The Power of STAT

OSD STAT T&E COE
www.AFIT.edu/STAT
937-255-3636 x 4736

What is STAT?

Scientific Test and Analysis Techniques (STAT) are the scientific and statistical methods and processes used to enable the development of efficient, rigorous test strategies that will yield defensible results.

The results of the **repeated use** of STAT is a more progressive sequential testing approach that carefully **leverages past test** information, along with **informing** the systems engineering process.

- STAT tools:
  - Design of Experiments
  - Quasi-Experimental designs
  - Natural Experiments
  - Hypothesis Testing
  - Regression Analysis
  - Data Analysis Techniques
  - Operating Characteristic curves
  - Optimization
  - Distribution fitting (historical data)
  - Reliability growth
  - Observational Studies
  - Surveys

The STAT in T&E COE

The goal of the STAT T&E COE is to **assist the program** in developing **rigorous, defensible** test strategies to more effectively **quantify and characterize** system performance and provide information that **reduces risk**.

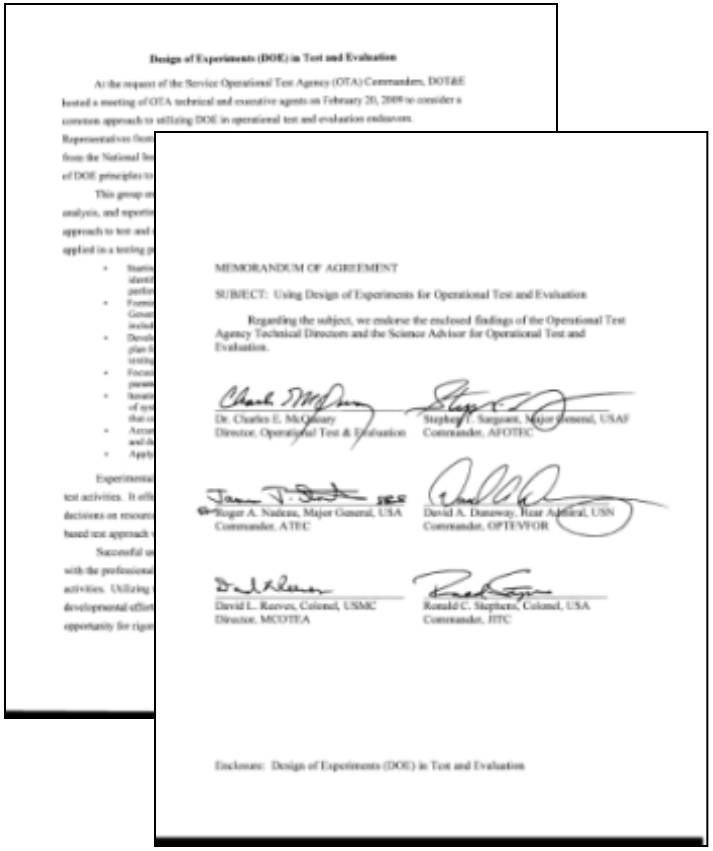
- Support TEMP development & test planning with scientific designs and processes.
- Leverage the state of the art in scientific testing.
- Engage and mentor T&E practitioners.
- Support test execution, data redux and analysis w/professional statistical software.
- Established July 2012 with HQ at AFIT, WPAFB, OH
  - Currently 25 Major Acquisition Programs; Director & 6 STAT Experts

The STAT T&E COE helps programs convert “technical to practical”

STAT in Test & Evaluation

Policy

Using Design of Experiments for Operational Test and Evaluation;
Memorandum of Agreement (MOA), Service Operational Test Agencies, April 2009



MEMORANDUM OF AGREEMENT

SUBJECT: Using Design of Experiments for Operational Test and Evaluation

Regarding the subject, we endorse the enclosed findings of the Operational Test Agency Technical Directors and the Science Advisor for Operational Test and Evaluation.

Dr. Charles E. McQueary
Director, Operational Test & Evaluation

Stephen P. Sargeant, Major General, USAF
Commander, AFOTEC

Roger A. Nadeau, Major General, USA
Commander, ATEC

David A. Dunaway, Rear Admiral, USN
Commander, OPTEVFOR

David L. Reeves, Colonel, USMC
Director, MCOTEA

Ronald C. Stapleton, Colonel, USA
Commander, ITC

Enclosure: Design of Experiments (DOE) in Test and Evaluation

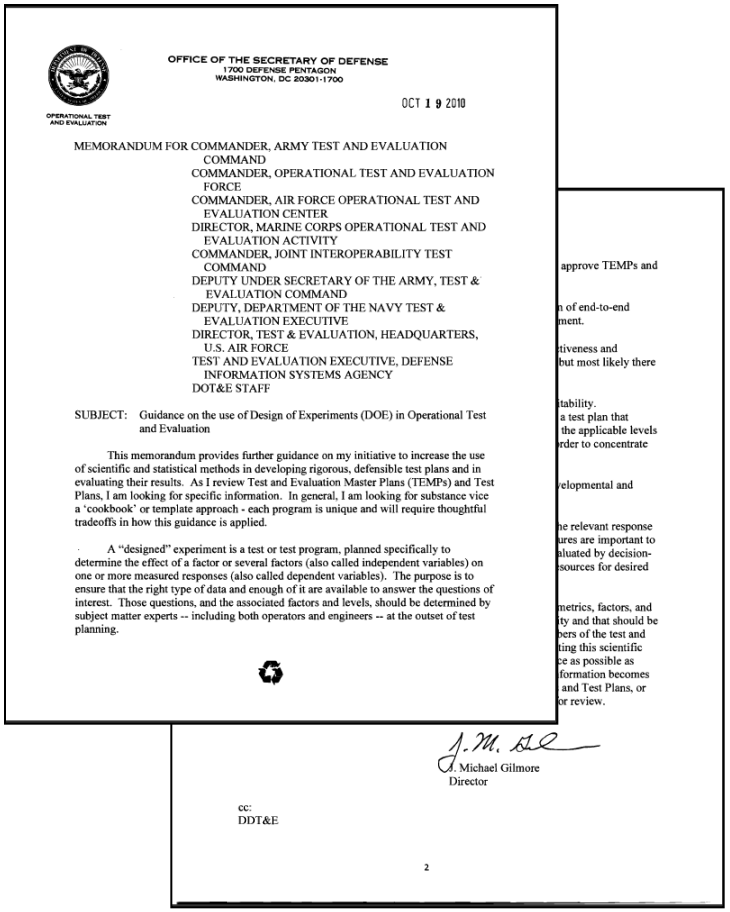
- Experimental design provides a valuable tool to identify and mitigate risk in all test activities.
- It offers a framework from which test agencies may make well-informed decisions on resource allocation and scope of testing required for an adequate test.
- Using DOE in all phases of program developmental and operational testing affords the opportunity for rigorous systematic improvement in test processes.

