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

OPERATIONAL TEST
AND EVALUATION

JAN 19 2017

MEMORANDUM FOR USERS OF THE DIRECTOR, OPERATONAL TEST AND EVALUATION (DOT&E) TEST AND EVALUATION MASTER PLAN (TEMP) GUIDEBOOK

SUBJECT: DOT&E TEMP Guidebook 3.1

DOT&E TEMP Guidebook 3.1 updates the Design of Experiments (DOE), Scientific Test and Analysis Techniques (STAT), Mission-focused Metrics, Operational Evaluation Framework (OEF), Modeling and Simulation (M&S) for Test and Evaluation, Defense Business Systems, Cybersecurity, and Software-Intensive Systems sections of the DOT&E TEMP Guidebook 3.0. Version 3.1 also contains a new Survey Design guidance section consistent with the 6 January 2017 DOT&E Survey Pre-testing and Administration memo.

The DOE and STAT sections now highlight the importance of justifying resources, especially long-lead items, using experimental design techniques at Milestone A and in TEMPs supporting Requests for Proposals. The Mission-focused Metrics section was renamed Quantitative Mission-focused Measures. The terms "metrics," "measures," and "response variables" have been used interchangeably in the Mission-focused Metrics, DOE, and STAT sections – which has led to some confusion among readers. Thus where it makes sense, we replaced the terms "metrics" and "response variables" with "measures." We added the term "quantitative" to Mission-focused Measures to highlight the importance of evaluating systems with quantitative measures as opposed to qualitative measures.

The cybersecurity guidance includes a statement reminding readers that a Cooperative Vulnerability and Penetration Assessment and Adversarial Assessment are normally required as part of an operational test or assessment supporting a fielding decision. The command and control cybersecurity example now contains a detailed table describing cyber defenders' roles and responsibilities.

Minor content updates were made in the Software-intensive Systems section to align the discussion with Department of Defense Instruction (DoDI) 5000.02 as compared to the interim DODI 5000.02.

That STAT and Defense Business Systems sections now include DOT&E survey content expectations in TEMPs, along with dedicated survey guidance. The Defense Business Systems section also now discusses the importance of describing software change requests when explaining the defect tracking process.

The measures of merit discussion in the OEF guidance was incorporated into the STAT section.

The M&S section was updated to reflect the 14 March 2016 and 17 January 2017 DOT&E guidance memoranda on the validation of M&S used in operational test and live fire assessments.

Program Managers will use the TEMP as the primary planning and management tool for all test activities starting at Milestone A. Program Managers will prepare and update the TEMP as needed and to support acquisition milestones or decision points. The TEMP should be specific to the program and tailored to meet program needs. Accordingly, the guidance in this guidebook, in DoDI 5000.02, and in the TEMP format are provided to assist in developing the appropriate TEMP format and content for each program. Strict or immediate adherence to the new TEMP format is not required. Use common sense to apply the guidance to fit your program. Evaluation of TEMP adequacy is based on the TEMP's content, not the format.

Questions or suggestions about this guidebook should be addressed to Dr. Catherine Warner, catherine.w.warner.civ@mail.mil, 703-697-3655.

J. Michael Gilmore
Director

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

OPERATIONAL TEST
AND EVALUATION

NOV 16 2015

MEMORANDUM FOR USERS OF THE DIRECTOR, OPERATIONAL TEST AND EVALUATION (DOT&E) TEST AND EVALUATION MASTER PLAN (TEMP) GUIDEBOOK

SUBJECT: DOT&E TEMP Guidebook 3.0

This new version of the DOT&E TEMP Guidebook complements the January 2015 version of DoDI 5000.02 by illustrating with selective guidance and examples how to develop and document an adequate test and evaluation (T&E) strategy. The Program Manager will use the TEMP as the primary planning and management tool for all test activities starting at Milestone A. Best practices outlined in this TEMP Guidebook should be applied to all versions of the TEMP, including the Development Request for Proposal (RFP) TEMP.

The Program Manager will prepare and update the TEMP as needed and to support acquisition milestones or decision points. The TEMP should be specific to the program and tailored to meet program needs. Accordingly, the guidance in this guidebook, in DoDI 5000.02, and in the TEMP format guide are provided to assist in developing the appropriate TEMP format and content for each program. Strict or immediate adherence to the new TEMP format is not required. Use common sense to apply the guidance to fit your program. Evaluation of TEMP adequacy is based on the TEMP's content, not the format.

Summary of the TEMP and TEMP Guidebook Format

The TEMP format has been changed as illustrated below. The previous TEMP format on the left explained in sentences and paragraphs what DOT&E required for adequacy. TEMP Guidebook 2.1 added colored callout boxes with links to the DOT&E Guidebook guidance and examples.

The new TEMP format on the right enumerates in bullets what should be considered for inclusion in each paragraph/section of the TEMP. Callouts with links to DOT&E guidance in the Guidebook 3.0 are in bold blue font.

Previous TEMP and Guidebook 2.1 Format

1.2. Mission Description. Briefly summarize the mission need described in the program capability requirements documents in terms of the capability it will provide to the Joint Forces Commander. Describe the mission to be accomplished by a unit equipped with the system using all applicable CONOPS and Concepts of Employment. Incorporate an OV-1 of the system showing the intended operational environment. Also include the organization in which the system will be integrated as well as significant points from the Life Cycle Sustainment Plan, the Information Support Plan, and Program Protection Plan. Provide links to each document referenced in the introduction. For business systems, include a summary of the business case analysis for the program.

1.3. System Description. Describe the system configuration. Identify key features and subsystems, both hardware and software (such as architecture, system and user interfaces, security levels, and reserves) for the planned increments within the Future Years Defense Program (FYDP).

1.3.1. **System Threat Assessment.** Succinctly summarize the threat environment (to include cyber-threats) in which the system will operate. Reference the appropriate DIA or component-validated threat documents for the system.

1.3.2. Program Background. Reference the Analysis of Alternatives (AoA), the APB and the materiel development decision to provide background information on the proposed system. Briefly describe the overarching Acquisition Strategy (for space systems, the Integrated Program Summary (IPS)), and the Technology Development Strategy (TDS). Address whether the system will be procured using an incremental development strategy or a single step to full capability. If it is an evolutionary acquisition strategy, briefly discuss planned

Threat Representation
Guidance
Example for ¶ 1.3.1

Information Assurance (Cybersecurity)
Guidance
Examples for ¶ 1.3.1

New TEMP and Guidebook 3.0 Format

1.2. MISSION DESCRIPTION

1.2.1 Mission Overview

- Summarize the mission need described in the program capability requirements documents in terms of the capability the system will provide to the Warfighter.
- Describe the mission to be accomplished by a unit that will be equipped with the system.
- Incorporate an Operational View (OV-1) of the system showing the intended operational environment.
- Include significant points from the Life Cycle Sustainment Plan, the Information Support Plan, and the Program Protection Plan.
- For business systems, include a summary of the business case analysis for the program.

1.2.2 Concept of Operations

- Reference all applicable Concepts of Operations and Concepts of employment in describing the mission. Describe test implications.
  - *CONOPS Guidance and Examples*

1.2.3 Operational users

- Describe the intended users of the system, how they will employ the system, and any important characteristics of the operational users (e.g., experience level, training requirements, area of specialization, etc.).

The callouts have been placed throughout TEMP Guide 3.0 at locations where DOT&E and other applicable policies apply. Keep in mind that the examples are notional and apply to a specific or notional system, not to every system. In preparing your TEMP, you should apply 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 the underlying policy.

Summary of Milestone A TEMP Requirements in the January 2015 DoDI 5000.02

The Milestone A TEMP should address all major sections of the TEMP outline, but some of the details in the TEMP format may not be mature until Milestone B. The Milestone A TEMP should be complete enough to estimate and plan for the major resources required for adequate test and evaluation. Other specifics that should be included in the Milestone A TEMP include:

- Operational rationale for requirements. A link or reference to the capabilities development document (CDD) or similar document that provides rationale for requirements would be sufficient.
- For software acquisitions, an analysis of operational risk to mission accomplishment covering all planned capabilities or features in the system. The analysis will include commercial and non-developmental items.
- All planned T&E for phase completion. Major test events should have test entrance and test completion criteria.
- A table of independent variables (or "conditions," "parameters," "factors," etc.) that may have a significant effect on operational performance.
- Strategy and resources for cybersecurity T&E.

Summary of Milestone B and Subsequent TEMP Requirements in the January 2015 DoDI 5000.02

Regarding operational and live fire testing, the Milestone B and subsequent TEMPs should be updated to address all plans of the T&E strategy until system deployment. The

detailed focus of each TEMP should be on plans for the Developmental Test and Evaluation (DT&E), Live Fire Test and Evaluation (LFT&E), and Operational Test and Evaluation (OT&E) supporting the next major acquisition decision. In addition to updating the Milestone A content, the Milestone B and subsequent TEMPs should include:

- Expand on details of each LFT&E and OT&E phase/test to include cybersecurity testing.
- Expanded use of scientific and test analysis techniques to design effective and efficient testing.
- Reliability Growth Curves (RGCs) or Software Tracking metrics, updated RGCs (if applicable) that reflect test results to date, and a working link to the Failure Modes, Effects and Criticality Analysis (FMECA) data. A software defect or failure tracking database may replace the FMECA in software acquisitions.
- Operational evaluation framework that shows how the major test events and test phases link together to form a systematic, rigorous, and structured approach to evaluating mission capability across the applicable values of the independent variables.
- The updated table of variables will include the anticipated effects on operational performance, the range of applicable values (or "levels," "settings," etc.), the overall priority of understanding the effects of the variable, and the intended method of controlling the variable during test (uncontrolled variation, hold constant, or controlled systematic test design).
- Plans for Verification, Validation, and Accreditation if applicable.
- Appropriate cybersecurity measures to evaluate operational capability to protect, detect, react, and restore to sustain continuity of operation. The TEMP will document the threats to be used, which should be selected based on the best current information available from the intelligence community.
- Complete test resource requirements. Resources will reflect the best estimate for conducting all test activities. Resources will be mapped against the developmental and operational evaluation frameworks and schedule to ensure adequacy and availability. Ensure that resource estimates identified in the TEMP are matched against the schedule and justified by analysis.

Summary of the TEMP Outline from the January 2015 DoDI 5000.02

As before, the four major sections of the TEMP remain:

- Part I – Introduction
- Part II – Test Program Management and Schedule

- Part III – Test and Evaluation Strategy and Implementation
- Part IV – Resources Summary.
- Appendices may be added as needed for Scientific Test and Analysis Techniques, Cybersecurity, and Reliability.

Questions or suggestions about this guidebook should be addressed to Dr. Catherine Warner. She may be reached at Catherine.W.Warner.civ@mail.mil or (703) 697-3655.

J. Michael Gilmore
Director

Director, Operational Test and Evaluation (DOT&E) Test and Evaluation Master Plan (TEMP) Guidebook

Version 3.1

19 January 2017

TEMP Guidebook Table of Contents

PART 1 – Introduction

1.1 PURPOSE

1.2 MISSION DESCRIPTION

1.2.1 Mission Overview

1.2.2 Concept of Operations

CONOPS Guidance and Examples

1.2.3 Operational Users

1.3 SYSTEM DESCRIPTION

1.3.1 Program Background

1.3.2 Key Interfaces

1.3.3 Key Capabilities

1.3.4 System Threat Assessment

Threat Representation Guidance and Examples

Cybersecurity OT&E Guidance and Examples

1.3.5 Systems Engineering (SE) Requirements

Reliability Growth Guidance

1.3.6 Special Test or Certification Requirements

Threat Representation Guidance and Examples

Cybersecurity OT&E Guidance

1.3.7 Previous Testing

LFT&E Strategy Guidance

TEMP Guidebook Table of Contents

PART II – Test Program Management and Schedule

2.1 T&E MANAGEMENT

2.1.1 T&E Organizational Construct

LFT&E Strategy Guidance

2.2 COMMON T&E DATA BASE REQUIREMENTS

2.3 DEFICIENCY REPORTING

Defense Business Systems Guidance and Examples

2.4 TEMP UPDATES

2.5 INTEGRATED TEST PROGRAM SCHEDULE

Figure 2.1 – Integrated Test Program Schedule

TEMP Guidebook Table of Contents

PART III – Test and Evaluation Strategy and Implementation

3.1 T&E STRATEGY

Integrated Testing Guidance and Best Practices

3.1.1 Decision Support Key

3.2 DEVELOPMENTAL EVALUATION APPROACH

3.2.1 Developmental Evaluation Framework

3.2.2 Test Methodology

3.2.3 Modeling and Simulation

3.2.4 Test Limitations and Risks

Test Limitations Guidance and DT Examples

3.3 DEVELOPMENTAL TEST APPROACH

3.3.1 Mission-Oriented Approach

Integrated Testing Guidance and Best Practices

3.3.2 Developmental Test Events and Objectives

Integrated Testing Guidance and Best Practices

Software Algorithm Testing Guidance and Examples

Reliability Growth Guidance

Cybersecurity OT&E Guidance

3.4 CERTIFICATION FOR INITIAL OPERATIONAL TEST AND EVALUATION (IOT&E)

IOT&E Entrance Criteria Guidance and Examples

3.5 OPERATIONAL EVALUATION APPROACH

Mission Focused Evaluation Guidance and Examples

Baseline Evaluation Guidance with Best Practices

End-to-End Operational Testing Guidance and Examples

Integrated Testing Guidance and Best Practices

Integrated Survivability Assessment Guidance and Best Practices Force Protection Evaluation Guidance

Cybersecurity OT&E Guidance

Survey Design and Administration Guidance

3.5.1 Operational Test Events and Objectives

Realistic Operational Conditions Guidance and Examples

OT of Software Intensive Systems Guidance and Examples

3.5.2 Operational Evaluation Framework

Operational Evaluation Framework Guidance with Examples

Test Instrumentation Guidance and Examples

Software Evaluation Guidance with Examples

Quantitative Mission Focused Measures Guidance with Examples

Scientific Test and Analysis Techniques Guidance with Examples

Production Representative Test Articles Guidance and Examples

Test Resources Guidance and Examples

3.5.3 Modeling and Simulation

M&S for OT&E Guidance

3.5.4 Test Limitations

Test Limitations Guidance and OT Examples

3.6 LIVE FIRE EVALUATION APPROACH

LFT&E Strategy Guidance

Integrated Survivability Assessment Guidance and Best Practices

Force Protection Evaluation Guidance

3.6.1 Live Fire Test Objectives

3.6.2 Modeling and Simulation

M&S for LFT&E Guidance

3.6.3 Test Limitations

Test Limitations Guidance and LFT&E Examples

3.7 OTHER CERTIFICATIONS

3.8 FUTURE TEST AND EVALUATION

TEMP Guidebook Table of Contents

PART IV – RESOURCE SUMMARY

4.1 **INTRODUCTION**

Test Resources Guidance and Examples

4.2 **TEST RESOURCE SUMMARY**

4.2.1 Test Articles

Production Representative Test Articles Guidance and Examples

4.2.2 Test Sites

4.2.3 Test Instrumentation

Test Instrumentation Guidance and Examples

4.2.4 Test Support Equipment

4.2.5 Threat Representation

Threat Representation Guidance and Examples

4.2.6 Test Targets and Expendables

4.2.7 Operational Force Test Support

4.2.8 Models, Simulations, and Test Beds

4.2.9 Joint Operational Test Environment

4.2.10 Special Requirements

4.3 **FEDERAL, STATE, AND LOCAL REQUIREMENTS**

4.4 **MANPOWER / PERSONNEL TRAINING**

4.5 **TEST FUNDING SUMMARY**

Test Funding Guidance and Examples

APPENDIX A – BIBLIOGRAPHY

APPENDIX B – ACRONYMS

APPENDIX C – POINTS OF CONTACT

APPENDIX D – SCIENTIFIC TEST AND ANALYSIS TECHNIQUES

APPENDIX E – CYBERSECURITY

APPENDIX F – RELIABILITY GROWTH PLAN

APPENDIX G – REQUIREMENTS RATIONALE

ADDITIONAL APPENDIXES AS NEEDED

1. PART I - INTRODUCTION

1.1. PURPOSE

- State the purpose of the Test and Evaluation Master Plan (TEMP).
- Identify if this is an initial or updated TEMP.
- State the Milestone (or other) decision the TEMP supports.
- State if the program is listed on the DOT&E Oversight List or is an MDAP, MAIS, or USD(AT&L)-designated special interest program.

1.2. MISSION DESCRIPTION

1.2.1 Mission Overview

- Summarize the mission need described in the program capability requirements documents in terms of the capability the system will provide to the Warfighter.
- Describe the mission to be accomplished by a unit that will be equipped with the system.
- Incorporate an Operational View (OV-1) of the system showing the intended operational environment.
- Include significant points from the Life Cycle Sustainment Plan, the Information Support Plan, and the Program Protection Plan.
- For business systems, include a summary of the business case analysis for the program.

1.2.2 Concept of Operations

- Reference all applicable Concepts of Operations and Concepts of Employment in describing the mission. Describe test implications.
    - ***CONOPS Guidance and Examples***

1.2.3 Operational Users

- Describe the intended users of the system, how they will employ the system, and any important characteristics of the operational users (e.g., experience level, training requirements, area of specialization, etc.).
    - ***Cybersecurity OT&E Guidance and Example***

1.3 SYSTEM DESCRIPTION

- Describe the system configuration.
- Identify key features and subsystems, both hardware and software (such as architecture, system and user interfaces, security levels, and reserves) for the planned increments within the Future Years Defense Program (FYDP).

1.3.1. Program Background

- Reference the Analysis of Alternatives (AoA), the Acquisition Program Baseline (APB), the Materiel Development Decision (MDD), and the last Milestone decision (including Acquisition Decision Memorandum (ADM)) to provide background information on the proposed system.
- Briefly describe the overarching Acquisition Strategy. Address whether the system will be procured using an incremental development strategy or a single step to full capability.
- If it is an evolutionary acquisition strategy, discuss planned upgrades, additional features and expanded capabilities of follow-on increments. The main focus must be on the current increment with brief descriptions of the previous and follow-on increments to establish continuity between known increments.
- Describe the nomenclature used for increments, waves, releases, etc.

1.3.2. Key Interfaces

- Identify interfaces with existing or planned systems' architectures that are required for mission accomplishment.
- Address integration and modifications needed for commercial items. Include interoperability with existing and/or planned systems of other Department of Defense (DoD) Components, other Government agencies, or Allies.
- Provide a DoD Architectural Framework (DoDAF) that shows the different system interfaces, e.g., SV2, SV6, etc., from the Capability Development Document (CDD) or Capability Production Document (CPD).

1.3.3. Key Capabilities

- Identify the Key Performance Parameters (KPPs), Key System Attributes (KSAs), Critical Technical Parameters (CTPs), and additional important information for the system. For each listed parameter, provide the threshold and objective values from the CDD / CPD/ Technical Document and reference the CDD / CPD/ Technical Document paragraph.
- Identify Critical Operational Issues (COIs).
    - COIs should identify key elements for operationally effectiveness, operationally suitability, and survivability; they represent a significant risk if not satisfactorily resolved.
    - COIs should be few in number and reflect operational mission concerns. Existing documents such as capability requirements documents, Business Case Analysis, AoA, APB, warfighting doctrine, validated threat assessments and CONOPS may provide useful insights in developing COIs.

1.3.4. System Threat Assessment

- Describe the threat environment (to include cyber-threats) in which the system will operate. Reference the appropriate Defense Intelligence Agency (DIA) or component-validated threat documents for the system.
    - ***Threat Representation Guidance and Examples***
    - ***Cybersecurity OT&E Guidance and Example***

1.3.5. Systems Engineering (SE) Requirements

- Describe SE-based information and activities that will be used to develop the test and evaluation plan. Examples include hardware reliability growth and software maturity growth strategies. Selected Technical Performance Measures (TPMs) from the Systems Engineering Plan (SEP) should be included to show desired performance growth at various test phases.
    - ***Reliability Growth Guidance***

- Reference the SEP and ensure alignment to the TEMP.

1.3.6. Special Test or Certification Requirements

- Identify unique system characteristics or support concepts that will generate special test, analysis, and evaluation requirements.
- Identify and describe all required certifications, e.g., cybersecurity, Risk Management Framework (RMF), post deployment software support, resistance to chemical, biological, nuclear, and radiological effects; resistance to countermeasures; resistance to reverse engineering/exploitation efforts (Anti-Tamper); development of new threat simulation, simulators, or targets.
    - ***Threat Representation Guidance and Examples***
    - ***Cybersecurity Guidance***

1.3.7. Previous Testing

- Discuss the results of any previous tests that apply to, or have an effect on, the test strategy.
    - ***LFT&E Strategy Guidance***

Ensure that the narrative in Part I is consistent with the schedule in Part II, the T&E strategy in Part III, and allocated resources in Part IV. This will require iterative coordination between sub-workgroups and the T&E WIPT.

2. PART II – TEST PROGRAM MANAGEMENT AND SCHEDULE

2.1. T&E MANAGEMENT

- Discuss the test and evaluation roles and responsibilities of key personnel and organizations such as:
    - Program Office
    - Chief Developmental Tester.
    - Lead DT&E Organization
    - Prime Contractor
    - Lead OTA
    - User representative

2.1.1. T&E Organizational Construct

- Identify the organizations or activities (such as the T&E Working-level Integrated Product Team (T&E WIPT) or Service equivalent, LFT&E IPT, etc.) in the T&E management structure, to include the sub-workgroups, such as a Modeling and Simulation; Survivability; Transportability; MANPRINT/Human System Integration; Environmental, Safety, and Occupational Health (ESOH); or Reliability.
    - ***LFT&E Strategy Guidance***
- Provide sufficient information to adequately understand the functional relationships.
- Reference the T&E WIPT charter that includes specific responsibilities and deliverable items for detailed explanation of T&E management. These items include TEMPs and Test Resource Plans (TRPs) that are produced collaboratively by member organizations.

2.2. COMMON T&E DATABASE REQUIREMENTS

- Describe the provisions for and methods of accessing, collecting, validating, and sharing data as it becomes available from contractor testing, Government Developmental Testing (DT), Operational Testing (OT), and oversight organizations, as well as supporting related activities that contribute or use test data.
- Describe how the pedigree of the data will be established and maintained. The pedigree of the data refers to understanding the configuration of the test asset, and the actual test conditions under which the data were obtained for each piece of data.
- Describe the data acquisition and management approach.
- State which organization will be responsible for maintaining the data. For a common T&E database, a single organization is preferred.
- In the case where multiple organizations require separate databases, briefly justify their requirement and describe how data will be synchronized among the databases and which database will be the data of record.
- Describe how users of test data will access the data. Describe any special permissions or authorizations needed. Describe if any special tools or software are needed to read and analyze the data.
- Reference a data dictionary or similar document that clearly describes the structure and format of the database.

2.3. DEFICIENCY REPORTING

- (Post MS A TEMP) Describe the processes for documenting and tracking deficiencies identified during system development and operational testing. Relate this to the Failure Reporting, Analysis, and Corrective Action System (FRACAS) in the SEP. Describe any deficiency rating system. Describe how the deficiency reporting database is different from the common T&E database, if appropriate.
- Describe how the information is accessed and shared across the program, to include all applicable T&E organizations. The processes should address problems or deficiencies identified during both contractor and Government test activities. The processes should also include issues that have not been formally documented as a deficiency (e.g., watch items).
    - ***Defense Business System Guidance and Examples***

2.4. TEMP UPDATES

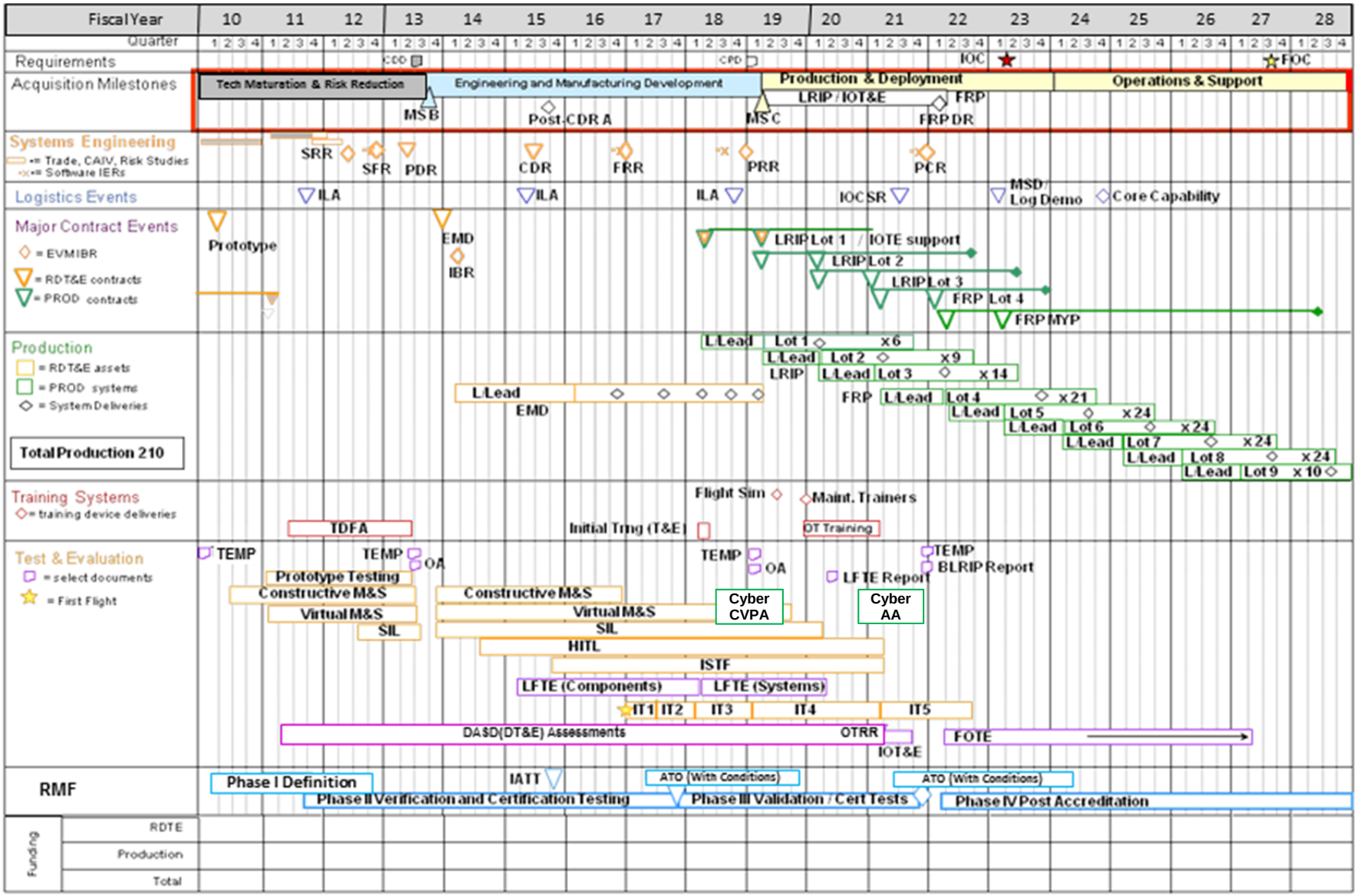
- Reference instructions for complying with DoDI 5000.02 required updates or identify exceptions to those procedures if determined necessary for more efficient administration of document.
- Provide procedures for keeping TEMP information current between updates. For a Joint or Multi-Service TEMP, identify references that will be followed or exceptions as necessary.

2.5. INTEGRATED TEST PROGRAM SCHEDULE

- Display (***see Figure 2.1***) the overall time sequencing of the major acquisition phases and milestones. Include the test and evaluation major decision points, related activities, and planned cumulative funding expenditures by appropriation by year. Ensure sufficient time is allocated between significant test events to account for test-analyze-fix-test and correction of deficiencies, assessments, and reporting.
- Include event dates such as major decision points as defined in DoD Instruction 5000.02, e.g., developmental and operational assessments, preliminary and critical design reviews, test article availability; software version releases; appropriate phases of DT&E; LFT&E; Cybersecurity testing; Joint Interoperability Test Command (JITC) interoperability testing and certification date to support the MS-C and Full-Rate Production (FRP) Decision Review (DR).
- Include significant Cybersecurity event sequencing, such as Interim Authorization to Test (IATT) and Authorization to Operate (ATO).
- Include operational test and evaluation; Low-Rate Initial Production (LRIP) deliveries; Initial Operational Capability (IOC); Full Operational Capability (FOC); and statutorily required reports such as the Live-Fire T&E Report and Beyond Low-Rate Initial Production (B-LRIP) Report.
- Provide a single schedule for multi-DoD Component or Joint and Capstone TEMPs showing all related DoD Component system event dates.

Ensure that the schedule in Part II is consistent with the narrative in Part I, the T&E strategy in Part III, and allocated resources in Part IV. This will require iterative coordination between sub-workgroups and the T&E WIPT.

Figure 2.1 SAMPLE Integrated Program Test Schedule



3. PART III – Test and Evaluation Strategy and Implementation

3.1 T&E STRATEGY

- Introduce the program T&E strategy by briefly describing how it supports the acquisition strategy as described in Section 1.3.1.
- The discussions should focus on the testing for capabilities, and address testing of subsystems or components where they represent a significant risk to achieving a necessary capability.
- Describe the scientific approach to designing an efficient test program that will characterize system performance across the operational conditions anticipated to be encountered by users. Summarize with details referenced in the appropriate appendix.
- The strategy should address the conditions for integrating DT and OT tests.
    - ***Integrated Testing Guidance and Best Practices***
- Evaluations shall include a comparison with current mission capabilities using existing data, so that measurable improvements can be determined.
    - Describe the strategy for achieving this comparison and for ensuring data are retained and managed for future comparison results of evolutionary increments or future replacement capabilities.
    - If such evaluation is considered costly relative to the benefits gained, the PM shall propose an alternative evaluation strategy.
- To present the program's T&E strategy, briefly describe the relative emphasis on methodologies (e.g., Modeling and Simulation (M&S), Measurement Facility (MF), Systems Integration Laboratory (SIL), Hardware-In-the-Loop Test (HILT), Installed System Test Facility (ISTF), Open Air Range (OAR), and Live, Virtual, and Constructive (LVC)).
- Describe the evaluation products.
    - Describe how the products will be linked.
    - Identify the organization that is providing the products and to whom they are being provided.
    - Identify the decision being supported by the products.
    - Ensure sufficient time is allocated for analysis of the products.

3.1.1. Decision Support Key

- Connect key test events to the acquisition decisions they support. Describe the information required to support such decisions.

3.2. DEVELOPMENTAL EVALUATION APPROACH

- Describe the developmental evaluation approach that will be used to support technical, programmatic, and acquisition decisions.
- Identify how the government intends to evaluate the design and development of technologies, components, subsystems, systems, and systems of systems as applicable in order to assess programmatic and technical risk.
- Describe the integrated testing approach and how it will support the overall evaluation strategy.

3.2.1. Developmental Evaluation Framework

- Embed a Developmental Evaluation Framework (DEF) in the form of a table or spreadsheet. Describe the contents of the developmental evaluation framework, including descriptions of columns and the origin of information contained. Include instructions to the reader on the use of the table or spreadsheet and its contents.
- Arrange the table or spreadsheet to show time-phased, iterative test progression toward the achievement of performance goals and measures.
- Include elements (columns, rows, or cells) bearing the following essential information:
    - Functional evaluation area. Categorical groupings of functional areas brought forward or derived from baseline documentation.
    - Decision supported. The significant program decision points where data and information gathered during testing will be used to make decisions or give program direction.
    - Decision support question. Key question related to performance, reliability, cybersecurity, or interoperability that when answered determines the outcome of an evaluation for the decision supported.
    - Key system requirements and T&E measures (one or more fields of requirements identification and performance measurement).
        - Technical requirements document reference.
        - Description.
        - Technical measures. CTP, TPM, Metrics.
    - Method (technique, process, or verification method).
    - Test Event.
    - Resources. Brief reference may appear here.
    - 3Cross-Reference. Used to refer to related requirements, capabilities, and line items to aid in requirements traceability, precedence, interdependency, and causality.

3.2.2. Test Methodology

- For each capability and key functional area, address a test methodology that:
    - Verifies achievement of critical technical parameters and the ability to achieve key performance parameters, and assess progress toward achievement of critical operational issues.
    - Measures the system's ability to achieve the thresholds prescribed in the capabilities documents.
    - Provides data to the Program Manager to enable root cause determination and to identify corrective actions.
    - Measures system functionality.
    - Provides information for cost, performance, and schedule tradeoffs.
    - Assesses system specification compliance.
    - Identifies system capabilities, limitations, and deficiencies.
    - Assesses system safety.
    - Assesses compatibility with legacy systems.
    - Stresses the system within the intended operationally relevant mission environment.
    - Supports cybersecurity assessments and authorizations.
    - Supports the interoperability certification process.
    - Documents achievement of contractual technical performance and verifies incremental improvements and system corrective actions.
    - Provides DT&E data to validate parameters in models and simulations.
    - Assesses the maturity of the chosen integrated technologies.

3.2.3. Modeling and Simulation (M&S)

- Describe the key models and simulations and their intended use. Include the developmental test objectives to be addressed using M&S to include any approved operational test objectives.
- Identify who will perform M&S verification, validation, and accreditation.
- Identify data needed and the planned accreditation effort.
- Identify how the developmental test scenarios will be supplemented with M&S, including how M&S will be used to predict the Sustainment KPP and other sustainment considerations.
- Identify and describe LVC requirements.
- Identify developmental M&S resource requirements in Part IV.

3.2.4. Test Limitations and Risks

- Discuss any developmental test limitations that may significantly affect the evaluator's ability to draw conclusions about the maturity, capabilities, limitations, or readiness for dedicated operational testing.
- Address the impact of these limitations as well as resolution approaches.
- Discuss any known test risks at the time the TEMP is being written. These are risks that may prevent or delay the satisfactory execution of the test events. Any test risks that are included in the program-level risk management database should be included. Include a risk mitigation plan for the identified test risks.
    - ***Test Limitations Guidance and DT Examples***

3.3. DEVELOPMENTAL TEST APPROACH

3.3.1. Mission-Oriented Approach

- Describe the approach to test the system performance in a mission context, i.e., how the system will actually be employed.
- Discuss how developmental testing will reflect the expected operational environment to help ensure developmental testing is planned to integrate with operational testing.
- Describe the use of actual user subjects to support human factors engineering assessments and NET development.
    - ***Integrated Testing Guidance and Best Practices***

3.3.2. Developmental Test Events (Description, Scope, and Scenario) and Objectives

- For each developmental test event shown in the schedule and the DEF, prepare a subparagraph that summarizes: Who is the lead test organization; the objectives of the test event, the test event's schedule; other associated test events, location(s), etc.
- Summarize the planned objectives and state the methodology to test the system attributes defined by the applicable capability requirement document (CDD, CPD, CONOPS) and the CTPs that will be addressed during each phase of DT. Subparagraphs can be used to separate the discussion of each phase.
- For each DT phase, discuss the key test objectives to address both the contractor and Government developmental test concerns and their importance to achieving the exit criteria for the next major program decision point. If a contractor is not yet selected, include the developmental test issues addressed in the Request for Proposals (RFPs) or Statement of Work (SOW).
- Address measurable exit/entrance criteria for each major T&E phase and milestone decision points.
- Discuss how developmental testing will reflect the expected operational environment to help ensure developmental testing is planned to integrate with operational testing.
    - ***Integrated Testing Guidance and Best Practices***
    - ***Software Algorithm Testing Guidance and Examples***
- Include key test objectives related to logistics testing.
- Summarize the developmental test events, test scenarios, and the test design concept.
- Quantify the testing sufficiently (e.g., number of test hours, test articles, test events, test firings) to allow a valid cost estimate to be created.
- Identify and explain how models and simulations, specific threat systems, surrogates, countermeasures, component, or subsystem testing, test beds, and prototypes will be used to determine whether or not developmental test objectives are achieved.
- Identify the DT&E reports required to support decision points/reviews and OT readiness.
- Address the system's reliability growth strategy, goals, and targets and how they support the Developmental Evaluation Framework. Detailed developmental test objectives should be addressed in the System Test Plans and detailed test plans (Provide specific details in Appendix F – Reliability Growth Plan).
    - ***Reliability Growth Guidance***
- Discuss plans for interoperability and cybersecurity testing, including the use of cyber ranges for vulnerability and adversarial testing (Provide specific details in Appendix E – Cybersecurity).
    - ***Cybersecurity OT&E Guidance and Example***

3.4. CERTIFICATION FOR INITIAL OPERATIONAL TEST AND EVALUATION (IOT&E)

- Explain how and when the system will be certified safe and ready for IOT&E.
- Explain who is responsible for certification and which decision reviews will be supported using the lead Service's certification of safety and system materiel readiness process.
- List the DT&E information (i.e., reports, briefings, or summaries) that provides predictive analyses of expected system performance against specific COIs and the key system attributes – measures of effectiveness (MOE) and measures of suitability (MOS).
- Discuss the entry criteria for IOT&E and how the DT&E program will address those criteria.
    - ***IOT&E Entrance Criteria Guidance and Examples***

3.5. OPERATIONAL EVALUATION APPROACH

- Summarize the mission focused evaluation methodology and supporting test strategy, including the essential mission and system capabilities that contribute to operational effectiveness, suitability, and survivability.
    - ***Mission Focused Evaluation Guidance and Examples***
    - ***Baseline Evaluation Guidance with Best Practices***
    - ***End-to-End Operational Testing Guidance and Examples***
    - ***Cybersecurity OT&E Guidance and Example***
    - ***Survey Design and Administration Guidance***
- Summarize the operational test events, key threat simulators and/or simulation(s) and targets to be employed, and the type of representative personnel who will operate and maintain the system.
- Summarize integrated testing strategy to include:
    - Developmental test data that will be used for operational evaluation
    - Conditions on data pedigree and test conduct to make data suitable for operational evaluation
    - ***Integrated Testing Guidance and Best Practices***
    - ***Integrated Survivability Assessment Guidance and Best Practices***
    - ***Force Protection Evaluation Guidance***

3.5.1 Operational Test Events and Objectives

- Identify the key operational test objectives for each test event and test phase
- Outline the approach for characterizing the COIs and important MOEs/MOSs across relevant operational conditions.
    - ***Realistic Operational Conditions Guidance and Examples***
    - ***OT of Software Intensive Systems Guidance and Examples***

3.5.2 Operational Evaluation Framework The evaluation framework should identify and link:

- The goal of the operational test within a mission context
- The mission-oriented response variables, the factors that affect those variables, and he required test resources
- (Post MS A TEMP) The test designs for strategically varying the factors across the operational envelope
- ***Operational Evaluation Framework Guidance with Examples***
- ***Test Instrumentation Guidance and Examples***
- ***Software Evaluation Guidance with Examples***

- The evaluation framework should focus on the subset of mission-oriented measures critical for assessing operational effectiveness, suitability, and survivability.
    - ***Mission Focused Metrics Guidance with Examples***
- (Post MS A TEMP) Use a systematic, rigorous, and structured approach to link major test events and phases to quantitatively evaluate system capabilities across relevant operational conditions.
- (Post MS A TEMP) Describe the statistical test design strategy and corresponding statistical measures of merit (e.g., confidence and power).
    - ***Scientific Test and Analysis Techniques Guidance with Examples***
- Identify planned sources of information (e.g., developmental testing, testing of related systems, modeling, simulation) that may be used to supplement operational test and evaluation.
- Describe the scope of the operational test by identifying the test mission scenarios and the resources that will be used to conduct the test.
    - ***Production Representative Test Articles Guidance and Examples***
    - ***Test Resources Guidance and Examples***

3.5.3 Modeling and Simulation (M&S)

- If described in either the DT&E or Live Fire sections, do not repeat. Just reference and hyperlink. Only discuss what is unique to OT&E.
- Describe the key models and simulations and their intended use.
- Include the operational test objectives to be addressed using M&S.
- (Post MS A TEMP) Identify who will perform the M&S verification, validation, and accreditation.
- (Post MS A TEMP) Identify data needed and the planned accreditation effort.
- Identify how the operational test scenarios will be supplemented with M&S.
- Identify operational M&S resource requirements in Part IV.
    - ***M&S for OT&E Guidance***

3.5.4 Test Limitations

- Discuss test limitations including threat realism, resource availability, limited operational (military; climatic; Chemical, Biological, Nuclear, and Radiological (CBNR), etc.) environments, limited support environment, maturity of tested systems or subsystems, safety, that may impact the resolution of affected COIs.
- Describe measures taken to mitigate limitations.
- Indicate if any system contractor involvement or support is required, the nature of that support, and steps taken to ensure the impartiality of the contractor providing the support according to Title 10 U.S.C. §2399.
- Indicate the impact of test limitations on the ability to resolve COIs and the ability to formulate conclusions regarding operational effectiveness and operational suitability.
- Indicate the COIs affected in parentheses after each limitation.
    - ***Test Limitations Guidance and OT Examples***
    - ***Cybersecurity OT&E Guidance and Example***

3.6. LIVE FIRE TEST AND EVALUATION APPROACH

- If live fire testing is required, describe the approach to evaluate the survivability/lethality of the system, and (for survivability LFT&E) personnel survivability of the system's occupants.
  - ***LFT&E Strategy Guidance***
  - ***Integrated Survivability Assessment Guidance and Best Practices***
  - ***Force Protection Evaluation Guidance***
- Include a description of the overall live fire evaluation strategy to influence the system design (as defined in Title 10 U.S.C. § 2366), critical live fire evaluation issues, and major evaluation limitations.
- Discuss the management of the LFT&E program, to include the shot selection process, target resource availability, and schedule.
- Discuss a waiver, if appropriate, from full-up, system-level survivability testing, and the alternative strategy.

