D-4548 non-standard
H12-000110 OED

Statistically Based T&E Using Design of Experiments

79th Executive Committee of Range Commanders Conference

7 February 2012

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

IDA

Approved for public release; distribution is unlimited.

IDA

Outline

- Why Design of Experiments?
- History of Design of Experiments in T&E
- DOT&E Guidance
- Implications for Testing
- When to apply DOE?
- Training Implications
- Conclusions & Next Steps

Approved for public release; distribution is unlimited.

IDA

Rationale for DOE

- **The purpose of testing is to provide relevant, credible evidence with some degree of inferential weight to decision makers**
  - DOE provides a framework for the argument and methods to help us do that systematically
- **Four Challenges faced by any test**
  1. *How many? Depth of Test*
  2. *Which points? Breadth of Testing* – spanning the operational envelope
  3. *How to execute? Order of Testing*
  4. *What conclusions? Test Analysis* – drawing objective, robust conclusions while controlling noise
- **DOE in conjunction with operational expertise effectively addresses all these challenges!**
- **DOE Provides:**
  - the most powerful allocation of test resources for a given number of tests.
  - a scientific, structured, objective way to plan tests.
  - an approach to integrated test.
  - a structured, mathematical analysis for summarizing test results.

DOE changes "I think" to "I know"

Approved for public release; distribution is unlimited.

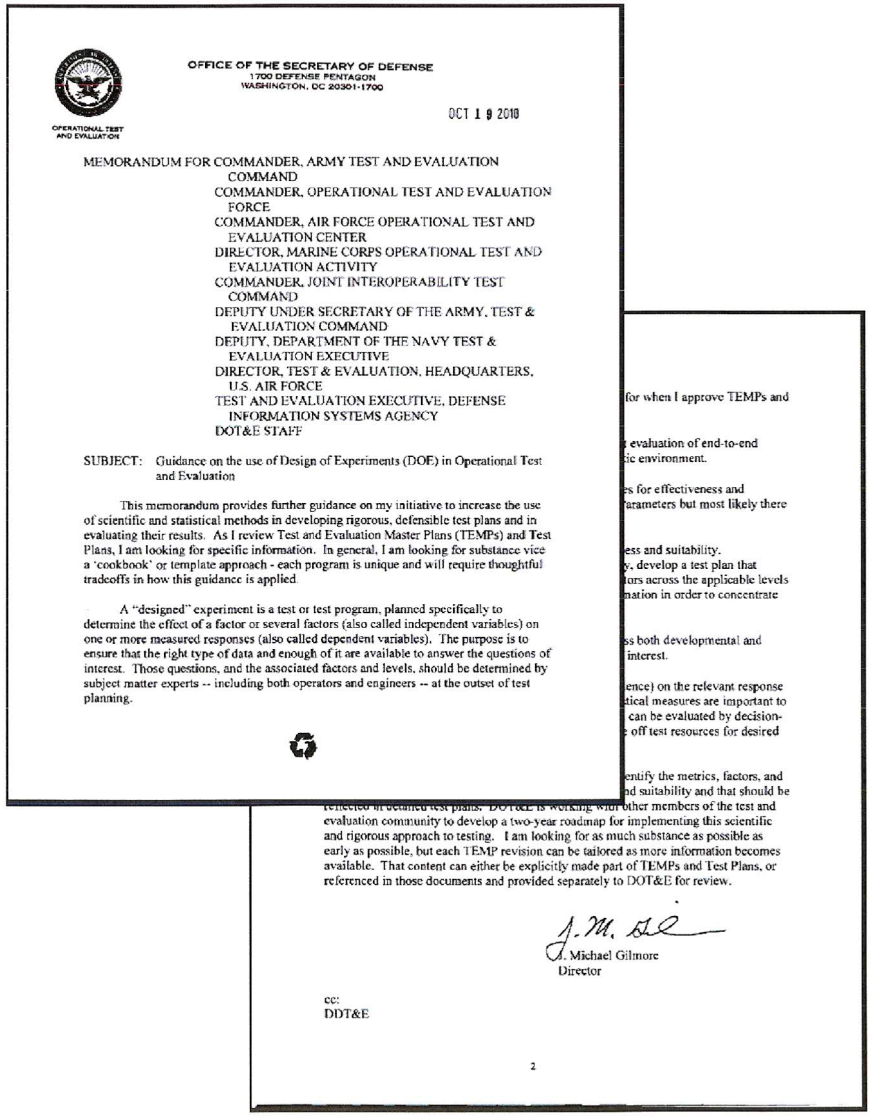
A Brief and Selective History of DOE in T&E

- **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."

Approved for public release; distribution is unlimited.