STAT in Test & Evaluation

Policy

Guidance on the use of Design of Experiments (DOE) in Operational Test and Evaluation; Director, Operational Test and Evaluation (DOT&E); 19 Oct 2010.

- Provides guidance on the use of scientific and statistical methods for developing rigorous, defensible test plans and in evaluating their results.
- Sets expectations on elements of experimental design that must be addressed in the Test & Evaluation Master Plan (TEMP)
  - The goal of the experiment, which should reflect evaluation of end-to-end mission effectiveness in an operationally realistic environment.
  - Quantitative mission-oriented response variables for effectiveness and suitability.
  - Factors that affect those measures of effectiveness and suitability and a good breath of coverage of their associated levels.
  - A method for strategically varying factors across both developmental and operational testing with respect to responses of interest.
  - Statistical measures of merit on the response variables for which it makes sense.



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.

J. Michael Gilmore
Director

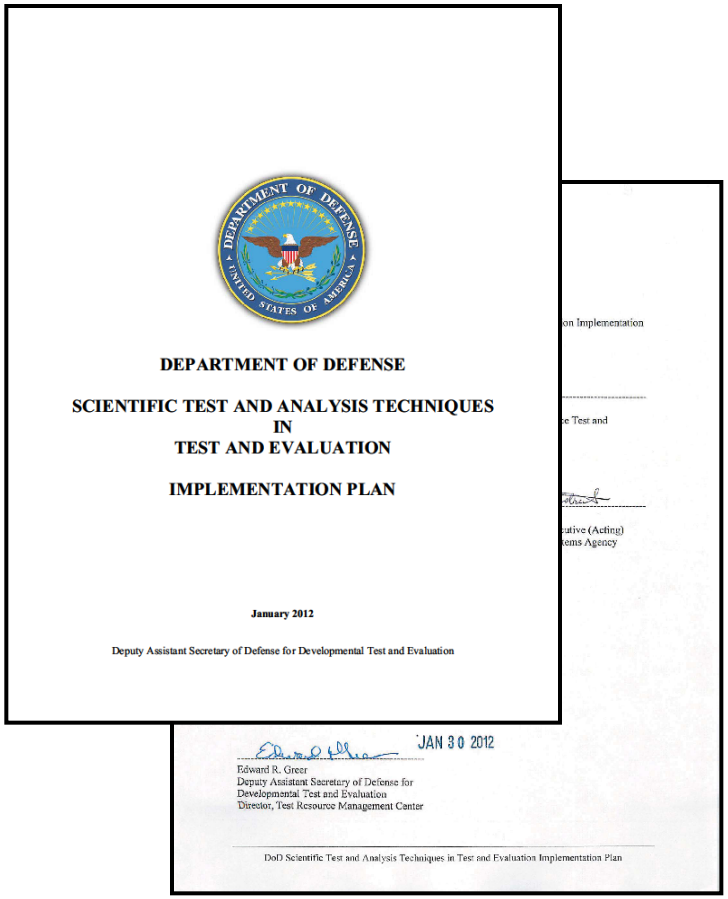
cc:
DDT&E

2

STAT in Test & Evaluation

Policy

Department of Defense Scientific Test and Analysis Techniques in Test and Evaluation; Implementation Plan; Deputy Assistant Secretary of Defense for Developmental Test and Evaluation (DASD(DT&E)); 30 Jan 2012.



DEPARTMENT OF DEFENSE

SCIENTIFIC TEST AND ANALYSIS TECHNIQUES
IN
TEST AND EVALUATION

IMPLEMENTATION PLAN

January 2012

Deputy Assistant Secretary of Defense for Developmental Test and Evaluation

JAN 30 2012

Edward R. Greer
Deputy Assistant Secretary of Defense for
Developmental Test and Evaluation
Director, Test Resource Management Center

DoD Scientific Test and Analysis Techniques in Test and Evaluation Implementation Plan

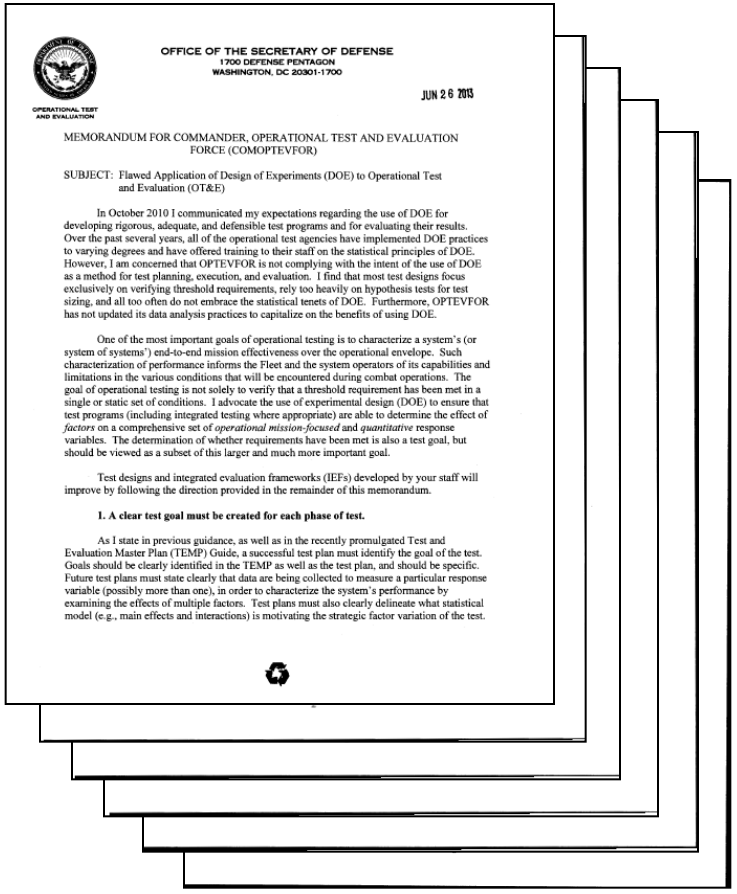
- Provides guidance for increasing the scientific and statistical rigor within the T&E planning, design, execution, and analysis phases throughout the life cycle.
- Intent is to apply the appropriate methods for the planning, design, execution, and analysis of tests.
- End State
  - Increase T&E efficiencies and effectiveness
  - Enable PMs to make better informed decisions based on acceptable risk thresholds.
- The STAT Center of Excellence (STAT COE) is a “Key” element of the overall STAT in T&E Implementation plan.
- Signed by DOT&E, DASD(DT&E) and Service T&E Executives.