3.6.1. Live Fire Test Objectives

- State the key live fire test objectives for realistic survivability or lethality testing of the system.
- Include a matrix that identifies all tests within the LFT&E strategy, their schedules, the issues they will address, and which planning documents will be submitted for DOT&E approval and which will be submitted for information and review only.
- Identify whether full-up, system-level testing will be conducted, or whether a waiver will be required from such testing. If a waiver will be required from full-up, system-level testing, describe the key features of the alternative LFT&E plan, including the planned levels of test realism to support the evaluation of survivability or lethality.
- Quantify the testing sufficiently (e.g., number of test hours, test articles, test events, test firings) to allow a valid cost estimate to be created.

3.6.2. Modeling and Simulation (M&S)

- Only discuss what is unique to live fire.
- Describe the key models and simulations and their intended use.
- If M&S is to be used for test planning, describe how M&S will be used as a basis for decisions regarding test scope or test conditions.
- If M&S is to be used for prediction of test results, identify which tests will have predictions based on M&S, and which models will be used for such predictions.
- If M&S is to be used for evaluation of critical LFT&E issues, summarize the degree of reliance on M&S, and identify any evaluation issues that will be addressed solely by M&S.
- Include the LFT&E test objectives to be addressed using M&S to include operational test objectives.
- (Post MS A TEMP) Identify who will perform M&S verification, validation, and accreditation
- (Post MS A TEMP) Identify data needed and the planned accreditation effort.
- Identify how the test scenarios will be supplemented with M&S.
- Identify and describe LVC requirements.
- Identify M&S resource requirements in Part IV.
  - ***M&S for LFT&E Guidance***

3.6.3. Test Limitations

- Discuss any test limitations that may significantly affect the ability to assess the system's vulnerability and survivability.
- Also address the impact of these limitations, and resolution approaches.
  - ***Test Limitations Guidance and LFT&E Examples***

3.7. OTHER CERTIFICATIONS

- Identify key testing prerequisites and entrance criteria, such as required certifications (e.g. DoD Risk Management Framework (RMF), Authorization to Operate, Weapon Systems Explosive Safety Review Board (WSERB), flight certification, etc.)

3.8. FUTURE TEST AND EVALUATION

- Summarize all remaining significant T&E that has not been discussed yet, extending through the system life cycle.
  - Significant T&E is that T&E requiring procurement of test assets or other unique test resources that need to be captured in the Resource section.
  - Significant T&E can also be any additional questions or issues that need to be resolved for future decisions.
- Do not include any T&E in this section that has been previously discussed in this part of the TEMP.

Ensure that the T&E strategy in Part III is consistent with the narrative in Part I, the schedule in Part II, and allocated resources in Part IV. This will require iterative coordination between sub-workgroups and the T&E WIPT.

4. PART IV-RESOURCE SUMMARY

4.1. INTRODUCTION

- In this section, specify the resource elements, both government and contractor, necessary to plan, execute, and evaluate a test event or test campaign.
    - ***Test Resources Guidance and Examples***
- Resource elements include test articles, models, simulations, test facilities, manpower for test conduct and support, and other items that are described below.
- Resource estimates must be quantifiable and defensible, derived from STAT methodologies (identified in the evaluation framework and included in the STAT section or appendix) and where appropriate, based on test experience.
- Testing will be planned and conducted to take full advantage of existing DoD investment in ranges, facilities, and other resources wherever practical. Justify use of non-government facilities.
- Along with each resource element, include an estimate of element quantity, when the elements will be used (consistent with figure 2.1 schedule), the organization responsible for providing them, and their cost estimate (if available).
- Include long-lead items for the next increment if known.
- Callout any shortfalls, their impact on planned T&E, and describe an appropriate mitigation.

Use of tables to more accurately convey information for each of the sub-paragraphs below is encouraged. See TEMP Guide for real world TEMP examples.

4.2. TEST RESOURCE SUMMARY

4.2.1. Test Articles

- Identify the actual number of and timing requirements for all test articles, including key support equipment and technical information required for testing in each phase of DT&E, LFT&E, and OT&E.
    - ***Production Representative Test Articles Guidance and Examples***
- If key subsystems (components, assemblies, subassemblies or software modules) are to be tested individually, before being tested in the final system configuration, identify each subsystem in the TEMP and the quantity required. Specifically identify when prototype, engineering development, or production models will be used.

4.2.2. Test Sites

- Identify the specific test ranges/facilities and schedule to be used for each type of testing.
- Compare the requirements for test ranges/facilities dictated by the scope and content of planned testing with existing and programmed test range/facility capability.
- Summarize the results of a cost benefit analysis (CBA) in those cases where government test facilities are not used.
- Test Facilities may include the following and other test venues:
    - Digital Modeling and Simulation Facility (DMSF).
    - Measurement Facility (MF).
    - System Integration Laboratory (SIL).
    - Hardware-in-the-Loop (HWIL) Facility.
    - Installed System Test Facility (ISTF).
    - Open Air Ranges (OAR).
    - Cyber Ranges.
    - Distributed Live, Virtual, and Constructive (DLVC) Environments.

4.2.3. Test Instrumentation

- Identify instrumentation that must be acquired or built specifically to conduct the planned test program
    - ***Test Instrumentation Guidance and Examples***
- Identify the specific data classes that the instrumentation will capture and relate it to the DEFM.
- Identify any special tools or software that analysts or evaluators will need to read the data from the instrumentation.

4.2.4. Test Support Equipment

- Identify test support equipment and schedule specifically required to conduct the test program. Anticipate all test locations that will require some form of test support equipment.
- This may include test measurement and diagnostic equipment, calibration equipment, frequency monitoring devices, software test drivers, emulators, or other test support devices that are not included under the instrumentation requirements.
- Identify special resources needed for data analysis and evaluation.

4.2.5. Threat Representation

- Identify the type (actual or surrogates, jammers, opposing forces, air defense systems, cyber), number, availability, fidelity requirements, and schedule for all representations of the threat (to include threat targets) to be used in testing.
- Include the quantities and types of units and systems required for each of the test phases. Appropriate threat command and control elements may be required and utilized in both live and virtual environments. The scope of the T&E event will determine final threat inventory.
    - ***Threat Representation Guidance and Examples***
    - ***Cybersecurity OT&E Guidance and Example***

4.2.6. Test Targets and Expendables

- Specify the type, number, availability, and schedule for all test targets (actual and surrogates) and expendables, (e.g. targets, weapons, flares, pyrotechnics, chaff, sonobuoys, smoke generators, countermeasures) required for each phase of testing.
- Include threat targets for LFT&E lethality testing and threat munitions for vulnerability testing.

4.2.7. Operational Force Test Support

- Identify doctrinally-representative systems and trained operators necessary to execute a test event.
- For each test and evaluation phase, specify the type and timing of aircraft flying hours, ship steaming days, and on-orbit satellite contacts/coverage, and other operational force support required.
- Include supported/supporting systems that the system under test must interoperate with if testing a system-of-systems or family-of-systems.
- Include size, location, and type unit required.

4.2.8. Models, Simulations, and Test-Beds

- For each test and evaluation phase, specify the models, simulations, any hybrid tool (e.g. simulation over live system) and simulations to be used, including computer-driven simulation models and hardware/software-in-the-loop test beds.
- Identify opportunities to simulate any of the required support.
- Include the resources required to verify, validate, and accredit the models, simulations, and hybrid tool usage.
- Identify the resources required to validate and accredit their usage, responsible agency and timeframe.

4.2.9. Joint Operational Test Environment

- Describe the live, virtual, or constructive components or assets necessary to create an acceptable environment to evaluate system performance against stated joint requirements.
- Describe how both DT and OT testing will utilize these assets and components.
- Describe distributed testing events. The Joint Mission Environment Test Capability (JMETC) should be considered as a resource for distributed testing.

4.2.10. Special Requirements

- Identify requirements and schedule for any necessary non-instrumentation capabilities and resources such as: special data processing/data bases, unique mapping/charting/geodesy products, extreme physical environmental conditions or restricted/special use air / sea / landscapes.
- Briefly list any items impacting the T&E strategy or government test plans that must be put on contract or which are required by statute or regulation. These are typically derived from the JCIDS requirement (i.e., Programmatic Environment, Safety and Occupational Health Evaluation (PESHE) or Environment, Safety and Occupational Health (ESOH)).
- Identify frequency management and control requirements
- Include key statements describing the top-level T&E activities the contractor is responsible for and the kinds of support that must be provided to government testers.

4.3. FEDERAL, STATE, AND LOCAL REQUIREMENTS

- All T&E efforts must comply with federal, state, and local environmental regulations. Current permits and appropriate agency notifications will be maintained regarding all test efforts.
- Specify any National Environmental Policy Act documentation needed to address specific test activities that must be completed prior to testing and include any known issues that require mitigations to address significant environmental impacts.
- Describe how environmental compliance requirements will be met.

4.4. MANPOWER / PERSONNEL AND TRAINING

- Include T&E personnel numbers for the program office, lead DT&E organization, OTA, SME analysts, and other evaluators (e.g. JITC, DISA, cybersecurity assessment teams).
- Include contractor personnel and specify the kinds of support that they must provide to government testers.
- Specify manpower/personnel and training requirements and limitations that affect test and evaluation execution.
- Identify how much training will be conducted with M&S.
- Identify TDY and travel costs.

4.5. TEST FUNDING SUMMARY

- Summarize cost of testing by FY separated by major events or phases and within each Fiscal Year (FY) DT and OT dollars.
    - ***Test Funding Guidance and Examples***
- When costs cannot be estimated, identify the date when the estimates will be derived.
- Funding should be aligned with the most current Congressional budget justifications, e.g., R2s, R3s, TE-1s, etc.

Ensure that the allocated resources in Part IV is consistent with the narrative in Part I, the schedule in Part II, and the T&E strategy in Part III. This will require iterative coordination between sub-workgroups and the T&E WIPT.

Baseline Evaluation – Guidance

Summary

The primary objective of Defense acquisition is to acquire quality products that satisfy user needs with measurable improvements to mission capability and operational support, in a timely manner, and at a fair and reasonable price.

One way to determine "measurable improvements" is through comparative or baseline evaluation, which compares unit mission accomplishment when equipped with the new system to unit mission accomplishment when equipped with the legacy system. This comparison is in addition to assessing a new system's achievement of its required performance characteristics.

Typically, many uncontrollable variables are present during operational testing, especially in force-on-force exercises. Areas where commonality should be sought between trials in order to enable valid comparisons include: the mission to be accomplished; the size, organization, and capability of the enemy force; the terrain (or environment) where the test is conducted; the size, organization, and capability of the Blue forces; and time available to accomplish the mission.

Best Practices

Conduct a side-by-side operational test, as during the Stryker IOT&E, with a unit equipped Stryker and another unit equipped with the legacy system.

In the M2A3 Bradley IOT&E, the M2A3 Bradley unit conducted operations against a M2A1 Bradley unit for a head-to-head comparison.

In the Apache Block III IOT&E, mission performance of an Air Weapons Team (AWT) with Apache Block III was compared to mission performance of an AWT with legacy Block 2 Apache. The operational effectiveness of improved Block III flight performance was decisively demonstrated when the AWT with legacy Apache could not successfully accomplish a mission in high, hot, windy conditions that was successfully accomplished by the Block III AWT with power to spare.

The Task Force XXI Advanced Warfighting Experiment at the National Training Center used three NTC rotations to establish a baseline for normal unit performance.

Analysis of Alternatives can be helpful in determining the factors and levels to be examined, and also for estimating baseline force performance in field trials.

The Navy made effective use of hardware-in-the-loop (HWIL) M&S to support the evaluation of heavyweight torpedoes. The OT objective was to assess a form-fit-functional replacement of the weapon's Guidance and Control section running a rehosted version of the tactical software. The HWIL simulation allowed testers to run both the legacy and upgraded systems through a series of identical scenarios and compare the results. A limited number of in-water trials were conducted to validate the model and verify system suitability. This M&S

approach provided a large, well-controlled data sample to compare the performance of the two variants in similar conditions.

Reference

Test and Evaluation Policy Revisions, DOT&E, December 22, 2007

CONOPS – Guidance

Guidance

To prepare an adequate Test and Evaluation Strategy, T&E practitioners must understand how the system will be employed and the anticipated employment environment. Every system should have a written concept of operations (CONOPS), operational mode summary (OMS) / mission profile (MP), field manual, table of organization and equipment, tactical operations manual, or tactics, techniques, and procedures manual. All TEMPS, to include the MS A TEMP should reference these documents. Any aspects of the CONOPS/OMS/MP that may require significant consideration for testing, such as specialized units, target sets, ranges, threat emulators, or long production lead times should be highlighted. The number of system units to be employed by the user in the context of an operational scenario (e.g., number of systems in a company), are identified to help scope the test program's resources. If the new system capability is intended to be applicable to a joint force, the joint aspects of the test program should be described.

The CONOPS need not be replicated in the TEMP.

Example

1.2.2 Concepts of Operations. The Chinook supports the Army's requirement to be strategically responsive across the full spectrum of operations. The Chinook enhances the Army's ability to support the rapid response capability necessary for forcible and early entry contingency missions and the tactical and operational noncontiguous, simultaneous or sequential operations, which will be characteristic of future operations. The Chinook provides a heavy lift capability that enables the force to accomplish critical tasks across the Battle Functional Areas of maneuver, maneuver support and maneuver sustainment by conducting air assault, air movement, mass casualty evacuation, aerial recovery, and aerial resupply across the full spectrum of operations. The Chinook provides the means to continue the time sensitive transport of personnel, equipment, and supplies not available from other transportation systems. The High Level Operational Concept graphic, OV-1, which is in Figure 1 below, depicts the Chinook mission environment. OV-1 provides a description of the interactions between the Chinook and its operational environment and highlights the importance and complexity of interoperability for successful Chinook employment.

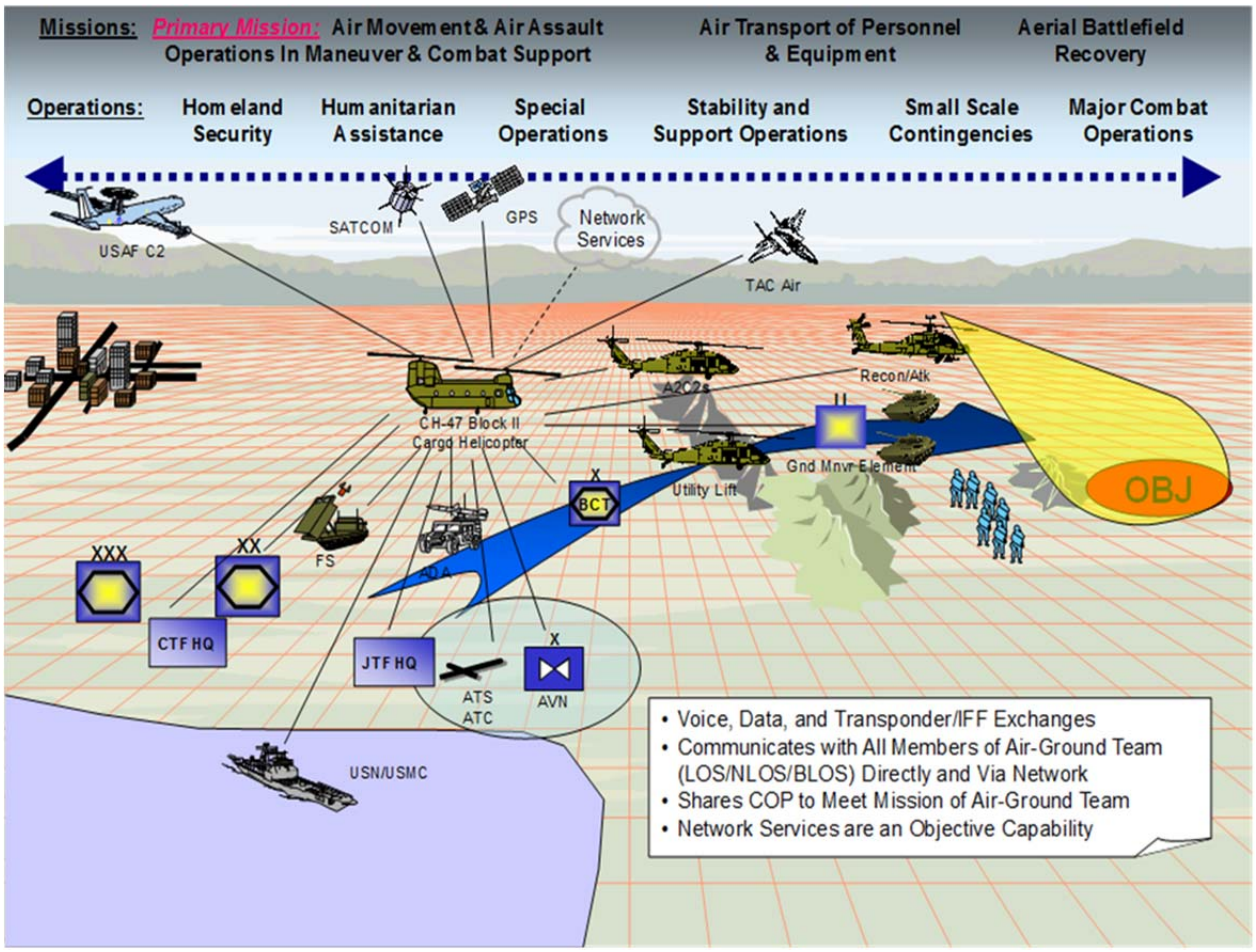


Figure 1 - Chinook Operational Concept Graphic (OV-1)

Cybersecurity OT&E – Guidance

General Guidance

The TEMP should describe a test and evaluation strategy for cybersecurity that uses relevant data from all sources and includes testing production representative systems in an operationally representative environment. Data sources may include, but are not limited to, information security assessments, inspections, component and subsystem level tests, and system of system tests. As needed, the TEMP can provide details on the cybersecurity test and evaluation strategy in Appendix E.

The purpose of testing cybersecurity during operational testing is to assess the ability of the system to enable operators to execute critical missions and tasks in the expected operational environment. Testing of cybersecurity during OT&E includes the representative users and an operationally representative environment that may include hardware, software (including embedded software and firmware), operators, maintainers, operational cyber/network defense, end users, network and system administrators, help desk, training, support documentation, Tactics, Techniques, and Procedures, cyber threats, and other systems that exchange information with the system under test.

In the memorandum, "Procedures for Operational Test and Evaluation of Cybersecurity in Acquisition Programs" (April 3, 2018), henceforth referred to as DOT&E 2018, DOT&E specifies that OTAs should evaluate cybersecurity in OT&E via two assessments: a Cooperative Penetration and Vulnerability Assessment (CVPA) and an Adversarial Assessment (AA). The OTA should design the CVPA and AA to identify cyber vulnerabilities, examine attack paths, evaluate operational cyber defense capabilities, and establish the operational mission effects (loss of critical operational capability) in a cyber threat environment while conducting operational missions.

A CVPA characterizes the cybersecurity and resilience of a system in an operational context and provides reconnaissance information about the system in support of the tests providing information for the AA. If possible, the OTA should conduct the events providing data for the CVPA far enough in advance of the AA to enable mitigation of vulnerabilities before proceeding to the AA, but close enough to remain a relevant input to AA planning. The CVPA requires information from events that include an operationally and production representative system unless specific differences are defined and approved by DOT&E prior to execution. The events can be a standalone test events, a series of events (separate from or embedded in other tests), or an operational component of integrated test.

The AA characterizes the operational effects to critical missions caused by threat-representative cyber activity against a unit training and equipped with a system, as well as the effectiveness of defensive capabilities. The AA requires information from events that include a production-representative and operationally configured system with representative operators, users, and cyber defenders, an operational network configuration, and representative missions. Missions can include military, business, command and control, and cyber tasks. The AA events

should include third party or external defenders including those responsible for defending networks connecting to the system under test. The scope of defensive capabilities and extent of defender roles should match the operational deployment and concept of operations for the system. The AA requires information from events conducted in concert with other operational testing, but might require closed environments, cyber ranges, or other operationally representative tools to demonstrate mission effects. OTAs will ensure verification, validation, and accreditation of these closed environments, cyber ranges, or tools according to Service standards.

For information systems that manage financial/fiscal/business activities or funds, OTAs should conduct a Cyber Economic Vulnerability Assessment (CEVA), see the DOT&E memorandum, "Cyber Economic Vulnerability Assessments (CEVA)" dated January 21, 2015.

Cybersecurity Information for the Body of the TEMP

The TEMP describes the cybersecurity OT&E strategy in the following paragraphs:

- Paragraph 1.3. System Description. Describe the operational configuration and the concept of operations for deployment and operation of the system. Specify and identify the cyber defense responsibilities of the system users, any dedicated system cyber defenders, and the cyber defenders supporting the networks and enclaves on which the system will connect and operate. Identify whether the system has specialized components such as cross-domain solutions, industrial control systems, non-internet data transfers, and data transfer via alternate media such as radio frequency and data links.
- Paragraph 1.3.4. System Threat Assessment. Describe the cyber portion of the complete threat environment in which the system will operate. An advanced cyber threat is appropriate for all systems. Reference the applicable available threat documents, including but not limited to the most recent Defense Intelligence Agency Computer Network Operations Capstone Threat Assessment and component-validated threat documents for the system.
- Paragraph 2.5. Integrated Test Program Schedule. Show on the Integrated Program Test Schedule all events (tests, inspections, cyber table tops, demonstrations, etc.) that will provide information for the CVPA, AA, and, if required, the CEVA (Figure 2.1).
- Paragraph 3.3.2. Developmental Test Events. Identify any developmental and integrated test events that will provide data to support the OT&E assessments, identify the OTA for the event, and confirm plans to obtain DOT&E approval of (1) any specific differences from an operationally and production representative system, and the (2) the operational test plan for the integrated event.
- Paragraph 3.5. Operational Evaluation Approach. Describe the strategy for using the cybersecurity results to inform the overall evaluation of operational effectiveness, suitability, and survivability. Confirm that the OT&E strategy and design will examine operational resilience, including the key attributes of prevent, mitigate, and recover.

- Paragraph 3.5.1 Operational Test Events and Objectives. Describe the strategy (single event, multiple events, etc.) for providing the information needed for the CVPA, the AA, and, if required, the CEVA. List the critical issues and measures for cybersecurity.
- Paragraph 3.5.1.1 Events Supporting the Cooperative Vulnerability and Penetration Assessment. For each event providing information for the CVPA, provide the schedule, identify the involved organizations, list the complete set of limitations and constraints with implications for the assessments, describe the architecture of the system under test including anticipated differences from an operationally and production representative system, describe the operational environment, identify activities (e.g., system and network scans, penetration tests, access control checks, physical inspections, personnel interviews, and reviews of system architecture), and describe expected information (which portions of the information in Attachment B of DOT&E 2018). If planning multiple events, then confirm that the set of events combined will provide the complete set of information in Attachment B of DOT&E 2018.
- Paragraph 3.5.1.2 Events Support the Adversarial Assessment. For each event providing information for the AA, provide the schedule, identify the involved organizations, list the complete set of limitations and constraints with implications to the assessments, describe the architecture of the system under test including any anticipated differences from an operationally and production representative system, describe the operational environment, identify activities (e.g., attacks to create observable mission effects, assessment of mission effects using closed environments, cyber ranges, or other tools, "white cards,", etc.), and describe expected information (which portions of Attachment C of DOT&E 2018). If planning multiple events, then confirm that the set of events will provide the complete set of information in Attachment C of DOT&E 2018.
- Paragraph 3.5.1.3 Cyber Economic Vulnerability Assessment (if required). Identify the test teams that will support the CEVA; this should include both a cyber team and an accounting firm. Name the system and economic subject matter experts who will assist in the Cyber Economic Threat Analysis and assess the mission effects of exploitation, and discuss their qualifications for these roles.
- Paragraph 3.5.1.4. Cybersecurity Test Architecture. Either include the diagrams or provide references to available documents that provide diagrams (e.g., DoD Architectural Framework, system specifications, etc.) with the following information:
  - Major sub-systems (e.g., guidance and communication)
  - Connections between the subsystems including their protocols (e.g., target identification receives input from both Link 16 and the fire control radar via a 1553 data bus)
  - External connections, direct (e.g., NIPRNet, SIPRNet, or JWICS) or indirect (e.g., maintenance laptop, Mission Planning System data transfer devices)

  - Physical access points (e.g., operator consoles) and removable media ports (e.g., USB ports, CD/DVD drives)
  - Other systems to which the system will connect (e.g., SATCOM)

- Paragraph 3.5.2.1. Cybersecurity Critical Issues. Identify the critical issues affected by cybersecurity and describe the cybersecurity evaluation criteria.
- Paragraph 3.5.4 Test Limitations. Identify common restrictions that apply to all test events and discuss how these restrictions affect the efficacy or realism of the CVPA, AA, or CEVA (e.g., safety restrictions on altering system data during operations) and any associated mitigations (e.g., white cards, validated laboratory environment).
- Paragraph 4.2.5 Resources for Cybersecurity Tests. Identify resources required to conduct the tests providing information for the CVPA, AA, and CEVA to include funding, organizations, participants (system operators, inherent cyber defenders, external cyber defenders, test teams, etc), test assets (tools, software, data collection, closed environments, cyber ranges, etc.), and related efforts such as verification, validation, and accreditation efforts. Specifically identify capabilities that the cyber teams do not already possess and must develop (attack tools for non-Internet Protocol buses, threat-representative attack capability against specialized components such as cross-domain solutions) and any system-specific capability (closed environments, cyber ranges, or other tools).

Cybersecurity OT&E Information for Appendix E

Appendix E provides an opportunity for expanded discussion of details not already stated in the body of the TEMP. Detailed system architectures and diagrams are an example of the type of information for the appendix. The TEMP only needs a cybersecurity appendix if the main body does not provide references for or includes all needed information.

Examples

TBD

References

Procedures for Operational Test and Evaluation of Cybersecurity in Acquisition Programs, DOT&E, 3 April 2018

Cyber Economic Vulnerability Assessments (CEVA), DOT&E, 21 January 2015

Cybersecurity Test and Evaluation Guidebook, DoD, 1 July 2015

Operational Test Conditions for Cross Domain Solutions, 1 August 2016

This page intentionally left blank.

Cybersecurity – TEMP Body Example

1.3. System Description

(...) A unit equipped with TGVS performs armed reconnaissance missions and provides operators with sensors and weapons to observe and engage enemies. TGVS uses the Single Channel Ground and Airborne Radio System (SINCGARS) and Force XXI Battle Command Brigade and Below (FBCB2) systems to communicate digitally with other TGVSs and tactical vehicles on the battlefield.

The TGVS comprises the ground vehicle with its integrated sensors, weapons, computers, displays, controls, external data links, and other networked devices hosted on board the vehicle. Systems that connect with the TGVS vehicle include the maintenance support device and the remote computer display unit. Communications include IP and Controller Area Network (CAN) data bus traffic. External data sources including NIPRNet provide data used by the maintenance components of TGVS. Units equipped with the TGVS perform cyber defense functions interoperating with the U.S. Army Cyber Command (ARCYBER) Regional Cyber Centers (RCCs).

1.3.4. System Threats (...) A full range of cyber adversaries with nascent, limited, moderate, and advanced capabilities will target the Tactical Ground Vehicle System (TGVS). Adversaries will attempt to compromise the system; exfiltrate, infiltrate, or corrupt data; disrupt system operations; and, if possible, physically destroy equipment. Additional information on cyber threats to the TGVS is provided in the TGVS System Threat Assessment Report (STAR) and the Computer Network Operations Capstone Threat Assessment (IO Capstone, Volume 10) (CORRECTED), 2nd Edition, May 2013, DIA-08-1209-908.A. (…)

3.5. Operational Evaluation Approach

(...) The OTA will use the results of TGVS cybersecurity testing, in part, to determine its operational effectiveness, suitability, and survivability. These evaluations should take into account the results of any bench testing.

3.5.1. Cybersecurity Operational Test Events and Objectives. The Operational Test Agency (OTA) will perform cybersecurity testing as part of OT&E for the Tactical Ground Vehicle System (TGVS) in accordance with 1 Aug 2014 DOT&E guidance. Prior to these tests, TGVS will have a signed Authority to Operate. The overall schedule of cybersecurity testing events is shown in Figure 3-1. *<If the CVPA and AA scheduling is not already denoted in the integrated test schedule in the body of the TEMP >*

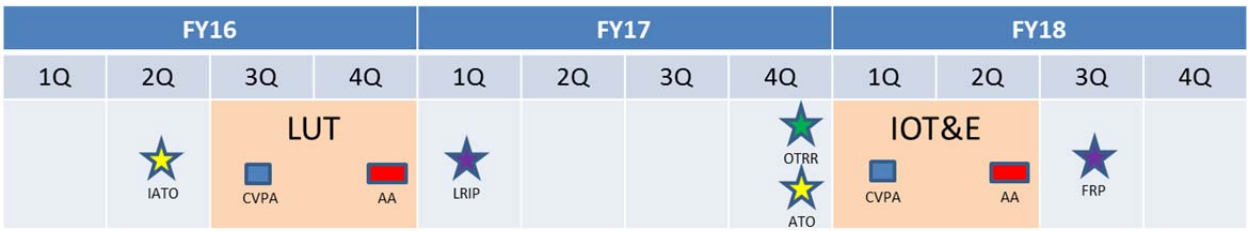


Figure 3-1. TGVS Cybersecurity Test Schedule

3.5.1.1. Cooperative Vulnerability and Penetration Assessment (CVPA). The OTA will employ the Army Research Laboratory Survivability/Lethality Analysis Directorate (ARL/SLAD) to perform Cooperative Vulnerability and Penetration Assessments (CVPAs) during both the LUT and the IOT&E prior to Adversarial Assessments. ARL/SLAD will perform the CVPAs on an operationally representative TGVS, including the use of local cybersecurity defenders such as system operators, maintainers, and system administrators to support data collection (e.g., through interviews), while the TGVS is in the motor pool with all systems present and powered. ARL/SLAD will execute vulnerability and penetration testing using their accredited tools and processes, which include automated scans and manual inspection. The TGVS will have all external interfaces active, and ARL/SLAD will conduct assessment activities from the insider, outsider, and nearsider postures; the proposed test boundary is shown in Figure 3-2. ARL/SLAD will collect and report, at a minimum, the data in Attachments A and B of DOT&E guidance. ARL/SLAD will provide a full report and all data to DOT&E within 45 days of the assessment. Resources required for this test can be found in Table 4-1. The OTA will submit the CVPA test plan to DOT&E for approval 90 days prior to execution.

3.5.1.2. Adversarial Assessment (AA). The OTA will conduct Adversarial Assessments (AAs) during both the LUT and the IOT&E using the Army Threat Systems Management Office (TSMO) to portray the cyber threat. TSMO is an NSA-certified, USCYBERCOM-accredited cyber threat team. TSMO will execute the AAs using their accredited tools and processes to portray a representative cyber threat (insider, nearsider, and outsider) in accordance with the TGVS STAR, the DIA Computer Network Operations Capstone Threat Assessment, and the TGVS Computer Network Operations (CNO) Annex to the Threat Test Support Package. The OTA will conduct the assessment in the context of TGVS mission operations, with representative data sources, network traffic, and external interface connectivity; the proposed test boundary is shown in Figure 3-2. The assessment will include operationally representative network defense, including local operator, maintainer and administrator defense functions and will measure the detect and react abilities of a unit equipped with the TGVS and interoperating with the Tier 2 CNDSP, the ARCYBER 2nd RCC.

During the Adversarial Assessment the OTA will collect and report, at a minimum, the data in Attachment C of the DOT&E guidance, which requires cyber-trained protect, detect, react, and restore (PDRR) data collectors located in both the local and Tier 2 network defense locations. Where allowed by crew safety or equipment damage concerns, the OTA will directly

measure mission effects; otherwise, the OTA will evaluate mission effects using independent subject matter experts and the details of the attacks performed during the Adversarial Assessment. These subject matter experts will consider the effect of the attacks and any demonstrated cyber defender responses on the execution of mission threads and associated system performance parameters.

In the event that the network defenders do not detect malicious network activity, the OTA will inject one or more detection scenarios (white cards) in order to evaluate the reaction and response chain of events.

The OTA will submit the Adversarial Assessment plan for DOT&E approval 90 days prior to execution, and provide a report from the cyber test team along with the data collected in accordance with Attachment C of DOT&E Guidance within 45 days of the end of the assessment.

3.5.1.3. Cybersecurity Test Architecture. The architecture, proposed test boundary for the CVPA and AA, and external interfaces of the TGVS are shown in Figure 3-2.

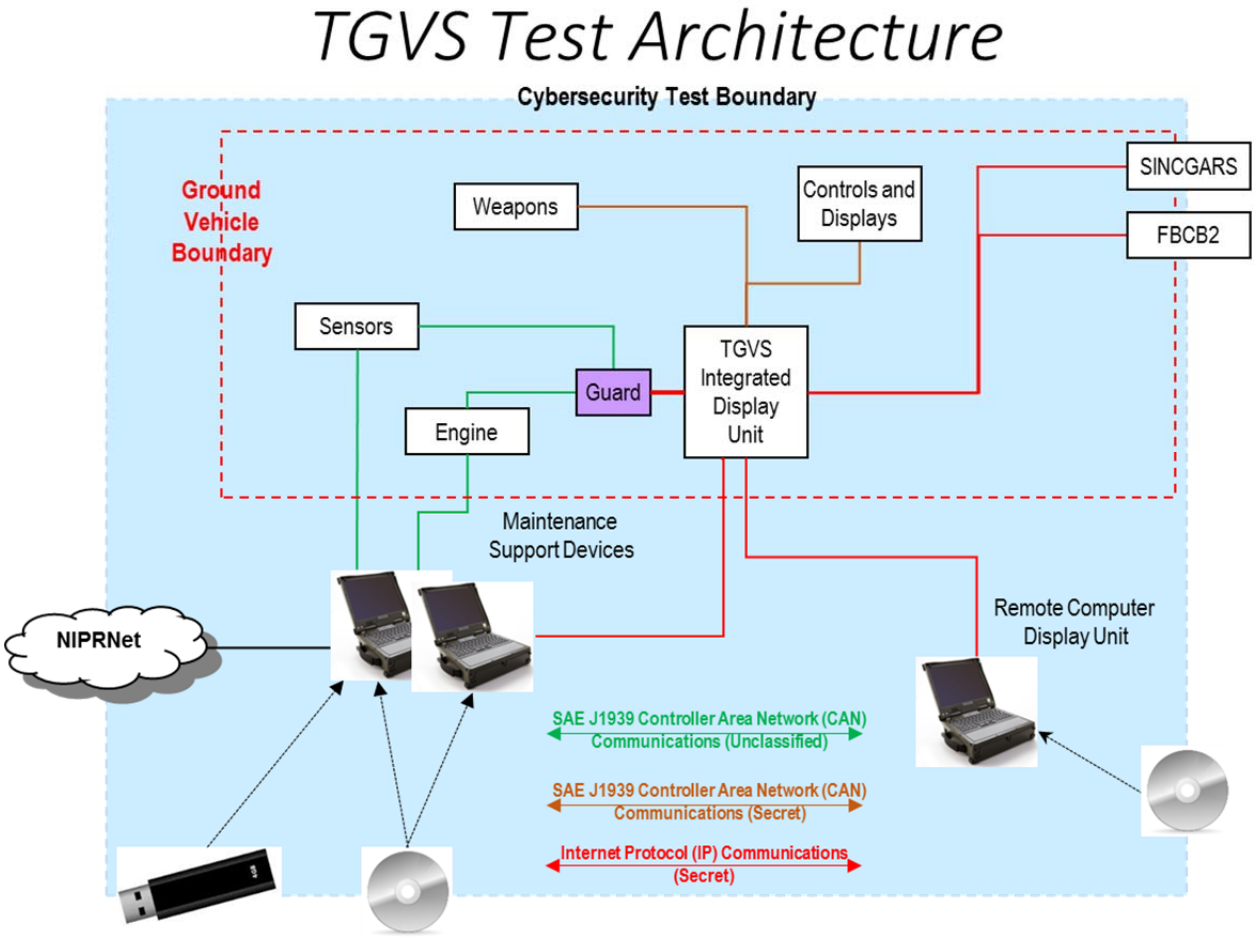


Figure 3-2. TGVS Test Architecture

Cybersecurity – TEMP Main Body Example

In typical operations, cyber defense for the TGVS is provided locally (Tier 3) by the system operators, maintainers, and system administrators, including a contingent of sustainment support from the development contractor. The Tier 2 Computer Network Defense Service Provider (CNDSP)[1] for the TGVS is the U. S. Army Cyber Command (ARCYBER) Regional Cyber Center (RCC). (…)

3.5.2.1. Cybersecurity Critical Operational Issue. The OTA will assess cybersecurity under Critical Operational Issue X using the following evaluation criteria:

Table 3-1: TGVS Cybersecurity Critical Operational Issue Evaluation Criteria

Criterion	Standard	Minimum Data Required
CyberX.1: Ability to Protect Information and Information Systems	Do the Vulnerabilities and Exploitations discovered during cybersecurity testing of the system put the unit's ability to conduct missions at risk?	DOT&E 2014 Attachments A, B, C
CyberX.2: Ability to Detect Cyber Threat Activity and Malfunctions	Are the accuracy of detections by the TGVS-equipped unit and their defenders during cybersecurity testing sufficient to identify cyber threat activity or malfunctions that put the unit's ability conduct missions at risk?	DOT&E 2014 Attachments A and C
CyberX.3: Ability to React to Cyber Threat Activity and Malfunctions	Are the mitigation actions provided by the TGVS-equipped unit and their defenders during cybersecurity testing sufficient to ensure the unit's ability to conduct missions following cyber threat activity or malfunctions?	DOT&E 2014 Attachment C
CyberX.4: Ability to Restore System after Cyber Threat Activity or Malfunction	Has the TGVS-equipped unit and their defenders demonstrated the ability to restore normal system operation and conduct missions following cyber threat activity or malfunctions?	DOT&E 2014 Attachments A and C
CyberX.5: Ability to Conduct Missions	Can a TGVS-equipped unit conduct their missions in the presence of malicious cyber threat activity or when encountering malfunctions?	DOT&E 2014 Attachment C
CyberX.6: Ability to Perform Reliably and Be	Can the TGVS-equipped unit perform its mission reliably and perform	DOT&E 2014 Attachments A, B,

[1] Sometimes called Cybersecurity Defense Service Provider (CDSP)

Maintained while also being Secure from Cyber Threat Activity	maintenance in the operational context with a degraded cyberspace environment?	and C
CyberX.6: Ability to Preserve System Physical Integrity and the Safety of Operators from Cyber Threat Activity and Malfunctions	In the presence of malicious cyber activity or following a malfunction, is the TGVS able to preserve its own physical integrity and the physical safety of its operators?	DOT&E 2014 Attachments B and C

3.5.4. Test Limitations. (…) Because the unit equipped with the system normally operates in a team with other identically-equipped units that are not resourced for the AA, the scope of mission threads the operators will execute for supporting mission effects data collection may be reduced. Also, TSMO will not knowingly launch cyber attacks that could affect control of the vehicle while it is in motion.

If equipment damage concerns preclude the evaluation of any systems connected to the CAN bus, independent laboratory testing of these systems will be performed. This data will be included in the CVPA report and cyber exploitations based on the findings will be white-carded in the AA. (…)

4.2.5. Threat Representation. (…) Resources required for TGVS cybersecurity testing are found in Table 4-1. The figures for the Army Research Lab include funds for developing advanced cyber exploits against the system, e.g. for the subsystems on the CAN bus. (…)

Table 4-1. TGVS Cybersecurity Test Resources

SUPPORTING UNITS	FY16	FY17	FY18
ARL/SLAD CVPA Team	$x1		
TSMO AA Team			$x2
ARL/SLAD AA PDRR Data Collection			$x3
OTA Cybersecurity Testing Support	$x4		$x5
Instrumentation			$x6
Army Research Lab Testing Support	$x7		$x8

Cybersecurity – Appendix E Command and Control System Example

<The following information is provided in Appendix E only if not already included in the body of the TEMP. The cybersecurity information need not be duplicated in both places.>

The Operational Test Agency (OTA) will perform cybersecurity testing as part of OT&E for the Warfighter Command and Control System (WC2S) in accordance with 1 Aug 2014 DOT&E guidance. Prior to these tests, WC2S system will have a signed Authority to Operate.

