and Evaluation is the Principal Staff Assistant and advisor to the Secretary of Defense on all matters pertaining to operational test and evaluation within the US DoD.

DOT&E 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



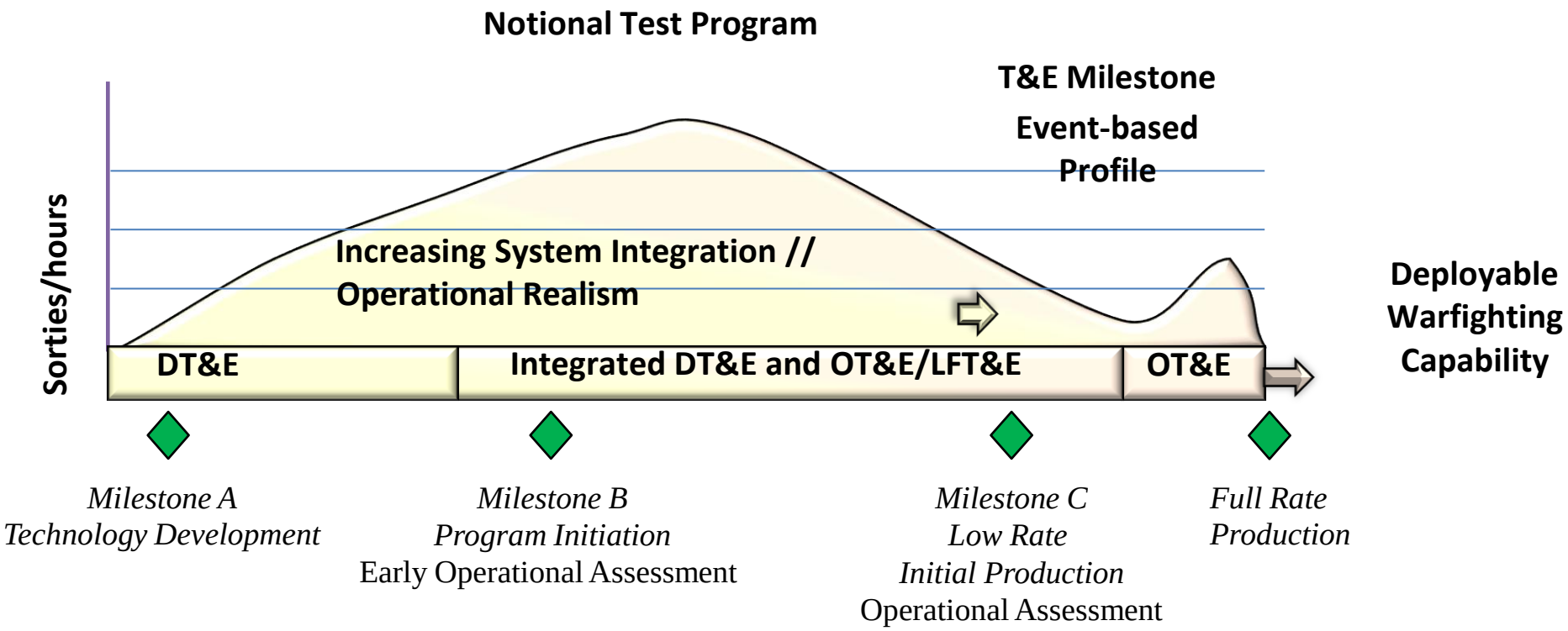


DOT&E Responsibilities

- Prescribe Department of Defense policy for:
  - Operational Test & Evaluation (OT&E)
  - Live Fire Test & Evaluation (LFT&E)
- Provide guidance on all OT&E and LFT&E matters
- Monitor & review all OT&E and LFT&E
- Report annually to Congress on OT&E and LFT&E
- Member of Defense Acquisition Board
- Approve test plans for OT & LF oversight programs
- Report on programs, before full-rate production decision to the Secretary, OSD, Services, & four congressional committees:
  - Adequacy operational and live fire testing
  - Operational Effectiveness
  - Operational Suitability
  - Survivability and Lethality