STAT in Test & Evaluation

Policy

Flawed application of design of experiments (DOE) to Operational Test and Evaluation (OT&E); Director, Operational Test and Evaluation (DOT&E); 26 June 2013.

- Clear test goals
- Mission oriented metrics
- Do not limit factors
- Strategically control factors
- Avoid confounding of factors
- Avoid single hypothesis tests
- Consider all factors



OFFICE OF THE SECRETARY OF DEFENSE
1700 DEFENSE PENTAGON
WASHINGTON, DC 20301-1700

OPERATIONAL TEST AND EVALUATION

JUN 26 2013

MEMORANDUM FOR COMMANDER, OPERATIONAL TEST AND EVALUATION FORCE (COMOPTEVFOR)

SUBJECT: Flawed Application of Design of Experiments (DOE) to Operational Test and Evaluation (OT&E)

In October 2010 I communicated my expectations regarding the use of DOE for developing rigorous, adequate, and defensible test programs and for evaluating their results. Over the past several years, all of the operational test agencies have implemented DOE practices to varying degrees and have offered training to their staff on the statistical principles of DOE. However, I am concerned that OPTEVFOR is not complying with the intent of the use of DOE as a method for test planning, execution, and evaluation. I find that most test designs focus exclusively on verifying threshold requirements, rely too heavily on hypothesis tests for test sizing, and all too often do not embrace the statistical tenets of DOE. Furthermore, OPTEVFOR has not updated its data analysis practices to capitalize on the benefits of using DOE.

One of the most important goals of operational testing is to characterize a system's (or system of systems') end-to-end mission effectiveness over the operational envelope. Such characterization of performance informs the Fleet and the system operators of its capabilities and limitations in the various conditions that will be encountered during combat operations. The goal of operational testing is not solely to verify that a threshold requirement has been met in a single or static set of conditions. I advocate the use of experimental design (DOE) to ensure that test programs (including integrated testing where appropriate) are able to determine the effect of *factors* on a comprehensive set of *operational mission-focused* and *quantitative* response variables. The determination of whether requirements have been met is also a test goal, but should be viewed as a subset of this larger and much more important goal.

Test designs and integrated evaluation frameworks (IEFs) developed by your staff will improve by following the direction provided in the remainder of this memorandum.

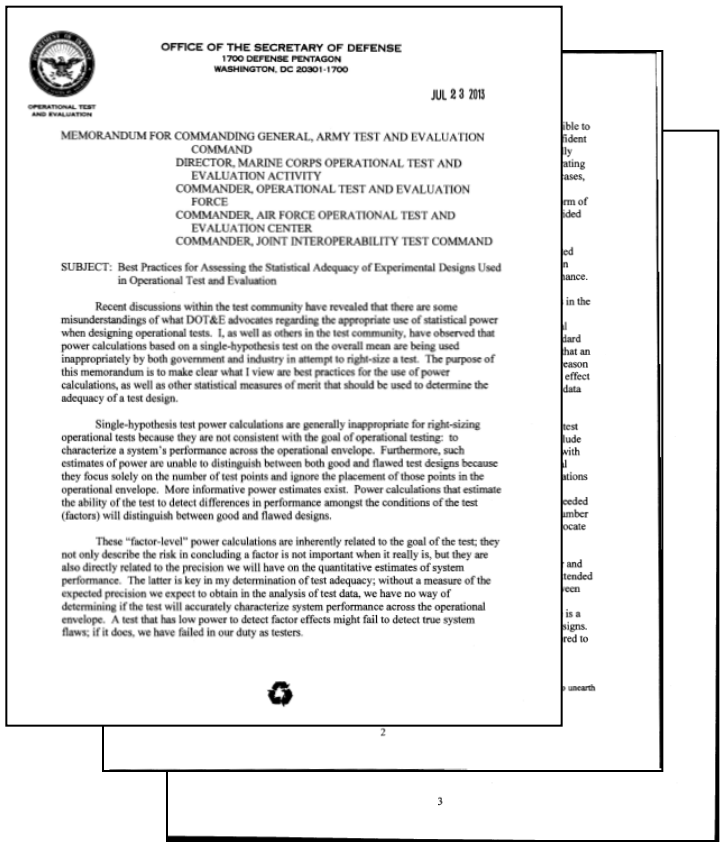
1. A clear test goal must be created for each phase of test.

As I state in previous guidance, as well as in the recently promulgated Test and Evaluation Master Plan (TEMP) Guide, a successful test plan must identify the goal of the test. Goals should be clearly identified in the TEMP as well as the test plan, and should be specific. Future test plans must state clearly that data are being collected to measure a particular response variable (possibly more than one), in order to characterize the system's performance by examining the effects of multiple factors. Test plans must also clearly delineate what statistical model (e.g., main effects and interactions) is motivating the strategic factor variation of the test.

STAT in Test & Evaluation

Policy

Best practices for assessing the statistical adequacy of experimental designs used in Operational Test and Evaluation; Director, Operational Test and Evaluation (DOT&E); 23 July 2013.