E.1. System Description A unit equipped with WC2S is able to communicate between the Joint Warfighting Command and deployed Joint Warfighting Units. WC2S allows commanders at the Joint Warfighting Command to receive and synthesize intelligence from unclassified and classified sources, and to issue orders in those domains. WC2S also hosts database services at all classification levels. Units equipped with WC2S perform cyber defense functions interoperating with the Defense Information Systems Agency (DISA) Global Support Center – North America (GNSC-NA) for unclassified and secret networks and the Defense Intelligence Agency (DIA) Regional Support Center (RSC) for the JWICS network.

E.2. System Threats A full range of cyber adversaries with nascent, limited, moderate, and advanced capabilities will target WC2S. Adversaries will attempt to compromise the system; exfiltrate, infiltrate, or corrupt data; disrupt system operations; and, if possible, physically destroy equipment. Additional cyber threat information for the WC2S is provided in the System Threat Assessment Report (STAR) and the Computer Network Operations Capstone Threat Assessment (IO Capstone, Volume 10) (CORRECTED), 2nd Edition, May 2013, DIA-08-1209-908.A.

E.3. WC2S Architecture and Test Boundary

WC2S comprises servers hosted at the Joint Warfighting Command Headquarters with unclassified, secret, and TS/SCI enclaves (see Figure E-1). In all three enclaves, there are database servers, and infrastructure and customer-facing services. On the unclassified enclave, WC2S receives and delivers data via NIPRNet, including web applications, and physical media devices. The unclassified enclave transfers information to the secret enclave via an approved cross-domain solution and connects via Ethernet (RJ-45) to the legacy system that WC2S is replacing.

In addition to the unclassified data that arrives via the cross-domain solution, the WC2S secret enclave receives data via the SIPRNet and physical media devices. WCS2 has a web-based interface for SIPRNet users, similar to the NIPRNet version, to allow those users to query the secret database. The TS/SCI database consists of the data transferred from the secret and unclassified enclaves via the attached cross-domain solution and JWICS data. JWICS users can use a virtual private network (VPN) to connect and query the WC2S database.

Finally, commanders can push appropriately-tagged intelligence products and tactical messages from the TS/SCI and secret enclaves down to the lower-classification enclaves via the cross-domain solutions.

The architecture, proposed test boundary for the CVPA and AA, and external interfaces of the WC2S are shown in Figure E-1.

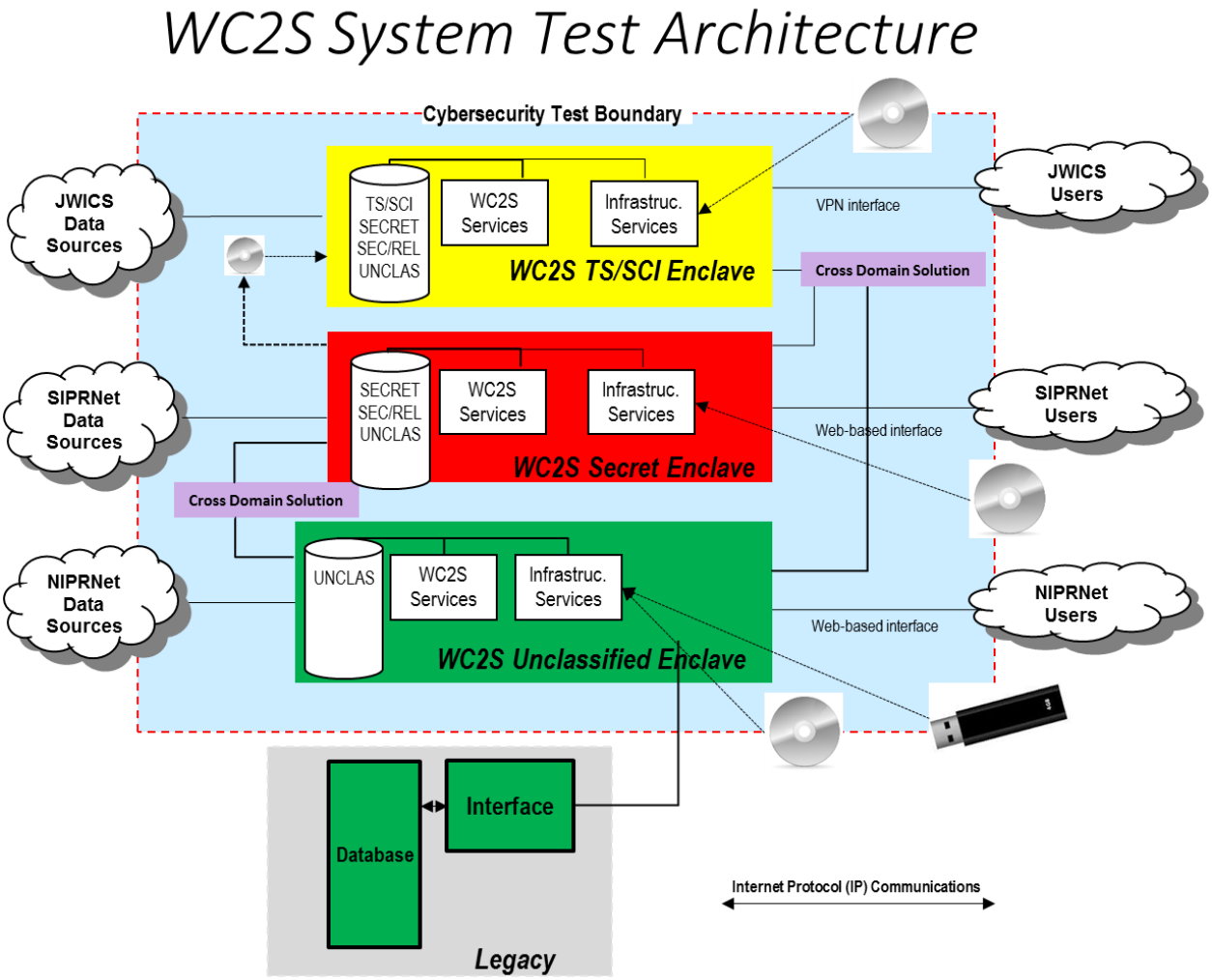


Figure E-1. WC2S System Test Architecture

In typical operations, cyber defense for the WC2S is provided locally (Tier 3) by the system operators and system administrators, including a contingent of sustainment support from the development contractor. The Tier 2 Computer Network Defense Service Provider (CNDSP)[1] for the unclassified and secret portions is the Defense Information Systems Agency (DISA) Global Support Center – North America (GNSC-NA) in Columbus, Ohio. The JWICS Tier 2 CNDSP is the Defense Intelligence Agency (DIA) Regional Support Center (RSC). See Table E-1 for each organizations cyber defense and test responsibilities.

[1] Sometimes called Cybersecurity Defense Service Provider (CDSP)

Table E-1. WC2S Cyber Defenders' Roles and Responsibilities

Cyber Tier	Role	Cyber Defense Responsibility	Test Responsibility
Local Subscribers and Defenders (Tier 3)	Commander WC2S Operations Center (Network AO/Owner)	Ensure that the network is maintained and available to support operations.	The Network AO/Owner is responsible for identifying personnel for testing support under their control (example, Network or System Administrators) and ensuring the personnel availability for supporting the test efforts. <*The roles can be combined if the Network AO/Owner is the same as the Facility Owner/Ops or System Program Manager/Owner.*>
	Commander WC2S Operations Center Facility Owner/Ops	Establishes physical security for networks operating within the facility.	The Facility Owner/Ops is responsible for identifying personnel for testing support under their control (example, Network or System Administrators) and ensuring the personnel availability for supporting the test efforts. <*The roles can be combined if the Facility Owner/Ops is the same as the Network AO/Owner or System Program Manager/Owner.*>
	WC2S Program Office (System Program Manager/Owner)	Designs and implements the system with cyber security as a priority. Creates patches to identified vulnerabilities in a timely manner. Identifies and publishes mitigation techniques to known vulnerabilities until patches are implemented.	The System Program Manager/Owner is responsible for identifying personnel for testing support under their control (example, Network or System Administrators) and ensuring the personnel availability for supporting the test efforts. <*The roles can be combined if the System Program Manager/Owner is the same as the Facility Owner/Ops or Network AO/Owner.*>
	Vandenberg Base Network Administrator (Network Administrator)	Ensures that the Network is patched and only accessed by authorized users. Implements actions to mitigate known vulnerabilities. Configures Host Based Security Systems. Monitors the system for unauthorized and malicious activity. Reports anomalies to the Information Assurance Manager.	Responsible for providing network assistance and troubleshooting to the Red team for access needed to execute the events. This will include assisting with placement of remote access devices or virtual machines employed on the network infrastructure.
	WC2S Local System Administrator	Ensures that the system is patched and only accessed by authorized users. Implements	Responsible for providing system level assistance and troubleshooting to the Red

Cyber Tier	Role	Cyber Defense Responsibility	Test Responsibility
	(System Administrator)	actions to mitigate known vulnerabilities. Monitors the system for unauthorized and malicious activity. Reports anomalies to the Network Administrator or the Information Assurance Manager.	team for access needed to execute the events. This will include assisting with troubleshooting issues with the system, passwords, or access management.
	Vandenberg Air Force Base IAM (Information Assurance Manager)	Ensures that information systems are compliant with the Information Assurance Vulnerability Management Program and all applicable Security Technical Implementation Guides. Ensure security incidents are reported and corrective action taken. The IAM operates the Tier 3 Help Desk.	Trusted Agent responsible for assisting with deconfliction of events if needed and assist in ensuring that the test is executed in a secure posture. Assist in data collection and providing information needed for the report from this Tier Level and participating in any post-test events as needed.
Unclassified & SIPRNET Tier 2	DISA Global Support Center (Cyber Network Defense Service Provider)	Certified and accredited by US Cyber Command. Provides component attack detection, malware protection, situational awareness, and incident response and analyses. The CNDSP coordinates the reporting flow between Tier 1 and Tier 3 and operates Tier 2 Help Desk.	Trusted Agent responsible for assisting with deconfliction of events if needed and assist in data collection or providing information needed for the report.
JWICS Tier 2	DIA RSC (Cyber Network Defense Service Provider)	As directly above.	As directly above.
Tier 1	Joint Force Headquarters – Department of Defense Information Network Joint Operations Center	Centrally coordinates and directs cyber network defense that affect more than on DoD Component. Coordinates with law enforcement and counter-intelligence operations.	Trusted Agent responsible for assisting with deconfliction of events if needed and assist in ensuring that the test is executed in a secure posture. Assist in data collection and providing information needed for the report from this Tier Level and participating in any post-test events as needed.

E.4. Cooperative Vulnerability and Penetration Assessment (CVPA) The OTA will employ the Army Research Laboratory Survivability Lethality Analysis Directorate (ARL/SLAD) to perform the Cooperative Vulnerability and Penetration Assessment (CVPA) during the OA.

ARL/SLAD will perform the CVPA on the operationally representative WC2S, including the use of local cybersecurity defenders such as system operators and system administrators to support data collection (e.g., through interviews). ARL/SLAD will use accredited tools and processes, which include automated scans and manual inspection and will execute their activities from the insider, nearsider, and outsider postures. All external interfaces to the WC2S will be active and accessible; the proposed test boundary is shown in Figure E-1. ARL/SLAD will collect, at a minimum, the data in Attachments A and B of DOT&E guidance. ARL/SLAD will provide a full report and all data will be provided to DOT&E within 45 days of the assessment. Resources required for this test can be found in Table E-2. The OTA will submit the CVPA test plan to DOT&E for approval 90 days prior to execution.

E.5. Adversarial Assessment (AA) The OTA will conduct an Adversarial Assessment (AA) during the IOT&E using the Army Threat Systems Management Office (TSMO) to portray the cyber threat. TSMO is an NSA-certified, USCYBERCOM-accredited cyber threat team. TSMO will execute the AA using their accredited tools and processes and portray a representative cyber threat (insider, nearsider, and outsider) in accordance with the WC2S STAR, the DIA Computer Network Operations Capstone Threat Assessment, and the W2CS Computer Network Operations (CNO) Annex to the Threat Test Support Package. TSMO will obtain any and all special authorizations from DIA needed to operate on JWICS. The OTA will conduct the assessment in the context of WC2S mission operations, with representative data sources, network traffic, and external interface connectivity; the proposed test boundary is shown in Figure E-1. The assessment will include operationally representative network defense, including local user and administrator functions and will measure the detect and react abilities of a unit equipped with the WC2S and interoperating with the Tier 2 CNDSPs, the DISA GNSC-NA and DIA RSC. Because of the complexity of the system and the extent of the cyber defense capabilities to be exercised, an extended assessment period is planned (see schedule below.)

During the Adversarial Assessment, the OTA will collect and report, at a minimum, the data in Attachment C of the DOT&E guidance, which requires cyber-trained protect, detect, react, and restore (PDRR) data collectors located in both the local and Tier 2 network defense locations. Direct measurement of mission effects will be made; however, if such a demonstration would interfere with real world operations, the OTA will evaluate mission effects using independent subject matter experts and the details of the attacks performed during the Adversarial Assessment. These subject matter experts will consider the effect of the attacks and any demonstrated cyber defender responses on the execution of mission threads and associated system performance parameters.

In the event that the network defenders do not detect malicious network activity, the OTA will inject one or more detection scenarios (white cards) in order to evaluate the reaction and response chain of events.

The OTA will submit the Adversarial Assessment plan for DOT&E approval 90 days prior to execution, and provide a report from the cyber test team along with the data collected in accordance with Attachment C of DOT&E Guidance within 45 days after the assessment.

E.6 Test Limitations

To avoid interfering with real-world operations, system operators will execute mission threads using simulation data sources to support mission effects data collection during the AA.

E.7 Schedule *<If the CVPA and AA schedules are not already denoted in the integrated test schedule in the body of the TEMP, they should be included in the Appendix. Multiple CVPA and AA events may be required to support milestone/production decisions.>*

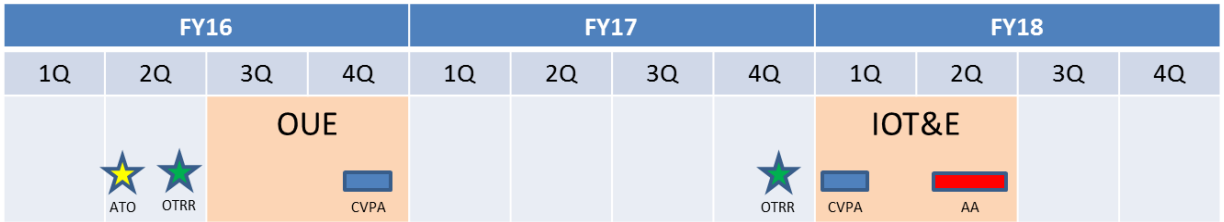


Figure E-2. WC2S Cybersecurity Test Schedule

E.8 Resources Resources required for WC2S cybersecurity testing are found in Table E-2. The figures for ARL include funds for developing advanced cyber exploits against the system; e.g., for bridging air-gaps.

Table E-2. WC2S Cybersecurity Test Resources

SUPPORTING UNITS	FY16	FY17	FY18
ARL/SLAD CVPA Team	$x1		
TSMO AA Team			$x2
ARL/SLAD AA PDRR Data Collection			$x3
OTA Cybersecurity Testing Support	$x4		$x5
Simulation & Instrumentation			$x6
Army Research Lab Testing Support	$x7		$x8

E.9 Evaluation Structure The OTA will use the results of WC2S cybersecurity testing, in part, to determine its operational effectiveness, suitability, and survivability. These evaluations should take into account the results of any bench testing. The OTA will assess cybersecurity under Critical Operational Issue X using the following evaluation criteria:

Table E-3: WC2S Cybersecurity Critical Operational Issue Evaluation Criteria

Criterion	Standard	Minimum Data Required
CyberX.1: Ability to Protect Information and Information Systems	Do the Vulnerabilities and Exploitations discovered during cybersecurity testing of the system put the unit's ability to conduct missions at risk?	DOT&E 2014 Attachments A, B, C

CyberX.2: Ability to Detect Cyber Threat Activity and Malfunctions	Are the accuracy of detections by the WC2S-equipped unit and their defenders during cybersecurity testing sufficient to identify cyber threat activity or malfunctions that put the unit's ability to conduct missions at risk?	DOT&E 2014 Attachments A and C
CyberX.3: Ability to React to Cyber Threat Activity and Malfunctions	Are the mitigation actions provided by the WC2S-equipped unit and their defenders during cybersecurity testing sufficient to ensure the unit's ability to conduct missions following cyber threat activity or malfunctions?	DOT&E 2014 Attachment C
CyberX.4: Ability to Restore System after Cyber Threat Activity or Malfunction	Has the WC2S-equipped unit and their defenders demonstrated the ability to restore normal system operation and conduct missions following cyber threat activity or malfunctions?	DOT&E 2014 Attachments A and C
CyberX.5: Ability to Conduct Missions	Can a WC2S-equipped unit conduct their missions in the presence of malicious cyber threat activity or when encountering malfunctions?	DOT&E 2014 Attachment C
CyberX.6: Ability to Perform Reliably and Be Maintained while also being Secure from Cyber Threat Activity	Can the WC2S-equipped unit perform its mission reliably and perform maintenance in the operational context with a degraded cyberspace environment?	DOT&E 2014 Attachments A, B, and C
CyberX.7: Ability to Preserve System Physical Integrity and the Safety of Operators from Cyber Threat Activity and Malfunctions	In the presence of malicious cyber activity or following a malfunction, is the WC2S able to preserve its own physical integrity and the physical safety of its operators?	DOT&E 2014 Attachments B and C

Cybersecurity – Appendix E Shipboard Example

<The following information is provided in Appendix E only if not already included in the body of the TEMP. The cybersecurity information need not be duplicated in both places.>

The Operational Test Agency (OTA) will perform cybersecurity testing as part of OT&E for the Shipboard Integrated Mission System (SIMS) in accordance with 1 Aug 2014 DOT&E guidance. Prior to these tests, SIMS will have a signed Authority to Operate.

E.1. System Description A unit equipped with SIMS is able to employ multiple systems from integrated control operator consoles. SIMS consoles have access to both NIPRNet and SIPRNet. The consoles provide a human interface to sensors, weapons, and systems required to safely operate the ship, including network accessible Programmable Logic Controllers (PLCs) and other industrial controls systems for propulsion and electrical distribution. Units equipped with SIMS perform cyber defense functions interoperating with the Navy Cyber Defense Operations Command (NCDOC) for both unclassified and secret networks.

E.2. System Threats A full range of cyber adversaries with nascent, limited, moderate, and advanced capabilities will target the SIMS. Adversaries will attempt to compromise the system; exfiltrate, infiltrate, or corrupt data; disrupt system operations; and, if possible, physically destroy equipment. Additional cyber threat information on the SIMS is provided in the SIMS System Threat Assessment Report (STAR) and the Computer Network Operations Capstone Threat Assessment (IO Capstone, Volume 10) (CORRECTED), 2nd Edition, May 2013, DIA-08-1209-908.A.

E.3. SIMS Architecture and Test Boundary

SIMS comprises servers, computers / consoles, and other networked devices hosted aboard a ship with unclassified and secret enclaves (see Figure E-1). In both enclaves, there are servers for databases, SIMS services, and SIMS operator-facing control consoles. The unclassified SIMS enclave includes connectivity to NIPRNet, various sensors, systems, and physical media devices, and provides data transfer capability via SIMS consoles. The unclassified enclave also has connectivity to the secret enclave via an approved cross-domain solution.

In addition to the unclassified data that arrives via the cross-domain solution, the secret enclave receives data via the SIPRNet, connected sensors and systems, and physical media devices. Like the unclassified version, the secret enclave has consoles to enable secret processing and telecommunication.

The architecture, proposed test boundary for the CVPA and AA, and external interfaces of the SIMS are shown in Figure E-1.

In typical operations, cyber defense for the SIMS is provided locally (Tier 3) by the system operators and system administrators, including a contingent of sustainment support from the development contractor. The Navy Cyber Defense Operations Command (NCDOC) in

Norfolk, Virginia is the Tier 2 Computer Network Defense Service Provider (CNDSP)[1] for both the unclassified and secret networks.

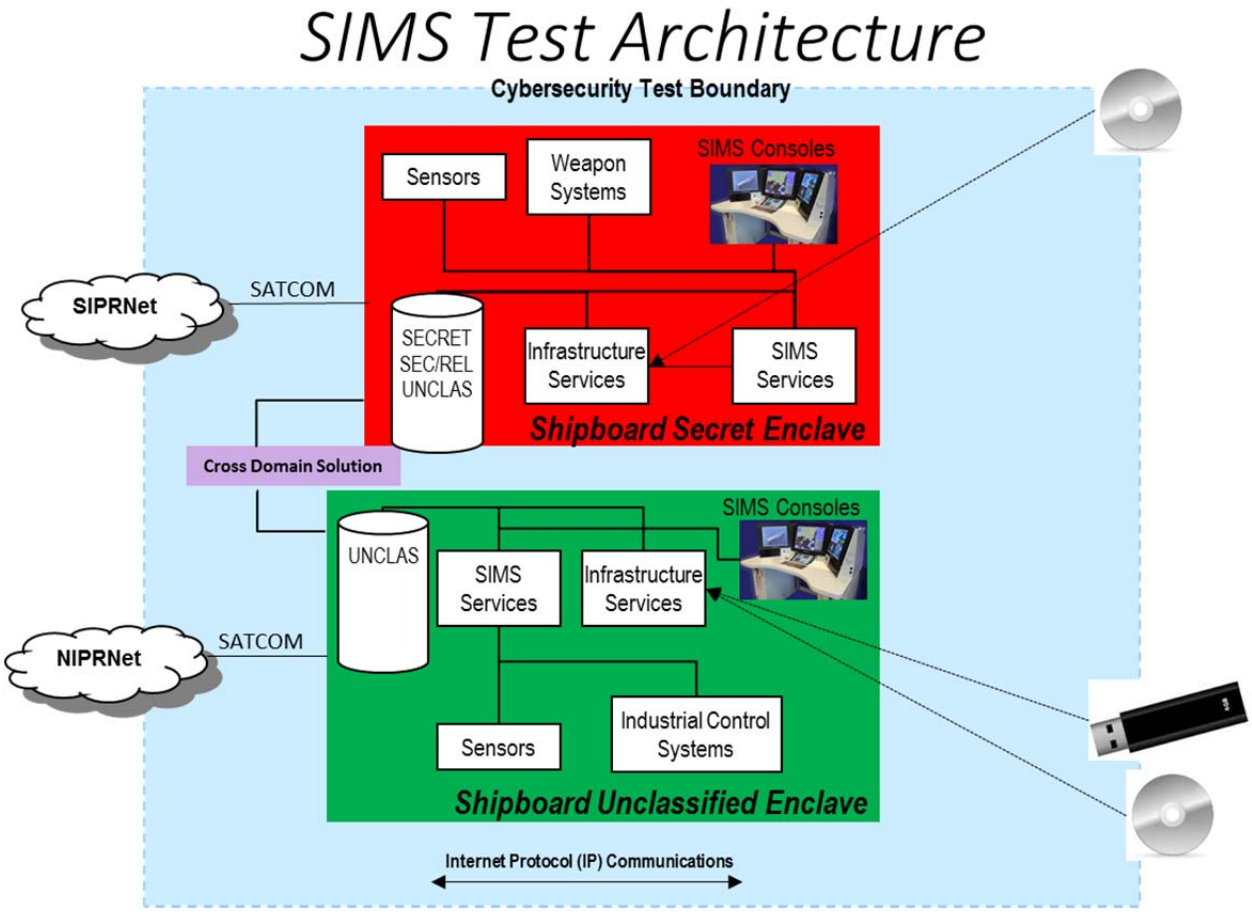


Figure E-1. SIMS Test Architecture

E.4. Cooperative Vulnerability and Penetration Assessment (CVPA). The OTA will employ a combined Navy Information Operations Command (NIOC) and Commander Operational Test and Evaluation Force (COMOPTEVFOR) cyber team to perform the Cooperative Vulnerability and Penetration Assessment (CVPA) during the OA. NIOC/COMOPTEVFOR will perform the CVPA on an operationally representative SIMS, including local cybersecurity defenders such as system operators and system administrators to support data collection (e.g., through interviews), while the ship is in port during a pre-deployment time period when all ship systems will be present and powered. NIOC/COMOPTEVFOR will execute vulnerability and penetration testing using their accredited tools and processes, which include automated scans and manual inspection. The SIMS will have all external interfaces active, and NIOC/COMOPTEVFOR will conduct assessment activities from the insider, outsider, and nearsider postures; the proposed test boundary is shown in Figure E-1. NIOC/COMOPTEVFOR will collect, at a minimum, the data in Attachments A and B of DOT&E guidance. NOIC/COMOPTEVFOR will provide a full

[1] Sometimes called Cybersecurity Defense Service Provider (CDSP)

report and all data to DOT&E within 45 days of the assessment. Resources required for this test can be found in Table E-1. The OTA will submit the CVPA test plan to DOT&E for approval 90 days prior to execution.

E.5. Adversarial Assessment (AA). The OTA will conduct an Adversarial Assessment (AA) during the IOT&E using a combined NIOC/COMOPTEVFOR cyber team led by NIOC, who will portray the cyber threat. NIOC is an NSA-certified, USCYBERCOM-accredited cyber threat team. NIOC/COMOPTEVFOR will execute the AA using their accredited tools and processes and portray a cyber threat (insider, nearsider, and outsider) in accordance with the STAR and the DIA Computer Network Operations Capstone Threat Assessment. The OTA will conduct the assessment in the context of SIMS mission operations, with representative data sources, network traffic, and external interface connectivity; the proposed test boundary is shown in Figure E-1. The assessment will include operationally representative network defense, including the local user and administrator functions, and will measure the detect and react abilities of a unit equipped with the SIMS and interoperating with the Tier 2 CNDSP, NCDOC. Because of the complexity of the system and the extent of the cyber defense capabilities to be exercised, an extended assessment period is planned (see schedule below.)

During the Adversarial Assessment, the OTA will collect and report, at a minimum, the data in Attachment C of the DOT&E guidance, which requires cyber-trained protect, detect, react, and restore (PDRR) data collectors located in both the local and Tier 2 network defense locations. Where allowed by crew safety or equipment damage concerns, the OTA will directly measure mission effects; otherwise, the OTA will evaluate mission effects using independent subject matter experts and the details of the attacks performed during the Adversarial Assessment. These subject matter experts will consider the effect of the attacks and any demonstrated cyber defender responses on the execution of mission threads and associated system performance parameters.

In the event that the network defenders do not detect malicious network activity, the OTA will inject one or more detection scenarios (white cards) in order to evaluate the reaction and response chain of events.

The OTA will submit the Adversarial Assessment plan for DOT&E approval 90 days prior to execution, and provide a report from the cyber test team along with the data collected in accordance with Attachment C of DOT&E Guidance within 45 days after the assessment.

E.6 Test Limitations

Both the CVPA and AA will be conducted in-port, as the testing will necessarily decertify the platform. Ship's crew will be executing mission threads using simulation data sources to support mission effects data collection during the AA.

If crew safety or equipment damage concerns preclude the evaluation of any systems (e.g., industrial control systems such as PLCs) while onboard the ship, independent laboratory testing of these systems will be performed. This data will be included in the CVPA report and cyber exploitations based on the findings will be white-carded in the AA.

E.7 Schedule *<If the CVPA and AA schedules are not already denoted in the integrated test schedule in the body of the TEMP, they should be included in the Appendix. Multiple CVPA and AA events may be required to support milestone/production decisions.>*

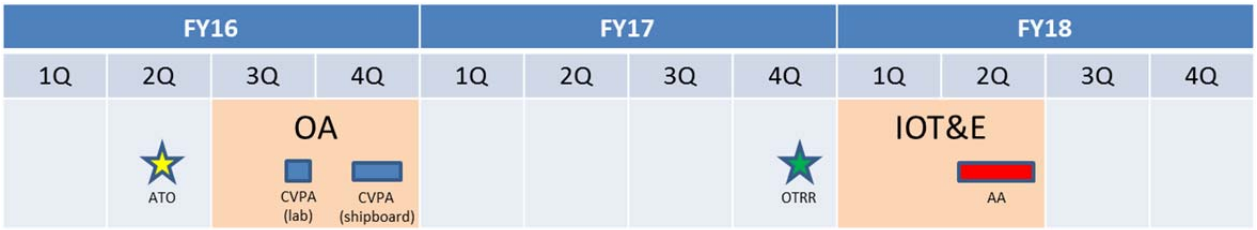


Figure E-2. Cybersecurity Test Schedule

E.8 Resources Resources required for SIMS cybersecurity testing are found in Table E-1. The figures for the NIOC/COMOPTEVFOR CVPA Team and the Naval Research Laboratory include funds for developing advanced cyber exploits against the system, e.g. for PLCs.

Table E-1. SIMS Cybersecurity Test Resources

SUPPORTING UNITS	FY16	FY17	FY18
NIOC/COMOPTEVFOR CVPA Team	$x1		
NIOC/COMOPTEVFOR AA Team			$x2
OTA AA PDRR Data Collection			$x3
OTA Cybersecurity Testing Support	$x4		$x5
Simulation & Instrumentation			$x6
Naval Research Lab Testing Support	$x7		$x8

E.9 Evaluation Structure. The OTA will use the results of SIMS cybersecurity testing, in part, to determine its operational effectiveness, suitability, and survivability. These evaluations should take into account the results of any bench testing. The OTA will assess cybersecurity under Critical Operational Issue X using the following measures:

Table E-2: SIMS Cybersecurity Critical Operational Issue Evaluation Criteria

Criterion	Standard	Minimum Data Required
CyberX.1: Ability to Protect Information and Information Systems	Do the Vulnerabilities and Exploitations discovered during cybersecurity testing of the system put the unit's ability to conduct missions at risk?	DOT&E 2014 Attachments A, B, C
CyberX.2: Ability to Detect Cyber Threat Activity and Malfunctions	Are the accuracy of detections by the SIMS-equipped unit and their defenders during cybersecurity testing sufficient to identify cyber threat activity	DOT&E 2014 Attachments A and C

	or malfunctions that put the unit's ability to conduct missions at risk?	
CyberX.3: Ability to React to Cyber Threat Activity and Malfunctions	Are the mitigation actions provided by the SIMS-equipped unit and their defenders during cybersecurity testing sufficient to ensure the unit's ability to conduct missions following cyber threat activity or malfunctions?	DOT&E 2014 Attachment C
CyberX.4: Ability to Restore System after Cyber Threat Activity or Malfunction	Has the SIMS-equipped unit and their defenders demonstrated the ability to restore normal system operation and conduct missions following cyber threat activity or malfunctions?	DOT&E 2014 Attachments A and C
CyberX.5: Ability to Conduct Missions	Can a SIMS-equipped unit conduct their missions in the presence of malicious cyber threat activity or when encountering malfunctions?	DOT&E 2014 Attachment C
CyberX.6: Ability to Perform Reliably and Be Maintained while also being Secure from Cyber Threat Activity	Can the SIMS-equipped unit perform its mission reliably and perform maintenance in the operational context with a degraded cyberspace environment?	DOT&E 2014 Attachments A, B, and C
CyberX.7: Ability to Preserve System Physical Integrity and the Safety of Operators from Cyber Threat Activity and Malfunctions	In the presence of malicious cyber activity or following a malfunction, is the SIMS able to preserve its own physical integrity and the physical safety of its operators?	DOT&E 2014 Attachments B and C

Cybersecurity – Appendix E Tactical Aircraft Example

<The following information is provided in Appendix E only if not already included in the body of the TEMP. The cybersecurity information need not be duplicated in both places.>

The Operational Test Agency (OTA) will perform cybersecurity testing as part of OT&E for the Tactical Air Vehicle System (TAVS) in accordance with 1 Aug 2014 DOT&E guidance. Prior to these tests, TAVS will have a signed Authority to Operate.

E.1. System Description A unit equipped with TAVS performs armed reconnaissance missions and provides operators with multiple sensors and weapons to observe and engage various enemies. In-flight digital communications are performed using multiple external data links, which are detailed below. Units equipped with the TAVS perform cyber defense functions interoperating with the 24th Air Force.

E.2. System Threats A full range of cyber adversaries with nascent, limited, moderate, and advanced capabilities will target the TAVS. Adversaries will attempt to compromise the system; exfiltrate, infiltrate, or corrupt data; disrupt system operations; and, if possible, physically destroy equipment. Additional information on cyber threats to the TAVS is provided in the TAVS System Threat Assessment Report (STAR) and the Computer Network Operations Capstone Threat Assessment (IO Capstone, Volume 10) (CORRECTED), 2nd Edition, May 2013, DIA-08-1209-908.A.

E.3. TAVS Architecture and Test Boundary

TAVS comprises the air vehicle with its integrated sensors, weapons, propulsion systems, computers, various displays, controls, external data links (RF, SATCOM), and other networked devices hosted on-board the air vehicle (see Figure E-1). Systems that connect with the TAVS include mission planning and maintenance systems. Communications include IP and 1553 data bus traffic and some components have connectivity through both. External data sources including NIPRNet provide data used by the maintenance and mission planning components of TAVS.

The architecture, proposed test boundary for the CVPA and AA, and external interfaces of the TAVS are shown in Figure E-1.

In typical operations, cyber defense for the TAVS is provided locally (Tier 3) by the system operators, maintainers, and system administrators, including a contingent of sustainment support from the development contractor. The 24th Air Force in San Antonio, Texas is the Tier 2 Computer Network Defense Service Provider (CNDSP)[1] for TAVS.

[1] Sometimes called Cybersecurity Defense Service Provider (CDSP)

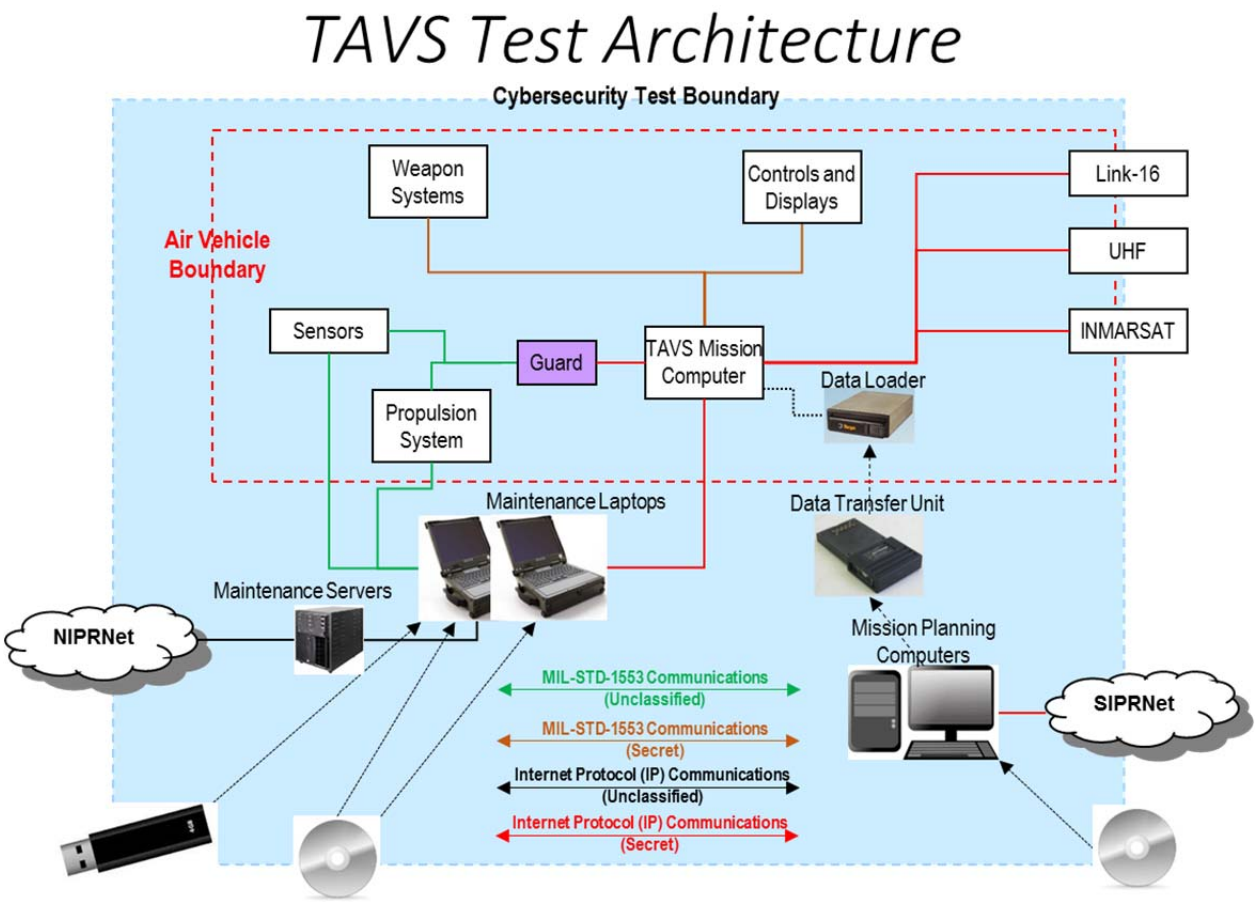


Figure E-1: TAVS Test Architecture

E.4. Cooperative Vulnerability and Penetration Assessment (CVPA). The OTA will employ the 92d Information Operations Squadron (92 IOS) cyber team to perform the Cooperative Vulnerability and Penetration Assessment (CVPA) during the OA. The 92 IOS will perform the CVPA on an operationally representative TAVS, including local cybersecurity defenders such as system operators, maintainers, and system administrators to support data collection (e.g., through interviews), while the aircraft is on the flight line with all systems present and powered. The 92 IOS will execute vulnerability and penetration testing using their accredited tools and processes, which include automated scans and manual inspection. The TAVS will have all external interfaces active, and the 92 IOS will conduct assessment activities from the insider, outsider, and nearsider postures; the proposed test boundary is shown in Figure E-1. The 92 IOS will collect and report, at a minimum, the data in Attachments A and B of DOT&E guidance. 90 IOS will provide a full report and all data to DOT&E within 45 days of the assessment. Resources required for this test can be found in Table E-1. The OTA will submit the CVPA test plan to DOT&E 90 days prior to execution.

E.5. Adversarial Assessment (AA). The OTA will conduct an Adversarial Assessment (AA) during the IOT&E using a 177th Information Aggressor Squadron (177 IAS) to portray the cyber threat. The 177 IAS is an NSA-certified, USCYBERCOM-accredited cyber threat team. The

177 IAS will execute the AA using their accredited tools and processes to portray a cyber threat (insider, nearsider, and outsider) in accordance with the TAVS STAR and the DIA Computer Network Operations Capstone Threat Assessment. The OTA will conduct the assessment in the context of TAVS mission operations, with representative data sources, network traffic, and external interface connectivity; the proposed test boundary is shown in Figure E-1. The assessment will include operationally representative network defense, including local user, maintainer, and administrator defense functions, and will measure the detect and react abilities of a unit equipped with the TAVS and interoperating with the Tier 2 CNDSP, 24th Air Force.

During the Adversarial Assessment the OTA will collect and report, at a minimum, the data in Attachment C of the DOT&E guidance, which requires cyber-trained protect, detect, react, and restore (PDRR) data collectors located in both the local and Tier 2 network defense locations. Where allowed by equipment damage concerns, the OTA will directly measure mission effects; otherwise, the OTA will evaluate mission effects using independent subject matter experts and the details of the attacks performed during the Adversarial Assessment. These subject matter experts will consider the effect of the attacks and any demonstrated cyber defender responses on the execution of mission threads and associated system performance parameters.

In the event that the network defenders do not detect malicious network activity, the OTA will inject one or more detection scenarios (white cards) in order to evaluate the reaction and response chain of events.

The OTA will submit the Adversarial Assessment plan for DOT&E approval 90 days prior to execution, and provide a report from the cyber test team along with the data collected in accordance with Attachment C of DOT&E Guidance within 45 days after the assessment.

E.6 Test Limitations

Both the CVPA and AA will be conducted with the aircraft on the ground to ensure physical safety. Flight safety concerns related to later flight ops will not limit testing as the platform will be reimaged and recertified after both the CVPA and the AA (this process will support data collection for the Restore evaluation). System operators will be executing mission threads using simulated data to support data collection on mission effects during the AA.

If equipment damage concerns preclude the evaluation of any systems on the aircraft (e.g., avionics), independent laboratory testing of these systems will be performed. This data will be included in the CVPA report and cyber exploitations based on the findings will be white-carded in the AA.