Acquisition Timeline

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



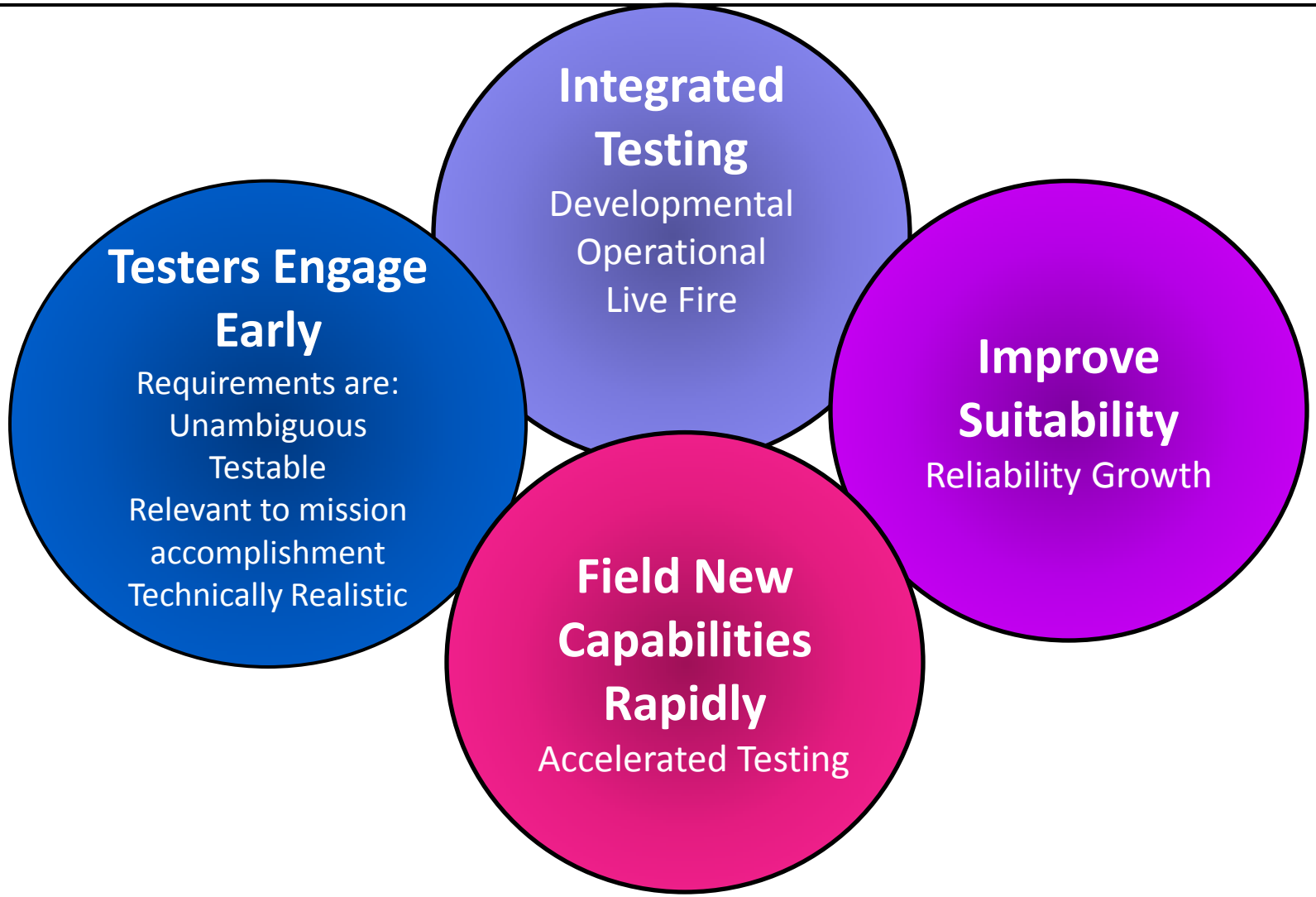
Why is a Program Placed on DOT&E OT Oversight?