OFFICE OF THE SECRETARY OF DEFENSE
1700 DEFENSE PENTAGON
WASHINGTON, DC 20301-1700

JUL 23 2013

OPERATIONAL TEST AND EVALUATION

MEMORANDUM FOR COMMANDING GENERAL, ARMY TEST AND EVALUATION COMMAND
DIRECTOR, MARINE CORPS OPERATIONAL TEST AND EVALUATION ACTIVITY
COMMANDER, OPERATIONAL TEST AND EVALUATION FORCE
COMMANDER, AIR FORCE OPERATIONAL TEST AND EVALUATION CENTER
COMMANDER, JOINT INTEROPERABILITY TEST COMMAND

SUBJECT: Best Practices for Assessing the Statistical Adequacy of Experimental Designs Used in Operational Test and Evaluation

Recent discussions within the test community have revealed that there are some misunderstandings of what DOT&E advocates regarding the appropriate use of statistical power when designing operational tests. I, as well as others in the test community, have observed that power calculations based on a single-hypothesis test on the overall mean are being used inappropriately by both government and industry in attempt to right-size a test. The purpose of this memorandum is to make clear what I view are best practices for the use of power calculations, as well as other statistical measures of merit that should be used to determine the adequacy of a test design.

Single-hypothesis test power calculations are generally inappropriate for right-sizing operational tests because they are not consistent with the goal of operational testing: to characterize a system's performance across the operational envelope. Furthermore, such estimates of power are unable to distinguish between both good and flawed test designs because they focus solely on the number of test points and ignore the placement of those points in the operational envelope. More informative power estimates exist. Power calculations that estimate the ability of the test to detect differences in performance amongst the conditions of the test (factors) will distinguish between good and flawed designs.

These "factor-level" power calculations are inherently related to the goal of the test; they not only describe the risk in concluding a factor is not important when it really is, but they are also directly related to the precision we will have on the quantitative estimates of system performance. The latter is key in my determination of test adequacy; without a measure of the expected precision we expect to obtain in the analysis of test data, we have no way of determining if the test will accurately characterize system performance across the operational envelope. A test that has low power to detect factor effects might fail to detect true system flaws; if it does, we have failed in our duty as testers.

2

3

- Re-emphasizes the importance of statistical power in test design.
- Links the importance of a clear test goal and the appropriate tests designs
  - Screening
  - Characterization
  - Optimization
  - Comparison
- Advocates different statistical methods for assessing test design, such as
  - Empirical model strength
  - Collinearity
  - Power and Confidence
  - Scaled Prediction Variance
  - Optimality Criteria

STAT in Test & Evaluation

Leadership Expectations

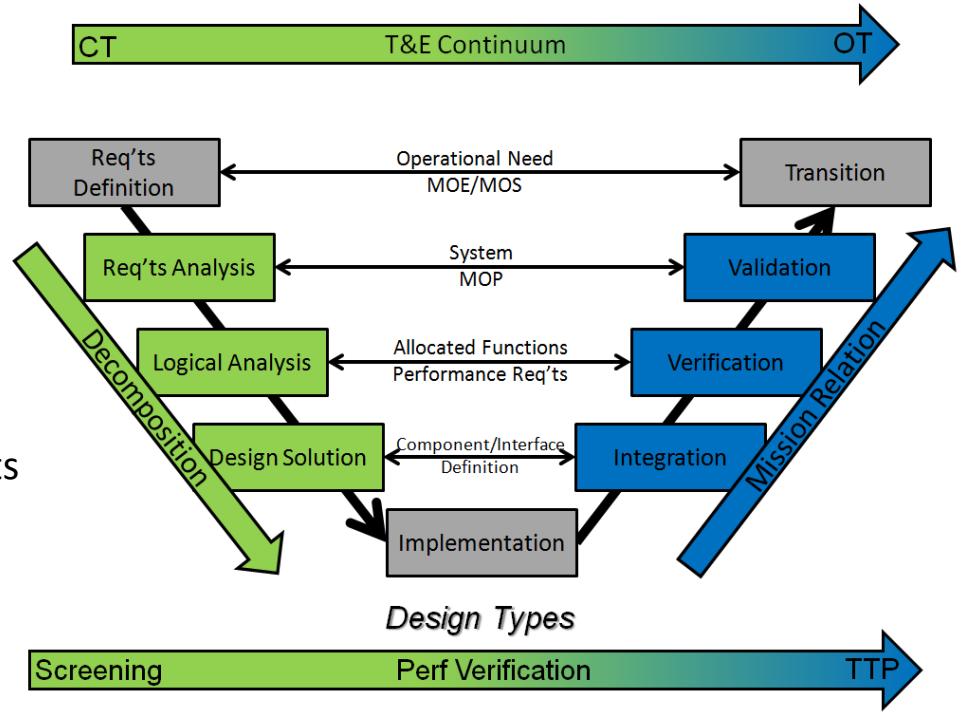
- Improve all aspects of the T&E strategy: test planning, test design, test execution, data analysis and assessment, and reporting
  - Produce defensible test plans
  - Ensure scientifically rigorous coverage of the test space
  - Optimize test execution
  - Add statistical rigor to the analysis
  - Improve the understanding of risk to support programmatic decisions
- Provide a better understanding of the technical and operational capabilities of a system and the true state of the technology
- Comply with DOT&E Guidance (TEMP Guide Annex, Feb 27 2012)

STAT in T&E improves the outcome of acquisition decisions.

Decomposition and Mission Relation for Test Planning

Process Principles

- Decompose **system components**
  - Defines low level testable elements
  - Screen factors early in DT
  - Develop models in DT (not just point check)
- Decompose **mission segments**
  - Assign mission-related (derived) requirements
  - Include significant DT factors
- Tailor test space for OT
  - Critical operational regions or DT risk areas
  - OT events can verify DT model predictions