E.7 Schedule *<If the CVPA and AA schedules are not already denoted in the integrated test schedule in the body of the TEMP, they should be included in the Appendix. Multiple CVPA and AA events may be required to support milestone/production decisions.>*

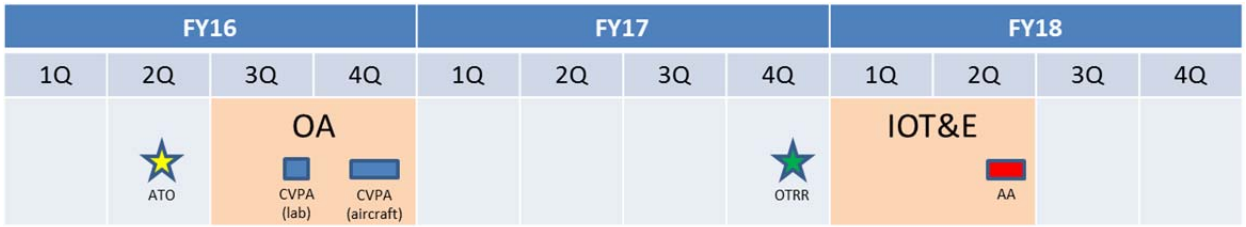


Figure E-2. TAVS Cybersecurity Test Schedule

E.8 Resources Resources required for TAVS cybersecurity testing are found in Table E-1. The figures for the 92 IOS CVPA Team and the Air Force Research Lab include funds for developing advanced cyber exploits against the system, e.g. for the subsystems on the 1553 bus.

Table E-1. TAVS Cybersecurity Test Resources

SUPPORTING UNITS	FY16	FY17	FY18
92 IOS CVPA Team	$x1		
177 IAS AA Team			$x2
OTA AA PDRR Data Collection			$x3
OTA Cybersecurity Testing Support	$x4		$x5
Simulation & Instrumentation			$x6
Air Force Research Lab Testing Support	$x7		$x8

E.9 Evaluation Structure. The OTA will use the results of TAVS cybersecurity testing, in part, to determine its operational effectiveness, suitability, and survivability. These evaluations should take into account the results of any bench testing. The OTA will assess cybersecurity under Critical Operational Issue X using the following evaluation criteria:

Table E-2: TAVS Cybersecurity Critical Operational Issue Evaluation Criteria

Criterion	Standard	Minimum Data Required
CyberX.1: Ability to Protect Information and Information Systems	Do the Vulnerabilities and Exploitations discovered during cybersecurity testing of the system put the unit's ability to conduct missions at risk?	DOT&E 2014 Attachments A, B, C
CyberX.2: Ability to Detect Cyber Threat Activity and Malfunctions	Are the accuracy of detections by the TAVS-equipped unit and their defenders during cybersecurity testing sufficient to identify cyber threat activity	DOT&E 2014 Attachments A and C

	or malfunctions that put the unit's ability to conduct missions at risk?	
CyberX.3: Ability to React to Cyber Threat Activity and Malfunctions	Are the mitigation actions provided by the TAVS-equipped unit and their defenders during cybersecurity testing sufficient to ensure the unit's ability to conduct missions following cyber threat activity or malfunctions?	DOT&E 2014 Attachment C
CyberX.4: Ability to Restore System after Cyber Threat Activity or Malfunction	Has the TAVS-equipped unit and their defenders demonstrated the ability to restore normal system operation and conduct missions following cyber threat activity or malfunctions?	DOT&E 2014 Attachments A and C
CyberX.5: Ability to Conduct Missions	Can a TAVS-equipped unit conduct their missions in the presence of malicious cyber threat activity or when encountering malfunctions?	DOT&E 2014 Attachment C
CyberX.6: Ability to Perform Reliably and Be Maintained while also being Secure from Cyber Threat Activity	Can the TAVS-equipped unit perform its mission reliably and perform maintenance in the operational context with a degraded cyberspace environment?	DOT&E 2014 Attachments A, B, and C
CyberX.7: Ability to Preserve System Physical Integrity and the Safety of Operators from Cyber Threat Activity and Malfunctions	In the presence of malicious cyber activity or following a malfunction, is the TAVS able to preserve its own physical integrity and the physical safety of its operators?	DOT&E 2014 Attachments B and C

Defense Business Systems – Guidance

Summary

Reliability, maturity, and sustainment measures for business systems acquisitions rely heavily on configuration management, defect tracking, and automated regression testing. This section of the guidebook provides related examples of text from previously approved TEMPs for business systems that have successfully prepared for developmental and operational testing.

Processes for developing and managing information technology software are provided in IEEE Standard 12207.2, Guide for Information Technology - Software Life Cycle Processes - Implementation Considerations, dated April 1998.

The TEMP should describe the acquisition program's configuration management and configuration control framework. Testers will need accurate configuration information to understand the system and to determine the system's adherence to effectiveness, suitability, and cybersecurity requirements.

Defect tracking should be conducted during all phases of test and evaluation, using a clearly-defined process that is explained in the TEMP. Generally, as a defect is discovered, the developer or tester will document it through a deficiency report (DR). A Deficiency Review Board (DRB) will assign a DR level as defined by IEEE 12207.2 Annex J, and track the status of each defect, over time, as to which are open, closed, or resolved. During regression testing or as part of another test event, testers will validate that identified deficiencies have been resolved. Additionally, software change requests (SCR) for capabilities that are needed but not inherent in the software baseline should be generated and assigned severity levels using the IEEE definitions based on impacts to mission accomplishment.

As a rule, test measures for business systems should be specified in terms of the types of data that can automatically be logged and reported by the system. Measures used for testing will typically be the same measures as those that operators and system managers will use over the course of a system's lifecycle to gauge acceptable performance or service degradation. Accordingly, automated logging and reporting of performance data should be included in the core system design. When possible, automated approaches to data collection should be used versus less accurate manual methods (e.g., relying on a stopwatch to measure system response times). Objective human performance measures, such as human error rates and the amount of time it takes the operator to complete a task, should be used to evaluate human factors, as appropriate for the system under test. Surveys should be used sparingly when human performance measures are not feasible or to supplement these measures. When used, surveys should comply with DOT&E's guidance on survey design and administration.

References

DoDI 8500.01, Cybersecurity, dated March 14, 2014 incorporates guidance from the now obsolete DoDI 8500.2, Procedures for the Operational Test and Evaluation of Information Assurance.

DoDI 8501.02, Risk Management Framework (RMF) for DoD Information Technology (IT), dated 12 March 2014, specifies the use of the National Institute of Standards and Technology (NIST) Risk Management Framework (RMF). RMF replaced the now-defunct DoD Information Assurance Certification and Accreditation Process (DIACAP).

Guidance on the Use and Design of Surveys in Operational Test and Evaluation (OT&E), DOT&E, 22 December 2014

Discussion on the Use and Design of Surveys, DOT&E, 24 February 2015

Discussion on Including Neutral Responses on Survey Questions, DOT&E, 2 April 2015

Survey Pre-Testing and Administration in Operational Test and Evaluation, 6 January 2017

Defense Business Systems Examples

Example 1

2.3. Deficiency Reporting

The EBS Workbench tool is used to document deficiencies (defects) detected during testing and tracks all steps in the defect resolution. The EBS Workbench uses the IEEE Standard 12207.2, (Annex J, dated April 1998), as the source for deficiency priority definitions.

Defect analysis will be conducted during all phases of Test and Evaluation. The developer should assign each problem in software products or activities to one of the priorities in Table 2-2.

Table 2-2: Priorities to be Used When Classifying Problems

(IEEE Standard 12207.2, Annex J, April 1998)

Priority	Applies if a problem could
1	a) Prevent the accomplishment of an essential capability b) Jeopardize safety, security, or other requirement designated "critical"
2	a) Adversely affect the accomplishment of an essential capability and no work-around solution is known b) Adversely affect technical, cost, or schedule risks to the project or to life cycle support of the system, and no work-around solution is known
3	a) Adversely affect the accomplishment of an essential capability but a work-around solution is known b) Adversely affect technical, cost, or schedule risks to the project or to life cycle support of the system, but a work-around solution is known
4	a) Result in user/operator inconvenience or annoyance but does not affect a required operational or mission-essential capability b) Result in inconvenience or annoyance for development or maintenance personnel but does not prevent the accomplishment of the responsibilities of those personnel
5	Any other effect

Priority/state changes and reworks are also tracked on a daily basis. DLA EProcurement Management reserves the right to change the EProcurement requirements if it appears that excessive rework will be needed to resolve a defect. The following data has been tracked on a daily basis since November 2009:

- Total Defects Created
- Total Critical and High Defects Created
- Total Medium and Low Defects Created
- Total Defects Closed
- Total Critical and High Defects Closed
- Total Medium and Low Defects Closed

- Total Defects Open
- Total Critical and High Defects Open
- Total Medium and Low Defects Open
- Average Days from Creation to Submit for Resolution
- Average Days from Lead Approval to Assigned to Developer
- Average Days to Resolve Defects
- Average Days from Resolution to Close
- Average Days from Creation to Close
- Open Defects by State
    - Defect Drafted
    - Resolution in Progress
    - Clarification Required
    - Ready for Retest

An example of Management's daily tracking report is shown in Table 2-3 below.

Table 2-3: EProcurement Daily Defect Aging Summary

	1 Mo	2 Mo	3 Mo	Older	TOTAL
CRITICAL	0	0	0	0	0
HIGH	3	0	0	0	3
MEDIUM	21	0	1	3	25
LOW	1	0	1	0	2
TOTAL	25	0	2	3	30

Once a deficiency/problem is detected during testing, a "Defect Report" is entered into Workbench. The Development Team Leads reviews the Defect and determine its validity and probable cause. If valid, the Defect is assigned to the appropriate developer for resolution. Once resolved, the Defect is assigned to the appropriate tester for validation of the fix.

All valid defects and their resolutions are stored in a repository for future use in testing. The Development Team records the user actions that lead to the validated defect. The recorded user actions are then used by EBS Workbench and installed in the script library.

Production Deficiency Reporting

Once the system goes live, it enters the sustainment phase of the program. At this point the system is no longer in development. During the sustainment phase it is managed by our J6 Sustainment Operations Division in Columbus, OH. They manage the production deficiency reporting process by listing the identification, investigation, and resolution activities in workbench.

Defense Business Systems – Examples

Identification Phase

- Sustainment POC receives Remedy ticket to research an issue that is occurring in the production environment. If the incident requires a defect to be created, the sustainment or build POC will create the defect in the EBS Workbench.
- If it's determined that a Remedy ticket requires a code fix/configuration change, the assigned person creates a defect in Workbench citing the ticket number. The assignee creates a defect in the workbench using the ticket information. The Workbench is then used to track the flow of work for ultimate transport to production. Information contained in the Workbench documents are functional specification documentation, test results data, table views.
- Defect should be updated to "Team Lead Submit" state
- Sustainment Functional Lead assigns EProcurement (SRM) defects to appropriate POC. Development Lead assigns the defect to the appropriate developer for investigation

Investigation Phase

- Developer performs necessary modifications in the development environment and documents the resolution details in the defect
- If a production issue requires a code or configuration change, the developer test the specific functionality in question, including any inputs, outputs, and dependent tasks before migrating the changes to PRD.

Resolution Phase

- Sustainment Configuration Control team migrates the changes to the System Test environment
- Once migrated, the Tester will get an email notification (automated)
- Once migrated to System Test (S*1), testers will determine if the defect has been resolved; if so, testers will document test results in the defect, update any associated regression test plans & cases with additional test steps/data/expected results to validate the defect scenario during future test efforts, and set it to 'Ready for Production Approval' state
    - If the defect has not been fixed or resolved, the tester updates the existing defect with retest details and assigns the defect back to the developer
    - If a new issue/problem emerges as a result of the defect fix/resolution, a new defect should be created
- Once in 'Ready for Production Approval' state, the Sustainment Functional Leads will start the administrative approval process for release to Production
- When the code is released to production, the assigned person goes into Remedy to annotate that the ticket is completed/closed.

Example 2

2.3. Deficiency Reporting

- Discrepancy Report (DR) status: Each DR written against KMI developed software was prioritized into 5 levels as defined by the IEEE 12207.2 specification. Each DR was initially assigned a level by the sub-contractor developing that particular software. The prime integrator and the Government Program Office performed an independent analysis and redefined levels accordingly. Graphs were maintained showing the number of open, closed and resolved (fixed but not tested) statistics over time, categorized by priority level. Example data is shown in Figures 2-4 and 2-5.

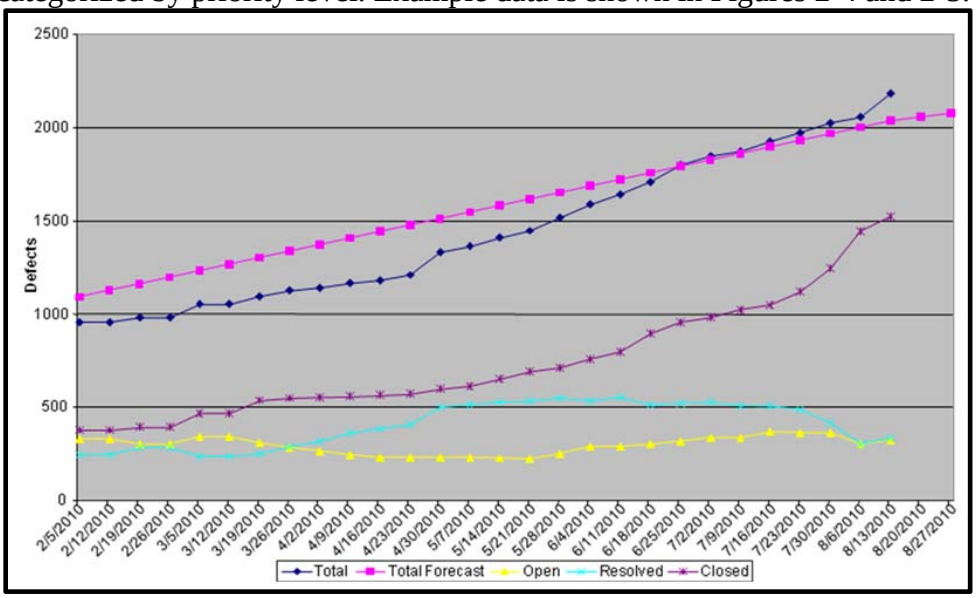


Figure 2-4: DR volume tracking (all priorities)

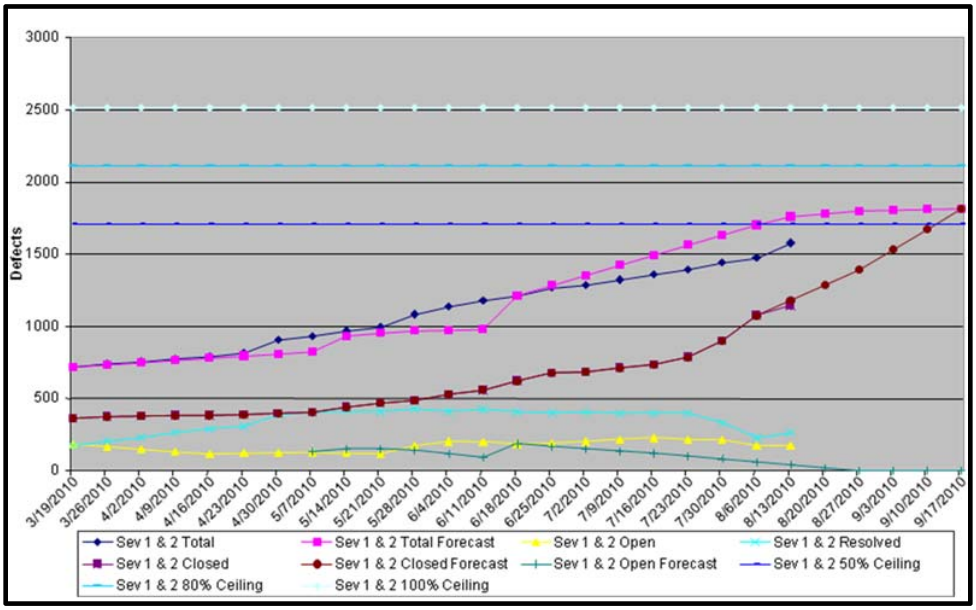


Figure 2-5: DR volume tracking (Priorities 1 and 2)

- DR Aging: DRs at each priority level were also tracked to show how many of each level were open for a particular timeframe. The timeframes were separated into 30 day increments, up to a column for >120 days. Example data is shown in Table 2-6.

Table 2-6: Sample DR Aging Metric

	Assigned and Submitted Defects - Days Open				
	0-30	31-60	61-90	91-120	>120
Severity					
1	4	2	7	0	0
2	99	21	11	7	9
3	38	28	11	8	16
4	3	6	6	1	5
5	3	0	2	3	6
Total	147	57	37	19	36

- Commercial Off the Shelf (COTS) DRs: The ageing statistic described above was also maintained for issues found with commercially purchased equipment, such as routers, servers, etc. Example data is shown in Exhibit 2-6.

The Management Strategy for fixing software and hardware failures is as follows. Every DR will be analyzed to determine the effect of the failure. Using this information, a determination will be made as to the severity of the problem (a.k.a Priority, as defined by the IEEE 12207.2 specification). All failures that rate a Priority 1 and 2 will be fixed prior to entering the next phase of testing.

Design of Experiments – Guidance

General

Design of Experiments (DOE) is a statistical methodology for planning, conducting, and analyzing a test. Any program that applies DOE principles should begin early in the test planning process. The test planners should assemble a group of subject matter experts who can identify the primary quantitative mission-focused measures (in DOE parlance: response variables) of interest that will characterize the performance of the system in the context of a mission-oriented evaluation. The test planners should identify environmental and operational factors that are expected to drive the performance of the system, as well as the levels of these factors (i.e., the various conditions or settings that the factors can take). A master test strategy should include the resources needed, the concept for early tests (including component tests), and the use of the results of early tests to plan further testing. One goal of the test strategy should be to ensure adequate coverage of all important factors while evaluating the quantitative mission-focused measures through planned testing. The testing strategy should be iterative in nature to ensure an adequate Initial Operational Test and Evaluation (IOT&E). The testing strategy should accumulate evidence that the system performs across its operational envelope before and during IOT&E. The test planners should apply DOE at each test iteration.

Elements of DOE for the TEMP

A brief overview of the design philosophy should be outlined in Section 3 of the TEMP. The information content may vary depending on the Milestone that the TEMP is supporting. Table 1 outlines information content that is appropriate for each milestone. Systems with legacy data will be expected to include more detail and have more robust test designs. The details of each of the test designs should be provided in a supporting appendix to the TEMP. Elements of experimental design should include the following:

- The goal of the test (experiment). See Mission Focused Evaluation Guidance.
- Quantitative mission-focused measures for effectiveness, suitability, and survivability. See Quantitative Mission-focused Measures Guidance.
- Factors that affect those measures of effectiveness, suitability, and survivability. See Integrated Survivability Evaluation Guidance.
- A method for strategically varying factors across developmental, operational, and live fire testing with respect to responses of interest.
- Statistical measures of merit (power and confidence) and corresponding effect sizes on the relevant quantitative mission-focused measures (i.e., those for which doing so 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.

These elements include all of the planning steps for designing an experiment, with the exception of execution order. Standard statistical designs assume the test point execution order

can be randomized. Designs including blocking and/or split-plot techniques should be noted. The execution of the test, including run plans/order, should be discussed in the Test Plan.

Commonly, the system under test (SUT) is a complex system with multiple missions and functionalities. The test design should reflect the complexity of the system. Often, multiple test designs will be necessary to fully characterize SUT mission performance. This might also require multiple experimental designs to capture all stages or aspects of mission execution.

Table 1: DOE Information Content for the TEMP

	Information Content
A	Responsibilities of T&E WIPT for test design purposes The goal(s) to be addressed at each stage of testing Quantitative mission-focused measures for each goal/question Initial listing of factors for each quantitative mission-focused measure Language for the overall testing strategy, including: • Screening experiments to ensure important factors are considered in operational testing • Sequential experimentation Test designs to support resourcing for limited user tests (LUT), operational assessments (OA), and IOT&Es While test designs may be very preliminary at MS A it is essential that DOE (or some other scientific test design technique) be used to estimate resources for a Request For Proposal adequately near MS A. Therefore, special attention should be paid towards making sure adequate resources are allocated for long lead items (e.g., number of targets, weapons, specialize range capabilities, etc.).
B	Identify responsibilities of T&E WIPT for test design purposes The goal(s) to be addressed at each stage of testing Quantitative mission-focused measures for each goal/question Refined listing of factors and levels for each quantitative mission-focused measure Test designs to support resourcing for limited user tests (LUT), operational assessments (OA), and IOT&E • Test Designs should be updated from MS A to account for any new information. Language for the overall testing strategy, including: • Screening experiments to ensure important factors are considered in operational testing • Sequential experimentation

C	Identify responsibilities of T&E WIPT for test design purposes The goal(s) to be addressed at each stage of testing, focusing on IOT&E Quantitative mission-focused measure for each goal/question Refined listing of factors and levels, based on prior testing and the operational mission, for each quantitative mission-focused measure. Details on how the factors and levels will be varied and controlled during each stage of testing Complete test designs to support resourcing for IOT&E Language for the overall testing strategy, including: • How previous knowledge is being used to inform IOT&E test planning. Analysis plans to support power calculations

References

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

Montgomery, D. C. (2009), *Design and Analysis of Experiments,* John Wiley and Sons

Myers, R. H., and Montgomery, D. C. (2002), *Response Surface Methodology: Process and Product Optimization Using Designed Experiments,* John Wiley and Sons.

TEMP Body Examples

Precision Guided Weapon Example Appendix

Artillery Example Appendix

Software Example Body and Appendix

Design of Experiments – TEMP Body Example

3.4.2.# Design of Experiments (Subpara to 3.4.2 Operational Evaluation Framework)

Design and Analysis of Experiments will be used to develop test plans for the developmental, integrated, and operational testing of system XYZ. The T&E WIPT will identify the following components of the experimental design: (1) goals, (2) metrics, (3) factors and levels that impact the outcome of the test, (4) a strategic method for varying those factors and levels across all tests, and (5) appropriate statistical power and confidence levels for important responses for which they make sense.

Note: Table 3.1, Top-Level Evaluation Framework Matrix, should capture the key test goals and metrics/measures that are discussed in the test design section of the TEMP.

The T&E WIPT will use a sequential approach in test planning, meaning that screening of factors will occur in DT and integrated test events, only factors that are deemed significant or of particular operational interest will be investigate in OT. The overarching test strategy outlined in this TEMP is adequate to support the OTA's evaluation plan. Tables 3.X1 – 3.XX provide the overall DOE strategy for each test objective. The overarching test strategy may change after the initial test events are conducted to allow for increased information on the effect of the factors on the critical responses. See the DOE Appendix for supporting information on the statistical qualities of the experimental design (factor selection, process diagrams, exact designs, and power/confidence levels).

Table 3.X: Overview of DOE Strategy for Test Objective 1

		Test Phase			
		DT	MS	IT	IOT
Critical Responses (Only MOE's, MOP's, KPP's, MOS's that relate to the current test objective should be included)		Select MOE, MOP, MOS, KPP	Select MOE, MOP, MOS, KPP	Select MOE, MOP, MOS, KPP	Select MOE, MOP, MOS, KPP
Factors	Factor Levels				
Factor 1	Categorical 2 levels	SV*	SV	SV	Record*
Factor 2	Continuous	HC*	HC	SV	SV
Factor 3	Continuous	SV	SV	SV	SV

Factor 4	Categorical 6 levels	SV	SV	SV	SV, Demo 2 levels

*In Table 3.X there are three common factor management strategies used (1) systematically vary (SV) the factor by including the factor in the experimental design, (2) hold constant (HC) at a fixed level during testing to minimize its impact on the test outcome, (3) record the level of the factor. Additionally, there are two levels of the fourth factor that will only be demonstrated (demo) in operational testing because of the cost associated with testing those levels.

Best Practices for Table 3.X:

Note 3.X can be replicated as many times as needed to ensure that all major test objectives are captured. These tables should not be exhaustive; instead they should capture the major test objectives, the primary measures (or response variables), and the factors that will be considered in test planning.

Recordable factors across all test phases should only be included in the DOE strategy table if they are expected to have a large impact on the outcome of the test objective. Other recordable factors can be included in a footnote and documented in more detail in the test plan.

It is also possible to have a factor or levels of a factor that will be systematically varied during a test but not in a statistically defensible fashion. These conditions are sometimes necessary to demonstrate (demo) in tests for safety, cost, or simply the fact that they rarely occur in regular operation of the system

Design of Experiments – Artillery Howitzer Example

DESIGN OF EXPERIMENTS (for a Milestone B Artillery Howitzer)

Design of Experiments (DOE) Overview

The purpose of this appendix is to provide a framework for the OTA's Design of Experiments (DOE) methodology in support of a howitzer acquisition. The OTA will plan and conduct both the LUT/OA/OA and the IOT using DOE principles. This method of assessment will provide a systematic approach to assess the effects of pre-determined factors on key performance aspects of the howitzer. The design goal is to vary key factors that affect measurable system characterizations such as timeliness and accuracy. Table D.1 below shows how the factors and factor levels will be controlled during each test event.

Table D.1: DOE Campaign Strategy

Factors	**Factor Levels**	**Test Events**	
		LUT /OA	IOT
Ammo-Lethal	Projectile 1(P1), Projectile 2(P2)	SV	SV
Ammo-Non Lethal	Smoke, Illum	Non-Lethal limited # missions	Non-Lethal limited # missions
Time	Day, Night	SV	SV
Range Band	C1 + C2, C3, C4, C5	SV	SV
Traverse	0-15, 15-45, Out of Sector	SV (0-15, 15-45), Out of Sector (limited # missions)	SV (0-15, 15-45), Out of Sector (limited # missions)
Angle	Low, High	SV	SV
Fuze	Time Delay (TD), Point Detonation(PD), Multi-option fuse (MOF)	SV	SV
MOPP	0, IV	HC-MOPP 0, MOPP IV limited # missions	HC-MOPP O, MOPP IV limited # missions
Test Elements	# of test elements	HC (1 Element)	SV (3 Elements)
IA	None, Red team	None	HC-None, Red team excursion at end of test

Notes/Definitions:

*HC-Held Constant *SV – Systematically Varied *C1-MACS 1 or equivalent

*C2-MACS 2 or equivalent *C3-MACS 3 or equivalent *C4-MACS 4 or equivalent

*High Angle of fire – Above maximum range Quadrant of Elevation(>~800 mils)

*Low Angle of Fire – Below maximum range Quadrant of Elevation(<~800mils)
*IA – Information Assurance

LUT /OA:

The objectives of the LUT/OA shall be to evaluate the howitzer interoperability, fire mission accuracy and responsiveness and automotive performance as well as mobility and reliability in support of combat operations. Table D.2 shows critical responses.

Table D.2: Critical Responses

Critical Responses	Accuracy (Miss Distance in meters, CEP)
	Timeliness (Time to Complete Mission in seconds)
	Reliability (Mean Time between Failure)

This phase of the operational testing will follow a D-optimal split-plot design of experiments approach with some of the hard to control factor systematically controlled to balance DOE and operational realism from the OMS/MP. Table D.3 lists the factors and levels for the two responses: accuracy and timeliness.

Table D.3: Factors and Levels

Factor	Levels	Control
Projectile	P1, P2	Hard, Systematic
Time	Day, Night	Hard, Systematic
Range Band	C1 + C2, C3, C4, C5	Hard, Systematic
Traverse Angle	0-15, 15-45	Hard
Angle of Fire	Low, High	Easy
Fuze Type	TD, PD, MOF	Hard

If a factor it systematically controlled it was organized in an operationally realistic manner yet based on a D-optimal design. Projectile, Time, and Range were organized so that it followed a scenario where it starts on closest range bands (C1 + C2) and then moves to the C5 range band over the first two 24-hour periods before returning to the initial bands over the next two 24-hour periods. If a factor was hard to control, these factors were randomized over whole plots (blocks of time where the time, Projectile, range band, traverse, and fuze could randomly be assigned). Angle is an easy to control so it could be randomly assigned to the individual missions or within the blocks. The DOE consists of 96 missions, but to meet the reliability requirements, 160 missions are necessary. These additional missions are distributed between special case requirements (Non-Lethal, emergency firings, MOPP IV, Out of Sections, and other long range missions to meet the OMS/MP. These additional missions will be injected into the DOE run matrix at the discretion of the Test Officer to ensure operational realism. For example,

all the Out of Sector and Emergency missions will be conducted right after tactical moves. Table D.4 shows the breakout by mission.

Table D.4: Factor Breakout By Mission

	Range	Charge	P1 Missions	P2 Missions	Illum Missions	Smoke Missions	Total Missions
DOE	4 - 9 KM	1/2L	16	0	-	-	16
	9-12 KM	3H	16	0	-	-	16
	12-15 KM	4H	16	20	-	-	36
	16.4 - 20 KM	5H	-	28	-	-	28
Non-Lethal	TBD	TBD	-	-	3	3	6
Emergency firings	16.4 - 20 KM	5H	-	12	-	-	12
MOPP IV	16.4 - 20 KM	5H	-	8	-	-	8
Additional Long range for RAM	16.4 - 20 KM	5H	-	26	-	-	26
Out of Sector	TBD	TBD	-	12	-	-	12
Total	-	-	48	108	3	3	160

The D-Optimal Split-Split Plot design permits the ability to estimate all main effects, all 2-way interactions with time, and the following additional interactions: range band and traverse, traverse and angle, angle and fuze, traverse and fuze, and projectile and angle. The run matrix, which it the required order that these runs must follow, is shown in table D.5 below.

Table D.5: LUT/OA D-Optimal Split-Split Plot Run Matrix

Day	Time	Projectile	Range Band	Traverse	Angle	Fuze
1	Day	P1	C1 + C2	0-15	High	TD
1	Day	P1	C1 + C2	0-15	Low	TD
1	Day	P1	C1 + C2	0-15	Low	TD
1	Day	P1	C1 + C2	0-15	High	TD
1	Day	P1	C1 + C2	0-15	High	PD
1	Day	P1	C1 + C2	0-15	Low	PD
1	Day	P1	C1 + C2	0-15	Low	PD
1	Day	P1	C1 + C2	0-15	High	PD

Design of Experiments – Artillery Howitzer Example

Day	Time	Projectile	Range Band	Traverse	Angle	Fuze
1	Day	P1	C3	30-45	Low	PD
1	Day	P1	C3	30-45	High	PD
1	Day	P1	C3	30-45	Low	PD
1	Day	P1	C3	30-45	High	PD
1	Night	P1	C3	0-15	High	TD
1	Night	P1	C3	0-15	High	TD
1	Night	P1	C3	0-15	Low	TD
1	Night	P1	C3	0-15	Low	TD
1	Night	P1	C4	30-45	High	TD
1	Night	P1	C4	30-45	High	TD
1	Night	P1	C4	30-45	Low	TD
1	Night	P1	C4	30-45	Low	TD
1	Day	P1	C4	0-15	Low	MOF
1	Day	P1	C4	0-15	Low	MOF
1	Day	P1	C4	0-15	High	MOF
1	Day	P1	C4	0-15	High	MOF
2	Day	P2	C4	30-45	High	MOF
2	Day	P2	C4	30-45	Low	MOF
2	Day	P2	C4	30-45	Low	MOF
2	Day	P2	C4	30-45	High	MOF
2	Day	P2	C4	30-45	Low	TD
2	Day	P2	C4	30-45	High	TD
2	Day	P2	C4	30-45	Low	TD
2	Day	P2	C4	30-45	High	TD
2	Night	P2	C5	30-45	Low	MOF
2	Night	P2	C5	30-45	Low	MOF
2	Night	P2	C5	30-45	Low	MOF
2	Night	P2	C5	30-45	Low	MOF
2	Night	P2	C5	30-45	Low	PD

Design of Experiments – Artillery Howitzer Example

Day	Time	Projectile	Range Band	Traverse	Angle	Fuze
2	Night	P2	C5	30-45	Low	PD
2	Night	P2	C5	30-45	Low	PD
2	Night	P2	C5	30-45	Low	PD
2	Night	P2	C5	0-15	Low	TD
2	Night	P2	C5	0-15	Low	TD
2	Night	P2	C5	0-15	Low	TD
2	Night	P2	C5	0-15	Low	TD
2	Night	P2	C5	30-45	Low	TD
2	Night	P2	C5	30-45	Low	TD
2	Night	P2	C5	30-45	Low	TD
2	Night	P2	C5	30-45	Low	TD
3	Day	P2	C5	0-15	Low	MOF
3	Day	P2	C5	0-15	Low	MOF
3	Day	P2	C5	0-15	Low	MOF
3	Day	P2	C5	0-15	Low	MOF
3	Day	P2	C5	30-45	Low	PD
3	Day	P2	C5	30-45	Low	PD
3	Day	P2	C5	30-45	Low	PD
3	Day	P2	C5	30-45	Low	PD
3	Day	P2	C5	0-15	Low	TD
3	Day	P2	C5	0-15	Low	TD
3	Day	P2	C5	0-15	Low	TD
3	Day	P2	C5	0-15	Low	TD
3	Day	P2	C4	0-15	High	PD
e	Day	P2	C4	0-15	High	PD
3	Day	P2	C4	0-15	Low	PD
3	Day	P2	C4	0-15	Low	PD
3	Night	P2	C4	0-15	Low	MOF
3	Night	P2	C4	0-15	High	MOF

Design of Experiments – Artillery Howitzer Example

Day	Time	Projectile	Range Band	Traverse	Angle	Fuze
3	Night	P2	C4	0-15	Low	MOF
3	Night	P2	C4	0-15	High	MOF
3	Night	P2	C4	0-15	Low	PD
3	Night	P2	C4	0-15	High	PD
3	Night	P2	C4	0-15	Low	PD
3	Night	P2	C4	0-15	High	PD
3	Night	P1	C4	0-15	High	PD
3	Night	P1	C4	0-15	Low	PD
3	Night	P1	C4	0-15	High	PD
3	Night	P1	C4	0-15	Low	PD
4	Day	P1	C4	30-45	Low	MOF
4	Day	P1	C4	30-45	High	MOF
4	Day	P1	C4	30-45	High	MOF
4	Day	P1	C4	30-45	Low	MOF
4	Day	P1	C3	30-45	Low	TD
4	Day	P1	C3	30-45	Low	TD
4	Day	P1	C3	30-45	High	TD
4	Day	P1	C3	30-45	High	TD
4	Night	P1	C3	30-45	High	MOF
4	Night	P1	C3	30-45	High	MOF
4	Night	P1	C3	30-45	Low	MOF
4	Night	P1	C3	30-45	Low	MOF
4	Night	P1	C1 + C2	30-45	High	PD
4	Night	P1	C1 + C2	30-45	Low	PD
4	Night	P1	C1 + C2	30-45	Low	PD
4	Night	P1	C1 + C2	30-45	High	PD
4	Night	P1	C1 + C2	0-15	Low	MOF
4	Night	P1	C1 + C2	0-15	High	MOF
4	Night	P1	C1 + C2	0-15	High	MOF

Day	Time	Projectile	Range Band	Traverse	Angle	Fuze
4	Night	P1	C1 + C2	0-15	Low	MOF

The power of the tests to illustrate how the factors influence the responses are listed below in Table D.6:

Table D.6: Power Effect on Factors and Responses

Effect	Variance	Power (90% Confidence, S:N=2)	Power (80% Confidence, S:N=1)
Intercept	0.228	0.994	0.789
Time	0.303	0.974	0.701
Range Band 1	0.333	0.963	0.671
Range Band 2	0.245	0.991	0.767
Range Band 3	0.180	0.999	0.855
Traverse	0.305	0.974	0.699
Angle	0.018	1.000	1.000
Fuze 1	0.208	0.997	0.816
Fuze 2	0.194	0.998	0.836
Projectile	0.390	0.937	0.624
Time*Range Band 1	0.559	0.842	0.524
Time*Range Band 2	0.273	0.984	0.733
Time*Range Band 3	0.147	1.000	0.906
Time*Traverse	0.208	0.997	0.816
Time*Angle	0.016	1.000	1.000
Time*Fuze 1	0.095	1.000	0.974
Time*Fuze 2	0.269	0.985	0.738
Time*Projectile	0.464	0.897	0.574
Range Band*Traverse 1	0.299	0.976	0.705
Range Band*Traverse 2	0.257	0.988	0.752
Range Band*Traverse 3	0.222	0.995	0.797

Design of Experiments – Artillery Howitzer Example

Effect	Variance	Power (90% Confidence, S:N=2)	Power (80% Confidence, S:N=1)
Traverse*Angle	0.016	1.000	1.000
Angle*Fuze 1	0.016	1.000	1.000
Angle*Fuze 2	0.014	1.000	1.000
Traverse*Fuze 1	0.145	1.000	0.908
Traverse*Fuze 2	0.182	0.999	0.852
Projectile*Angle	0.018	1.000	1.000

IOT:

The objective of the IOT shall be to evaluate the howitzer interoperability, rate of fire, fire mission accuracy, responsiveness and automotive performance as well as mobility and reliability in support of combat operations. The test results shall support a full rate production decision.

The IOT will follow the same DOE philosophy and have the same factors and levels as the LUT/OA except it will be larger. A split plot design will be created based on the same set of factors and levels. Similarly the factors will be controlled in the same manner with the missions starting out close moving to the C5 ranges and the returning to the initial range bands over the course of the three 96-hour scenarios. Due to the increased number of missions, number of rounds fired and length of the test in the IOT compared to the LUT/OA, more interactions can be estimated, to include main effects and second order interactions. IOT design will ensure a similar balance between statistical capabilities and operational coverage. Similar to the LUT/OA, the IOT will consist of a smaller subset of the total number of required missions compared to the DOE missions. The overall ratio of the DOE to the total number of missions will be the same or very similar. Thus all the non-lethal, emergency firings, out of sector missions, and additional C5 missions needed to meet the OMS/MP, which would again follow tactical moves, and additional C5 missions will be injected into the matrix at the discretion the Test Officer to ensure operational realism.

Red Team excursions will be conducted at the discretion of the IOT Test Officer. These excursions will support Information Assurance evaluation requirements in an operational environment at a system of systems level. Additional information relating to Red Team excursions can be found in paragraph 4.3.2.5 “IOT Events, Scope of Testing and Scenarios” of the TEMP.

DESIGN OF EXPERIMENTS (for a Precision Guided Weapon)

D.1 Design of Experiments (DOE) Definitions

This appendix uses terminology specific to DOE; the following definitions should be applied while reading.

- Initial Factor – A factor determined to potentially impact the performance of the precision guided weapon system in which the weapon system operates. Initial factors are pulled from the test design framework developed by the Operational Test Activity (OTA) or from subject matter expert inputs. Initial factors are accepted on their own, combined with other initial factors and accepted, placed in recordable status, determined to be a demo item, or eliminated from consideration for the DOE design.
- Accepted Factor – a factor accepted as a standalone from an initial factor or through the combination of multiple initial factors. Accepted factors were input into JMP[1] software to create the DOE. Accepted factors are given levels.
- Level – the regions or levels that would be input into JMP software to create the DOE tables. Each accepted factor has a minimum of two levels.
- Recordable (Non-DOE) factor – a factor for which data are recorded during testing, but is not included in the DOE design. Factors that cannot be controlled, but might impact the performance the weapon system are placed into this category. These factors and their values will be recorded and compared against the performance of the weapon system to determine the impact they may have on the system.
- Demo Items – a factor or particular capability that will be tested against but is not incorporated into the DOE design created with JMP software. Demo items will be tested in standalone events if deemed to impact response variable, or incorporated into the DOE events when deemed to not impact response variable.
- Strike Warfare (STW) – the precision guided weapon system when used against Stationary Land Targets (SLT).
- Surface Warfare (SUW) – the precision guided weapon system when used against Moving Maritime Targets (MMT).

D.2.0 Overarching DOE Strategy

The precision guided weapon system effectiveness will depend on its ability to conduct two primary missions:

- Surface Warfare (SUW) against MMTs, and

[1] JMP (http://jmp.com/) is the registered trademark for a statistical software package that can assist with experimental design. Design Expert (http://www.statease.com/dx8descr.html), can also be used for DOE.