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
rameters but most likely there

ess and suitability.
y, develop a test plan that
urs 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
d suitability and that should be
reflected in defined 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

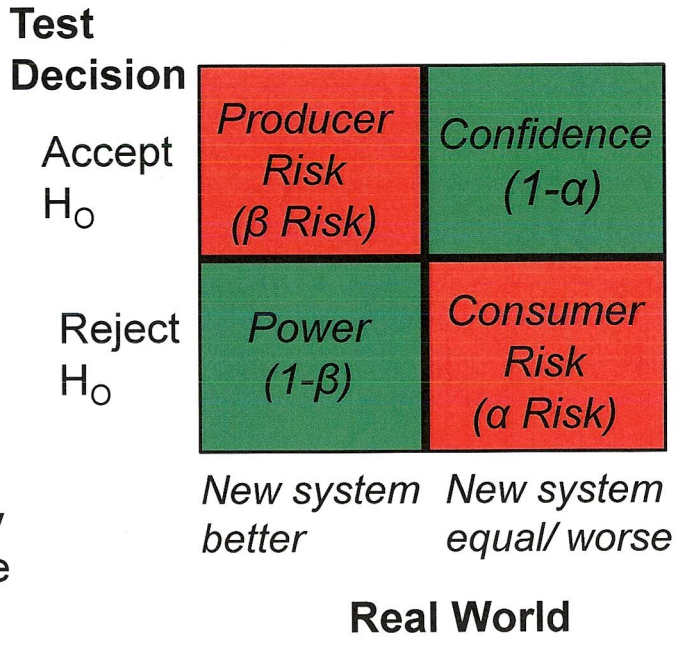
- **The goal of the experiment**. This should reflect evaluation of end-to-end mission effectiveness in an operationally realistic environment.
- Quantitative mission-oriented **response variables** for effectiveness and suitability. (These could be Key Performance Parameters but most likely there will 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.

Approved for public release; distribution is unlimited.

IDA

Power and Confidence

- **DOD 5000:** "acquire quality products that satisfy user needs with **measurable improvements** to mission capability and operational support"
- **Statistical Hypothesis Test:**
  - H_O: New system equal to or worse than the legacy system
  - H_A: New system **better** than the legacy system
- **Confidence**
  - Confidence Level – the probability we make the right decision based on the test data. Typically confidence tells us the probability a test concluded a systems is bad, when it truly is a bad system.
- **Power**
  - Similar to confidence level, power is the probability we make the right decision. Typically, power is the probability that a test concluded a system is good, when it truly is a good system.



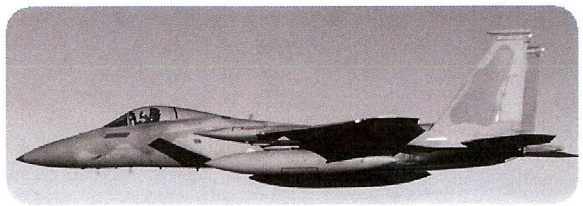
We need to understand risk!

Approved for public release; distribution is unlimited.

IDA

Implications for Testing
Operational Response Variables

- **Response Variables: maximize information obtained in a given test**
- **Example: Measuring the effectiveness of an onboard aircraft countermeasures system.**
  - Hypothetical requirement: probability of hit is less than 10%
  - Possible response variables:
    - » Missile hit/miss – hit is defined as the missile lethality radius intersects aircrafts path
    - » Missile miss distance
- **We can leverage the relationship between hit/miss and miss distance!**
  - Hit occurs if:
    - » Miss distance < lethality radius



Miss Distance

Approved for public release; distribution is unlimited.

IDA

Implications for Testing
Operational Response Variables

- **Factors impacting missile effectiveness:**
  - Countermeasures – yes/no
  - Aircraft altitude
  - Aircraft speed
  - Aircraft maneuvering – yes/no
- **A simple experimental design: 2^4 full-factorial**
  - Single replicate = 32 engagements