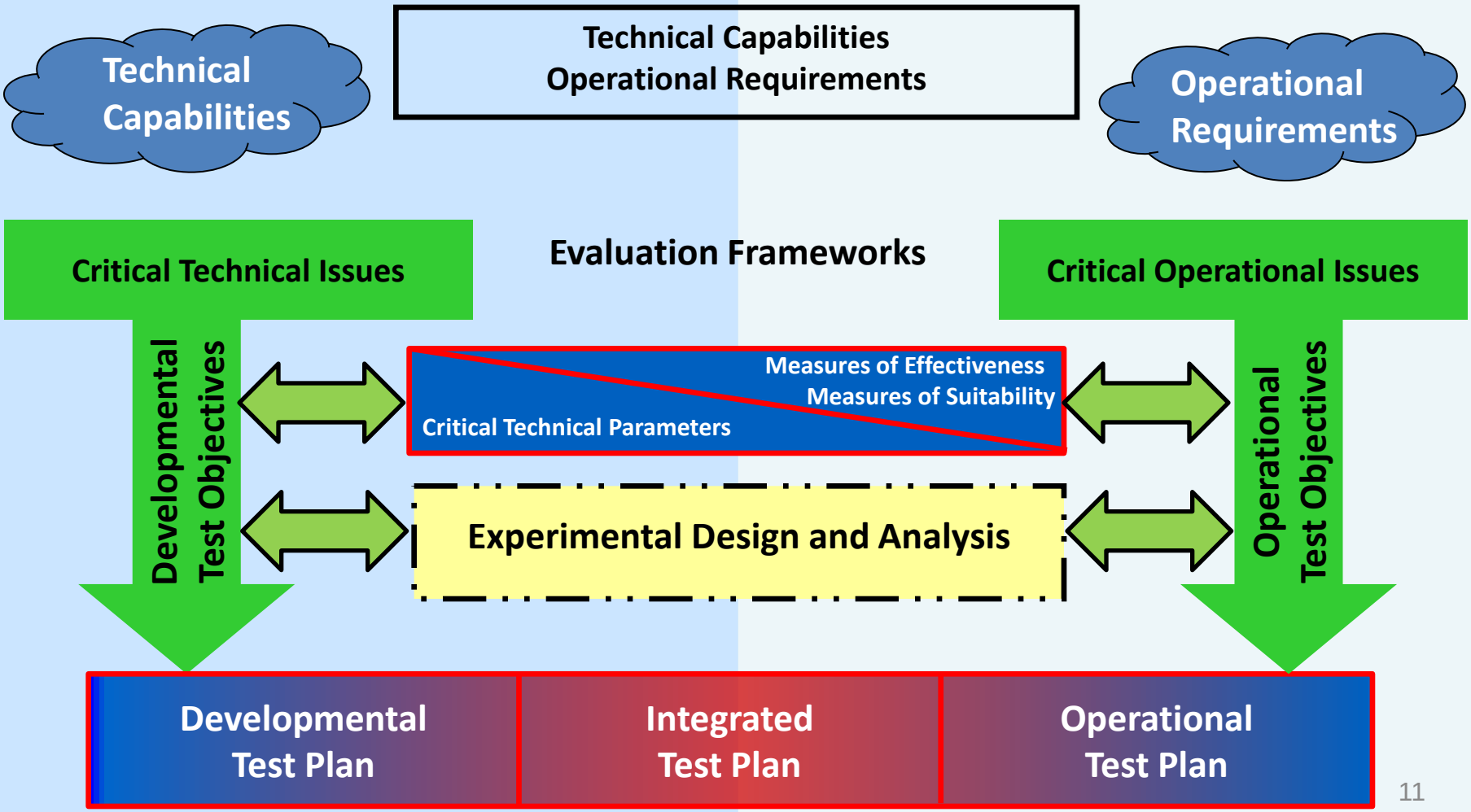


Decomposition helps define DT scope and allows assignment of **mission related requirements** for OT

STAT in Test & Evaluation

A Place for Experimental Design and Analysis

STAT T&E COE: Scientia Prudentia et Valor



STAT in Test & Evaluation

Tradeoffs

STAT T&E COE: Scientia Prudentia et Valor

Risk of rejecting a "good" system

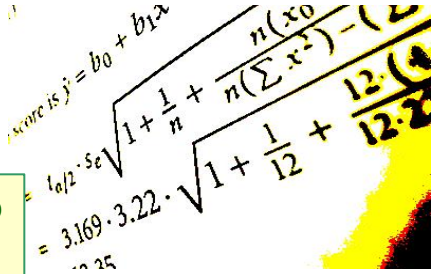
β

n

Adequate number of representative samples

δ

Performance effects to detect



Risk of accepting a "bad" system

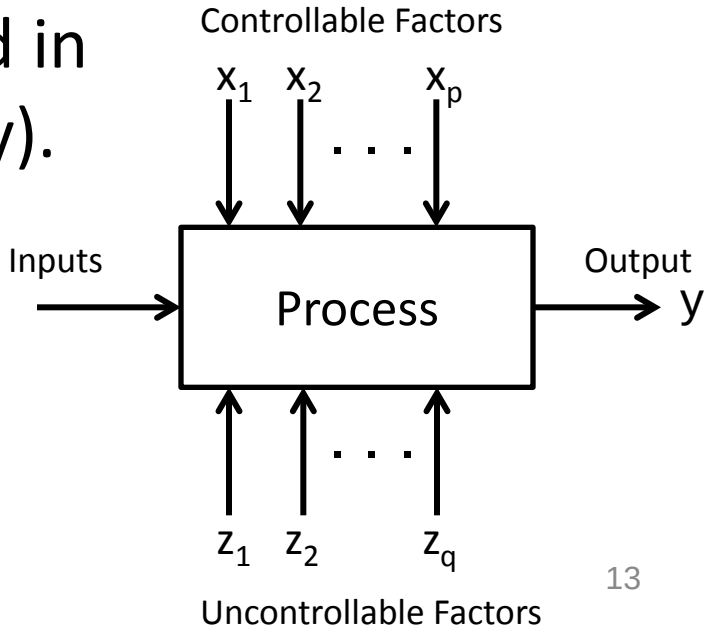
α

The *Central Problem of Test* is to determine the true nature of the system, in all possible scenarios, with a finite number of samples that yield valid conclusions while minimizing the risk of error.

Fundamentals of Experimental Design

Experiment - Definition

- A test or a series of tests in which purposeful changes are made to the input variable of a process or system to observe and exploit the relationships between those changes and changes that may be observed in the output response (Montgomery).