Design of Experiments – Precision Guided Weapon Example

- Strike Warfare (STW) against SLTs

Design of Experiments was used to develop the DT&E, integrated test events, and the IOT&E. A significant amount of data from previous testing of this precision guided weapon system exists, which helped to refine the test design. Captive carry testing will be used to execute the majority of the testing. The captive carry testing uses a precision guided weapon system digital simulation consists of high fidelity guidance and electronics unit (GEU) and seeker models coupled with a target scene generator. The scene generator creates a perspective projection of the infrared target scene as presented to the seeker optics; the scenes are developed from empirical data and incorporate environmental effects such as time of day, sea state, humidity, and atmospheric conditions. Seeker imagery and GEU performance data captured during previous captive carry flight testing has been used to successfully validate the all digital precision guided weapon system simulation. The T&E WIPT consisting of the Technical Program Office, Lead Test Engineers, Systems Engineers, OTA testers, and DOE Subject Matter Experts determined that the appropriate response variables for evaluating the effectiveness of the system are:

- ***Aim point delta***: the distance between seeker aimpoint and the preplanned aimpoint at the final seeker aimpoint refinement. This response variable applies to both the captive carry (CC) and free flight (FF) live fire tests.
- ***Miss distance***: the distance between the preplanned aimpoint and the actual impact point for FF live fire shots.

Additionally, the T&E WIPT determined and defined the initial set of factors selected for both SUW and STW missions. These factors were then ranked based on their predicted impact to the response variable and their intended use in the design. Tables D.1 – D.2 provide the overall DOE strategy for each test objective (assessing weapon system effectiveness for SUW Missions and STW Missions).

Table D.1: Overview of DOE Strategy for Surface Warfare (SUW) Against Moving Maritime Targets (MMT)

		Test Phase		
		DT	**IT**	**IOT**
Critical Responses		**Aim Point Delta**	**Aim Point Delta**	**Aim Point Delta** **Miss Distance**
Factors	**Factor Levels**			
Sun Elevation	4 Levels	SV*	SV	SV
Target Type	4 Levels	SV	SV	SV

Target Range	Continuous	Record	Record	SV
Target Aspect	4 Levels	SV	SV	SV
Location Defenses	Maneuvering, RFCM, GPS Jamming	SV (Target Maneuver only)	SV(Target Maneuver only)	SV
Seeker Defenses	IRCM, Camouflage, Shipping Presence	Demo	Demo	SV

Table D.2: Overview of DOE Strategy for Surface Warfare (STW) Against Stationary Land Targets

		Test Phase		
		DT	**IT**	**IOT**
Critical Responses		**Aim Point Delta**	**Aim Point Delta**	**Aim Point Delta**
Factors	**Factor Levels**			
Terrain	4 Levels	Operational Testing will be used solely to determine system performance against the less challenging STL		SV
Target Orientation	4 Levels			SV
Contrast	Continuous			SV
Sun Elevation	4 Levels			SV
Defenses	Camouflage, IRCM, GPS Jamming			Demo

D.3.0 Developmental and Integrated Testing

Developmental and integrated testing will focus on the prioritized surface warfare (SUW) scenario against moving maritime targets (MMTs). The factors investigated in DT&E and IT are highlighted in more detail in table D-3 below.

D.3.1 DT/IT Power, Confidence, and Matrix for DOE Runs (MMT)

Using the accepted factors and assuming a normal distribution, the test design was created with JMP software for MMT using a D-optimal design for main effects and two-way interaction estimates. The matrix created includes 60 runs and using 80% confidence and provides sufficient a power to test for main effects. The power for detecting a 2 sigma shift difference in the response for Target Type is 80 percent, for Target Aspect is 63 percent, for Target Manuver is 98 percent, and for Sun Elevation is 51.5 percent. The lower power for Sun Elevation is due to the five levels of the factor and acceptable because it is expected that not all five levels will result in significantly different performance. The data will be collected during 60 captive carry runs. In addition to these 60 (30 DT&E, 30 IT&E) data runs, there will be 8 (4

DT&E, 4 IT&E) captive carry dress rehearsals and 4 (2 DT&E, 2 IT&E) free flight live fire runs where the data will be recorded during the MMT DT/IT testing.

Table D-3. MMT DOE for DT&E and IT&E

MMT DOE FACTORS (DT/IT)		
INITIAL FACTORS	**ACCEPTED FACTORS**	**LEVELS**
Thermal Contrast Day/Night Glint	Sun Elevation	≤ 1/2 Peak Rising - 1 > 1/2 Peak Rising - 2 > 1/2 Peak Setting - 3 ≤ 1/2 Peak Setting - 4 Night - 5
Target Speed Target Size	Target Type	Small (≤ ft) & Slow (≤ knots) Small (≤ ft) & Fast (> knots) Large (> ft) & Slow (≤ knots)
Target Aspect	Target Aspect	Head (0) Beam (90/270) Qtr (45/135/225/315) Tail (180)
TGT Maneuvering	TGT Maneuver	Evasive S Turn Non-maneuvering (constant course and speed)
RECORDABLE (NON-DOE)		
Sea State	Thermal Crossover	Humidity
DEMO ITEMS		
Multi Weapons Weapon Datalink	Datalink Source IRCM	Search Altitude WPN/Datalink RNG

The overall average miss distance will be compared against threshold values for the system to support the evaluation of the precision guided weapon system CPD requirements. ANOVA and regression analysis will also be performed based on the results. The analysis will provide additional evaluation understanding of overall system capabilities and limitations.

D.4.0 Operational Test DOE Development

In order to better evaluate precision guided weapon system performance in the STW and SUW operational environments, two distinct mission-based DOEs were developed: one for engaging stationary land targets (SLT) and one for engaging MMTs. Since the STW and SUW missions and requirements for precision guided weapon system employment are so different, one combined DOE would not adequately test the system.

STW requires the delivery platform to fly to the release point and launch the precision guided weapon system with prelaunch coordinates entered into the weapon. When the weapon approaches the target, the seeker will refine the flight profile to ensure the precision guided

weapon system strikes the desired impact point on a stationary target. The precision guided weapon system incorporates a new seeker design.

SUW requires the delivery platform to detect the target with either a radar or targeting sensor, fly to the release point, and launch the precision guided weapon system. The delivery platform provides IFTU support to get the precision guided weapon system as close as possible to the MMT. As the weapon approaches the MMT, the seeker takes over, refining the flight profile in the final miles to ensure the precision guided weapon system strikes at the desired impact point on a moving target. These two distinct missions are described in detail below.

D.4.1 Operational Test DOE (STW)

Using DOE, the OT team leveraged the knowledge base from previous precision guided weapon system testing in developing the streamlined STW test design. The following assumptions provided the foundation for selecting the factors and levels for the test design:

- the weapons procedures for employment against SLT remained unchanged from the legacy precision guided weapon system;
- the weapon Launch Area Region (LAR), release and separation characteristics from the launch aircraft, and warhead capabilities remained the same;
- the new seeker capabilities and limitations will be compared against the legacy precision guided weapon system seeker; and
- the same target set will be used for the comparison of seeker performance data as much as possible.

The DOE factors considered known capabilities and limitations of the legacy precision guided weapon system seeker.

The precision guided weapon system test design was created primarily for Captive Carry (CC) runs. Replication was used to increase the understanding of the effects size and variability of data for specific test runs while increasing the statistical power and confidence of the test. The breadth of the design, coupled with the ease of performing multiple CC runs in a short period of time against SLTs in STW scenarios, facilitated replication in a cost efficient matter. With targets grouped together in a target area it is possible to fly against three or four different targets during an event, but not possible to transit to a new area during the course of one flight. It was deemed effective and efficient to fly three runs against each target in the target area, allowing nine runs or greater to be performed during each flight.

Outside of the primary DOE for CC runs, a robust test against Global Positioning System (GPS) jamming and Infra-red Countermeasures (IRCM) was also developed. This test will be used to demonstrate the specific effects of GPS denial, IRCM, and camouflage on the precision guided weapon system seeker. The performance of the precision guided weapon system will be compared directly against the legacy system in this same environment.

In addition to the CC STW DOE matrix and the CC test against GPS jamming/IRCM described above, data from two Free Flights (FF)/live fire (performed in IT) will be evaluated

and compared with the results from the CC runs. Each of the FF/live fire shots will have CC dress rehearsal runs performed prior to the weapon release. These CC dress rehearsal runs will occur on a flight prior to the actual FF event to run through the FF scenario and ensure pilot familiarization with the event. The data gathered during the CC dress rehearsal and the CC runs just prior to the launch will also be used to compare with previous data gathered during the CC DOE and CC test against GPS jamming.

Table D-4 presents the factors for STW during OT&E. Table D-5 and D-6 provide the test matrix.

Table D-4. OT&E Factors and Levels for STW

STW DOE FACTORS (OT)		
INITIAL FACTORS	ACCEPTED FACTORS	LEVELS
Terrain	Terrain	Desert Mountain Urban Littoral
Target Orientation	Target Orientation	Horizontal Face Vertical Face
Clutter Civil Structures Snow	Contrast	High Low
Thermal Contrast	Sun Elevation	<1/2 peak AM or PM >1/2 peak AM or PM
RECORDABLE (NON-DOE)		
Thermal Crossover		Humidity
DEMO ITEMS		
IRCM	Camouflage Day/Night	GPS jamming

D.4.1.1 Operational Test Power, Confidence, and Matrix for DOE Runs (STW)

Using the factors above and assuming a normal distribution, the design was created with JMP for STW using a full factorial design for main effects and two-way interaction estimates. The matrix created includes 32 runs, which will each be replicated three times, for a total of 96 runs. The replications are a result of efficient use of flight sortie time by repeating runs rather than repeating flights. This design used 80 percent confidence level and yielded a power of test of greater than 95 percent to detect a 1 sigma change in performance across all main effects and greater than 85 percent power for all two-factor interactions. The runs are displayed in Table D-3.

Table D-5. OT&E STW Run Matrix

OT STW Matrix Full Factorial						
High Humidity Det						
Run	Sun Elevation	Orientation	Contrast	Humidity	Terrain	Actual Target
1-3	<1/2 max	Horizontal	Low	High	Littoral	Corpus Christi Command Center Wall
4-6	<1/2 max	Horizontal	High	High	Littoral	Corpus Christi Hangar
7-9	<1/2 max	Vertical	Low	High	Littoral	Corpus Christi Small Building on Pier
10-12	<1/2 max	Vertical	High	High	Littoral	Corpus Christi Tower
13-15	<1/2 max	Horizontal	High	High	Urban	Orange Grove Roof of NE Bldg
16-18	<1/2 max	Horizontal	Low	High	Urban	Orange Grove Airfield Arresting gear building
19-21	<1/2 max	Vertical	Low	High	Urban	Orange Grove ILS Radar
22-24	<1/2 max	Vertical	High	High	Urban	Target TBD
25-27	>1/2 max	Horizontal	Low	High	Littoral	Corpus Christi Command Center Wall
28-30	>1/2 max	Horizontal	High	High	Littoral	Corpus Christi Hangar
31-33	>1/2 max	Vertical	Low	High	Littoral	Corpus Christi Small Building on Pier
34-36	>1/2 max	Vertical	High	High	Littoral	Corpus Christi Tower
37-39	>1/2 max	Vertical	High	High	Urban	Orange Grove Roof of NE Bldg
40-42	>1/2 max	Horizontal	Low	High	Urban	Orange Grove Airfield Arresting gear building
43-45	>1/2 max	Vertical	Low	High	Urban	Orange Grove ILS Radar
46-48	>1/2 max	Horizontal	High	High	Urban	Target TBD
Low Humidity						
Run	Sun Elevation	Orientation	Contrast	Humidity	Terrain	Actual Target
49-51	<1/2 max	Horizontal	High	Low	Mountain	Independence Courthouse Multi level Building
52-54	<1/2 max	Horizontal	Low	Low	Mountain	Independence Jailhouse Large building
55-57	<1/2 max	Vertical	Low	Low	Mountain	Independence Microwave Tower
58-60	<1/2 max	Vertical	High	Low	Mountain	Target TBD
61-63	<1/2 max	Horizontal	Low	Low	Desert	Trona Large Yellow Building
64-66	<1/2 max	Horizontal	High	Low	Desert	Trona Movie Theater
67-69	<1/2 max	Vertical	High	Low	Desert	Trona Post Office Wall
70-72	<1/2 max	Vertical	Low	Low	Desert	Ballarat Radar/R2508
73-75	>1/2 max	Horizontal	High	Low	Mountain	Independence Courthouse Multi level Building
76-78	>1/2 max	Horizontal	Low	Low	Mountain	Independence Jailhouse Large building
79-81	>1/2 max	Vertical	Low	Low	Mountain	Independence Microwave Tower
82-84	>1/2 max	Vertical	High	Low	Mountain	Target TBD
85-87	>1/2 max	Horizontal	Low	Low	Desert	Trona Large Yellow Building
88-90	>1/2 max	Horizontal	High	Low	Desert	Trona Movie Theater
91-93	>1/2 max	Vertical	High	Low	Desert	Trona Post Office Wall
94-96	>1/2 max	Vertical	Low	Low	Desert	Ballarat Radar/R2508

The overall average miss distance will be compared against threshold values for the system to support the evaluation of the precision guided weapon system CPD requirements. ANOVA and regression analysis will be performed as well, based on the results. The analysis will provide additional understanding of overall system capabilities and limitations.

D.4.1.2 Matrix for Demo and Countermeasure Runs (STW)

The STW demonstration items (IRCM, GPS jamming, GPS availability, and camouflage) will be demonstrated during the following 30 runs, which are displayed in Table D-6.

Twelve runs versus GPS jamming in mountainous terrain (six against co-altitude jamming)

Twelve runs in R-2505 versus multiple countermeasures in the White Sands area

Six runs in R-2505 versus multiple IR countermeasures.

Table D-6. OT&E STW Demo Run Matrix

Advanced Countermeasures								
Run	Sun Elevation	Orientation	Contrast	Humidity	Terrain	Actual Target	Jamming Profile	Countermeasure
1	>1/2 max	Vertical	High	Low	Mountain	GPS Jamming Parrot Peak Radar dish	25K to 20 degree	
2	>1/2 max	Vertical	High	Low	Mountain	GPS Jamming Parrot Peak Radar dish	25K to 20 degree	
3	>1/2 max	Vertical	High	Low	Mountain	GPS Jamming Parrot Peak Radar dish	25K to 20 degree	
4	<1/2 max	Vertical	High	Low	Mountain	GPS Jamming Parrot Peak Radar dish	Co altitude	
5	<1/2 max	Vertical	High	Low	Mountain	GPS Jamming Parrot Peak Radar dish	Co altitude	
6	<1/2 max	Vertical	High	Low	Mountain	GPS Jamming Parrot Peak Radar dish	Co altitude	
7	>1/2 max	Horizontal	High	Low	Mountain	GPS Jamming Parrot Peak Building roof	25K to 20 degree	
8	>1/2 max	Horizontal	High	Low	Mountain	GPS Jamming Parrot Peak Building roof	25K to 20 degree	
9	>1/2 max	Horizontal	High	Low	Mountain	GPS Jamming Parrot Peak Building roof	25K to 20 degree	
10	<1/2 max	Horizontal	High	Low	Mountain	GPS Jamming Parrot Peak Building roof	Co altitude	
11	<1/2 max	Horizontal	High	Low	Mountain	GPS Jamming Parrot Peak Building roof	Co altitude	
12	<1/2 max	Horizontal	High	Low	Mountain	GPS Jamming Parrot Peak Building roof	Co altitude	
13	<1/2 max	Vertical	Low	Low	Desert	2505 Sams Town T-Building	Point	Multiple/White Sands
14	<1/2 max	Vertical	Low	Low	Desert	2505 Sams Town T-Building	Point	Multiple/White Sands
15	<1/2 max	Vertical	Low	Low	Desert	2505 Sams Town T-Building	Point	Multiple/White Sands
16	>1/2 max	Vertical	Low	Low	Desert	2505 Sams Town T-Building	Point	Multiple/White Sands
17	>1/2 max	Vertical	Low	Low	Desert	2505 Sams Town T-Building	Point	Multiple/White Sands
18	>1/2 max	Vertical	Low	Low	Desert	2505 Sams Town T-Building	Point	Multiple/White Sands
19	<1/2 max	Horizontal	Low	Low	Desert	2505 Sams Small Building 1 Story	Point	Multiple/White Sands
20	<1/2 max	Horizontal	Low	Low	Desert	2505 Sams Small Building 1 Story	Point	Multiple/White Sands
21	<1/2 max	Horizontal	Low	Low	Desert	2505 Sams Small Building 1 Story	Point	Multiple/White Sands
22	>1/2 max	Horizontal	Low	Low	Desert	2505 Sams Small Building 1 Story	Point	Multiple/White Sands
23	>1/2 max	Horizontal	Low	Low	Desert	2505 Sams Small Building 1 Story	Point	Multiple/White Sands
24	>1/2 max	Horizontal	Low	Low	Desert	2505 Sams Small Building 1 Story	Point	Multiple/White Sands
25	<1/2 max	Vertical	Low	Low	Desert	2505 POL Coles Flat	Point	Laser CM and Flames
26	<1/2 max	Vertical	Low	Low	Desert	2505 POL Coles Flat	Point	Laser CM and Flames
27	<1/2 max	Vertical	Low	Low	Desert	2505 POL Coles Flat	Point	Laser CM and Flames
28	>1/2 max	Vertical	Low	Low	Desert	2505 POL Coles Flat	Point	Laser CM and Flames
29	>1/2 max	Vertical	Low	Low	Desert	2505 POL Coles Flat	Point	Laser CM and Flames
30	>1/2 max	Vertical	Low	Low	Desert	2505 POL Coles Flat	Point	Laser CM and Flames

D.4.2 Operational Test DOE (SUW)

Using DOE, the OT team extensively leveraged the knowledge base from previous precision guided weapon system testing in developing the streamlined SUW test design. The following assumptions provided the foundation for selecting the factors and levels for the precision guided weapon system SUW test design:

- the weapon Launch Area Region (LAR), release and separation characteristics from the launch aircraft, and warhead capabilities remained the same;
- the new seeker capabilities and limitations will be compared against the legacy precision guided weapon system seeker.

The DOE factors included limitations of the legacy precision guided weapon system seeker.

The precision guided weapon system SUW test design was created primarily for CC runs. Replication was not used due to the large number of factors to be tested against and the difficulty in performing each run.

In addition to the CC SUW DOE matrix, data from two FF/live fire shots being performed in IT and data from two FF/live fire shots being performed in OT will be evaluated and compared with the results from CC runs. Each of the FF/live fire shots will have CC runs performed prior to the weapon release. These CC dress rehearsal runs will occur on a flight prior to the actual FF event. During the event for the FF/live fire shot, the profile will be flown CC a few times to ensure everything is working properly. The data gathered during the dress rehearsal and the CC runs prior to the launch will also be compared with previous data gathered during the CC DOE matrix.

Table D-7 presents the factors for SUW during OT&E.

Table D-7. OT&E Factors and Levels for SUW

SUW DOE FACTORS (OT)		
INITIAL FACTORS	ACCEPTED FACTORS	LEVELS
Thermal Contrast Day/Night Glint	Sun Elevation	≤ 1/2 Peak Rising - 1 > 1/2 Peak Rising - 2 > 1/2 Peak Setting - 3 ≤ 1/2 Peak Setting - 4 Night - 5
Target Speed Target Size	Target Type	Small (≤100 ft) & Slow (≤ 15 knots) Small (≤100 ft) & Fast (> 15 knots) Large (>100 ft) & Slow (≤ 15 knots) Large (>100 ft) & Fast (> 15 knots)
Threat WPN Range Target Slant Range	Target Range	≤ 40 nm > 40 nm
Target Aspect	Target Aspect	Head (0) Beam (90/270) Qtr (45/135/225/315) Tail (180)
TGT Maneuvering RFCM GPS Jamming	Location Defenses	Yes
IRCM Camouflage Shipping presence	Seeker Defenses	Yes No
RECORDABLE (NON-DOE)		
Sea State	Thermal Crossover Humidity	Glint
DEMO ITEMS		
Multi-Weapons	Datalink Source	Weapon Datalink

D.4.2.1 Operational Test Power, Confidence, and Matrix for DOE Runs (SUW)

Using these factors and assuming a normal distribution, the design was created with JMP for SUW using a D-optimal design for main effects and two-way interaction estimates. The matrix created includes 80 runs using 80 percent confidence and yields a power of test of 99 percent to detect a 2 sigma change in performance for Target Range, Location Defenses, and Seeker defenses. The power for Target Type and Target Aspect is 68 percent. The power for Sun Elevation is 56 percent. The lower powers for the OT SUW factors are acceptable because the DT&E and IT&E will provide amplifying information to the OT&E. If factors are deemed to

be insignificant in testing preceding the OT&E the test design will be revised to optimize power for the remaining factors in OT&E.

D.4.2.2 Additional SUW Runs

In addition to the 80 SUW test runs described above, a minimum of six CC runs will be conducted as dress rehearsal runs for the two free flight/live fire shots against MMT targets and then the two FF/live fire runs. The data will be recorded and compared to CC data. The specifics of these runs will be detailed in the Test Plan. See Table D-8.

Table D-8. OT&E SUW Free Flight

OT SUW Free Flight Matrix								
Run	Sun Elev.	Tgt Aspect	Tgt Type	Datalink Range	Humidity	Location Defenses	Seeker Defenses	Notes
65	2	Tail	Large/Slow	Long	Low	Yes	Yes	Dress
66	2	Tail	Large/Slow	Long	Low	Yes	Yes	Dress
67	2	Tail	Large/Slow	Long	Low	Yes	Yes	Dress
68	2	Tail	Large/Slow	Long	Low	Yes	Yes	Free Flight
69	3	Beam	Small/Fast	Short	Low	Yes	Yes	Dress
70	3	Beam	Small/Fast	Short	Low	Yes	Yes	Dress
71	3	Beam	Small/Fast	Short	Low	Yes	Yes	Dress
72	3	Beam	Small/Fast	Short	Low	Yes	Yes	Free Flight

D.4.3 Operational Test Data Analysis (STW & SUW)

The overall results of the response variable will be compared against threshold values for precision guided weapon system to support the resolution of COIs. ANOVA and regression analysis will be performed based on the results of the OT testing. This analysis will be utilized to understand system performance, the effects of the factors, and to provide tactical recommendations to the fleet operator in employment of precision guided weapon system.

Design of Experiments – Example for Software-Intensive System

(The following section would appear in the body of the TEMP for a Command and Control System at MS C. Appendix material begins on page 4.)

3.2 Test and Evaluation Framework

The Operational Test Activity (OTA) will accomplish the following during integrated testing:

- Determine if thresholds in the approved capabilities documents and COIs have been satisfied
- Determine Operational Effectiveness, Survivability, and Suitability of the system under realistic operational conditions
- Assess the contribution of the system to combat operations
- Provide additional information on the system's operational capabilities and limitations.

The OTA's evaluation plan creates a framework and methodology for evaluating the entirety of program data, obtained from late developmental testing, an operational assessment and IOT&E. The evaluation plan is intended provide a transparent, repeatable, and defensible approach to evaluation. The evaluation framework is captured in Table 3-1. The test team developed the test strategy by employing Design of Experiments (DOE) to ensure that a rigorous methodology supports the development and analysis of test results. DOE is used to design the tests to evaluate the data fusion KPP and the three COIs outlined in Table 3-1. A designed experiment is used to determine the effect of a factor or several factors (also called independent variables) on one or more measured responses (also called dependent variables). All COI DOEs are designed with mission-oriented response variables. Each design will include an estimation of the power of the test, which is included in the DOE Appendix. When gaps in the design are identified, these gaps will be listed as limitations, and a risk assessment will be provided in the appropriate Detailed Test Plan. In addition, the team will work with all appropriate parties to determine the most appropriate way to mitigate and/or manage the risks.

The OTA intends to exercise the command and control system during multiple training exercise (for a list of resources, see section 4.0) and dedicated test events. Real operators will be using the system for all tests where the data is considered in the evaluation of the COIs and data fusion KPP.

The Integrated test team has identified the response variables, factors and levels that will be exercised during each event in Table 3-2 to 3-5. The exact test size, experimental design, including expected trial replications, and confidence and power levels are outlined in the DOE Appendix. The identified confidence level and power are the maximums expected in a completely randomized event, due to restrictions in randomization. The major risk of not

completely randomizing the design is that some factors may become confounded with uncontrollable variables. The OTA will work to avoid any obvious confounding of variables. Data collected in training exercise will be supplemented by dedicated test events to mitigate any risks of data loss due to exercise objectives.

Table 3-2. Overview of DOE Strategy to Assess the Data Fusion KPP

		Test Phase		
		DT	**OA**	**IOT**
Critical Responses →		Track Accuracy, Timeliness, and Completeness	Track Accuracy, Timeliness, and Completeness	Track Accuracy, Timeliness, and Completeness
Factors	**Factor Levels**			
Connection	Categorical Factor with 5 levels: JREAP A/B/C, Link-16, CTN	SV*	SV	Record*
Number of Tracks	Low, Threshold, Objective	SV	SV	SV (simulated tracks in addition to live tracks)
Type of Track	Real time, Near real time, non-real time	SV	SV	Record

*Factors labeled systematically vary (SV) will be included in the DOE for data fusion. The data fusion DOE will be primarily executed in DT and the OA, IOT data will be used to confirm the results from DT and OT. If major configuration updates are made to the system between the OA and IOT, the factor management strategy for OT may need to be updated.

Tables 3-3 and 3-4 follow a similar format to Table 3-2 but are specific to each agency's respective mission.

Finally, a minimum of 3,000 hours of operation, equally spread across all three of the agencies employing the system are required to evaluate RAM and Ao requirements. These operation hours will be collect across late DT testing, the operational assessment, and the IOT&E. In order for the hours to count in the operational suitability assessment the system must be in a near final configuration and operated by operationally representative users.

Table 3-3. Overview of DOE Strategy to assess COI 1: System's ability to support mission of agency 1.

		Test Phase		
		DT	**OA**	**IOT**
Critical Responses →		1.Response time for critical information download/upload. 2.Number of missions successfully controlled.	1.Response time for critical information download/upload. 2. Rating of ability to control aircraft. 3.Number of missions successfully controlled.	1.Response time for critical information download/upload. 2.Rating of ability to control aircraft. 3.Number of missions successfully controlled.
Factors	**Factor Levels**			
Mission Load	Standard, High	SV	SV	SV
Track density	Standard, High	SV	SV	SV (simulated tracks in addition to live tracks)
Mission Duration	Short (4 hours), 24 hour operations	SV	SV	SV
Configuration	Small, Medium, Large	HC (Small)	HC (Medium)	HC (Large)
Environment	Desert, Hot & Humid, Cold	HC (Desert)	HC (Hot & Humid)	HC (Desert)

Sample DOE Appendix – Design of Experiment for COIs and Data Fusion KPP

Data Fusion KPP

Response variables

The data fusion KPP will be evaluated using the following critical measures, which have threshold requirements:

- Track Accuracy
- Track Completeness
- Track Timeliness

Factors

The following factors were considered for the data fusion KPP:

- Connection Method (JREAP A/B/C, Link-16, CTN)
    - Connection methods will be used both independently and simultaneously to assess an interoperability issues that may result
- Number of tracks (Low, Threshold, Objective)
- Type of Tracks (Real time, Near real time, Non-real time)

Table D-1 below provides the experimental design along with replications for achieving high power at the 95% confidence level to detect significant differences in factor levels. The power for detecting differences in the outcome based on the connection method is 91%, the power for detecting differences in the outcome based on the number and type of track is 99%. This design will be executed between both the developmental testing and the operational assessment. Half of each of the four runs will be conducted in DT, the other half will be conducted in the operational assessment. If for any reason this testing is not completed in DT and the OA it will be completed in the OT.

Table D-1. Experimental Design for Data Fusion KPP

		Connection Method					
Number Tracks	**Track Type**	JREAP A	JREAP B	JREAP C	Link-16	CTN	All Links
Low	Real time	4	4	4	4	4	4

	Near-real	4	4	4	4	4	4
	Non-real	4	4	4	4	4	4
Threshold	Real time	4	4	4	4	4	4
	Near-real	4	4	4	4	4	4
	Non-real	4	4	4	4	4	4
Objective	Real time	4	4	4	4	4	4
	Near-real	4	4	4	4	4	4
	Non-real	4	4	4	4	4	4

Figure D-1 shows power as a function of the number of replicates for each condition. Four replicates provide adequate power at the 95% confidence level to assess the data fusion KPP across all test conditions.

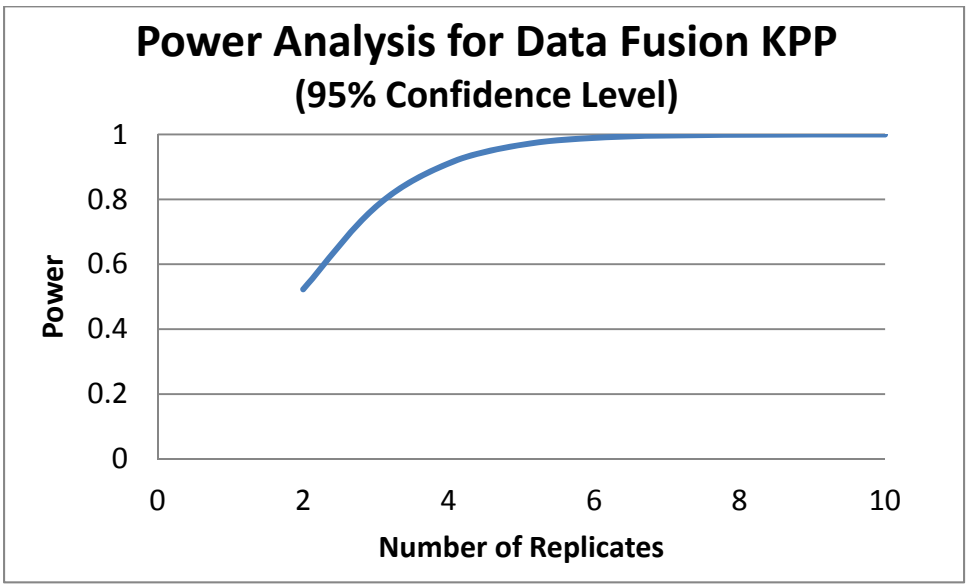


Figure D-1. Power Analysis for Data Fusion KPP

A similar discussion should follow for each of the additional COIs including the responses, factors, a proposed experimental design, and rational for the number of test points.

End-to-End Testing – Guidance

Guidance

End-to-end testing is the logical means to conduct a mission-based evaluation. End-to-end testing is easiest thought of as testing a mission thread. Mission threads result from a careful analysis of a unit's mission using the system and can be derived from the Joint Mission Essential Task List, from the Component-specific Mission Essential Task List, Concept of Employment (CONEMP), or the Army's Operational Mission Summary/Mission Profile (OMS/MP). The threads should make operational sense and evaluate the intended operational mission from beginning to end. The end-to-end evaluation of each mission thread should rely on testing that includes the entire thread in a single operational event. For example, a rocket or missile end-to-end test would include acquiring the target, passing the target information to a launch platform, firing the rocket or missile, hitting the target, and achieving the intended level of damage.

End-to-end testing is not just interoperability testing; it is simply not enough to verify that critical information can pass throughout the mission thread. The end-to-end evaluation must assess the quality and timeliness of the information as well as the success of mission outcomes. For example, the evaluation of a munition should address the ability of targeting systems to provide accurate and timely targeting data as well as evaluation of whether the intended target is hit and destroyed. The evaluation of a sensor platform should address the unit's ability to provide timely, accurate, and actionable information to the end user. The evaluation of a ship or aircraft should include the performance of all onboard and other supporting systems as well as evaluation of successful mission outcomes.

If it is not possible (due to cost or safety issues) to include all aspects of a mission in a single operational end-to-end test, separate portions of the mission threads can be included in multiple test events. Each of these events should include some overlap, so that the start of test B includes the end of Test A. Conditions affecting mission performance should be duplicated in overlapping events as much as possible. Each test of the thread parts should be operationally representative and all should represent similar operational environments and threats. If separate test events are used, the TEMP should explain why it is not possible to conduct the end-to-end mission in a single event; this is a test limitation, and the TEMP should discuss how this limitation is likely to affect the evaluation, and how the limitation will be mitigated.

For munitions, the end-to-end test can become a critical part of the LFT&E strategy. In an end-to-end test, the target aimpoint is selected operationally. Including this data increases the operational realism of the LFT&E. To be used as part of the LFT&E, full-up munitions must be used, targets must be realistic, and a damage assessment must be completed.

Systems often rely on other systems to complete missions. For these system-of-systems, the test and evaluation should address the impact of all systems to the mission, not just the system under test. It is possible that the system under test meets its requirements, yet cannot accomplish its mission due to the performance of another system.

End-to-End Testing – Guidance

For system-of-systems, end-to-end testing will involve systems other than the system under test. This can complicate test coordination when the additional systems are under the control of another program office. In these cases, DOT&E may require:

- That the availability of the critical system be included among the entrance criteria
- TEMP coordination signatures of the project office(s) responsible for the supporting system(s)

References

Reporting of Operational Test and Evaluation Results, DOT&E, January 6, 2010

Examples

End to End Testing – Examples

Cargo Aircraft Example

3.4 Operational Evaluation Approach. Operational testing of the C-100 cargo aircraft will employ the mission profiles as required by the CPD and described below. The missions will demonstrate delivery of time-sensitive/mission-critical supply items and/or personnel over operational/tactical distances to forward-deployed forces in remote and austere locations. Approximately 50 missions will demonstrate all variations of the mission profiles. Missions will include short notice logistical re-supply, casualty evacuation, troop movement, and aerial sustainment. The C-100 will operate to and from smaller, unimproved tactical landing strips and improved airfields up to the maximum cargo gross weight. The C-100 will be off-loaded to tactical rotary-wing aircraft and ground vehicles to demonstrate transloadability at Forward Operating Bases (FOBs) located near supported tactical units. The ability to rapidly reconfigure the C-100 will be evaluated. To evaluate adverse weather capability, the C-100 will conduct missions during day, night, night vision goggles (NVG), Visual Meteorological Conditions (VMC), and Instrument Meteorological Conditions (IMC).

The first three mission profiles (Mission profiles are described in an Annex) will be flown under day/night/NVG conditions to improved and unimproved runways, carrying various load configurations (463L pallets, troops, and vehicles), and will require 20 missions and approximately 64.0 flight hours.

Mission profiles 4 and 5 will include aircraft reconfiguration for aeromedical evacuation. Missions will be flown under day/night/NVG conditions to improved runways carrying various load configurations (463L pallets, troops, vehicles, and litter patients), and will require 16 missions and approximately 48.0 flight hours.

Mission profiles 6 and 7 will demonstrate single and multiple airdrops (four static line airlifts with door bundles and static line paratroop drops, and four military freefall airlifts). Airdrop missions will be flown under day/night/NVG conditions and will require eight missions and approximately 30 flight hours to demonstrate.

Mission profile 8 will demonstrate aerial sustainment under day/night/NVG conditions to improved runways, and will require approximately five missions and 34 flight hours.

Mission profile 9 will demonstrate self-deployment under day/night, visual flight rules/instrument flight rules (VFR/IFR), and will require one mission and approximately 40 flight hours.

Army Munition Example

3.4 Operational Evaluation Approach. The guided missile will be evaluated end-to-end. It is not possible to conduct the end-to-end mission in a single event due to availability of the unit, availability of real-time imagery of the test area, and delays between firing missions caused by the need to collect target data. Instead, the evaluation will be based on two operational events. The ground IOT&E will test the ability of a fire support unit to plan, target, and execute

guided missile missions. The flight IOT&E will test the unit's ability to fire guided missiles and examine the missile's effects on actual threat targets. During the ground phase, an operational unit will target and execute guided missile missions while executing other missions at an operational pace. Using satellite imagery of the actual test targets, the unit will mensurate the image using fielded equipment to estimate the target's location. Using fielded command and control equipment, the unit will determine the number of missiles and aimpoints. The mission information will be sent through the command and control chain to the launcher, which will dry-fire the missile. The flight phase will execute the missions generated during the ground phase. The test officer will digitally send a fire mission with aimpoints and number of missiles (determined in the ground IOT&E) to a battery command post. The battery will forward the fire missions to the launcher, which will move to a launch point and, after a brief safety delay, fire the missiles. The flight phase targets are threat-representative targets with threat-approved countermeasures. The Army Research Laboratory will conduct a damage assessment for each mission. The assessments are a critical component of the LFT&E strategy.

Details of the ground IOT&E, flight IOT&E, and LFT&E would be provided in other sections of the TEMP.

Force Protection and Personnel Casualties - Guidance

Summary

Force Protection attributes are those that contribute to the system's ability to protect its occupants and crew from the effects of threats likely to be encountered in combat. These threats often go beyond what is outlined in system requirements documents. For manned systems and systems designed to enhance personnel survivability on Live Fire Test and Evaluation (LFT&E) oversight, the critical LFT&E issues must include an evaluation of the vulnerability of its occupants to threats likely to be encountered in a combat environment. Personnel vulnerability should be addressed through dedicated measures of evaluation, such as "expected casualties" supported by specific details on the type and severity of injury, as well as the potential operational impact of such casualties on the ability of the platform to accomplish its mission after a threat engagement, when appropriate. Force protection must be addressed even in cases where the platform cannot survive.

Key Performance Parameters (KPPs) for force protection are required for all manned systems and systems designed to enhance personnel survivability, when those systems may be deployed in an asymmetric threat environment. Although force protection is a primary issue for programs on LFT&E oversight, evaluation of force protection may also be appropriate for programs that are not on LFT&E oversight. All Department of Defense (DoD) hard body armor acquisition programs under DOT&E oversight will execute, at a minimum, a DOT&E-approved protocol for testing that results in a decision to qualify a design for full-rate production (i.e., First Article Testing).

References

LFT&E Statute: 10 USC 2366

Policy for Updating Capabilities Documents to Incorporate Force Protection and Survivability Key Performance Parameters, The Joint Staff, 13 June 2005

Defense Acquisition Guidebook, Chapter 9

Standardization of Hard Body Armor Testing, DOT&E, 27 April 2010

Integrated Survivability Assessment – Guidance

Summary

The Developmental Test and Evaluation (DT&E), OT&E, and Live Fire Test and Evaluation (LFT&E) strategies should be integrated so that the full spectrum of system survivability is assessed in a consistent manner. For some systems, it might be appropriate for Critical Operational Issues (COIs) to address system and/or personnel survivability. Personnel survivability (force protection) must be addressed for systems under LFT&E oversight and should be integrated into the overall system evaluation of survivability.

Best Practices

The evaluation of survivability for many combat systems can be subdivided into assessment of susceptibility (probability of hit), vulnerability (probability of kill given a hit), force protection (measures or features to protect occupants), and recoverability as shown in Figure 1.

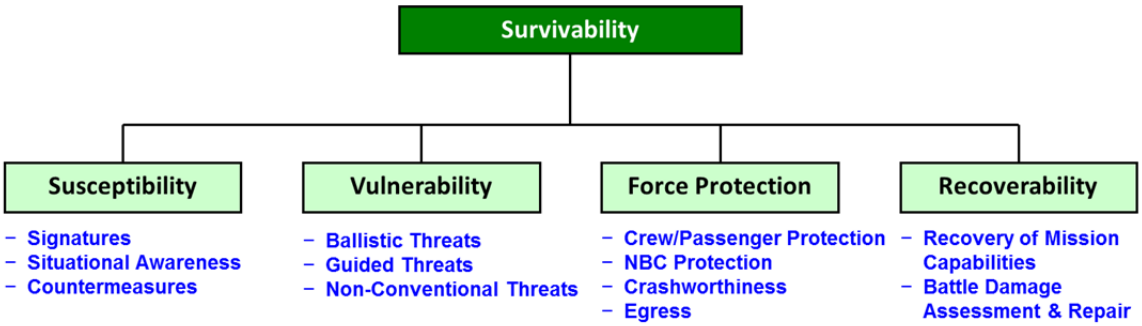


Figure 1. Survivability Evaluation Structure Example

An integrated survivability test strategy might include several operational scenarios or mission threads that guide the design of developmental testing of countermeasure systems, signature measurement, live fire testing of ballistic tolerance, vulnerable area analyses, and force protection assessments. The operational test might use real-time casualty assessment instrumentation to adjudicate force-on-force engagements and generate data on system-level survivability. The probabilities of kill given a hit built into the real-time casualty assessment should have a basis in LFT&E vulnerability assessments. Similarly, the shot lines and end game conditions investigated in LFT&E should have an identifiable basis in realistic threat engagement scenarios, such as those considered in OT&E. DT&E and OT&E testing might provide data on signatures, countermeasure performance, and tactics for use in LFT&E modeling and simulation of force protection analyses. An example overview of the various elements of survivability evaluation is provided below.