- Statutory requirement for Major Defense Acquisition Programs (MDAPs)
  - Designated by the Secretary of Defense, or
  - Estimated by the Secretary of Defense to require an eventual total expenditure for research, development, test, and evaluation of more than $300,000,000 (based on fiscal year 1990 constant dollars), or
  - Estimated to have an eventual total expenditure for procurement of more than $1,800,000,000 (based on fiscal year 1990 constant dollars).
- Congress or OSD has expressed a high level of interest
- Congress has directed DOT&E report as condition for production or progress
- Program requires joint or multi-Service testing
- Program has a close relationship or is key to a major program
- Militarily significant change to system

National Research Council Study (1998)

- Panel on Statistical Methods for Testing and Evaluating Defense Systems
  - Examine statistical techniques & make recommendations
- Statistics, Testing, and Defense Acquisition: New Approaches and Methodological Improvements
- Select conclusions & recommendations
  - Conclusion 3.1: Major advances can be realized by applying selected industrial principles and practices in restructuring the paradigm for operational testing…
  - Conclusion 4.1: 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.
  - Recommendation 4.2: All estimates of the performance of a system from operational test should be accompanied by statements of uncertainty through use of confidence intervals…
  - Recommendation 4.4: 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.
  - Recommendation 7.4: 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 still not been implemented 13 years later.

DOT&E Initiatives



DOT&E Initiatives & Increasing Test Rigor

- Integrated Testing
  - "… DOE provides the scientific and statistical methods needed to rigorously plan and execute tests and evaluate their results. … 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.
  - "DOE should allow DOT&E to make statements of the confidence levels we have in the results of the testing. 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."

Apply DOE Across Entire Acquisition Development Cycle

Design of Experiments (DOE)

- Test planning is a science
- DOT&E must evaluate test plan adequacy
  - TEMP
  - IOT&E
- Statistics equips us to determine:
  - Breadth of coverage
  - Power
  - Confidence
- 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 manipulate factors & levels (inputs) to influence the output.

Measures of test plan adequacy!

DOE is a scientific tool for developing robust test plans!

DOT&E Retrospective Case Studies

- Motivation: get beyond general concepts, look at real world examples
- Goal: provide baseline and highlight areas for improvement
- Conducted an analysis of select Beyond Low Rate Initial Production (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

Case Studies – BLRIP Reviews

- USS Virginia
- Guided Multiple Launch Rocket System (GMLRS) Unitary
- MH-60R and MH-60S
- **Stryker Mobile Gun System (MGS)**
- **Joint Chemical Agent Detector (JCAD)**
- DoN Large Aircraft Infrared Countermeasures (LAIRCM)
- EA-18G

Stryker 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

Stryker Mobile Gun System (MGS)

C pe

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

4 Factors: Mission Ty 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

Impact of Experimental Design

- 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: N	16	72	36	16
Confidence: (1-α)	Set to the same level across all 4 designs: Confidence = 95%			
1σ - Power (1-β)	8.1% - 28.0%	68.1% - 98.7%	35.0% - 81.5%	12.3% - 39.9%

- The case study suggests that 16 runs is far from 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 Case Design vs. D-Optimal Design