Top 5 Questions a Senior Leader Should Ask … About Test Designs

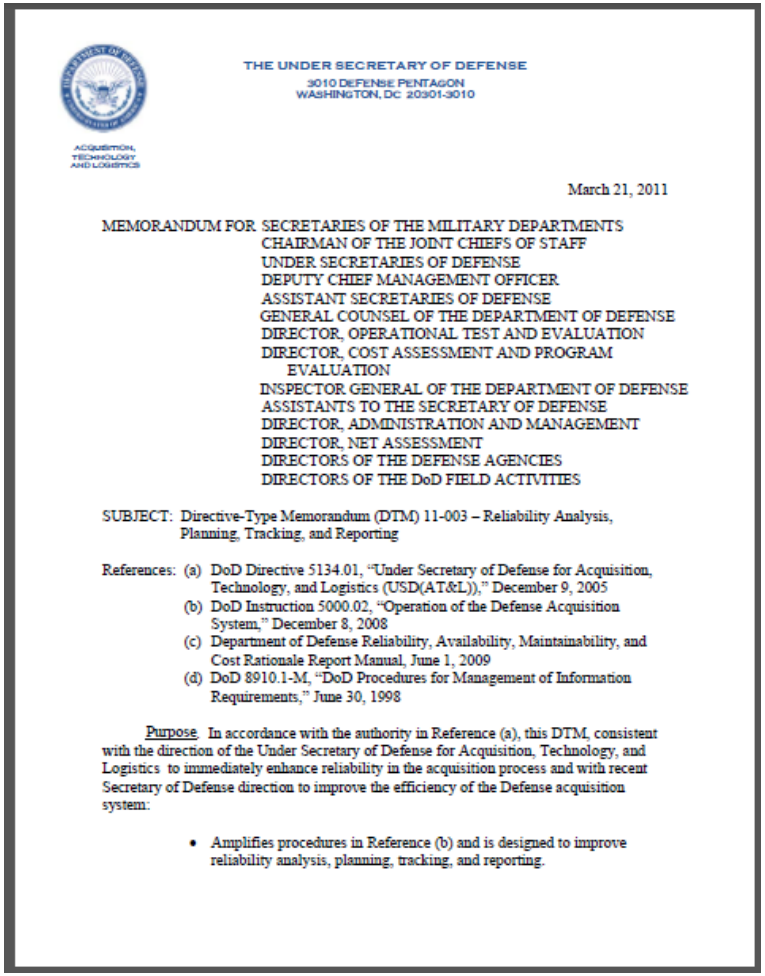
1. How is random variation accounted for?
   - What is the risk of passing a bad system?
   - What is the risk of failing a good system?
2. How are the three principles of DOE (randomization, replication, and blocking) implemented?
3. Has the DOE process been followed?
4. How will information obtained in DT be used to inform OT?
5. Will these tests result in data that informs decision makers?

Reliability

- Reliability is the ***probability*** that an item (component or system) will perform its ***intended function*** for a specified period of ***time*** when operated in its ***design environment***.

Decision Type Memorandum 11-003
Reliability Planning, Tracking, and Reporting

- 21 March 2011 DTM from OSD/AT&L amplifies 5000.02 reliability requirements to include reliability growth and reporting measures for all major acquisitions.
- PMs must use **Reliability Growth** (RG) and **Design for Reliability** (DfR) methods to achieve R&M requirements
- Specifically mentions critical reliability tasks
  - Block Diagrams and Predictions
  - Failure Definition/Scoring Criteria
  - FRACAS (Failure Reporting, Analysis and Corrective Action System) and FMECA (Failure Mode Effects and Criticality Analysis)
  - RG testing and Built-In Test Demonstrations



THE UNDER SECRETARY OF DEFENSE
3010 DEFENSE PENTAGON
WASHINGTON, DC 20301-3010

ACQUISITION, TECHNOLOGY AND LOGISTICS

March 21, 2011

MEMORANDUM FOR SECRETARIES OF THE MILITARY DEPARTMENTS
CHAIRMAN OF THE JOINT CHIEFS OF STAFF
UNDER SECRETARIES OF DEFENSE
DEPUTY CHIEF MANAGEMENT OFFICER
ASSISTANT SECRETARIES OF DEFENSE
GENERAL COUNSEL OF THE DEPARTMENT OF DEFENSE
DIRECTOR, OPERATIONAL TEST AND EVALUATION
DIRECTOR, COST ASSESSMENT AND PROGRAM EVALUATION
INSPECTOR GENERAL OF THE DEPARTMENT OF DEFENSE
ASSISTANTS TO THE SECRETARY OF DEFENSE
DIRECTOR, ADMINISTRATION AND MANAGEMENT
DIRECTOR, NET ASSESSMENT
DIRECTORS OF THE DEFENSE AGENCIES
DIRECTORS OF THE DoD FIELD ACTIVITIES

SUBJECT: Directive-Type Memorandum (DTM) 11-003 – Reliability Analysis, Planning, Tracking, and Reporting

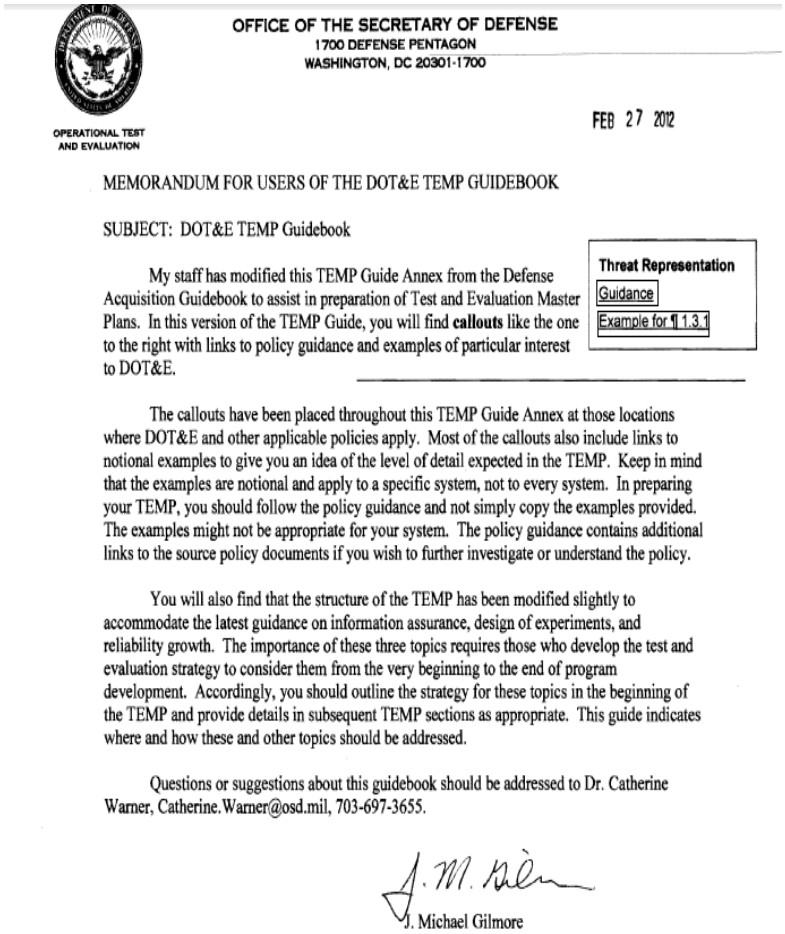
References: (a) DoD Directive 5134.01, "Under Secretary of Defense for Acquisition, Technology, and Logistics (USD(AT&L))," December 9, 2005
(b) DoD Instruction 5000.02, "Operation of the Defense Acquisition System," December 8, 2008
(c) Department of Defense Reliability, Availability, Maintainability, and Cost Rationale Report Manual, June 1, 2009
(d) DoD 8910.1-M, "DoD Procedures for Management of Information Requirements," June 30, 1998