Jammer DOE Matrix

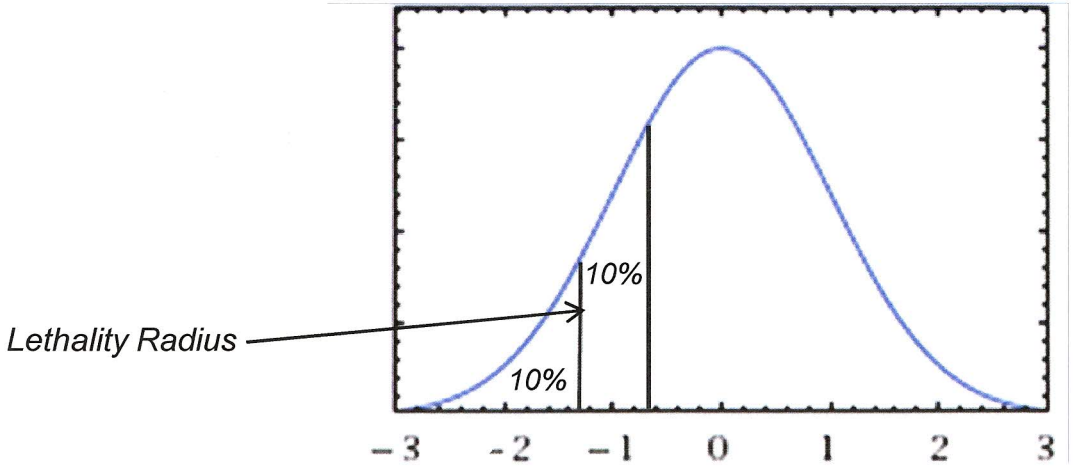
CM	Altitude	Non-maneuvering		Maneuvering		CM	Altitude	Non-maneuvering		Maneuvering	
		NM		M1	M2			NM		M1	M2
Yes	5000 ft	1	1	1	1	No	5000 ft	1	1	1	1
		1	1	1	1			1	1	1	1
	10000 ft	1	1	1	1		10000 ft	1	1	1	1
		1	1	1	1			1	1	1	1
Totals		4	4	4	4			4	4	4	4

Approved for public release; distribution is unlimited.

IDA

Implications for Testing
Operational Response Variables

- **What sample size is adequate?**
- **Power Analysis**
  - Hit/Miss: 10 replicates to determine a 10% change across factors
  - Miss distance: 6 replicates to determine an equivalent 0.44σ change across factors
  - 40% reduction in cost for the same information

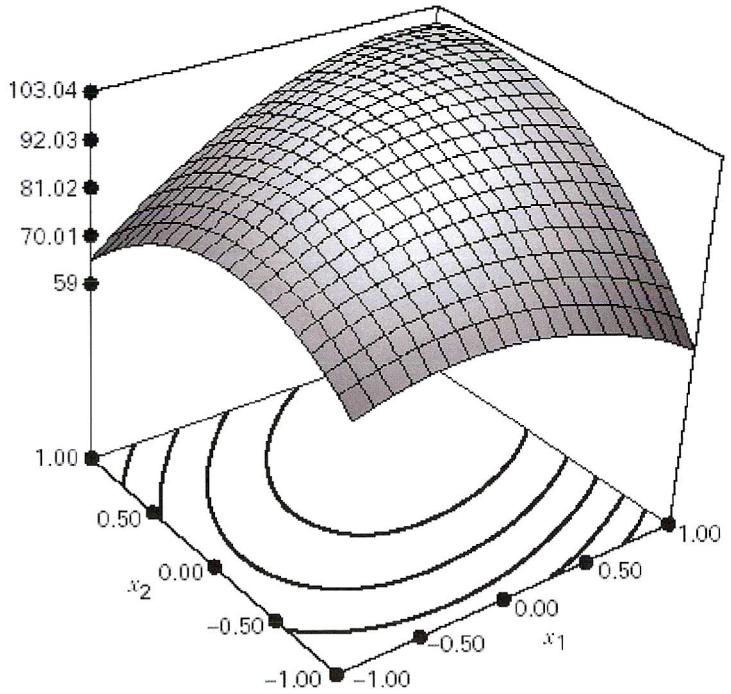


Approved for public release; distribution is unlimited.

IDA

Implications for Testing
Factors and Levels

- **Goal of testing changes under DOE**
  - Understand relationship between test conditions and outcomes
  - Not just a one number summary statistic
- **Strategic method for spanning the operational envelope**
  - Coverage of all levels of the operational envelope is key
- **Continuous factors are beneficial**
  - Allows for predictions at levels not tested
  - Supports efficient testing
- **Principles of DOE**
  - Randomization
    - » Ability to change test conditions
  - Replication
  - Local control of error
    - » Minimize measurement error
    - » Linking test conditions to responses



(a) Response surface

Approved for public release; distribution is unlimited.

IDA

When to Apply DOE

Test Objective	Testing Phase	Design Method	Analysis Method	Applications
Design for Reliability	Milestone A/B CT & early DT	Robust Parameter Design Taguchi Arrays, Orthogonal Arrays, Response Surface Methodology	Dual Response Models, Response Surface Models	Design for Reliability & designing products robust to varying operating environments.
Evaluation of material properties	Milestone A/B CT & early DT	Accelerated Life Tests, Mixture Designs	Accelerated Life Models, Mixture Models, Life Models	Accelerated Life Tests
Process optimization	Milestone A/B CT& early DT	Response Surface Designs, Optimal Designs	Response Surface Models	Trade Studies and Engineering Analyses
Test for problems	Milestone B/C DT, IT, OT	Combinatorial Designs, Orthogonal Arrays, Space Filling Designs	CART, Statistical Models if applicable (orthogonal array & space filling)	Software Testing Integration and Interoperability Testing
Screen for important factors	Milestone B/C DT, IT, OT	Super-Saturated Designs, Factorial and Fractional Factorial Designs, Optimal Designs	ANOVA, Regression, Response Surface Models	Characterizing Performance, Product design and development Screening factors in DT, IT to optimize and/or reduce OT testing.
Characterize a system or process over an envelope	Milestone C & Beyond IT & OT	Factorial and Fractional Factorial Designs, Response Surface Designs, Optimal Designs	ANOVA, Regression, Response Surface Models	Characterizing Performance

Approved for public release; distribution is unlimited.