Integrated Survivability Assessment – Guidance

Susceptibility-required
Vulnerability-required
Susceptibility- and Vulnerability-required

	Survivability			
	DT	M&S	OT	Live Fire Tests
Aircraft Signatures (RF, IR, Visual)		Susceptibility- and Vulnerability-required		
Mission Planning System Effectiveness			Susceptibility-required	
Off-Board Sensor Performance	Susceptibility-required			
RWR Performance	Susceptibility-required			
LAIRCM Declaration Performance	Susceptibility- and Vulnerability-required			
Aircraft Performance	Susceptibility-required			
TSAS Sensor Fusion Performance	Susceptibility-required			
Situational Awareness (OODA-loop)	Susceptibility-required		Susceptibility-required	
Aircrew TTPs			Susceptibility-required	
RF Threat miss distance		Susceptibility- and Vulnerability-required		
IR Threat Fly-out and Hit Points (HITL)	Susceptibility- and Vulnerability-required	Susceptibility- and Vulnerability-required		
OBIGGS Performance	Vulnerability-required			
Threat Tolerance (Vulnerability)		Vulnerability-required		Vulnerability-required
Force Protection	Vulnerability-required	Vulnerability-required		Vulnerability-required
Repairability				Vulnerability-required
Non-Conventional Threat Tolerance	Vulnerability-required	Vulnerability-required		

Weapons Effectiveness Data

To facilitate integrated survivability analyses and testing, each LFT&E oversight weapons program shall provide weapons effectiveness data to DOT&E for use in the Joint Munition Effectiveness Manuals. LFT&E oversight programs shall provide the data before the weapon achieves initial operational capability and shall prepare the data in coordination with the Joint Technical Coordinating Group for Munition Effectiveness.

Additional Guidance

Force Protection Guidance

References

LFT&E Statute: 10 USC 2366

Defense Acquisition Guidebook, Chapter 9

Integrated Testing – Guidance

Guidance

DOT&E and AT&L directives require the seamless integration of developmental and operational testing throughout the life cycle of a system under test. In their joint memo of 25 April 2008 DOT&E and AT&L defined integrated testing as follows:

"Integrated testing is the collaborative planning and collaborative execution of test phases and events to provide shared data in support of independent analysis, evaluation and reporting by all stakeholders particularly the developmental (both contractor and government) and operational test and evaluation communities."

Background

If planned and executed appropriately, integrated testing allows for a faster and more cost efficient T&E process that ultimately provides the Services with capable systems sooner and at a reduced cost as compared to sequential testing. As noted by DOT&E on 24 November 2009, integrated testing will never do away with the need for a dedicated operational test to confirm that systems will work in combat. The statutory requirements (USC 139, USC 2399) for dedicated operational testing are also clear. Fielding of past acquisition systems have been needlessly delayed by heel-to-toe segregation of developmental and operational testing. Such inefficient processes have been criticized by government studies.

Generally, technical performance measures that need only simple validation are candidates for measurement during integrated testing. If the data from these tests are to be used for the operational evaluation, the components and systems must be production representative. Measures that should be measured in dedicated OT&E include mission-dependent capabilities, CONOPS-related functions, scenario-dependent outcomes, and end-to-end or system-of-systems interactions or effects.

Integrated testing of production representative test articles may come about in two ways: (1) a developmental test incorporates characteristics of operational testing into the test, or (2) the data from developmental testing is accepted as adequate for the operational evaluation. The latter type of integrated test requires that the metrics being measured be equally valid under the conditions for developmental and operational testing. The description of plans for integrated testing should be documented throughout the TEMP. The relevant paragraphs are:

- **Paragraph 3.1, T&E Strategy of the TEMP:** The overarching T&E strategy section of the TEMP should address the conditions for integration of DT and OT testing, if planned.
- **Paragraph 3.2.1, Mission Oriented Approach Section of the TEMP:** Discuss when and how developmental testing will reflect the expected operational environment. This will help integrate developmental testing with operational testing.

- **Paragraph 3.2.3, Developmental Test Events Section of the TEMP:** Describe how selected developmental test events will reflect the expected operational environment. This will help integrate developmental testing with operational testing.
- **Paragraph 3.4, Operational Evaluation Approach Section of the TEMP:** Summarize integrated testing strategy to include: (1) developmental test data that will be used for operational evaluation and (2) conditions for data pedigree and test conduct that will make data suitable for use in the operational evaluation.

Best Practices

Measurement of cargo and storage capacity on a ship or aircraft does not depend on the conditions of the test. A developmental test that measures the capacity should provide adequate data for an operational evaluation. Developmental tests that are often used to provide this data include Marine Corps Certification Exercises and Navy In-service Inspections.

Integrated testing of production representative components and systems can be useful to measure satellite surveillance performance, communications range and throughput, storage capacity, network backup and restore capabilities, vehicle or sensor performance, weapons accuracy, countermeasure performance, bandwidth, sensitivity, user load simulations, or satellite autonomous operations.

For systems with high reliability requirement thresholds that would be impracticable to test fully in OT&E, it is sometimes possible to include data from operationally realistic integrated testing. This might be done for the whole system or for the important subsystems. In such cases, the TEMP should include rationale, applicability, and any limitations for including integrated test data in the evaluation of reliability. For systems adopting this approach, data from environmental testing (e.g. thermal, vacuum, vibration, rain, ice, sand, etc.) can assist with assessments of long-term reliability when combined with historical data and appropriate caveats.

Air Warfare Ship Self-Defense test events, particularly those conducted on the remotely controlled Self-Defense Test Ship,[1] are good examples of integrated tests where a developmental test is executed under conditions that are sufficiently operationally realistic. During Self-Defense Test Ship events, aerial targets are flown directly at the test ship. The combat system elements of the ship are operated by civilian experts via remote control. As a developmental test platform, the test ship provides a highly controlled environment for testing specific system metrics. By ensuring that the aerial targets are representative of actual anti-ship cruise missile threats and that the flight profile of the target is the same as the threat, the developmental test can be used as an integrated test.

---

1 The Self-Defense Test Ship is a former Spruance Class Destroyer that has been equipped with multiple modern-day anti-air warfare combat systems. The ship and its combat systems are both capable of being operated by remote control, thereby reducing the risk of mishap when engaging anti-ship cruise missiles and aerial targets.

IOT&E Entrance Criteria – Guidance

Guidance

The purpose of IOT&E Entrance Criteria is to ensure that the system under test is ready to commence IOT&E and the required resources are in place to support the test. The intent of this requirement is to ensure that systems do not enter IOT&E before they are sufficiently mature. Premature commencement of IOT&E could result in suspension or early termination because of technical problems that should have been resolved prior to the start of IOT&E. Suspension or early termination of IOT&E will result in an inadequate test and unnecessary waste of resources.

Best Practices for IOT&E Entrance Criteria:

- The system has demonstrated acceptable hardware and software performance during mission-focused DT conducted in operationally realistic environments with the hardware and software to be used in IOT&E.
- IOT&E test articles are production representative (as determined by DOT&E).
- Adequate reliability data are available to estimate the reliability of the system under test and the expected IOT&E reliability results.
- Threat surrogates and targets have been validated and approved by the DOT&E.
- All critical issues identified in developmental testing have been resolved or have an acceptable work-around.
- The required test ranges are ready to support all planned events as described in the IOT&E plan, including environmental, safety, and occupational health requirements.
- All required certifications and accreditations are in place.
- The manning for the system is consistent with Concept of Operations and training has been completed consistent with that planned for intended users.
- Pre-IOT&E M&S predictions are based on verified, validated, and accredited modeling and simulation.
- If DT data is required to support the evaluation, the required DT data have been provided to the OTA and DOT&E.
- The logistics system and maintenance manuals intended for use with the fielded system are in place for IOT&E.
- DOT&E has approved the Service-provided IOT&E plan.

References

Defense Acquisition Guidebook DoDI 5000.02, 7 January 2015

Examples

IOT&E Entrance Criteria – Examples

Example 1

3.3 Certification for IOT&E The Component Acquisition Executive (CAE) will evaluate and determine system readiness for Initial Operational Test and Evaluation (IOT&E). Prior to the CAE's determination of readiness for IOT&E, an independent Assessment of Operational Test Readiness will be conducted by OUSD (AT&L). It shall consider the risks associated with the system's ability to meet operational suitability and effectiveness goals and will be based on capabilities demonstrated during DT&E and OAs, as well as on the criteria described in this TEMP. The final report for DT will provide insight into the system's readiness for IOT&E.

3.3.1 DT&E Information Required Adequate test data will be collected during DT-IIG and DT-IIH to allow the Program Manager to assess and report the system's capabilities against the stated COIs using the MOE/MOS listed in this TEMP prior to IOT&E.

3.3.2 IOT&E Entrance Criteria

- All Milestone C exit criteria have been met.
- DOT&E has approved the IOT&E Test Plan.
- System is projected to meet or exceed the Mean Time Between System Abort threshold during IOT&E.
- Department of the Navy Criteria for Certification listed in Secretary of the Navy Instruction 5000.02 of December 8, 2008 have been satisfied and the system is certified for test.
- All deficiencies identified in previous testing have been resolved.
- All required targets have been accredited and the test range has been adequately surveyed.
- Production representative test articles are available to conduct IOT&E.
- Adversarial cyber security test team has been identified and is funded for testing.
- OTRR is completed and DOT&E concurs with proceeding to test.

Example 2

Table 3.3 Dakota Helicopter IOT&E Entrance Criteria

ENTRANCE CRITERIA	ASSESSMENT METHODOLOGY
Maneuver Flight Performance Hover Out-of-ground Effect (HOGE): with 3400-pound payload Range: 250 nautical miles Endurance: 2 hours 40 minutes	Characterize hover, speed, range, and endurance performance in developmental flight testing. Estimate aircraft performance at threshold atmospheric conditions (6,000 feet pressure altitude, 95 degrees Fahrenheit) through analysis.
Reliability Point estimate for system reliability for Mean Time Between Essential Maintenance Actions (MTBEMA) must be greater than 2.3 hours	Demonstrate in developmental flight testing. Limited User Test result was 2.6 hours MTBEMA.
Survivability 30-Minute Continued Safe Operation following a single hit by XXX Armor Piercing Incendiary Projectile. (Classified Requirement) Vulnerability Area for main rotor drive components and rotor blade damage size should not exceed XXX. (Classified Requirement)	Review LFT&E data and Service LFT&E report
New Mission Capability Demonstrate Remote Control of Unmanned Aircraft sensor	Demonstrate remote control in developmental testing with both aircraft in flight at operational ranges.
Software Maturity No Priority 1 or 2 software problem reports	Review developmental test data and reports.
Certifications of IOT&E aircraft by appropriate agencies.	Air Worthiness and Safety Release for flight operations with typical aircrew.
Successful completion of OTRR.	T&E WIPT Concurrence

LFT&E Strategy - Guidance

1.3 System Description

Identify survivability or lethality improvement related features, if applicable.

1.3.7 Previous LFT&E Testing

Any data sources that address Live Fire and can be mapped to the program's LFT&E critical issues should be listed, including data from other programs and contractor tests to the extent possible. Include rationale for the applicability of those data to the DT, OT, and LFT&E program.

2.1.1 T&E Organizational Construct

Among the stakeholder T&E organizations identified in this paragraph, include the LFT&E IPT and their specific responsibilities (such as test planning, provision of test articles, test support, data collection, reporting). Recommend that the user/operator participate in the LFT&E IPT. Provide sufficient information to adequately understand the functional relationships.

2.5 Integrated Test Program Schedule

Include the component- and system-level LFT&E test events in the Integrated Test Program Schedule. For Navy ship and submarine programs, the schedule should also include Component Shock Qualification for Contractor and Government Furnished Equipment (CFE/GFE). See Figure 2.1.

3.6 Live Fire Test and Evaluation Approach

TEMPs for systems covered by the LFT&E statute (Title 10 U.S.C. § 2366) must have a LFT&E strategy that supports a lethality/vulnerability/susceptibility evaluation of the munition/platform. If the LFT&E strategy is completed, it may be included in paragraph 3.6 or as an attachment. Prior to Milestone B, a Live Fire Plan must be developed for proper resourcing. Paragraph 3.6 of the TEMP should provide an overview of the system & Live Fire process, purpose of LFT&E, improvements/upgrades relevant to LFT&E, system description/variants, and pertinent background information. Some programs might decide to attach a LFT&E strategy to the TEMP if the strategy is detailed, classified, or not yet completed. Whether there is a LFT&E attachment or not, paragraph 3.6 of the TEMP should provide a LFT&E summary with the elements described below.

Discuss the scope of live fire testing, including design of experiment considerations, phases and building block approaches, pass/fail or scoring criteria, and evaluation methodology. The evaluation plan should be constructed so that vulnerability results are assessed in the context of overall system survivability and personnel survival. Discuss the authority of the Live Fire integrated product team in the selection process. Note: Typically, this section is substantial and is one of the prime areas for discussion and negotiation. It might be appropriate to put these details in a LFT&E appendix to the TEMP. This section should clearly articulate test and evaluation objectives, describe the testing, M&S, and engineering analysis required to support those objectives.

This section should include:

- A description of data requirements [metrics] to address the critical issues and to support the overall lethality/survivability and force protection evaluation
- Justification for the required test scope. Either DOE methods to justify the test design or other methods to assess the criticality of the required data to complete the evaluation. These details would justify the required test program and the accepted risk.
- A description of tests and identification of test ranges needed to generate the required data.
- A description of specific data analyses and M&S to support the overall evaluation. Describe any unique evaluation methodologies associated with each issue). Describe the required data needed for the M&S and how the data will be folded into final evaluation. M&S Section of the LFT&E Strategy could include these details.

See Integrated Survivability Assessment and Force Protection for additional guidance on LFT&E strategy approaches.

3.6.1 Live Fire Test Objectives

3.6.1.1 Critical LFT&E Issues

Develop Critical LFT&E Issues (CLI) in the form of questions that will be addressed. Consider the following generic CLIs and modify them as needed for the system under consideration:

- Susceptibility - What is the system susceptibility to threat weapons; what are the likely hit points to the platform from the threats selected for analysis?
- Vulnerability - What are the weapons effects and resulting platform degradation caused by the selected threats?

- Force Protection - What are the number and type of crew casualties resulting from the selected threats?
- Force Protection - Can the crew evacuate themselves and injured personnel from damaged compartments/areas of the platform?
- Recoverability - How will the crew limit the spread of secondary damage, restore ship's capabilities and systems, reconstitute mission, and treat casualties following damage from a weapon on the threat list?

See Aircraft System Example, Ground Combat System Example, or Ground Tactical System Example.

3.6.1.2 Lethality/Vulnerability Requirements

Summarize any requirements, specifications, or desired capabilities that are relevant to the LFT&E strategy, including (for vulnerability programs) any KKPs that address force protection or survivability against asymmetric threats. A target/threat matrix table should also be included in the Live Fire strategy and updated as required. The strategy should address all expected targets/threats, regardless of whether or not they are explicitly identified in the requirements. The System Threat Assessment Report (STAR) can be used to identify the targets/threats that will be addressed. (Example target/threat matrix)

3.6.1.3 Schedule, Funding and Resources

Identify schedule, funding and resources (targets/assets) pertaining to LFT&E. Include arena, coupon/component, exploitation/ballistic hull, or sled testing, along with the breakout of all integrated DT/OT tests that support LFT&E. Also include test ranges, targets, modeling, test/evaluation plan preparation, pre-shot predictions and reporting for each test phase. For Navy ship and submarine programs, include Component Shock Qualification for Contractor and Government Furnished Equipment (CFE/GFE). At a minimum, by Milestone B, there should be a Live Fire Plan that can be used for resourcing. See Adequate Test Resources Guidance and Examples.

3.6.1.4 Document Approval Matrix

Include a table of pertinent Live Fire documents, including pre-shot predictions, analysis/evaluation plans, test plans, and M&S VV&A documentation. The table should list who is responsible for originating/reviewing/signing each document. See Test Planning Documents Guidance and Examples.

3.6.2 Modeling & Simulation (M&S)

Identify whether M&S will be used to support test planning, pre-test prediction, and/or an evaluation and the M&S tools to be used. Indicate the anticipated inputs (test data) needed by the model(s), and the types of output expected to be provided to support test planning, pre-test prediction, and/or evaluation. If multiple models will be used, the overall M&S "flow" should be described (e.g., where the output of one model will be required as input for another). Discuss means of verification, validation and accreditation for models used and organizational responsibility. See M&S for T&E Guidance and M&S LFT&E Examples.

3.6.3 Test Limitations

List any test limitations and mitigations. See Test Limitations Guidance and LFT&E Test Limitation Examples.

Critical LFT&E Issues – Aircraft Example

3.6.1.1 Critical LFT&E Issues.

Issue Number	Critical LFT&E Issue	Evaluation Method PD: Prior Data: test, modeling, or combat EA/LF: Engineering Analysis MS/LF: Modeling & Simulation/LFT&E T: Testing D: Developmental Testing O: Operational Testing L: Live Fire Testing			
		PD	EA/LF	MS/LF	T
3.2	Susceptibility				
3.2.1	Take-off and Landing (MANPADS)				
3.2.1.1	Threat capabilities to target and hit JSTARS on take-off departure and landing approach.	TTPs, TOs		MOSAIC	
3.2.1.2	Effectiveness of susceptibility-reduction, take-off and landing procedures, and airfield protection TTPs in reducing threat engagements.	TTPs, TOs		MOSAIC or HITL	
3.2.1.3	Threat hit points on JSTARS in successful engagements.	TTPs, TOs		MOSAIC or HITL	
3.2.2	Mid-mission				
3.2.2.1	Capabilities of expected kinetic threats to detect, target, engage and hit JSTARS Recap across the range of likely mission sets.	Intel	X		
3.2.2.2	Capabilities of JSTARS Recap with supporting assets to avoid or escape threat engagements.	CONOPS, TOs, TTPs		X	OT
3.2.2.3	Effectiveness of JSTARS supporting assets in identifying threats to JSTARS, providing timely threat warning and directing other supporting assets to intercede.	CONOPS, TOs, TTPs		X	OT
3.2.2.4	Effectiveness of mission planning in preventing JSTARS from being engaged by threats on independent missions	CONOPS, TTPs	X		OT
3.2.2.5	Effectiveness of Broadcast Intelligence in providing JSTARS with real time threat information on independent missions	CONOPS, TTPs	X		OT
3.3	Vulnerability				
3.3.1	Direct threat induced damage.	X	X	X	
3.3.2	Major airframe structural component damage (wings, fuselage, empennage).	X		X	LF

Critical LFT&E Issues – Aircraft Example

Issue Number	Critical LFT&E Issue	Evaluation Method PD: Prior Data: test, modeling, or combat EA/LF: Engineering Analysis MS/LF: Modeling & Simulation/LFT&E T: Testing D: Developmental Testing O: Operational Testing L: Live Fire Testing			
		PD	EA/LF	MS/LF	T
3.3.3	Fuel system:				
3.3.3.1	Fuel tank damage.	x		x	
3.3.3.2	Fuel tank hydrodynamic-ram.	x		x	
3.3.3.3	Fuel tank ullage fire and explosion.	x		x	
3.3.3.4	Fuel tank dry bay fire.	x		x	
3.3.3.5	Fuel line damage, including aerial refueling lines	x		x	
3.3.3.6	Fuel starvation	x	x		
3.3.4	Propulsion system:				
3.3.4.1	Engine damage	x		x	
3.3.4.2	Uncontained engine debris damage.	x		x	LF
3.3.4.3	Engine nacelle damage (fuel & hydraulic lines).	x		x	
3.3.5	Other flight critical systems:				
3.3.5.1	Flight controls and flight control surfaces.	x		x	LF
3.3.5.2	Hydraulic systems (leak and fire).	x		x	LF
3.3.5.3	Avionics/electronic systems.	x		x	
3.3.5.4	Auxiliary power unit systems.		x		LF
3.3.6	Vulnerabilities associated with cascading damage to non-flight critical systems.				
3.3.6.1	Mission avionics/electronic systems.	x		x	
3.3.6.2	Installed and carry-on oxygen systems	x		x	
3.3.7	Nuclear, biological, and chemical (NBC) threat vulnerabilities				

Critical LFT&E Issues – Aircraft Example

Issue Number	Critical LFT&E Issue	Evaluation Method PD: Prior Data: test, modeling, or combat EA/LF: Engineering Analysis MS/LF: Modeling & Simulation/LFT&E T: Testing D: Developmental Testing O: Operational Testing L: Live Fire Testing			
		PD	EA/LF	MS/LF	T
3.3.7.1	Effectiveness of personnel protective gear in protecting crewmembers while allowing operational functions	x			DT
3.3.8	Cyber threat vulnerabilities		DT, OT		
3.3.9	Low-power laser threat vulnerabilities				
3.3.9.1	Crewmembers	x		x	
3.3.9.2	Sensor systems	x		x	
3.3.9.3	Effectiveness of crew protection systems	x	x		
3.3.10	Electromagnetic Pulse (EMP) vulnerabilities				
3.4	Force Protection				
3.4.1	Casualties due to direct exposure to threats	x	x		
3.4.2	Casualties due to loss-of-aircraft events	x	x		
3.5	Recoverability				
	None				

Critical LFT&E Issues – Ground Combat System Example

3.6.1.1 Critical LFT&E Issues.

Critical LFT&E Issue		Evaluation Strategy	Data Source				
			Existing Data	PIM LFT	BDAR/R	M&S	Eng. Analysis
1	What is the vulnerability of the combat loaded Paladin Integrated Management (PIM) and crew to the spectrum of current (Initial Operational Capability [IOC]) and future (IOC+10) threats?	Use all test data and modeling and simulation (M&S). Use engineering judgment to evaluate any synergistic effects that contribute to vehicle or crew vulnerability.	x	x	x	x	x
1.1	What are the major causes of crew and passenger casualties and incapacitation?	Use all test data and M&S. Use engineering judgment to evaluate any synergistic effects that contribute to vehicle or crew vulnerability.		x		x	x
1.2	How do stowed ammunition, supplies, and onboard equipment contribute to vulnerability?	Conduct Full-Up System Level (FUSL) testing. Use engineering judgment to supplement M&S and test data to evaluate.	x	x		x	x
1.3	How do vehicle subsystems contribute to vulnerability?	Conduct component testing, fuel subsystem testing, and FUSL test events. Use engineering judgment to supplement M&S and test data to evaluate.		x		x	x
1.4	How well does the PIM meet ballistic requirements?	(See Attachment 3 for threats corresponding to ballistic requirements.)	x	x		x	x
1.5	What is the penetration resistance of the armors?	(See Attachment 3 for direct fire, indirect fire, and improvised explosive device [IED] threats.)	x	x			

Critical LFT&E Issues – Ground Combat System Example

Critical LFT&E Issue		Evaluation Strategy	Data Source				
			Existing Data	PIM LFT	BDAR/R	M&S	Eng. Analysis
1.6	What are the behind armor debris characteristics following penetration?	(See Attachment 3 for overmatching threats.)		x		x	
1.7	What are the ballistic shock vulnerabilities of PIM components?	Conduct component testing and FUSL test events. Use engineering judgment to supplement test data to evaluate.		x			x
2	What components are mission critical? What is the vulnerability of these components and how do they impact the mission accomplishment?	Conduct component testing and FUSL test events. Conduct battle damage assessment and repair/recovery (BDAR/R) following FUSL test events. Use engineering judgment, Mission-Based Test and Evaluation (MBT&E), and M&S to supplement BDAR/R to evaluate.		x	x	x	x
3	Are there unexpected vulnerabilities or unexpected levels of vulnerabilities?	Use all test data and M&S. Use engineering judgment to supplement M&S and test data to evaluate.		x		x	x
3.1	What is the operational significance of the unexpected vulnerabilities?	Conduct BDAR/R following FUSL test events. Use engineering judgment and MBT&E to supplement BDAR/R to evaluate.			x		x
3.2	How can these vulnerabilities be reduced?	Conduct BDAR/R following FUSL test events. Use engineering judgment to supplement BDAR/R to evaluate. Use M&S to make recommendations for vulnerability reductions			x	x	x

Critical LFT&E Issues – Ground Combat System Example

Critical LFT&E Issue		Evaluation Strategy	Data Source				
			Existing Data	PIM LFT	BDAR/R	M&S	Eng. Analysis
4	What are the planned vulnerability reduction measures and how do they contribute to vehicle or crew survivability?	Conduct BDAR/R following FUSL test events. Use engineering judgment to supplement BDAR/R to evaluate. Use M&S to make recommendations for vulnerability reductions		x	x	x	x
5	How effective is BDAR/R in restoring the vehicle to functional combat capability and in recovering damaged vehicles following an attack?	Conduct BDAR/R following FUSL test events. Assessment by BDAR/R team.			x		
5.1	What design features facilitate or inhibit troubleshooting, repair or recovery?	Conduct BDAR/R following FUSL test events. Assessment by BDAR/R team.			x		
5.2	How effective and reliable are built-in diagnostic capabilities or the Vehicle Health Management System (VHMS) in supporting the BDAR process (if equipped)?	Conduct BDAR/R following FUSL test events. Assessment by BDAR/R team.			x		
5.3	Are BDAR manuals available and adequate?	Assessment by BDAR/R team.			x		
5.4	Is BDAR training, doctrine and provisioning adequate to facilitate the repair of battle damage vehicles?	Conduct BDAR/R following FUSL test events. Assessment by BDAR/R team.			x		
5.5	Does vehicle design allow expedient and safe recovery with existing recovery equipment?	Conduct BDAR/R following FUSL test events. Assessment by BDAR/R team.			x		

3.6.1.1 Critical LFT&E Issues.

Critical LFT&E Issues		Evaluation Strategy	Data Sources					
			Existing Data	LFT	BDAR	M&S	EA	MBT&E
CI 1. What are the expected/unexpected vulnerabilities of the crew/occupants of the combat		Utilize all test data and M&S. Use engineering judgment to evaluate any synergistic effects that	x	x	x	x	x	
1.1	What are the major causes of crew and passenger casualties	Utilize all test data and M&S. Use engineering judgment to evaluate any	x	x		x	x	
1.2	Does the JLTV meet Force Protection	(See Attachment 7 for threats corresponding to	x	x		x		
1.3	To what levels do the opaque and transparent	(See Attachment 7 for direct fire, indirect fire, and	x	x				
	What are the Behind Armor Debris (BAD)	(See Attachment 7 for overmatching threats.)	x	x		x		
1.4	To what levels do vulnerabilities affect the mission capabilities?	Conduct BDAR/R following FUSL test events. Use engineering judgment and		x	x		x	x
1.5	What are the potential vulnerability reductions?	Conduct BDAR/R following FUSL test events. Use engineering judgment to supplement BDAR/R to		x	x	x	x	
CI 2. What subsystems contribute, both directly and indirectly, to crew/occupant		Conduct FUSL test events. Use M&S and engineering judgment to evaluate.	x	x		x	x	
2.1	To what level do stowed ammunition or other energetics (e.g., Lithium Ion batteries), supplies,	Conduct FUSL test events. Use M&S and engineering judgment to evaluate.		x		x	x	x
2.2	To what level are mobility, firepower, and communication retained	Conduct FUSL test events. Conduct BDAR/R following FUSL test events. Use		x	x	x	x	x
2.3	To what level are crew/occupants able to ingress and egress following a ballistic or	Conduct IED events against ballistic cabs. Conduct FUSL test events. Conduct BDAR/R following		x	x		x	

Critical LFT&E Issues – Ground Tactical System Example

2.4	To what level are the Automatic Fire Extinguishing System (AFES) and other fire mitigation technologies effective?	Test the AFES effectiveness using a fireball generator. Conduct FUSL test events with threat focused on the fuel tank. Should any test event result in a fire, instrumentation will capture	x	x	x			
CI 3. To what level does Battle Damage Assessment and Repair/Recovery		Conduct BDAR/R following FUSL test events.			x			
3.1	What design features facilitate or inhibit	Conduct BDAR/R following FUSL test events.			x			
3.2	To what level are BDAR/R manuals and	Conduct BDAR/R following FUSL test events.			x			
3.3	To what level are the built-in diagnostic capabilities to support	Conduct BDAR/R following FUSL test events.			x			
3.4	To what level does the vehicle design allow expedient and safe recovery with existing recovery equipment and like-vehicle recovery?	Conduct BDAR/R following FUSL test events.			x			

LFT&E Threat/Target Matrix - Examples

Example 1 – Ground Vehicle Vulnerability LFT&E Threat Matrix

Threat System	Munition
Indirect fire	Dual Purpose Improved Conventional Munitions 152mm HE fragmenting artillery 120mm and 82mm HE fragmenting Mortars Smart munitions (EFP & hit-to-kill Terminally Guided Submunition Rockets
Mines / IEDs	IEDs (as improvised conventional ordinance) Explosively Formed Penetrators Anti-tank (various), Anti-personnel (various), Scatterable
Direct Fire	Rocket Propelled Grenades (RPG) (unitary, tandem, and thermobaric) Small arms (5.56mm and 7.62mm) including armor piercing (AP) and non-AP Heavy machine guns (12.7mm, 14.5mm) Sniper and anti-materiel rifles (12.7mm, 14.5mm, 20mm) Anti-armor and blast hand grenades Thermobaric/Flame weapons
Light-armor-fired	30mm AP and HEI
Aircraft-fired (fixed wing and rotary wing)	Projectiles, Rockets, Missiles
Tank-fired	Kinetic Energy, Chemical Energy, Anti Tank Guided Munitions

Example 2 – Munition Lethality LFT&E Threat Matrix

Threat Category	Target
Hard	Communications Facility (reinforced concrete) Aircraft Bunker (typical SWA theater)
Industrial	POL Refinery (one fractionating unit) POL, Large Partially Underground Tank Specialized Repair Complex Transformer
Soft Surface	SATCOM Antenna EW/GCI Radar GRILL PAN Radar FLOGGER fighter aircraft (revetted) SCUD Missile (on launcher) BM-21 Rocket Launcher
Lightly Armored Ground Combat System	152mm Towed Field Gun/Howitzer (stationary) Anti-Aircraft Artillery (stationary)

M&S for Test and Evaluation - Guidance

The Modeling and Simulation (M&S) sections of the TEMP should address how M&S will be employed in the overall test strategy and how the M&S will be verified, validated, and accredited (VV&A). Specifically, the TEMP should list any M&S expected to be used, the intended use, the data requirements, the test objectives to be addressed and/or how test scenarios will be supplemented with M&S, the planned VV&A effort, and who will conduct the VV&A effort (DoDI 5000.61). The TEMP should list any specific test events required for VV&A of the M&S. The resources for the VV&A test events should be included in Part IV.

DOT&E requires all OT&E and LFT&E test agencies to accredit models used to support OT&E and LFT&E. The accrediting test agency will establish the acceptability criteria for M&S use, and the accreditation must be based on a verification and validation approach that is tailored for the specific intended use of the model or simulation. This means that the OTA will conduct their own assessment to accredit M&S for their use in OT. DOT&E must review and concur with the OTA's accreditation plan before the plan is executed.

Testers should focus the validation on the full system or environment being evaluated in addition to validating sub-components of the model. While validating every subcomponent that comprises the full system is ideal, it is more important to ensure that the integration of all the components and the environment together adequately represents the real-world system relevant to the intended use of the model.

Before using M&S results to support an operational test or live fire evaluation, it is essential to understand and characterize the usefulness and limitations of the M&S capability. Thus, any M&S used to support OT&E should not be accredited until a rigorous comparison of live data to the model's predictions is done (if possible), and those predictions are found to have replicated live results with sufficient accuracy for the intended evaluation in the intended domain. In addition to these direct quantitative comparisons between live data and M&S output, a comprehensive strategy should also assess M&S output across the entire operational domain for which the M&S will be accredited.

Validation of M&S should include the same rigorous statistical and analytical principles that are used to design live tests. The principles and techniques that comprise statistical test design methodologies, including design of experiments and other formal statistical tests, should be employed as part of the process of determining what and how much live data are needed for a live versus simulation comparison, and in the process of determining how well the models/simulations reflect reality. Empirical models should be used to understand M&S outcomes across the operational space and assist in the uncertainty quantification in areas where there are no live data. All results, but especially those in the operational space where no live data is available, should be discussed in the context of limitations.

Statistical techniques (e.g. hypothesis testing and regression techniques) should also be used to rigorously compare live data to simulated output. This methodology can and should be supported by

other elements of validation, including face validation, documentation review, SME evaluation, and comparison to other models. If there are extraordinary circumstances prohibiting these statistical principles from being used, the reasons why they are not used must be clearly articulated and alternative approaches used to justify validation and accreditation must be cogently explained.

As stated in the March 14, 2016 and January 17, 2017 memorandums, a discussion of the following elements is required:

- Quantitative mission-focused measures
- The range of conditions over which the M&S will be validated
- The plan for collecting the necessary live and simulation data for M&S validation, including both the points to be used for direct comparison, and those used to evaluate the entire M&S space of interest
- An analysis of statistical risk
- The validation methodology

This information can be presented directly in the TEMP, or the TEMP can reference other relevant documents, such as validation or accreditation plans. If the necessary detail is not yet available while the TEMP is being drafted, the information may be conveyed to DOT&E via a standalone M&S concept briefing (similar to a Test Concept) or in an Operational Test Plan, so long as the overall strategy is socialized with DOT&E as early as possible.

In addition to applying rigorous statistical and analytical principles, some other important criteria for M&S accreditation are:

- Documentation which summarizes the purpose, development background, assumptions, and application domains and provides a complete and accurate description of M&S capabilities and limitations.
- Sound approaches for M&S capability acquisition, validation, and use. M&S capabilities used for T&E should be planned and resourced early. The M&S capabilities to be used, the T&E aspects of the system evaluation that these M&S capabilities will address, and the approach for assessing credibility of these models and simulations should all be described in the TEMP.

Establishing M&S Credibility for T&E: Additional Details and Definitions

Under DoDI 5000.61, each M&S capability must complete a verification, validation, and accreditation (VV&A) process to establish its credibility for a specific intended use. Some M&S capabilities associated with T&E have special validation requirements. For example, to validate that a non-US forces or threat weapon is appropriately represented in a model, the Director, Defense Intelligence Agency is the final validation authority for oversight systems. DOT&E, through the T&E Threat Resource Activity (TETRA), is the approval authority for threat representation validation reports used for T&E. OTAs accredit threat representation models for use in OT. The Defense Acquisition Guidebook, Section 9.7.3, Validation of Threat Representations (targets,

threat simulators, or M&S) provides guidance and references on validating M&S capabilities associated with threats and targets.

Existing M&S capabilities previously accredited for other applications must complete another VV&A process and be accredited for each new intended use. However, previous VV&A may simplify the process because the previous efforts have been documented and the new VV&A effort typically can focus on the changes.

Verification determines whether the M&S accurately represents the developer's specifications. The M&S is expected to add two numbers; does it add two numbers? Validation determines whether the model is an accurate representation of specific aspects of the real world or threat system. The M&S is expected to add two numbers; does it provide the correct sum? Accreditation is the official certification that the M&S and its associated data are acceptable for an intended use.

For accreditation, the intended use is important because an M&S capability useful in one application may not be useful in another due to limitations inherent in the M&S capability, existing validation data, or a prior VV&A process. The accreditation will explicitly state the intended use, such as: "The Big Weapon Model will be used to estimate the miss distance between the weapon and the target in support of developmental test DT-II." It also should acknowledge any significant limitations: "The Big Weapon Model does not include threat countermeasures, and consequently all scenarios are simulated in a clear environment."

The scope of the accreditation effort and VV&A process are functions of how each M&S capability will be used. For example, high level or conceptual models are often used early in a program (e.g., a spreadsheet model used to estimate system performance) that require limited data for validation and accreditation. Frequently, M&S capabilities used in prior similar programs can be used and pre-existing VV&A artifacts and analysis can simplify or streamline the VV&A process for the new application. At the other extreme are high-fidelity models an evaluator might use to assess measures of effectiveness, suitability, or survivability; these must undergo a rigorous VV&A process. In general, the more important the M&S results are to the final evaluation, the more rigorous the VV&A process must be. Wherever possible, design of experiments techniques should be leveraged to ensure that test data supporting the VV&A clearly defines the performance envelope of the model or simulation, and corresponding statistical analysis techniques should be employed to analyze the data and identify factors that influence the validity of the M&S.

Some common pitfalls in using M&S for T&E that should be avoided are:

- Faulty assumptions in developing or using M&S such as assuming independence between events that actually have some type of dependency or relationship.
- Using M&S results outside their validation domain which are uncharacterized and include unknown uncertainties.

- Improper use of data for M&S development or validation such as relying solely on heart-of-the-envelope performance data or using specification values instead of actual performance data when the latter is available.
- Averaging validation results across conditions rather than discussing where the M&S is valid and where it isn't.

References

DODI 5000.61: DoD Modeling and Simulation (M&S) Verification, Validation, and Accreditation (VV&A), 9 December 2009

Guidance on the Validation of Models and Simulation used in Operational Test and Live Fire Assessments, 14 March 2016

Clarifications on Guidance on the Validation of Models and Simulations used in Operational Test and Live Fire Assessments, 17 January 2017

Mission Focused Evaluation – Guidance

While the test and evaluation strategy should provide opportunities to determine whether a system meets documented requirements, the ultimate purpose of the test and evaluation strategy is to demonstrate the operational effectiveness, suitability, and survivability of the system in its expected operational environment. Operational effectiveness is defined as the overall ability of the system to support successful mission accomplishment, when used by representative operators in the intended environment. This definition takes into account the interplay of the system under test, the operators, and interrelated or supporting systems. In many cases, the system performance specifications in the requirements document will assist in the assessment of mission accomplishment, but a mission focused evaluation will not be limited to these specifications.

To assist in early identification of system problems that might only be manifest in operational environments, developmental test planners should incorporate elements of the operational environment (typical users and maintainers, realistic operational conditions, representative threat systems, end-to-end missions, production representative test articles, weapons, secure communications gear, survivability equipment, interfacing systems and networks, etc.) into developmental testing whenever possible. However, the injection of operational realism into developmental testing does not obviate the need for operational testing. The purpose for mission-oriented developmental testing is to find and fix problems that are unique to operational environments before the system begins operational testing.

References

Reporting of Operational Test and Evaluation Results, DOT&E, January 6, 2010

Examples

Operational Evaluation Approach Example

Mission Focused Metrics Guidance with Examples

3.4 Operational Evaluation Approach

Evaluation of the XYZ Anti-Submarine Warfare (ASW) system will be completed in realistic at-sea scenarios using a production-representative system. This testing will assess whether the system meets the performance thresholds in the CPD but will primarily focus on the operational effectiveness of the system. The test ship will be tasked to conduct ASW as well as intelligence, surveillance, and reconnaissance (ISR) tactical missions. The ASW test platform will be directed to clear an area with a suspected hostile submarine; the test ship will search for, detect, report, and initiate engagement of hostile submarines up to, but not including launch of live ordnance. The test ship will also be tasked to conduct an ISR mission in a high-density surface contact environment. In both cases, the tasking will provide an element of surprise or uncertainty for the test ship; the test platform commander will be able to respond to the tactical situation as perceived when employing the XYZ system. Successful accomplishment of testing events will support an evaluation of system operational effectiveness, operational suitability, and a recommendation on fleet release of the system.

Quantitative Mission-focused Measures – Guidance

General Guidance

TEMPs should include quantitative mission-focused measures (also referred to as quantitative mission-oriented response variables and can be thought of mathematically as dependent variables) for effectiveness, suitability, and survivability. These measures are key to good test designs; poorly-chosen or poorly-defined measures, even if they are directly connected to Key Performance Parameters (KPPs) or Key System Attributes (KSAs), could result in a poorly designed test, and can lead to test results that are not relevant to the mission effectiveness, suitability, or survivability of the system.

Choosing Quantitative Mission-focused Measures

The selection of quantitative mission-focused measures is a critical part of the test design effort, and should occur as test planning begins. Step 1 is to identify the goal of the test. This should reflect evaluation of end-to-end mission effectiveness, suitability, or survivability in an operationally realistic environment. Once the test goal is known, testers should select appropriate measures of system performance and provide data for addressing the goal of the test. Ideally, the measures will be quantitative, mission oriented, relevant, informative, and not rigidly adhere to the narrowest possible interpretation of definitions in requirements documents. Measures should provide a criterion of mission accomplishment (not technical performance for a single subsystem), lend themselves to good test design (i.e. be continuous in nature), and in general comprehensively cover the reasons for procuring the system.

Although many measures can be used to characterize system performance in a given mission, it is desirable that a small number of quantitative mission-focused measures be identified to be the focus of the evaluation of operational effectiveness, suitability, or survivability and used in concert with statistical test design methodologies. Additional secondary measures are encouraged, and are necessary to characterize other aspects of system performance. For example, for test design, the hit success rate may be identified as the quantitative mission-focused measure, even though other measures are needed to characterize success in the dependent portions of the kill chain (e.g., detection, identification, time to engage, engagement range).

Exceptions to using CDD/CPD-defined Measures

The quantitative mission-focused measure identified for test design need not be the KPPs. Often KPPs are insufficient for measuring the mission effectiveness, suitability, and survivability of the system. See the Inspector General report dated May 15, 2015 for two examples. If the requirements cannot be revised to define those system characteristics most critical for providing an effective military capability, the TEMP must identify and define those characteristics. Examples of quantitative mission-focused measures that enable mission-focused test design include detection/classification range, miss distance, probability of hit, search rate, time to accomplish a successful mission, counter-detection range, and probability of successful intercept.