Purpose. In accordance with the authority in Reference (a), this DTM, consistent with the direction of the Under Secretary of Defense for Acquisition, Technology, and Logistics to immediately enhance reliability in the acquisition process and with recent Secretary of Defense direction to improve the efficiency of the Defense acquisition system:

- Amplifies procedures in Reference (b) and is designed to improve reliability analysis, planning, tracking, and reporting.

Reliability and Reliability Growth in Test and Evaluation

STAT T&E COE: Scientia Prudentia et Valor



OFFICE OF THE SECRETARY OF DEFENSE
1700 DEFENSE PENTAGON
WASHINGTON, DC 20301-1700

OPERATIONAL TEST AND EVALUATION

FEB 27 2012

MEMORANDUM FOR USERS OF THE DOT&E TEMP GUIDEBOOK

SUBJECT: DOT&E TEMP Guidebook

My staff has modified this TEMP Guide Annex from the Defense Acquisition Guidebook to assist in preparation of Test and Evaluation Master Plans. In this version of the TEMP Guide, you will find **callouts** like the one to the right with links to policy guidance and examples of particular interest to DOT&E.

Threat Representation
Guidance
Example for ¶ 1.3.1

The callouts have been placed throughout this TEMP Guide Annex at those locations where DOT&E and other applicable policies apply. Most of the callouts also include links to notional examples to give you an idea of the level of detail expected in the TEMP. Keep in mind that the examples are notional and apply to a specific system, not to every system. In preparing your TEMP, you should follow the policy guidance and not simply copy the examples provided. The examples might not be appropriate for your system. The policy guidance contains additional links to the source policy documents if you wish to further investigate or understand the policy.

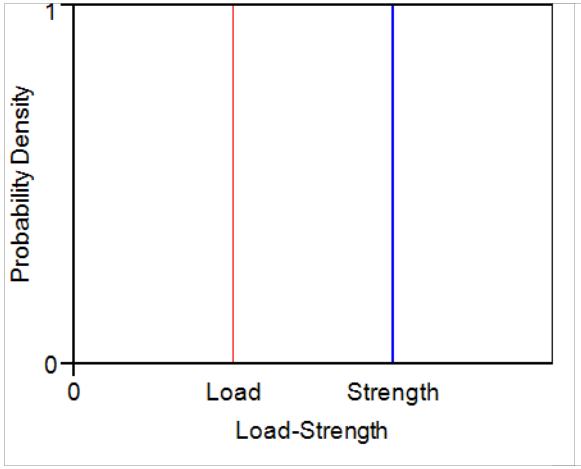
You will also find that the structure of the TEMP has been modified slightly to accommodate the latest guidance on information assurance, design of experiments, and reliability growth. The importance of these three topics requires those who develop the test and evaluation strategy to consider them from the very beginning to the end of program development. Accordingly, you should outline the strategy for these topics in the beginning of the TEMP and provide details in subsequent TEMP sections as appropriate. This guide indicates where and how these and other topics should be addressed.

Questions or suggestions about this guidebook should be addressed to Dr. Catherine Warner, Catherine.Warner@osd.mil, 703-697-3655.

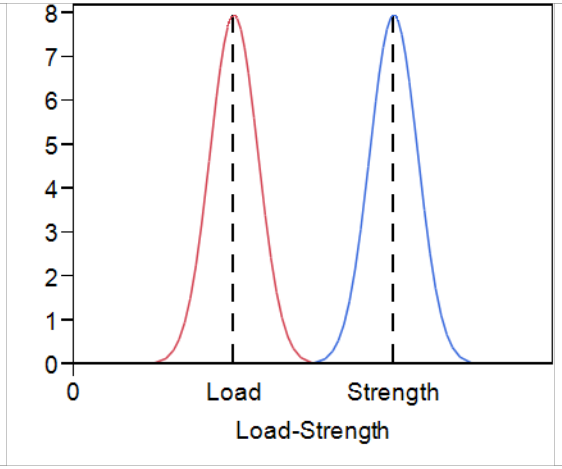
J. Michael Gilmore

- DOT&E issued reliability growth and DOE guidance for **TEMPs** for OT—forthcoming for DT.
- Callouts given for best practices, representative examples for Section 3 and Annexes now required.
- Provides considerations for software intensive systems and language to capture DfR activities to ensure reliability is sufficient to enter IOT&E.

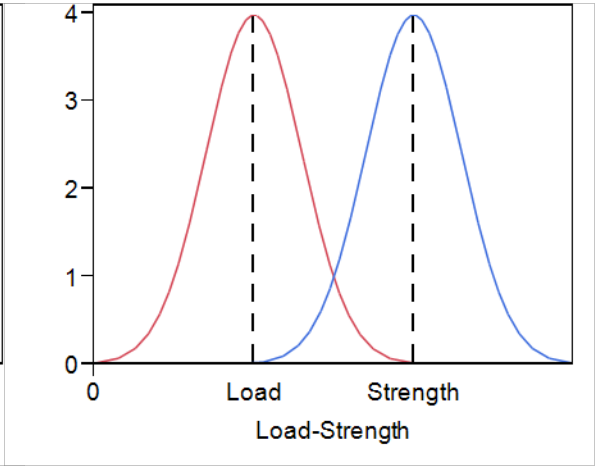
Probability's Close Relationship to Reliability