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

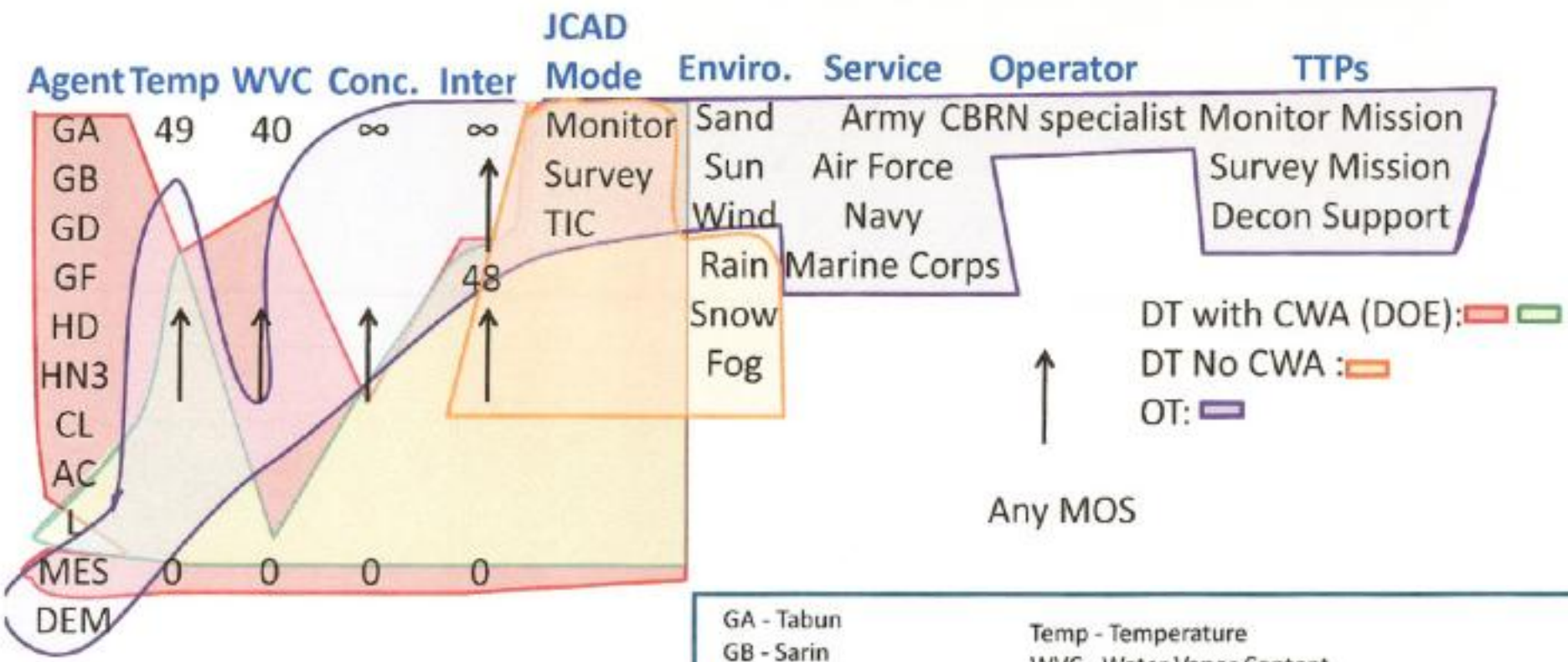
Statistical D-Optimal Design

		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

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)