When testers select these quantitative mission-focused measures, the resultant test design should ensure that adequate data will be collected to accomplish several goals:

- Provide adequate data to evaluate the effective military capability of the system
- Provide a meaningful measure of system performance across the operational envelope
- Provide sufficient data for the secondary measures needed to characterize system performance.

Types of Quantitative Mission-focused Measures

Quantitative mission-focused measures can be continuous or discrete, but continuous measures are always preferable. A continuous measure will almost always require a smaller sample size and fewer test resources for the risk levels chosen (confidence and power). Additionally, continuous measures often contain more information regarding the performance of the system, whereas a corresponding discrete measure will throw away information. For example, measuring detect/not detect provides no information about how close the sensor approached. Using the range at which detection occurred in concert with the closest point of approach in cases where no detection occurred provides a better characterization of sensor performance. The probability of detection over all ranges is the only quantity that can be calculated with the discrete data, but if the continuous variable (range) is measured, one can understand the distribution of detection ranges as well as the probability of detection as a function of range. Even if the requirements document defines a probability-based metric, great effort should be expended to find a related continuous measure on which to base the test design.

Examples of continuous measures include time to detect, miss distance, human error rate, time to complete task, and range of engagement. Examples of discrete measures include hit/miss, message complete/not complete, and detect/not detect.

Definitions of Quantitative Mission-focused Measures

The measures chosen must also be well-defined and meaningful. Testers and evaluators should consider example operational scenarios to ensure that the measure can be unambiguously measured (scored) and calculated in all cases. The following principles are critical:

- Formulas for the measures should not be ambiguous – TEMPs should provide amplifying information (explicit formulas and/or scoring criteria) if the CDD requirement is unclear
- Measures should be testable and not require unsafe or unexecutable test constructs or cost-prohibitive instrumentation
- Measures should accurately represent the desired performance of the system – Good scores should correspond to desired operational performance
- Measures should not lead to non-production representative modifications to the system or unrealistic tactics

Measure Selection for Survey Data

In operationally focused testing, the use of operator surveys and subject matter expert panels are needed and useful to aid in the characterization of system effectiveness, suitability, and survivability. This is particularly true when quantitative data is scarce due to expensive field testing or low sample sizes. However, before using them, other objective measures such as time to complete a task or human error rate should be explored.

Survey data, like physical measures, should be collected systematically for each test condition to facilitate statistical comparisons of survey responses across the operational test space. Imagine, for example, that one is interested in evaluating the performance of an aircraft under different light conditions, day versus night. Administering the same survey to pilots under both test conditions allows testers to determine how pilots' experiences shift under different operational conditions and if those shifts in experience explain observed differences in aircraft performance. Additionally, many important aspects of operational suitability are best addressed by survey data (e.g., human machine interface, operator workload). Ideally, survey data and subject matter expert panels should be used in concert with objective quantitative data.

Survey use should follow best practices, such as:

- Clearly identify survey objectives: TEMP should indicate which test goals will be addressed by survey data, the goal of the survey measure, and who will provide the survey data required to address the measure, for instance, operators or maintainers.
- Surveys should be pre-tested on an appropriate group to reveal if questions are confusing or if information is missing. The TEMP should include a plan for pre-testing custom-made surveys before IOT&E. This can often be accomplished during developmental testing or operational assessments occurring before IOT&E.
- Survey questions should be clear and unbiased (e.g., no leading questions)

Surveys should use quantitative (e.g., Likert-scale) and qualitative responses (open ended questions); quantitative data should be coded, compiled and summarized using statistical methods to aid in system characterization in concert with the measures employed in field testing.

References

Inspector General Report, May 15, 2015

Reporting of Operational Test and Evaluation (OT&E) Results, DOT&E, January 6, 2010

Test and Evaluation Policy Revisions, DOT&E, December 22, 2007

Guidance on the Use of Design of Experiments (DOE), DOT&E, October 19, 2010

Guidance on the Use and Design of Surveys in Operational Test and Evaluation (OT&E), DOT&E, June 23, 2014

Survey Pre-Testing and Administration in Operational Test and Evaluation, 6 January 2017

Operational Evaluation Framework – Guidance

Overview

The Operational Evaluation Framework (OEF) is a tool for communicating the entire OT plan and providing a basis for a decision maker to determine test adequacy. The OEF doesn't add information; it packages the plan for easy consumption.

The TEMP should be organized to present separate developmental and operational evaluation approaches. Part 3.2 should include the developmental evaluation methodology and framework. Part 3.4 should include the operational evaluation methodology and framework.

After the Developmental Evaluation Framework (DEF) and Operational Evaluation Framework (OEF) have been developed, the integrated test planning process can proceed. By comparing similar data requirements from the DEF and OEF, DT&E and OT&E planners can design integrated test events to generate the data needed for the independent evaluations. Scientific Test and Analysis Techniques (STAT) provide ideal tools for developing these integrated test events.

3.4.2. Operational Evaluation Framework

The OEF table summarizes the mission focused evaluation methodology and supporting test strategy, including the essential mission and system capabilities that contribute to operational effectiveness, suitability, and survivability. The table identifies the goal of the test (within a mission context), quantitative mission-focused measures (also referred to as quantitative mission-oriented response variables), factors that affect those measures, and test designs for strategically varying the factors across the operational envelope, test period, and test resources. The evaluation framework may also include standard measures of program progress including: key performance parameters, critical technical parameters, key system attributes, interoperability requirements, cybersecurity requirements, reliability growth, maintainability attributes, and others as needed. However, the framework should focus on (1) the subset of quantitative mission-focused measures critical for assessing operational effectiveness, suitability, and survivability and (2) resource, schedule, and cost drivers of the test program.

The OEF should show how the major test events and test phases link together to form a systematic, rigorous, and structured approach to quantitatively evaluate system performance across the operational envelope. The table should also be used to justify the resources necessary for an adequate test.

The operational evaluation framework should also support integrated testing by identifying opportunities for using DT data for OT evaluation. In cases where DT data supports OT evaluation, the evaluation framework table should link to the supporting developmental evaluation framework and summarize procedures for ensuring data collected in DT will be adequate for OT evaluation.

The evaluation framework table should mature as the system matures and be updated at each revision of the TEMP. The table may be inserted in Part III of the TEMP. Alternatively, the framework can be embedded as an Excel table/database, or provided as an appendix.

Table 3.X, provides the essential information to be included in the OEF. Below that table are hyperlinked examples (based on notational programs) of how an evaluation framework table can be organized. These examples should not be taken as a 'cookbook' or template – each program is unique and

will require thoughtful tradeoffs in how to apply this guidance. Equivalent Service-specific formats that identify the same relationships and information may also be used.

Table 3.X. Operational Evaluation Framework Essential Information

Goal of the Test	• Focus on an operational mission and/or capability being assessed. • Link each mission/capability to at least one quantitative mission-focused measure. • Identify the associated COI(s) or COIC(s), where applicable.
Quantitative Mission-focused Measures (Response Variables)	• Quantitative mission-focused test measures provide criteria for mission accomplishment (not technical performance for a single subsystem) and comprehensively cover the reasons for procuring the system (the need). • Also include the resource, schedule, and cost drivers of the test program.
Test Design	• Factors that affect the quantitative mission-focused measures during operation employment of the system. • Experimental design approach (e.g. Fractional Factorial or D-Optimal 2nd order model) for strategically varying the factors across the operational envelope. • Effect sizes for observing identified factors and their interactions where appropriate. • When not using an experimental design approach provide a brief description of the test design and further details on how the test method was selected in the STAT Appendix. The OEF should provide a brief summary; the STAT appendix should include the detailed test design, the corresponding statistical measures of merit (confidence and power), and effect sizes. •
Test Period	• Include all operational test periods when collecting data (e.g., LUT, OA, IOT&E, FOT&E, etc.)
Resources	• High level summary of the resources (time, people, places, and things) needed to execute an adequate test.

Operational Evaluation Examples (pdf files)

Operational Evaluation Framework Aircraft Example

Operational Evaluation Framework Space Observation Radar Example

Operational Evaluation Framework Clean Example

Downloadable Excel Files (These will take a few moments to download.)

Operational Evaluation Framework Aircraft Spreadsheet

Operational Evaluation Framework Space Observation Radar Spreadsheet

Operational Evaluation Framework Clean Spreadsheet

References

DoDI 5000.02

Guidance on the use of Design of Experiments (DOE) in Operational Test and Evaluation

DOT&E, October 19, 2010

Table 3.1. Top-Level Operational Test Evaluation Framework Matrix

Top-Level Operational Test Evaluation Framework (Assume 100% Test Efficiencies)					
Goal of the Test		**Mission-oriented Response Variable**	**Test Design**	**Resources**	**Test Period**
Operation/Capability	**COIs**	**Effectiveness / Survivability/Suitability[1]**	**STAT Methodology and Operational Context**	**People, Places, Things**	**e.g., LUT, OA, IOT&E**
Close Friendly Engagement	COI 1. Close Air Support	Time to employ weapons (KSA1) Aircrew rating workload (KSA2) Ground force rating of coordination	Flight test - DRY STRIKE test design (Table D.1) -- 77 test points (69 pts D-optimal of 2^6*3 + 8 demos)	19 dedicated sorties x 4 hrs	**IOT&E**
ID/Monitor Enemy Forces	**COI 1, COI 2. air interdiction COI 3. collateral CSAR/NTISR**	Range to identify target (KPP1)	Flight test - TARGET ID/MONITOR (Table D.2) -- 47 test points nested in DRY STRIKE sorties (39 pts D-optimal of 3^2*2^2*4 + 8 demos)	Nested in sorties above	IOT&E
ID/Monitor Friendly Forces	**1, 2, 3**	Range to identify target (KPP1)	Flight test - FRIENDLY ID/MONITOR (Table D.3) -- 47 test points nested in DRY STRIKE sorties (39 pts D-optimal of 3^2*2^2*4 + 8 demos)	Nested in sorties above	**IOT&E**
Direct Fire	**1, 2**	Miss distance/CEP Ability to correct fire (KPP2) Time to employ (KPP3) Time to reload	Flight test - 30MM LIVE SHOT (Table D.4) -- 16 test points (2^3 factorial x 2)	7 sorties nested in above 2 sorties nested in below for simultaneous demos 800 x 30mm	IOT&E
PGM Employment	**1, 2**	Miss distance Time to employ Time to impact Stand-off range (KPP4) Dual target engagement (KPP2)	Flight test - GRIFFIN LIVE SHOT (Table D.5) -- 22 test points (20 optimal split-plot 2^4*3 + 2 demo) Flight test - SDB LIVE SHOT (Table D.6) -- 18 test points (2^3 factorialx2 + 2 demo)	18 dedicated sorties on live fire range 18 x SDB 22 x Griffin	OA1 (50%, DT sorties) IOT&E (50%)
Net-centric Ops Supportable	**1, 2, 3**	Ability to support net-centric ops (KPP5) Aircrew rating of situational awareness Availability of ISR data	Surveys from all IOT&E sorties Flight test - VORTEX TRANSFER (Table D.7) -- 6 test points nested in sorties above	Nested in sorties above	IOT&E
Persistence	**1, 2, 3**	Compatibility of crew operating environment Loiter time	Surveys from all IOT&E sorties Paper analysis of stores	Existing sorties	IOT&E
Sortie Generation	**COI 4. mission taskings**	Time for Mx to generate aircraft Time for crew to preflight aircraft	Measured over existing IOT&E sorties -- Demo hot and cold-soak startups	McKinley Climate Lab for cold-soak startup	IOT&E
Materiel Reliability	**4**	Weapon system reliability (KSA10) Mission reliability	Measured over existing IOT&E sorties	Existing sorties. At least 151 flhrs required with <= 2 aborts.	IOT&E
Maintainability	**4**	Mean Time to Repair Mx crew rating of tech orders PSP integrated diagnostics	Measured over existing IOT&E sorties	Existing sorties	IOT&E
Materiel Availability	**COI 4, COI 5. operate globally**	Mission Capable Rate Aircraft Availability (KSA9) WRSK availability	Measured over existing IOT&E sorties	Existing sorties	IOT&E
Air Refuelable	**5**	Crew rating of refueling ops	Flight test: AERIAL REFUELING (Table D.5) -- 4 test points (2^2 factorial) during existing sorties	Existing sorties	IOT&E
Force Protection	**COI 5, COI 6. perform missions and survive**	Probability of casualty from specified ballistic threats (KPP7) Aircrew ability to use life support Aircrew egress time Size of security package	Ground test: Demo aircrew emergency egress x4 -- Day/night, with/without PPE	Existing sorties	IOT&E
Survivability	**6**	Pk of avoiding/defeating threat (KPP6) Probability of analyzing threat indications (correctly identifying) Probability of accomplishing avoidance tactics	Flight test: ELECTRONIC THREAT AWARENESS (Table D.6) Flight test: VISUAL THREAT AWARENESS (Table D.7) -- 20 test points each (2^3 factorial x2 + 4 demos) during existing sorties	Existing sorties. Approximately 10-14 sorties required over electronic warfare test range.	IOT&E

1. Label measures with a KPP or KSA identifer if the measure is associated with a KPP or KSA.

Mission: Provide Space Surveillance Data to Support the Space Control Mission

Goal of the Test	Quantitative Mission-Oriented Response Variables		Test Design		Scope		
	Effectiveness / Suitability	Thresholds	Factors (Levels)	Scientific Test and Analysis Techniques with Operational Context	Effort (24/7 ops)	Resources	Test Period
Evaluate Autonomous Surveillance	Uncued Observation Metric Accuracy	* Time < 1 s * Elevation, Azimuth <1.0 degree * Range <100 m * Range Rate < 200 m/s	Altitude (600 - 7k km) Inclination (8◦ - 172◦) Size (10 - 50 cm)	A. 1 Sample Variance Test with LASER-ranged targets * 95% power, 5% significance (α), 10% effect size B. 4x2x3 full factorial Analysis of Variance (ANOVA) design using live (satellite catalog (SATCAT)) and simulated tracks (Table D-1) * 95% power, 5% α, 10% effect size * Lowest power to differentiate between levels of a factor & their 1st order interactions is 96.1%, at 5% α and 0.5 S/N	5 days 25 days	Space Surveillance Network (SSN) radar and optical sensors NASA laser ranging AFSPC/A9 analysts M&S	DT_2 IOT&E
	Minimum Detectible Target (MDT) Size (KPP)	Objects * 10 cm: 600 ≤ x ≤ 4000 km * 50 cm: 4000 < x ≤ 7000 km	Inclination (8◦ - 172◦)	Logistic Regression Model with a 1x3 full factorial design using live 10cm SATCAT tracks (Table D-5). M&S needed for 50 cm tracks * Power to determine factor effects is 90% at 5% α, 10% effect size.	25 days		DT_2
	Uncued Probability of Track (KPP)	50% of objects be tracked if they pass through the radar's field of view (FOV)	Altitude (600 - 7k km) Inclination (8◦ - 172◦)	Logistic Regression Model w/ 3x3 full factorial design using live SATCAT tracks (Table D-2). * Lowest power to differentiate between levels of a factor is 90%, at 5% α, 10% effect size. 75% power for detecting the 1st order factor interaction	8 days		DT_2
	Track Coverage (KPP)	* 1 track/day in the altitude range 600-4000 km * 2 tracks/day in the altitude range 4000-7000 km	Inclination (8◦ - 172◦) Size (10 - 50 cm)	Logistic Regression Model w/ 3x2 full factorial design using live SATCAT and simulated tracks (Table D-3). * Lowest power to differentiate between levels of a factor & factor interactions is 90%, at 5% α, 10% effect size	23 days		DT_2
	Object Correlation	97% of previously detected objects must be correlated w/ SATCAT	Altitude (600 - 7k km) Inclination (8◦ - 172◦)	Logistic Regression Model with a 3x3 full factorial design using live SATCAT tracks (Table D-4) * Lowest power to differentiate between levels of a factor & factor interactions is 90%, at 5% α, 10% effect size	8 days		IOT&E
Evaluate Cued Operations	Cued Observation Metric Accuracy	* Time < 1s * Range, elevation, azimuth as specified in Figure 3-5. * Range Rate < 20 m/s	Altitude (600 - 7k km) Inclination (8◦ - 172◦) Size (10 - 50 cm)	A. Test of 1 Proportion with LASER-ranged targets * 95% power, 5% α, 10% effect size B. Logistic Regression Model with a 4x2x3 full factorial design using live SATCAT and simulated tracks (Table D-1) * Lowest power to differentiate between levels of a factor & factor interactions is 90%, at 5% α, 10% effect size	9 days 25 days	Resources listed above plus JSPOC	DT_2 IOT&E
	Probability of Track (KPP)	90% of objects be tracked if they pass through the radar's FOV	Altitude (600 - 7k km) Inclination (8◦ - 172◦)	Logistic Regression Model w/ 3x3 full factorial design using live SATCAT tracks (Table D-2). * Lowest power to differentiate between levels of a factor is 95%, at 5% α, 10% effect size. 85% power for detecting the factor interaction	8 days		DT_2

Goal of the Test	Quantitative Mission-Oriented Response Variables		Test Design		Scope		
	Effectiveness / Suitability	Thresholds	Factors (Levels)	Scientific Test and Analysis Techniques with Operational Context	Effort (24/7 ops)	Resources	Test Period
Evaluate Uncorrelated Target Processing	Initial Orbit Determination Accuracy	Reacquisition and correlation > 75% of objects w/in 24 hrs	Altitude (600 - 7k km)	Logistic Regression Model with a 1x3 full factorial design using live SATCAT tracks (Table D-8). * Power to determine factors effects is 90% at 5% α and 10% effect size.	20 days	SSN sensors JSPOC AFSPC/A9 analysts	IOT&E
Evaluate Narrow-band Space Object Identification	Radar Cross Section (RCS) Accuracy	RCS > (classified) dBsm	Size (10 - 50 cm)	One-way ANOVA with 1x3 factorial design using calibration spheres (Table D-10) * Power to determine factors effects is 90% at 5% α and 10% effect size.	21 days	JSPOC NASIC SSN sensors	DT_2
Evaluate space event detection and processing	Data Timeliness	data latency to end user must be < 2 min 99% of the time		Tolerance Intervals using the SATCAT (Table D-9) * 90% power, 5% α, 10 effect size	1 day	JSPOC M&S SSN sensors	IOT&E
	Flexible Coverage	* 0.5 cm: 600 ≤ x ≤ 1000 km * 5 cm: 1000 < x ≤ 2000 km * 8 cm: 2000 < x ≤ 4000 km * 15 cm: 4000 < x ≤ 12K km	Inclination (8◦ - 172◦)	A. Logistic Regression Model w/ 1x3 full factorial design using live SATCAT and simulated tracks (Table D-6). B. Logistic Regression Model with a 1x3 full factorial design using NaK debris (Table D-7). M&S needed for inclinations < 30 degrees	20 days 13 days		IOT&E IOT&E
	SOC Functionality	tasking and tracking 300 objects	# of objects (1 -400)	M&S of space events, such as ASATs, On-Orbit Maneuvers, New Launches, and Space Object Breakups.	10 days		IOT&E
Evaluate cybersecurity defenses	prevent, detect, react, and restore	See Cybersecurity section for detailed measures and thresholds		A. Cooperative Vulnerability and Penetration Assessment (CVPA) with the system in it's operational configuration. B. Fix C. Adversarial Assessment (AA) of the system in it's operational configuration against a nation-state cyber-threat	15 days 30 days 15 days	CVPA Team AA Team	DT_2 IOT&E
Evaluate DODIN interoperability	Net- Ready (KPP)	TBD		JITC certification	14-28 days	JTIC	DT
Evaluate Suitability	E3: On-Orbit E-Field	≤ 23 v/m, peak, RMS ≤ 8 v/m, avg, RMS	Range (100 -1k km)	Construct an RF profile based off the contractor design: RF antenna gain versus angle off-boresight. Refine based on: 1. Transmitter subassembly anechoic chamber testing 2. CONUS prototype testing - Near field: place RF detectors a precise distance from radar to determine the transmitted power to a known point - Ground: RF survey of the surrounding area 3. Fielded operational system - Repeat near field and ground RF survey tests - Perform atmospheric measurements with aircraft - Leverage in-band on-orbit assets to measure RF		land, air, and on-orbit E/M receivers	DT
	E3: Atmospheric Power Density	≤ 43.8 dBW/m2, peak, RMS ≤ 15.8 dBW/m2, avg, RMS	Range (10 m - 20 km)				
	E3: Ordnance	≤ 2,500 v/m, peak, RMS ≤ 220 v/m, average, RMS	Range (0- 5 km)				
	E3: Personnel	≤ 10 W/m2, averaged over 30 minutes	Range (0- 5 km)				
	Operational Availability	System Ao ≥ 95% SOC Ao Ao ≥ 98% System MTBCF ≥ 1000 hrs SOC MTBCF ≥ 1000 hrs		AMSAA-PM2 method for growth tracking and projection. (See Reliability Growth section.)	140 days	diesel fuel	DT & OT

Table 3.1. Top-Level Operational Test Evaluation Framework Matrix

Top-Level Operational Test Evaluation Framework (Assume 100% Test Efficiencies)					
Goal of the Test		**Mission-oriented Response Variable**	**Test Design**	**Resources**	**Test Period**
Operation/Capability	**COIs**	**Effectiveness / Survivability/Suitability[1]**	**STAT Methodology and Operational Context**	**People, Places, Things**	**e.g., LUT, OA, IOT&E**

1. Label measures with a KPP or KSA identifer if the measure is associated with a KPP or KSA.

Operational Testing of Software-Intensive Systems - Guidance

Summary

This guidance applies to software-intensive systems that are covered by the DoDI 5000.02, January 7, 2015, under Incrementally Deployed Software Intensive (Model 3), software-intensive Accelerated Acquisition (Model 4), and Hybrids Acquisition programs. The DOT&E policy, Guidelines for Operational Test and Evaluation of Information and Business Systems, 14 September 2010, especially applies to Model 3 systems. Model 3 systems are distinguished by the rapid delivery of capability through multiple acquisition increments, each of which provides part of the overall required program capability. Each increment may have several limited deployments; each deployment will result from a specific build and provide the user with a mature and tested sub-element of the overall incremental capability. Several builds and deployments will typically be necessary to satisfy approved requirements for an increment of capability. Software systems must also address cybersecurity testing, as required by DOT&E Procedures for Operational Test and Evaluation of Cybersecurity in Acquisition Programs, 01 August 2014 and further described in Appendix E, Cybersecurity.

OT&E for software acquisitions will be guided by the assessment of operational risks of mission failure. The DOT&E Guidelines should be used by the OTA to help determine the level of risk and the corresponding adequate level of OT&E for all capabilities that are to be deployed. There will be at least one full OT&E for every formal acquisition increment of a software intensive system unless waived by DOT&E. For software intensive systems on DOT&E oversight, DOT&E approval of the level of risk and adequate level of OT&E is also required. The degree of independent operational testing appropriate for each software increment or capability can be tailored by using the risk analysis described in the DOT&E Guidelines. The Guidelines also permit delegation of test plan approval using the same criteria.

Overall sustainment approaches should be adequately described in the Life Cycle Management Plan or similar document. A weak integrated logistics and sustainment approach can be a huge risk even if the system effectiveness and suitability are otherwise acceptable. There should be a documented, repeatable process whereby problems are documented at the help desk and problems that are fixed by any tier of help desk support are tracked to completion; those problems that the help desk system cannot resolve should be escalated through a well-defined process and IEEE 12207.2 priorities assigned as discrepancy reports (DRs). Then, each DR should go through a Configuration Control Board (CCB) process to verify operational impact and priority with the result being a plan to fix the problem. After fixes are implemented in projected releases, there needs to be a regression test procedure within the organization that provides the fix and a further CCB process to release into production the new version, with rollback procedures in case the new version fails. This aspect of risk directly relates to the operational impact if the problem were to be missed during testing and subsequently found during operational use, since it helps determine the fix process and appropriate regression testing.

The entire risk assessment and design/conduct of testing process should be a significant focus area for continuous improvement. Whenever significant risks are encountered after completion of testing, it must be assumed that the risk assessment process, operational test adequacy, and/or the test/fix/test process require significant improvement. A simple metric showing the cumulative number of Category I problems encountered, and cumulative Category I problems fixed, after completion of operational testing of the previous software release, should be shown as part of the risk assessment level of test package when submitted to DOT&E for approval.

References

DoDI 5000.02, 7 January 2015

DOT&E Guidelines for Operational Test and Evaluation of Information and Business Systems, 14 September 2010

DOT&E Procedures for Operational Test and Evaluation of Cybersecurity in Acquisition Programs, 01 August 2014

Directive-Type Memorandum (DTM) 11-009, Acquisition Policy for Defense Business Systems (DBS), 23 June 2011 with 9 Dec 2011 change, AT&L Directive

Software Maturity Criteria for Dedicated Operational Test and Evaluation of Software-Intensive Systems, DOT&E Memo, 31 May 1994

IEEE 12207.2

Examples

Operational Testing of Software Intensive Systems Example

OT of Software-Intensive Systems – Example

Example TEMP entries for Global Combat Support System - Joint:

The example shown below refers to Global Combat Support System – Joint (GCSS-J) which is an information system using Agile Software Development methodology and for which the DOT&E Guidelines apply. GCSS-J is a query-only web-based system accessing multiple databases. This program also utilizes a beta test site approach with significant emphasis on integrated testing. Examples have been shortened to convey only the most important information relating to the risk-based software testing approach and how it works with Agile Software Development processes, with TEMP paragraphs 3.1, 3.3, and 3.6 being most affected. The examples shown do not represent all the information suggested for these paragraphs.

Paragraph 3.1. T&E Strategy

As DISA becomes more agile in its development process, the intent of the Capability Test & Evaluation framework is to speed the delivery of capability to the warfighter. Adoption of a Capability Test & Evaluation framework will:

- Reduce risk and cost
- Eliminate duplication and improve data sharing between organizations
- Improve the quality of test results

The Capability Test & Evaluation model supports a "one team, one time, testing once under one set of conditions" process. Capability T&E concentrates test and certification activities into one test period, as early in the acquisition process as it is practical. The results, of which, then inform/satisfy the decision maker and all other testing stakeholders. Capability Test & Evaluation test designs are risk-based, mission-focused and do not limit the independence of the OTA or its ability to provide independent, objective evaluation of a capability's effectiveness and suitability. The OTA will conduct OT&E for releases based on the determined level of test based on an OTA-conducted risk analysis using the DOT&E Memorandum, "Guidelines for Operational Test and Evaluation of Information and Business Systems", 14 Sep 2010.

Paragraph 3.2. Developmental Evaluation Approach

The GCSS-J Developmental Test & Evaluation (DT&E) is designed to mitigate design risk and ensure compliance with system requirements. The DT&E risk analysis and risk mitigation efforts are an integral part of the overall Program Risk Management effort. Risks specific to testing will be included in the GCSS-J Program Risk Report. The status of risks and the progress of risk mitigation efforts are closely monitored by the PMO. DT&E will be conducted by employing a risk-based approach to identify test objectives, events, and personnel. The DT&E will also evaluate compliance with operational requirements to minimize risk and support certifying systems ready for dedicated OT.

DT&E will focus on risk assessment of functionality and the data gathered during DT will determine the appropriate scope and balance required to adequately test each increment.

OT of Software-Intensive Systems – Example

The testing strategy will utilize an integrated DT&E/OT&E approach to maximize the use of DT events and DT documentation that addresses specific functionality, issues, and criteria to reduce the scope of the OT&E events required. The intent is to reduce the scope of the OT&E events required by focusing only on those issues and criteria that need to be addressed in a purely operational environment. The DT strategy will include data gathering for independent certifications for required items (e.g., interoperability, security, etc.) and will assess compliance with the CDD/CPD specified functional and technical requirements and the CTP identified in this document.

Paragraph 3.4 Operational Evaluation Approach

The JITC serves as the Operational Test Agency (OTA) for GCSS-J. As the OTA, the JITC provides test directors and test personnel to support operational test events. The primary purpose of OT&E is to determine whether systems are operationally effective, suitable, and survivable for the intended use by representative users in a realistic environment before production or deployment. The JITC will conduct an OT&E for each of the planned releases (SIPRNet and NIPRNet) based on the determined level of test based on an OTA-conducted risk analysis using the Guidelines for Operational Test and Evaluation of Business and Information Systems. Each OT will be system-level and address the combined requirements and capabilities implemented during the version releases, to include regression testing of the existing system as appropriate.

Production Representative Test Articles – Guidance

Summary

Consistent statutory guidance and with the goal to “fly before buying” major systems for the Department of Defense, operational testing in support of Full-Rate Production decisions must be conducted with production systems or production-representative test articles. Whenever practicable, production systems are to be furnished from low-rate initial production (IOT&E) quantities. Through the TEMP, DOT&E can approve the use of production-representative test articles in lieu of production test articles. In evaluating whether systems are production-representative, DOT&E will consider whether the test articles were assembled using the parts, tools, and manufacturing processes intended for use in full-rate production. The system should also use the intended production versions of software. In addition, the logistics system and maintenance manuals intended for use on the fielded system should be in place. DOT&E must be provided detailed information describing any process differences in order to independently evaluate whether the differences are acceptable.

References

Title 10 USC 2399

Use of Production-Representative Test Articles for Initial Operational Test and Evaluation (IOT&E), DOT&E, October 18, 2010

Defense Acquisition Guidebook, Paragraph 9.3.2

DoDI 5000.02, 7 January 2015

Examples

Production Representative Test Articles – Examples

Example 1

3.4.2 Configuration Description. The IOT configuration will be a Dakota helicopter company with five LRIP Dakota aircraft and all authorized equipment, pilots, and maintenance personnel and support equipment.

Example 2

3.4.2 Configuration Description. The IOT configuration will be 15 production-representative Gemini missiles with complete capability as required by the CPD. The missiles are production systems with the exception of "white wires" in the guidance module used to fix a problem discovered late in developmental testing. In production, this "white wire" will be replaced by firmware circuitry. These missiles have been assembled at the production facility. Maintenance and support equipment is production representative.

Exceptions to the use of production test articles, if any, should be explained and will be subject to DOT&E approval.

Example 3

4.2.1 Test Articles. The test articles and testing sequence for the Dakota program are defined in Table 19, Test Article Matrix. See Chapter 3 for additional details on each test event in this table.

Test Article	Test Event	Quantity	Start Date	Source
Prototype aircraft	DT	2	FY07	Contract
Prototype aircraft with ASE	LUT	2	FY10	Contract
Spare Parts for flight testing	All	As Needed	FY07	Contract
LRIP aircraft	IOT&E	5	FY12	Contract
LFT&E Components	LFT&E	See LFT&E Strategy	FY11	USG/Contract

Table 19 - Test Article Matrix

Note confirmation in resources section of the TEMP that LRIP test articles are planned.

Realistic Operational Conditions - Guidance

General Guidance

Operational testing in support of Full-Rate production decisions shall be conducted under realistic operational conditions.

The Operational Test Agencies shall design the test to conform to the anticipated wartime operational tempo and provide detailed tactics, techniques, and procedures to the participating forces. Other considerations for realistic operational conditions include typical operators and maintainers, a mission focused evaluation, the use of production representative test articles, adequate threat representation, end-to-end testing and baseline evaluation when appropriate, cyber security testing, and selection of quantitative mission-focused measures in the design of experiments (DOE) analysis.

For each operational test, the TEMP will describe the resources, personnel, site selection, tactical considerations, and other factors intended to ensure appropriately realistic operational conditions. Specific resources and production representative test articles will be described in of the TEMP.

References

Title 10, U.S. Code, Section 139

Test and Evaluation Policy Revisions, DOT&E, December 22, 2007

Procedures for Operational Test and Evaluation of Cybersecurity in Acquisition Programs, DOT&E, August 1, 2014

Realistic Operational Conditions – Examples

Example TEMP entry for generic sonar system:

3.4.1 Operational Test Events and Objectives. OT will be conducted using an event driven and operationally realistic end-to-end scenario. Data gathered during previously completed IT and DT events with production-representative test articles will be considered in the evaluation. OT will be conducted using test events designed to assess all required capabilities of the sonar system and the ship's crew in operation of the system. The scenario will require the system to provide Undersea Warfare surveillance support to a Naval Strike Group. Within this scenario, the Blue Force test ship will sortie from port, conduct active, passive, and coordinated USW with friendly forces, and return to the port. USW operations will be conducted in deep, open ocean waters and Littorals against SSK and SSN threats executing validated threat tactics. Test sites will include representative levels of neutral shipping to provide realistic levels of interfering contacts. Threat forces will be tasked to aggressively pursue and attack the Naval Strike Group, and may preemptively engage the Blue Force test ship if possible.

Example TEMP entry for generic utility helicopter:

3.4.1 Operational Test Events and Objectives. The IOT will be conducted at a training center with appropriately equipped and trained pilots and maintainers. Vignettes will include company-level air movement, air assault and CASEVAC missions. Vignettes will be conducted with five LRIP Dakota digital and five Baseline analog aircraft side-by-side in wartime OPTEMPO as prescribed by the OMS/MP. Dakota aircraft will complete 150 hours of record test.

The IOT will start with a communications exercise to verify the aircraft communication systems meet interoperability requirements and to verify proper integration with the Tactical Internet. The Assault Helicopter Company leadership and portions of the battalion staff from the Assault Helicopter Battalion and the Aviation Unit Maintenance (AVUM) Company will participate in the test. An Infantry Company and an Artillery Battery will act as the supported unit.

Testing will focus primarily on vertical maneuver missions in an operational environment against appropriate validated threats. Representative threats (RF, IR, laser) will stimulate aircraft survivability equipment (ASE) to demonstrate proper integration and display, and evoke appropriate responses/flight maneuvers from flight crews. Brigade and below level operations orders will be provided to the headquarters staff for dissemination and execution by the lift element. Maintainers will employ the two level maintenance concept.

IOT Company level missions will emphasize navigational capability; day/night operations; interoperability (communication), situational awareness and other key performance parameters.

Simulated maneuver forces will be used to augment live maneuver forces to portray a realistic Common Operating Picture (COP). Operationally realistic command and control,

threat and friendly forces will be provided. A "Blue Cell" C2 element will perform the responsibilities of the higher level HQ and provide direction. A "Red Cell" will perform a similar function for threat forces operating within the scope of the Training and Doctrine Command (TRADOC) approved test vignettes. A "White Cell" consisting of Battlefield Operating Systems nodes not otherwise represented in other cells, or live, will serve to coordinate test matrix execution and perform test control functions. The simulation cells will serve to generate a real, partially correct, or false COP; stimulate communications; stress the command element; and, most importantly, to stimulate the aircrews.

Reliability Growth – Guidance

Summary

The majority of life cycle costs for DoD systems reside in the Operations and Sustainment (O&S) phase, where O&S costs are often driven by unreliability. The more reliable the system, the less it costs to operate and sustain in the field. With today's highly complex systems, a small decrease in reliability can mean additional, substantial cost, but a small investment in reliability growth can significantly decrease O&S costs.

A comprehensive reliability program, focusing on reliability growth is essential for developing and acquiring reliable systems. From the start, a program should formulate and document a comprehensive reliability, availability, and maintainability (RAM) program. The program should employ an appropriate reliability growth strategy to improve RAM performance until RAM requirements are satisfied. The reliability program should be documented in detail in the system engineering plan (SEP). In addition, key systems engineering and design activities needed for the test strategy should be included in the Test and Evaluation Master Plan (TEMP).

Elements of Reliability Program for the TEMP

The TEMP must provide an overview of the reliability program and testing needed to assess and monitor reliability growth, including design for reliability test and evaluation (T&E) activities. DOT&E is looking for a concise description of the following elements when reviewing the reliability portion of TEMPs:

- Key engineering activities supporting the reliability growth program including[1]:
  - reliability allocations to components and subsystems,
  - reliability block diagrams (or system architectures for software intensive systems) and predictions,
  - failure definitions and scoring criteria (FDSC),
  - failure mode, effects and criticality analysis (FMECA),
  - system environmental loads and expected use profiles,
  - dedicated test events for reliability such as accelerated life testing, and maintainability and built-in test demonstrations,
  - reliability growth testing at the system and subsystem level, and
  - a failure reporting analysis and corrective action system (FRACAS) maintained through design, development, production, and sustainment.
- The system's reliability growth program, including:

[1] The key engineering activities should be discussed in more detail in the appropriate supporting references. References to supporting information, such as the System Engineering Plan or the Reliability Program Plan, should be provided in the TEMP.

- initial estimates of system reliability and a description of how this estimates were arrived at,
- reliability growth planning curves (RGPC) illustrating the reliability growth strategy, and including justification for assumed model parameters (e.g. fix effectiveness factors, management strategy),
- estimates with justification for the amount of testing required to surface failure modes and grow reliability,
- sources of sufficient funding and planned periods of time to implement corrective actions and test events to confirm effectiveness of those actions,
- methods for tracking failure data (by failure mode) on a reliability growth tracking curve (RGTC) throughout the test program to support analysis of trends and changes to reliability metrics,
- confirmation that the FDSC on which the RGPC is based is the same FDSC that will be used to generate the RGTC
- entrance and exit criteria for each phase of testing, and
- operating characteristic (OC) curves that illustrate allowable test risks (consumer's and producer's risks) for assessing the progress against the reliability requirement. The risks should be related to the reliability growth goal. See the Reliability Test Planning Guidance for more information on OC curves.

- DOT&E has no default criteria for acceptable test risks. The rationale for the selection of test risks should derive from the specifics of each program.
- Resource requirements (including test articles and expendables) that reflect the best estimate for conducting all reliability T&E activities and are reflective of the allowable test risks

Reliability should be measured, monitored, and reported throughout the acquisition process. Reliability measurements and estimates should be recorded on the RGTC and compared to the RGPC. Systems not meeting entrance and exit criteria should revise the reliability growth strategy to reflect current system reliability. When necessary, reliability growth should continue after the full-rate production decision (FRP) and fielding until RAM requirements are met. Provisions should be made to monitor reliability even after requirements are met.

Figure 1 below shows a notional reliability growth curve. Key features include the idealized curve, important acquisition events, and corrective action periods (CAP). These CAPs are the projected periods during which the system will undergo changes to correct identified failure modes. Between the CAPs are test periods that should correspond to the test phases and acquisition decision milestones in the TEMP.

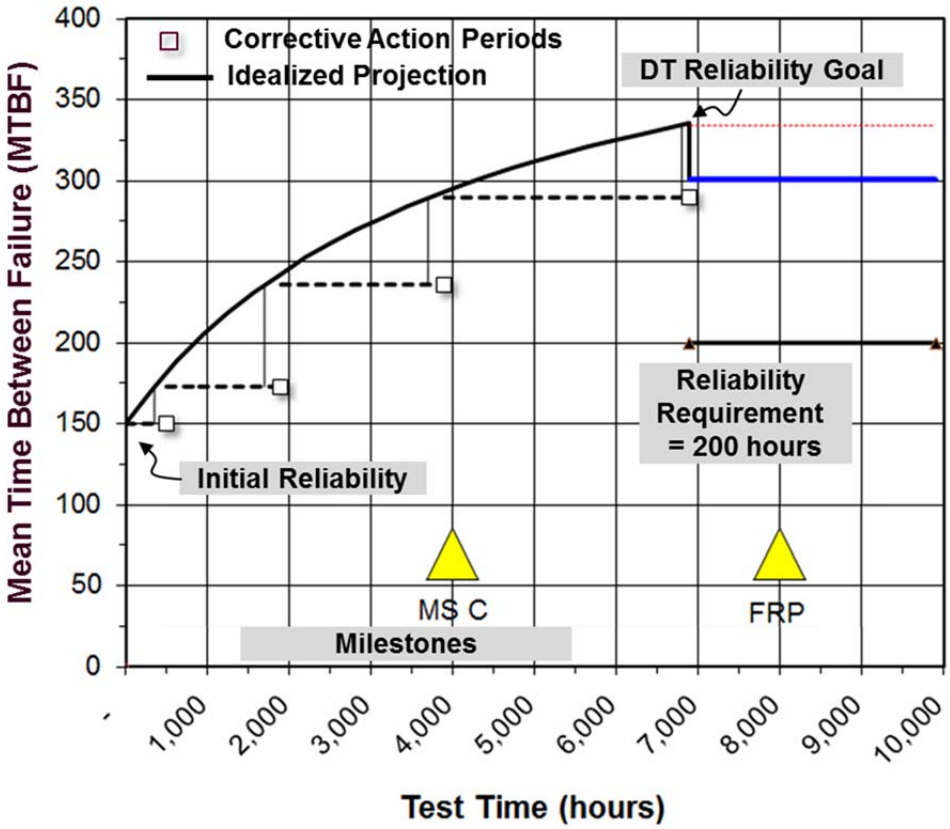


Figure 1: Notional Reliability Growth Curve

Note that the curve includes an adjustment from the DT reliability growth goal to the OT reliability growth goal, which is itself higher than the requirement. Design margins should be included to ensure that the requirement is met. Larger design margins increase the likelihood that the requirement can be demonstrated with statistical confidence during the program's IOT&E.