Perfect Reliability

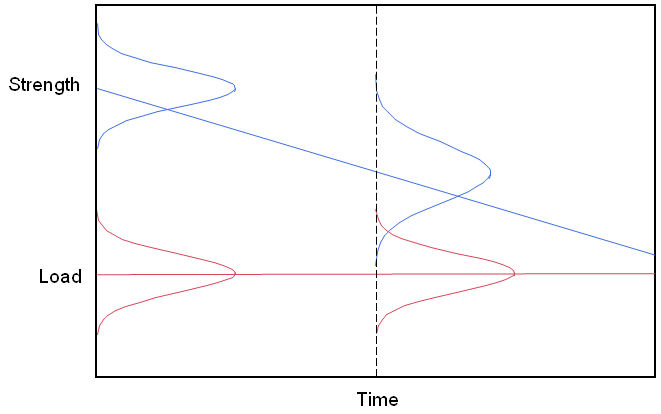


Good Reliability



Poor Reliability

Degradation



Reliability Growth

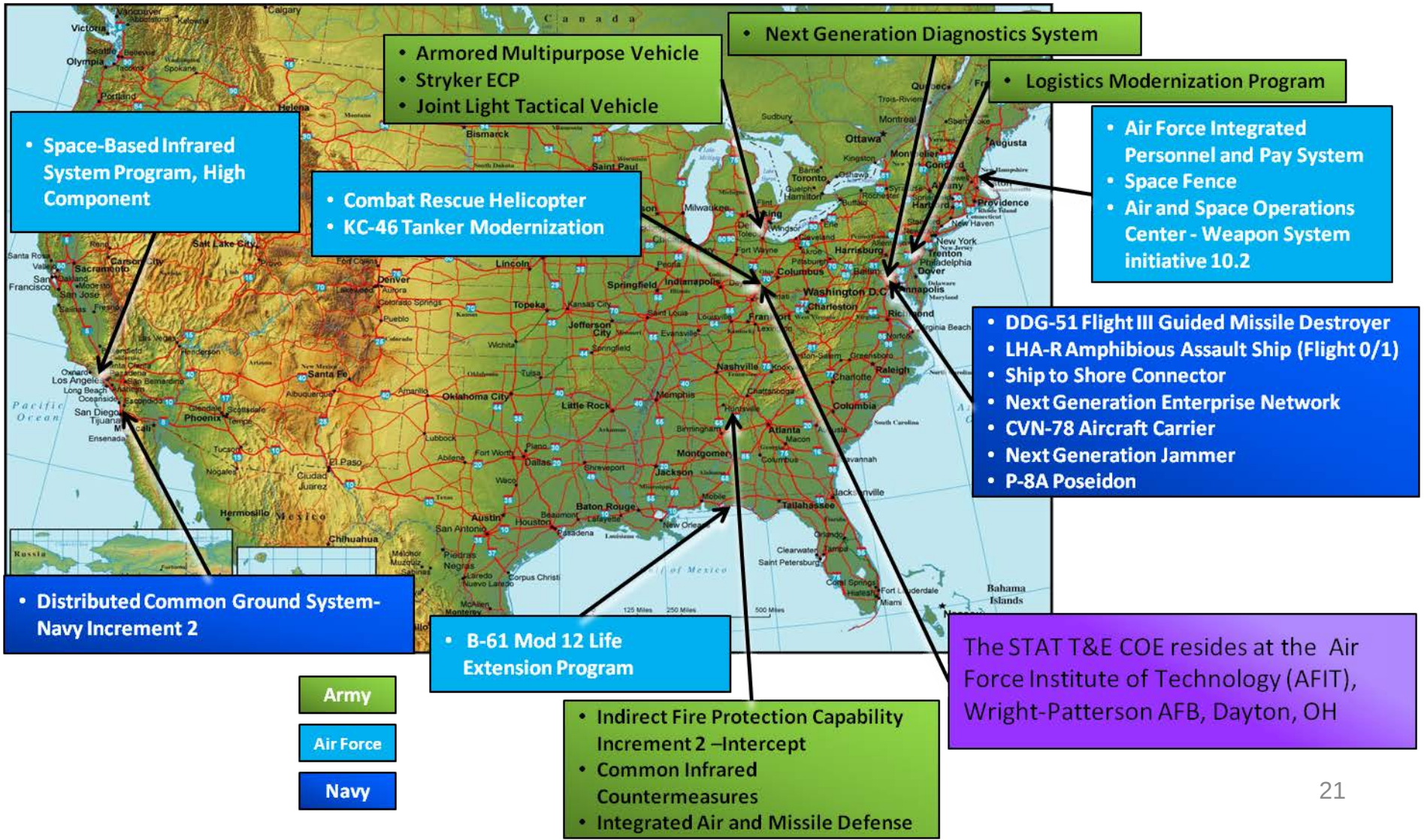
- ***Reliability Growth*** is the positive improvement in a reliability parameter over a period of time due to implementation of corrective actions to system design, operation or maintenance procedures, or the associated manufacturing process. (MIL-HDBK-189C 14 June 2011)

Top 5 Questions a Senior Leader Should Ask … About Reliability

1. Have all significant failure modes been identified?
2. Do proposed analyses reflect the complexity of defense systems?
   - What assumptions do proposed analyses make (i.e. growth curve, failure time distribution)? What is the basis for these assumptions?
3. How will information obtained in DT be used to inform OT?
4. Will these tests result in data that informs decision makers?
5. Does the level of required reliability relate to the operational requirement? What is the appropriate level of reliability?

STAT COE Supported Programs

STAT T&E COE: Scientia Prudentia et Valor



STAT COE Highlights

STAT T&E COE: Scientia Prudentia et Valor

- Acquisition programs beyond the initial 20 are seeking assistance as word spreads of the Center's effectiveness
- Engaged in STAT discussions with Program Managers, Operational Test Agencies, and Components
- Assisted in the STAT curriculum update within DAU test courses
- Director, Operational Test & Evaluation report to Congress highlights the STAT T&E COE
- Deputy Assistant Secretary of Defense(Developmental Test & Evaluation) report to Congress includes the COE as a major effort to accomplish the STAT Implementation Plan
- Director, Operational Test & Evaluation has endorsed the STAT COE and research consortium in the September issue of the ITEA Journal

Summary

- ***STAT*** offers many tools and processes to ensure more defensible test results.
- ***Design of Experiments*** involves planning, executing, and analyzing tests in a manner that allows the desired information to be obtained.
- ***Design for Reliability*** and ***Reliability Growth*** are two approaches for improving the reliability of systems.
- The ***STAT Center of Excellence*** enables STAT to begin at the program level.

Questions?

OSD STAT T&E COE

www.AFIT.edu/STAT

937-255-3636 x 4736