IDA

Implications for Training

- **Statistical Thinking**
  - Knowledge of the system lays the foundation
  - All work occurs in a system of interconnected processes
  - Variation exists in all processes
  - Understanding, (accounting for), and reducing variation are key for success (understanding performance, improving quality, making informed decisions, etc.)

- **Introductory Statistics & Design Course lay a solid foundation**
  - Execution is key for internalization of knowledge

- **Available Training**
  - Air Force classes available through AFIT, Edwards AFB, Eglin AFB and AFOTEC SAF AQ
  - Army – Introduction to Probability and Statistics
  - DAU – Probability and Statistics CLM
  - JITC – DOE for software intensive systems in coordination with NPS

Approved for public release; distribution is unlimited.

IDA

Moving Forward in 2012

- **DOT&E Goals**
  - Increase emphasis on Integrated Testing
  - Engage requirements community
  - Improve designs
    - » Continuous metrics and factors
    - » Expand beyond full factorial
  - Capitalize on all test data using advanced statistical analysis techniques (including Bayesian)

- **DASD (DT&E) Goals**
  - Initiate Scientific Test and Analysis Techniques (STAT) Implementation Plan
    - » Signed this February (By DASD(DT&E), DOT&E, and Component T&E Executives)
    - » Establishes a STAT in T&E Center of Excellence (Under AFIT)
    - » Multipronged approach, coordinated with DOT&E and Component 's
      - Education & Training
      - Guidance and Policy
      - Case Study Development

STAT in T&E is envisioned to provide a long-term T&E capability to the acquisition community . Increased Scientific and Statistical rigor within the T&E.

Approved for public release; distribution is unlimited.

IDA

Conclusions

- **DOT&E supports the use of DOE and statistical methods in test planning and analysis**
- **Implications for Test Ranges:**
  - Increased emphasis on informative data collection
  - Linking test conditions to outcomes
  - Data quality and precision
  - Testers need to understand statistical thinking
- **Moving forward:**
  - DOT&E Goals for 2012
  - DASD(DT&E) sponsored Scientific Test and Analysis Techniques (STAT) Implementation Plan and COE

Approved for public release; distribution is unlimited.

Backup Material

Approved for public release; distribution is unlimited.

Test Science Committee Charter

Overview

This document outlines the charter for the Committee to Institutionalize Scientific Test Design and Rigor in Test and Evaluation. The charter defines the problem, identifies potential steps in a roadmap for accomplishing the goals of the committee and lists committee membership. Once the committee is assembled, the members will revise this document as needed. The charter will be endorsed by DOT&E and DDT&E, once finalized.

Problem Statement

Modern computing and software improvements have made advanced statistical methods for test design and analysis increasingly accessible in recent years. These advances provide opportunities for improvement in robustness and efficiency of test and evaluation. The Service test and evaluation communities have not adopted a set of best practices to date.

Purpose

The purpose of this committee is to plan and help oversee the institutional adoption of scientific test design, including Design of Experiments (DOE), and scientific rigor throughout the DoD Test and Evaluation (T&E) communities. The committee is charged with addressing the following topics:

- Assess the current state of analytic capabilities within each of the services and OSD.
- Develop qualification guidelines for personnel performing test design and analytic services for different kinds of T&E organizations.
- Develop a roadmap for training, education, and other support that Services and Agencies will need to attain the required test design and analytic capabilities.
- Develop case studies of the implementation of scientific test design across the test program.
- Provide guidance for the documentation of test design and statistical rigor in TEMPs, Test Plans and Reports.
- Form a permanent Advisory Board to advise on future test plans on methods for incorporate statistical rigor and test plan science.

Membership

The committee will be chaired by the Science Advisor, DOT&E, and will represent the stakeholder community. The initial membership is listed below and will be adjusted as needed:

- Representatives from each of DOT&E and DDT&E, including FFRDC support
- Representative from each of the OTAs
- Representative from each of the Service T&E Executives
- Representative from participating military graduate schools (Navy Postgraduate School, Air Force Institute of Technology)

Approved for public release; distribution is unlimited.