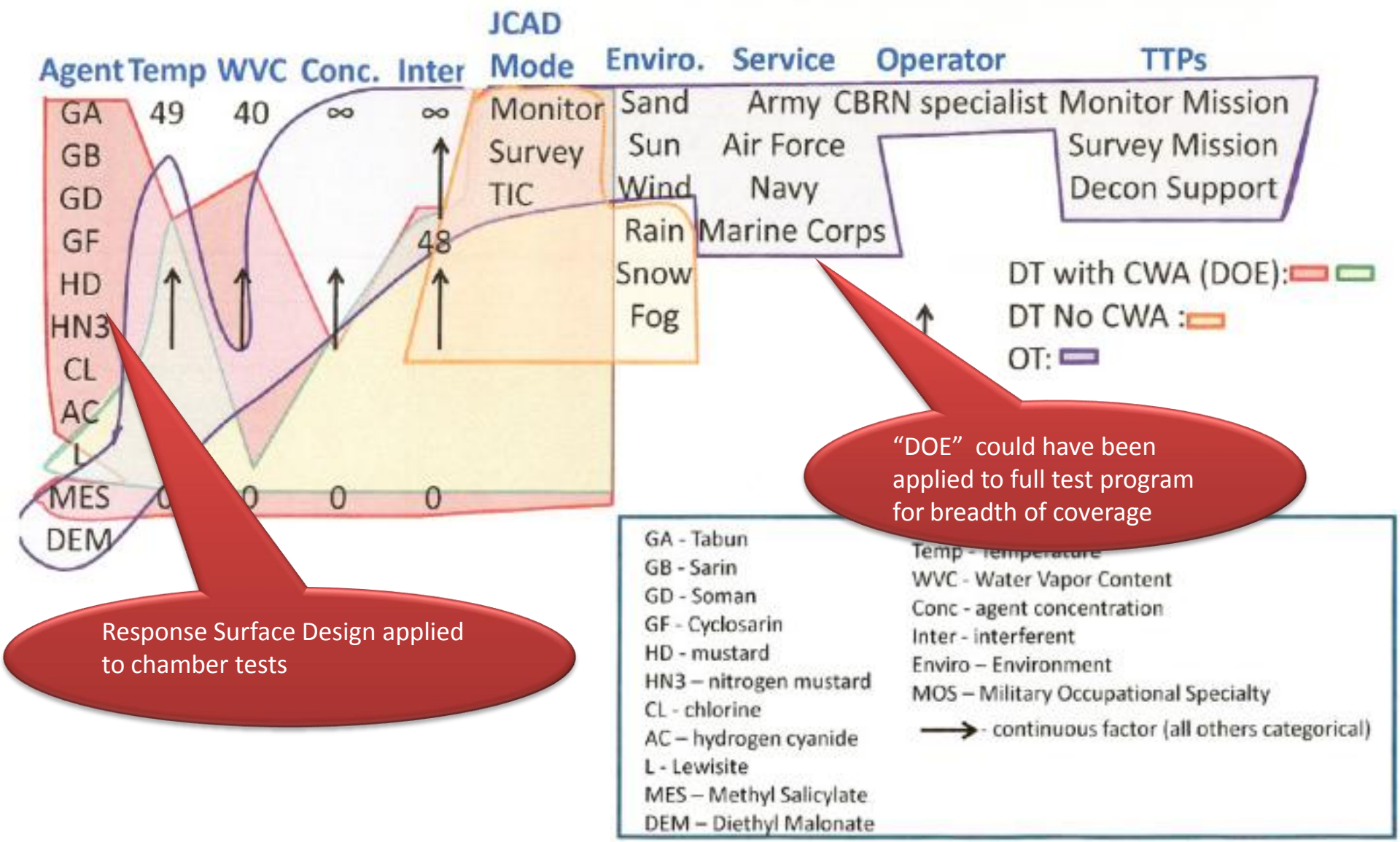
Joint Chemical Agent Detector (JCAD) Coverage of Operational Envelope



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



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

Lessons Learned – Case Studies

- There is no “one size fits all” approach for designing experiments.
- Review of past BLRIPS illustrated the need for a more rigorous approach to DT/OT incorporating statistical experimental design.
- Defining measurable, testable, important response variables is key to a good design.
- T&E already thinks about factors and levels (input variables); however we can improve on a scientific approach to varying factors & levels.
- DOE can be applied in both DT and OT. However, responses and factors may differ between the two.
- Confidence and power add important information to our assessment of the test adequacy and test results.

The Strength of Designed Experiments

- We face the following constraints in testing:
  - Limited Test Resources
    - Range Size
    - Forces (Manpower & Materiel)
  - Limited Test Time
    - Fielding & Production Schedules
    - Range Availability
  - Limited Test Articles
    - Unit Cost
    - Production Time
- Design of experiments allows us to understand the tradeoffs these constraints impose.
- Design of experiments can provide a statistically optimum allocation of our assets under given constraints.

DOT&E Guidance on Design of Experiments

- DOT&E Memorandum, October 19, 2010
- "Design of Experiments is a structured process to identify the metrics, factors, and levels that most directly affect operational effectiveness and suitability…"
- Elements of experimental design for TEMPs and Test Plans approval:
  - The goal of the experiment.
  - Quantitative mission-oriented response variables
    - effectiveness and suitability
  - Factors [and levels] that affect response variables
  - A method for strategically varying factors
    - across both developmental and operational testing
  - Statistical measures of merit
    - power and confidence on the relevant response variables for which it makes sense.
    - Important to understand "how much testing is enough?"

Challenges to Integrated Testing using DOE

- Current policy documents do not address the use of statistics during T&E planning or analysis.
  - No Policy → No funding
- Program Managers decide how much and where DT is done
  - Makes planning sequential experiments challenging.
- Developmental testing has not been sufficient or adequate
  - OT&E results indicate a Department-wide problem
  - Seeing more weapons systems not ready for IOT&E and combat.
- Congress recently created a Director of Developmental Test and Evaluation (DDT&E)
  - DDT&E has not yet provided guidance on the application of statistical methods in DT