Guidance for documentation of reliability growth in TEMPs is discussed below by grouping DoD systems into three general categories:

- Hardware only systems, which contain no software (bullets, personal protective equipment);
- Hybrid systems containing a combination of software, hardware, and human interfaces. Critical functionality is a combination of hardware and software sub systems (complicated ground combat vehicles, aircraft, and ships);
- Software-intensive systems characterized by built-in redundancies that result in high reliability for the hardware (or hardware is not a component of the system), leaving the software reliability as the limiting factor (safety critical systems, automated information systems, and some space systems).

Hardware Only and Hybrid Systems

System level reliability growth for hardware and hybrid systems can be planned for using the AMSAA Planning Model based on Projection Methodology (PM2) or the Crow-Extended Planning Model. Using these models, program management is able to establish a realistic reliability growth curve in relation to time (or distance, use cycles, etc.) that provides interim reliability goals and serves as a baseline against which reliability assessments can be compared.

Reliability Growth Planning Curves (RGPC) should be included in the TEMP and reflect the reliability growth strategy. A RGPC must be included in the TEMP beginning at Milestone B, and updated at each subsequent milestone. The RGPC should be stated in a series of intermediate goals and tracked using a suitable Reliability Growth Tracking Curve (RGTC) through fully integrated, system-level test and evaluation events until the reliability threshold is achieved. If a single curve is not adequate to describe overall system reliability, multiple curves should be provided for critical subsystems with rationale for their selection. In many cases, multiple curves should be constructed to track both top-level reliability metrics (e.g., system aborts) as well as lower-level metrics (e.g., essential function failures). Tracking lower-level metrics allows for a more granular assessment of reliability and given their more frequent occurrence, can give a higher resolution view of reliability growth early on in system testing.

Programs using quantitative time-based measures of mean time between failure (MTBF) metrics (or life units such as miles, cycles, rounds, operations, etc.) should calculate the reliability growth potential (the maximum life unit that can be attained with the current management strategy) to ensure that reliability thresholds are achievable. PMs should continue to track reliability on the RGTC after FRP, regardless of whether reliability requirements have been met.

Operational test events should be noted on the growth curve, and intermediate reliability goals should be associated with each OT event. These events may also include Reliability, Availability, and Maintainability (RAM)-based entrance and criteria. This could include demonstrating that the point estimate for system reliability is at or above the growth curve prior to entering test or that the system must achieve a certain level of reliability prior to proceeding beyond an acquisition milestone.

At Milestone C, RGPCs should be updated based on the current status results of the RGTC and the reliability program plan should be updated with current information (including the current reliability estimate). The TEMP should characterize key failure modes and their disposition. Post-Milestone C TEMPs must be updated as needed to continue reliability monitoring and reliability growth after fielding until terminated by the receiving Service.

For hybrid systems, in addition to the RGPC, the TEMP (or supporting documentation references in the TEMP) should outline a plan for categorizing hardware failures verses software failures, provide a plan for tracking software failures on the RGTC, and a clear plan for regression testing software failure fixes.

Software-intensive Systems

Software-intensive systems must address reliability growth by providing either a reliability growth planning curve (RGPC) or reliability growth tracking curve (RGTC). If a RGPC is appropriate for the program, then the TEMP should provide a RGPC based on an appropriate methodology. The Crow-Extended and the AMSAA Projection Methodology (PM2) models are two recommended reliability growth planning models. If using a RGTC, programs should follow the guidance for hybrid systems. For software-intensive systems that are primarily software, the RGTC may be more appropriate. The selection of the appropriate curve for inclusion in the TEMP should be reflective of the program.

If a RGTC is appropriate for the program, then the TEMP should outline a plan for categorizing software failures; a reliability tracking curve for software failures (plot of system faults over test time) should be provided once available and should be updated over time. Additionally, a plan for regression testing of software failure fixes should be discussed.

All software intensive systems, starting at Milestone A should describe the plan to track software reliability across the acquisition development life cycle with defined entrance and exit criteria for system reliability at critical decision points. Software reliability growth curves provide one rigorous methodology for defining reliability projections based on past test data. IEEE 1633™ - 2008, Recommended Practice on Software Reliability, Annex F, provides a three-step approach for applying software reliability growth models to plan, track, and project software reliability growth for software-intensive systems from detailed design and through design, analysis, coding, and testing. For more information on this methodology please see the DOT&E working group page of software reliability growth.

Reliability Growth for Ships

Guidance

New Ship Example

Mature Ship Example

Examples

Reliability Growth Example

Software Reliability Tracking – Example

Reliability Test Planning Guidance

References

Independent Operational Test and Evaluation (OT&E) Suitability Assessments

DoD Instruction 5000.02

Recommended Practice on Software Reliability, Annex F, IEEE 1633™ ,

MIL HDBK 189 C – Reliability Growth Management

DOT&E Working Group Software Reliability Growth

Reliability Growth - Example

3.3.2 Reliability Growth (or in Appendix F)

Dakota reliability growth will consist of positive improvement through systematic removal of failure modes by way of positive changes in design, material, or manufacturing. Dakota reliability growth will begin at program initiation and continue through production. Reliability growth will be achieved not only through lab and flight testing, but also by way of design analysis, production experience, and operational experience.

The reliability growth test program will accomplish its goals by: (1) finding reliability problems through testing, (2) establishing a Failure Reporting, Analysis, and Corrective Action System (FRACAS) to identify root causes of failure and corrective actions, (3) incorporating corrective actions when appropriate, and (4) continual monitoring of corrective actions and the system's reliability throughout all test phases.

Dakota Reliability, Availability, and Maintainability (RAM) performance will be continuously assessed using data from development flight testing, logistics demonstration, and operational testing. Dakota reliability growth will be tracked against a reliability growth curve that estimates reliability thresholds associated with program decision points. The focus of the Dakota reliability growth program will be on identification of new and existing failure modes and correction of hardware and software failures. A failure review board consisting of Government and contractor elements will convene monthly to discuss the FRACAS data and evaluate the root cause determination, proposed corrective actions, and the verification methodology. Once corrective actions are verified and incorporated, the corrective action will continue to be monitored for fix effectiveness to assess its impact on reliability growth.

RAM Scoring Conferences will be held quarterly. All RAM data will be scored using the approved Dakota Failure Definition/Scoring Criteria, which is in compliance with the DOT&E Guidance on Independent Operational Test and Evaluation (OT&E) Suitability Assessments. The RAM Scoring Conference voting members are the materiel developer, the combat developer, and the evaluator; however the final operational evaluation of Suitability will be based on the independent evaluators vote. Testers and technical support personnel may support the Scoring Conferences in an advisory capacity.

The goal for the reliability growth program is to demonstrate the 17-hour MTBF Full Rate Production requirement with 80 percent confidence using data from IOT&E. To provide evidence at Milestone C that reliability the reliability growth goal is achievable, the program will seek to demonstrate a MTBF of 20 hours during the Limited User Test (LUT). The development goals associated with this reliability growth program include addressing at least 80 percent of the initial failure intensity via corrective action with an average fix effectiveness factor of 70 percent.

Reliability Growth - Example

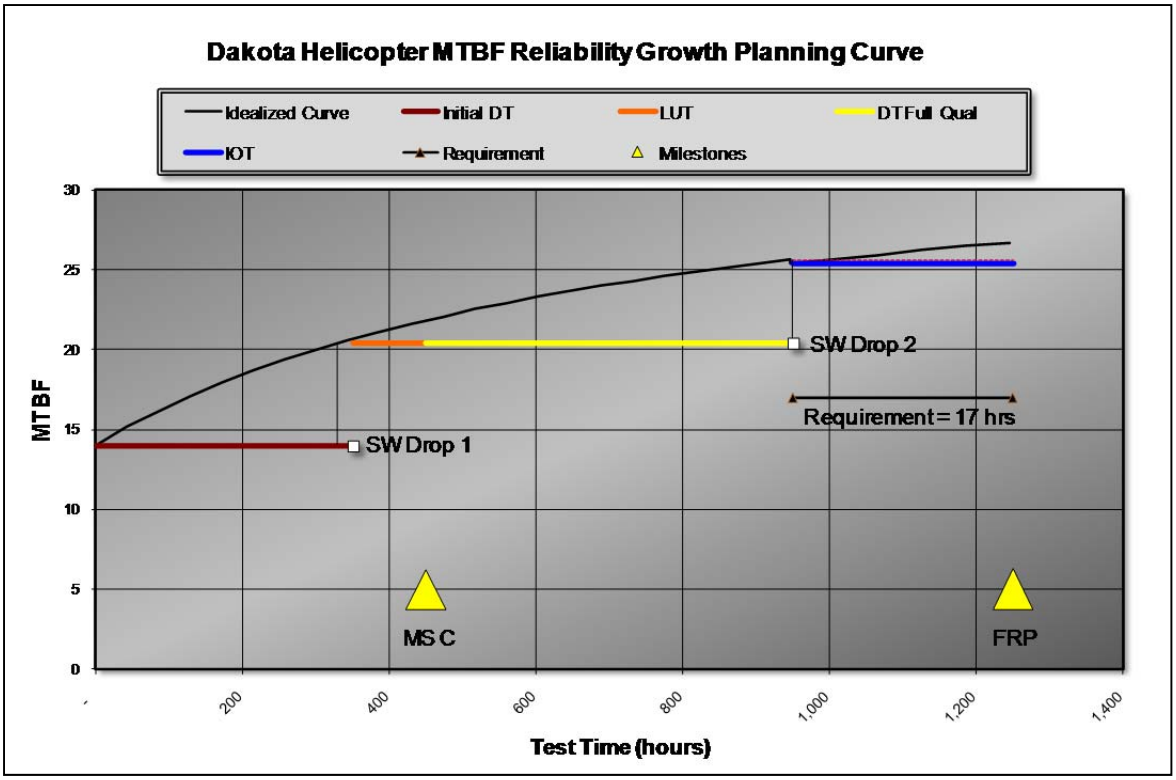


Figure 1. Reliability Growth Curve

The reliability growth plan consists of two corrective action periods for implementing corrective actions to reliability deficiencies observed during developmental test flights. Approximately nine B-mode failures are expected before the first CAP and an additional 5 are expected before the second CAP. There will be a major software release just prior to the LUT and another just prior to IOT&E. The majority of corrective actions discovered in developmental testing will be implemented in these software releases. If the true MTBF is 26 hours during the IOT, then there is a 73 percent chance Dakota will demonstrate its 17 hour requirement with 80 percent confidence.

Table 1. Projected Flight Hours Supporting Reliability Growth

Test	Test Flight Hours	Cumulative Flight Hours
Initial DT	350	350
LUT	100	450
DT Full Qualification	500	950
IOT	300	1250

3.2.3 Reliability Tracking (or Appendix F)

The software reliability tracking effort will start at the beginning of the software design effort in each of the nodes and/or components. Code design reviews will be held for each code module to ensure conformance with the particular contractors' standards and to identify and correct obvious errors. Beginning at the start of the Code and Unit Test (CUT) activity, quality metrics will be collected at all subcontractors for each of their coding efforts. For the Engineering, Manufacturing, and Development (EMD) phase of the program, collection and analysis will continue through all levels of code development, from CUT through Software Integration, Subsystem (node level) Integration, and System Integration.

3.2.3.1 Discrepancy Report (DR) Status

Each DR written against contractor-developed software will be prioritized into five levels as defined by the IEEE 12207 specification. Each DR will be initially assigned a level by the subcontractor developing that particular software. The prime integrator and the Government Program Office will perform an independent analysis and redefine levels accordingly. Graphs similar to Figures 1 and 2 will be maintained showing the number of open, closed, and resolved (fixed but not tested) statistics over time, by priority level.

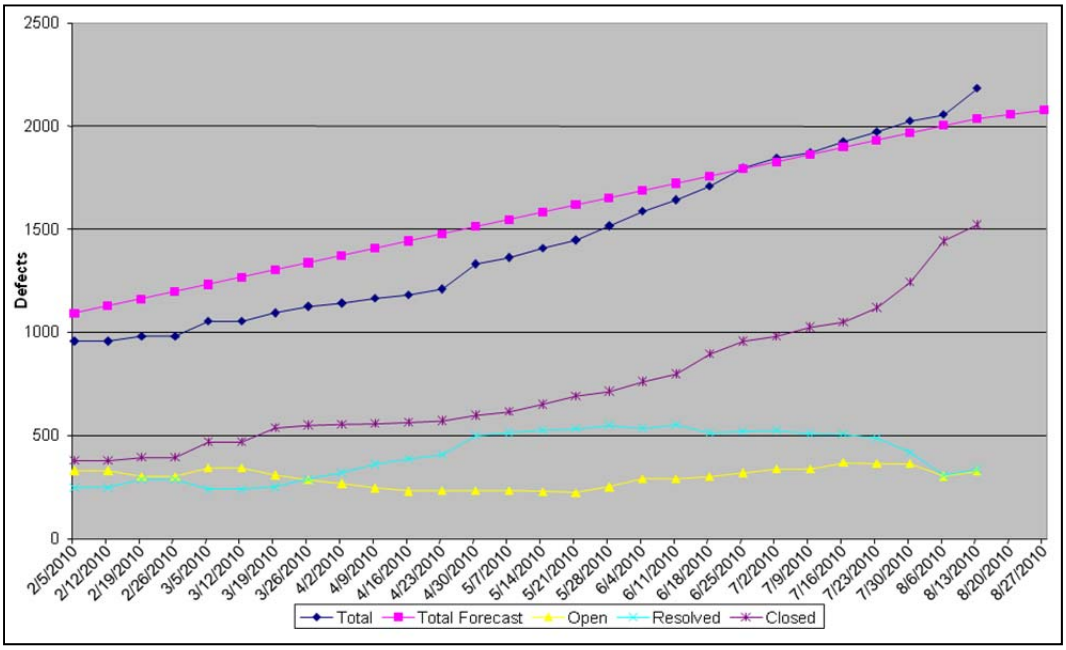


Figure 1. Example DR volume tracking (all priorities)

3.2.3.2 DR Aging

DRs at each priority level will be tracked to show how many of each level were open for a particular timeframe by priority. The timeframes will be separated into 30-day increments, up

to a column for >120 days. The values in parentheses reflect the status from the previous reporting period. Example data are shown in Table 1.

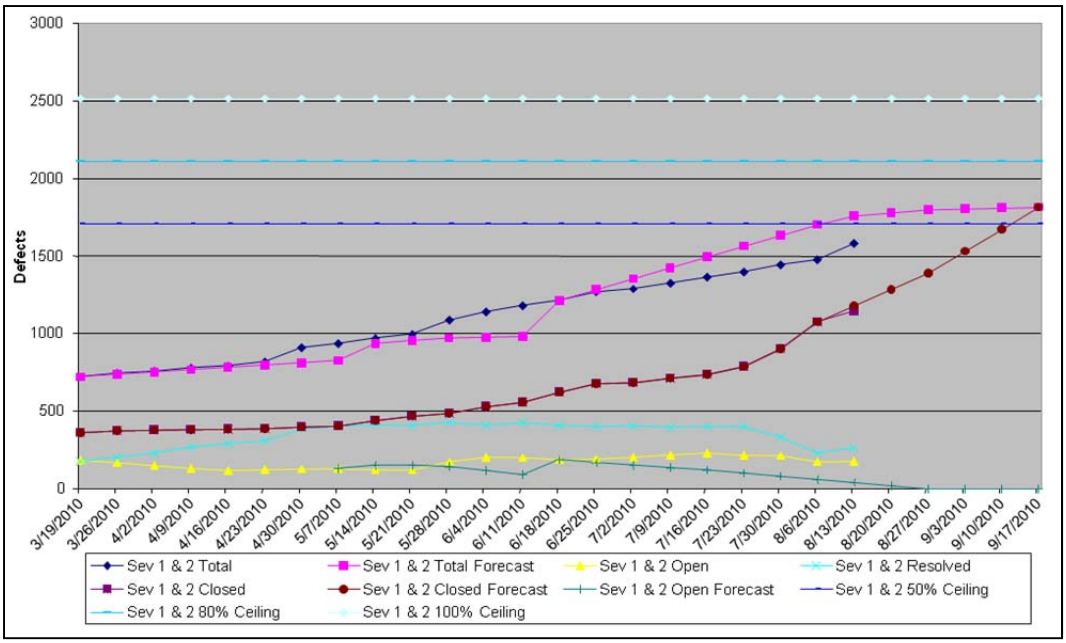


Figure 2. Example DR Volume tracking (Priorities 1 and 2)

Table 1: Sample DR Aging Metric

	Assigned and Submitted Defects – Days Open				
Severity	**0-30**	**31-60**	**61-90**	**91-120**	**>120**
1	9(18)	12(3)	1(1)	2(3)	4(2)
2	92(99)	41(28)	13(11)	5(8)	19(18)
3	48(45)	6(4)	8(11)	3(0)	16(18)
4	16(15)	3(3)	3(3)	1(1)	4(4)
5	0(1)	5(4)	0(1)	3(4)	4(3)
Total	165(178)	67(42)	25(27)	14(16)	47(45)

3.2.3.3 Commercial Off The Shelf (COTS) DRs

The ageing statistic described above will be maintained for issues found with commercially purchased equipment, such as routers, servers, etc.

3.2.3.4 Software Management Strategy

Every DR will be analyzed to determine the effect of the failure. Using this information, a determination will be made as to the severity of the problem (Priority, as defined by the IEEE 12207 specification). All failures that rate a Priority 1 or 2 will be fixed prior to entering the next phase of testing. These data will be collected and curves will be maintained throughout development and OT&E.

Reliability Test Planning – Guidance

Summary

Operational mission reliability is the ability of a system to perform a required function under given environmental and operating conditions and for a stated period of time. Operational testing provides the ability to assess mission reliability because the testing is conducted to evaluate how systems improve mission accomplishment under realistic combat conditions. Ideally, adequate data on the mission reliability will be collected during operational testing, using representative users under a range of operationally realistic conditions. In these cases operating characteristic curves should be used to assess the test adequacy for the assessment of mission reliability. Operating characteristic curves can be constructed for mission based reliability requirements or duration based requirements. In cases where it makes sense testers should use duration based versions of the requirements to maximize information.

Unfortunately, it is often not possible or cost effective to collect all of the data on system reliability in operational testing. In these cases, using a range of additional sources of information may provide a better assessment of the operational mission reliability. If additional information will be used in the reliability assessment the TEMP should outline the source of the additional information, the required fidelity to include operational conditions and scoring criteria for failures, finally the methodology for combining information should be outlined. Data from different test events should not be combined into one pool of data and used to calculate and average reliability, rather advanced analysis methodologies (See Bayesian Statistics Guidance) should be used to combine information from multiple tests.

Reliability Requirements

The duration of reliability testing depends on the form of the reliability requirement. Reliability requirements can be pass/fail in nature or time/duration based. Pass/fail reliability requirements are common for single-use systems:

Probability of a fuse igniting without failure in a weapon system > 90%

For repairable systems, reliability requirements might be specified in terms of mission duration and probability of mission completion:

A howitzer must have a 75% probability of completing an 18-hour mission without failure.

Alternatively, requirements might specify a mean time between failures:

A howitzer mean time between failures must exceed 62.5 hours.

One can translate between probabilistic mission duration requirements and the mean time between failures using the exponential distribution. The cumulative distribution function for the exponential distribution (cumulative probability of failure in a reliability context) is:

$$F(t) = 1 - e^{\frac{-t}{\theta}}$$

Where θ is the mean time between failures (MTBF) and *t* is the mission length. Note that by plugging the mission duration and the MTBF from the example howitzer requirement we obtain the probability based requirement:

$$F(18) = 1 - e^{\frac{-18}{62.5}} = 0.25$$

A system with a MTBF of 62.5 hours has a 25 percent chance of mission failure in an 18 hour mission, or a 75 percent chance of completing the mission successfully.

The mission reliability equation provides a useful translation between system reliability and operational implications, as long as it is reasonable to assume that the failure rate is exponentially distributed. In some cases, this equation can illustrate that MTBF requirements exceed the expected use of the system and are unnecessarily high. For example, consider a bomb that is employed by a fighter aircraft. As Table 1 illustrates, we would need a large MTBF requirement if we require a high probability of completing a standard 2-hour mission without an in-flight failure. Since an individual weapon will never exceed 50 hours of flight time, 200-hour MTBF requirement is unreasonable.

Table 1: Mission Reliability and MTBF

Probability of Mission Completion / Mission Duration	Mean Time Between Failure (MTBF)
99% (2-hour mission)	199 Hours
95% (2-hour mission)	39 Hours
95% (4-hour mission)	78 Hours

Operating Characteristic Curve – Planning an Adequate Test

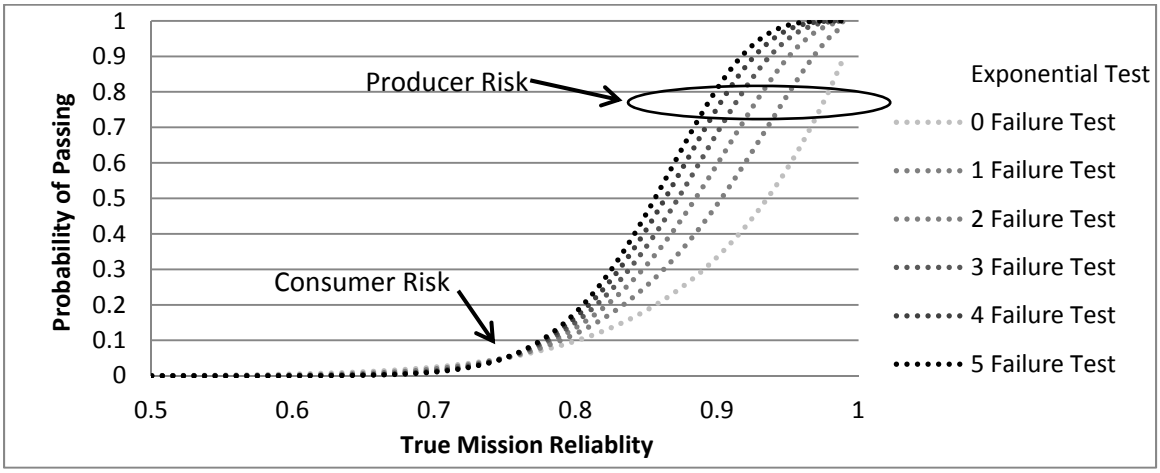
Operating characteristic curves are useful statistical tools for planning the length of a reliability test. OC curves illustrate the probability of passing the test as a function of the true mission reliability. In practice, we never know exactly how reliable the system is so it is important to select a test that balances risk and achieved reliability. Central to the development of these curves is the balancing of Consumer Risk and Producer Risk. Consumer Risk is defined as the probability that a bad system (below threshold reliability) will be accepted, whereas Producer Risk is the probability that a good system (above threshold reliability) will be rejected. The risks should be related to the reliability growth goal. An example of a generic OC curves is provided in Figure 1.

DOT&E has no default criteria for acceptable test risks, the rationale for the selection of test risks should derive from the specifics of each program.

The following curve shows an OC curve for determining the test length for a howitzer requirement of 75 percent probability of completing an 18-hour mission without

failure. OC curves should be provided for the selected test duration and corresponding risks highlighted in the TEMP.

Figure 1. Example Operating Characteristic Curve for Operational Test Planning



Incorporating Additional Information

Assessing the operational reliability of complex systems can often require the incorporation of multiple sources of data in order build a credible statistical analysis of system reliability. When additional sources of information are used in reliability assessments, all of the following should be specified in the TEMP:

- The conditions the data must be collected under to be acceptable for OT use.
- The methodology for scoring reliability data collected outside of an OT. If you plan to use developmental test data for operational evaluation, developmental test reliability failures must be scored by the same methods as the operational reliability data.
- The statistical models and methodologies for combining information. Data should not simply be pooled together and an average reliability calculated. The analysis should account for the conditions the reliability data were collected under to the extent possible. See Bayesian Methods for additional guidance on using various sources of information in analyses.
- The methodology for determining adequate operational test duration. Bayesian assurance testing can be used in place of traditional operating characteristic curves to determine adequate operational testing when prior information will be incorporated. Table 2 shows how Bayesian assurance testing can reduce required test time and control test risks.

Table 1: Bayesian versus OC Curve Reliability Test Planning

Failures Allowed	Bayesian Assurance Test Miles 10% Consumer Risk 5% Producer Risk	Classical OC Curve Miles 10% Consumer Risk Producer Risk Varies
1	2,940	7,780 – 58% Producer Risk
2	4,280	10,645 – 50% Producer Risk
3	5,680	13,362 – 43% Producer Risk
4	7,120	15,988 – 37% Producer Risk
5	8,580	18,550 – 32% Producer Risk

Requirements Rationale – Guidance

Guidance

At times, the Capability Development Document (CDD) or requirements documents do not provide an operational rationale for the requirements or their thresholds. To develop an adequate operational test and evaluation strategy, the operational testers and evaluators need an understanding for why the requirement exists and of the possible consequences of failing to meet the thresholds.

There have been cases when the Key Performance Parameters (KPPs) and Key System Attributes (KSAs) do not form a sufficient basis for evaluation of mission effectiveness. See the Inspector General report dated May 15, 2015 for two examples. If the requirements cannot be revised to define those system characteristics most critical for providing an effective military capability, the TEMP must identify and define those characteristics. See guidance on quantitative mission-focused measures.

If the key requirements are appropriate and their rationale documented in the requirements document is adequate to support test planning and evaluation, *no further clarification is necessary*. In cases where the requirement is derived or transformed for testability or the operational rationale is unclear, the TEMP should have an appendix that explains the operational rationale and/or the derivation of the metric as well as the chosen numerical thresholds.

If not adequately documented in the CDD or other requirement documents, add rationale to the TEMP. Here are three examples.

Example 1

Requirement: The Dakota Attack Aircraft must be capable of receiving full motion video from unmanned aircraft systems (UAS). The Dakota must be able to receive and display to the crew the following minimum information via Ku Band: encrypted and non-encrypted streaming and still video imagery, sensor platform position, sensor azimuth, target location, and range-to-target. The Dakota must be capable of storing and transmitting this data to other members of the Joint/Combined Arms air/ground maneuver team, including legacy Dakota aircraft. The acceptable level of communication performance must be such that a two-way, line-of-sight data link and appropriate upload/download data rate can be maintained between the Dakota and the unmanned aircraft at no less than 50 km (threshold) (100 km objective).

Rationale: Integration of information is critical for aircrew situational awareness. Ground maneuver commanders rely heavily on variety of data sources and types to develop courses of action and to initiate engagements. Displaying information in an accurate and organized manner reduces cockpit workload and enhances mission effectiveness and survivability. Additionally, Dakota-UAS interoperability supports the future Modular Force, Networked Lethality, and Networked Battle Command concepts, tactics, techniques, and procedures. It extends detection/targeting ranges; teams manned and unmanned aircraft systems for maximum synergy; and avoids placement of manned aircrews at unnecessary risk.

Example 2

Key Performance Parameter	Threshold	Objective
Mission Reliability	89%	90%

Mission Reliability. The mission reliability rate of 89% (Threshold, KPP), 90% (Objective) is required. Mission Reliability is the probability that the Heavy Lift Replacement (HLR) shall successfully support the USMC Ship to Objective Maneuver (STOM) concept. In the course of one period of darkness, an HLR unit of 28 aircraft must transport 73 external loads from ship to shore, a distance of 110 nautical miles. Each of the 73 external loads consists of various amounts of ammunition, fuel, water, supplies, or equipment weighing up to 27,000 pounds.

Example 3

Reliability. The threshold requirement for mean time between mission failure (MTBF(M)) is 20 flight hours (objective of 22 flight hours). The threshold requirement for mean time between essential maintenance action (MTBEMA) is 2.9 flight hours (objective of 3.1 flight hours). The specified reliability is needed to ensure a dependable level of aircraft performance and to ensure that operations and support costs of the current fleet of Dakota helicopters are reduced. Achieving the reliability thresholds will assure that the user obtains an aircraft with improved reliability performance and improved mission success capability. The reliability thresholds reflect a 20 percent improvement over the reliability performance of the current Dakota fleet, and the reliability objectives reflect a 30 percent improvement over the reliability performance of the current Dakota fleet.

References

DoDI 5000.02, 7 January 2015

Inspector General Report, May 15, 2015

Ship Reliability Growth – Guidance

Background

The necessity for a reliability growth program for Major Defense Acquisition Programs (MDAP) is well established. Despite this, it is often argued that Navy ship class programs are exempt from such requirements because the Navy's well established oversight of ship construction and pre-delivering testing makes it unlikely that ships will deliver with serious reliability problems. Additionally, some have argued that because new ship classes are often comprised of numerous, mature and reliable technologies (e.g. hull, mechanical, and propulsions systems) there is little risk that the ship will have poor reliability.

However, some recent ship-class IOT&Es have demonstrated that ship programs are subject to the same reliability problems, including reliability problems with mature systems, that other acquisition programs are subject to. Ships might be different from other types of acquisition programs, but they still need to be reliable. This guidance highlights the key aspects of a reliability growth program for ships that need to be documented in a TEMP.

Reliability Growth for New Ship Programs

For new ship class programs, the following steps should be included in the program's reliability growth plan:

1. Early-on, identify, in the context of the ship completing its primary missions, the ship's critical systems. This work is typically already done early during the detail design phase to support ship survivability studies.
2. Determine what the overall reliability and availability requirements for the ship imply about the required reliability of critical systems. This requires the construction of reliability block diagrams and modeling and simulation.
3. As construction begins, measure the reliability of critical systems at the factory, at the shipyard, or elsewhere in the fleet, to verify that the critical system reliability supports the overall ship reliability.
4. Record failures in a Failure Reporting, Analysis, and Corrective Action System, implement corrections as needed, and continue to monitor reliability.
5. At delivery, continue collecting reliability data and verify that the overall reliability is on track to meet its reliability requirements at IOT&E.
6. Confirm reliability at IOT&E and possibly rerun M&S with measured critical system reliability data instead of specification reliability data. Verification, validation, and accreditation of M&S should include a review of M&S assumptions to ensure that critical systems were not overlooked and to verify that reliability block diagrams are correct.

Reliability Growth for Mature Ship Programs

It is not uncommon to find a ship class program that pre-dates OSD's reliability growth requirement. In these instances, where there is no previous requirement, a strategy similar to the steps for a new ship program above should be implemented.

1. Map overall reliability requirements to critical system reliability using fleet standards to determine if system failures equate to ship failures (e.g., Status of Resources and Training System (SORTS) ratings). This analysis was likely done to support ship survivability studies.
2. Collect critical system reliability data wherever available (e.g., other ships using the same systems) and periodically review data collected with test and evaluation stakeholders.
3. When the ship is delivered, start collecting reliability data on critical systems and against overall reliability requirements whenever possible.
4. Correct reliability deficiencies before IOT&E.
5. Collect data through IOT&E and update M&S with observed component reliability to determine if ship meets its reliability requirements. Verification, validation, and accreditation of M&S should include a review of M&S assumptions to ensure that critical systems were not overlooked and to verify that reliability block diagrams are correct

TEMP Language

The TEMP must include language that describes the steps above and must include resources for the collection and analysis of reliability data. Additionally, the TEMP must include resources for the Verification, Validation, and Accreditation of whatever reliability M&S is used to assess requirements. If the ship has a reliability growth program, then it must be documented in the TEMP as it would for any other program. (See the Reliability Growth Section of this guide book and the included New Ship Example). The relevant TEMP language for an ongoing ship class program without a reliability growth program is provided as the Mature Ship Example.

Ship Reliability Growth – Mature Ship Example

Program Managers are responsible to provide fully capable Government Furnished Equipment (GFE) for installation aboard the ship. The GFE systems are Programs of Record and have completed OT. Upon shipboard installation, the ship program performs production and post-delivery testing to ensure the equipment and systems are properly integrated to support mission requirements.

A Reliability, Maintainability, and Availability (RM&A) analysis conducted on Propulsion and Electrical Distribution systems predicated that the ship will attain the ship Capability Development Document (CDD) A_o requirements. The analysis was conducted with the NAVSEA TIGER Computer Simulation Program. TIGER program is a Monte Carlo simulation technique used to provide the analyst with a generalized capability for determining system reliability, readiness, and availability estimates. The result of the analysis is provided in the Ship Hull, Mechanical & Electrical (HM&E) Systems RM&A Analysis, Naval Surface Warfare Center – Carderock Division (NSWC-CD) Report. The TIGER Model used a 180-day Design Reference Mission (DRM) developed by the Ship Program Office based on program documentation (CDD, CONOPs, etc.).

The TIGER Model identified four critical systems to achieve the Propulsion and Electrical Distribution A_o requirements of 0.85 (Threshold) and 0.95 (Objective):

- Main Propulsion System
- Auxiliary Propulsion System
- Ship Service Diesel Generators
- Machinery Control System

The Program Office will track the reliability of the four critical systems and three additional mission essential systems:

- Heating, Ventilation, and Air Conditioning (HVAC) system
- Refrigeration system
- Cargo and aircraft elevators

Comprehensive production testing is conducted on the Ship to confirm shipbuilder compliance with the contract reliability provisions and specifications. Additionally, the production testing will test for proper installation and integration of Government Furnished Equipment (GFE). Production testing during pre-acceptance test and evaluation will be conducted at the shipbuilder facility and witnessed by the government test team. Sea trials provide the first opportunity to observe full system operation for a sufficient length of time or number of cycles and will be used for the evaluation of the reliability metrics.

Ship Reliability Growth – Mature Ship Example

At sea testing will occur prior to the Navy accepting delivery and will continue through the post-delivery test and trial period. The accumulative hours at sea will not be sufficient to statistically validate Mean Time between Failures (MTBF). The shipbuilder is required to analyze and correct all premature failures during the warranty period. System and equipment discrepancies identified during the warranty period are entered and tracked via trial cards in the Technical Support Management (TSM) tool. After completion of acceptance trials conducted by the Navy Board of Inspection and Survey (INSURV) prior to ship delivery and upon correction of deficiencies, the Navy accepts delivery of the ship and assumes maintenance responsibility.

Upon delivery, all system and equipment discrepancies will continue to be entered and tracked via trial cards in TSM during the warranty period. Maintenance data is also entered into the Navy 3M maintenance system. Final Contract Trials (FCT) will be conducted by INSURV prior to the end of warranty period to confirms material readiness to support operational missions.

The ship is a modified variant of an existing ship and, as such, incorporates: (1) the existing hull design / electric plant modifications, and (2) fact of life modifications to Command, Control, Communications, Computers, and Intelligence (C4I) and Warfare Systems (each with an approved Program of Record). The ship program will track the reliability of select common (between the new ship class and the existing ship class) components and equipment via the OPNAV Material Readiness Database (MRDB,) maintained by Naval Surface Warfare Center Corona, and via data through the Open Architecture Retrieval System (OARS).

Design or equipment deficiencies identified on existing ship class are (and continue to be) evaluated; and where practical, design modifications are implemented on the new ship class. Upon delivery, the ship reliability will be similarly tracked. The data collection effort for the identification and evaluation of deficiencies will continue similarly for follow-on ships.

Reliability data will be collected and posted after each trial event in the Common T&E Data Repository on the Naval Sea Systems Command Corporate Document Management System (CDMS).

Data analysis working groups (scoring committees of subject matter experts (SME)) will convene, as required, to adjudicate and analyze reliability data to ensure a common set of data and mutual rules for data evaluation. SMEs will be nominated by the Program Office, PEO IWS, DOT&E, and COMOPTEVFOR.

Ship Reliability Growth – New Ship Example

The following example is for the USS *Reliable* (ABC 10) ship class. The ABC 10 class is the replacement class for the USS *Unreliable* (ABC 1) class ship.

ABC 10 Reliability Growth Strategy Overview

The ABC 10 reliability growth strategy was developed in accordance with MIL-HDBK-189C, DoD Handbook on Reliability Growth Management. The ABC 10 Reliability Growth Strategy was developed to capitalize on the lessons learned from the legacy ABC 1 program. Failure modes identified in ABC 1 have been identified and their fixes applied to the ABC 10. Additionally, the majority of the equipment that will be used to construct the ship has several years of demonstrated reliability.

The reliability growth strategy leverages critical equipment, integrated sub-systems, and ship-level testing to assess Reliability, Availability and Maintainability (RAM). These critical pieces of equipment are expected to be the primary reliability drivers for ABC 10 and include: main engines, propulsion subsystems, C4N hardware and software, auxiliary and electrical power generation subsystems. The reliability growth strategy will focus on these critical systems. Equipment level testing serves to identify and correct design weaknesses early in the program. Reliability block diagrams and simulation tools (Raptor Reliability Simulation Software) and were used to determine reliability requirements for selected critical equipment (main engines, APUs, etc). Equipment level reliability growth curves have been developed and will be utilized to monitor reliability growth during equipment level testing. It is expected that critical equipment will be responsible for 58% of the failures (reference the ABC 10 RAM Predictions and Analysis Report).

The Shipbuilders a robust RAM program is described in more detail in the reliability program plan. Key elements include:

- Development and analysis of component/system level RAM modeling
- Implementation of RAM predictions/allocation, to include quantitative RAM requirements in Shipbuilder/vendor procurement specifications
- Conduct a Failure Mode, Effects and Criticality Analysis (FMECA)
- Develop and apply operational and environmental life cycle loads when selecting equipment/components
- Perform maintainability demonstrations
- Implement a Failure Reporting, Analysis and Corrective Action System (FRACAS)
- Use a Government led Failure Reporting Board (FRB)
- Conduct equipment and ship-level reliability growth testing.

critical equipment

In order to adequately assess the reliability of the critical equipment, adequate testing was allocated for five ABC 10 critical systems. Table 1 shows the dedicated hours of reliability testing for each of the critical systems. Sufficient test time at the equipment level has been allocated to discover and fix equipment level failures.

Table 1. Hours of Reliability testing for each ship subsystem from predesign to IOT&E.

System	Cumulative System Hours Prior to Shipboard Installation		Quantity per ship	Cumulative Ship-Level Testing	
	Operating Hours from Prior Testing not under the ABC 10 program	System Testing at shipyard prior to ship installation		Contractor Test Hours	Government Test Hours
Main Engines	10,200	1,416	4	960	960
Propulsion System		104	2	480	480
C4N System		1,210	1	240	240
Auxiliary System	500	1,204	1	240	240
Electrical Generation	1,000	304	2	480	480

In order to develop a ship-level reliability growth model, equipment-level testing is used to determine the initial ship-level MTBF entering the Shipbuilder test phase of ship-level testing, the management strategy required for successful Shipbuilder and Government testing, and the ability to achieve the respective equipment-level MTBFs in support of the threshold MTBF requirement.

The goal is to grow to an effective ship-level MTBF of 32.5 hours, while ABC 10 is underway. Derivation of the effective ship-level MTBF (aka, threshold MTBF) underway is described below. Although the ship-level MTBF 32.5 hours for underway time will be used to measure the ship's reliability growth, reliability data will be recorded for all phases of testing.

MTBF While Underway Derivation

The six phases of the Design Reference Mission profile is described in Table 1. The most stressing mission phases from a reliability perspective are mission phases B and C where the ship is actually underway. Therefore, the underway periods will be used to derive a reliability underway requirement.

Table 2. ABC 10 Mission Phases and Reliability Predictions

Mission Phase	Predicted Mission Phase MTBF	Time in Phase	Predicted Reliability	Derived Required Reliability
Phase A: Mission Prep	481	1.88	0.996	0.996
Phase B/C: Transit with and without payload (aka., underway)	41.2	4.12	0.905	0.88
Phase D: Loiter	206	2.85	0.986	0.986
Phase E: Off-load	168	0.95	0.994	0.994
Phase F: On-Load	451	2.20	0.995	0.995
Total Mission Time		12.0	**0.88** (Product of above reliabilities)	**0.85** (Product of above reliabilities)

The effective ship-level MTBF is based on the threshold reliability requirement of 85% (0.85) for the 12-hour mission requirement. This overarching reliability requirement can be decomposed into reliability requirements for each phase. The predicted reliabilities in Table 1 are based on reliability block diagrams and critical system growth curves. The high predicted reliabilities (and agreement among all stakeholders that these predicted reliabilities are reasonable) for phases A, D, E, and F provide flexibility in an underway requirement. The system level requirement of 85% can be achieved with an underway (Phase B/C) reliability of 88%. Using the exponential distribution we can solve for a required underway MTBF of 32.5 hours:

$$\text{MTBF (underway)} = \frac{-4.12 \text{ hours}}{\ln(0.88)} = 32.5 \text{ hours}$$

Reliability Growth Planning Software Tool

ReliaSoft's RGA 7® software modeling tools were selected to develop the ABC 10 reliability growth plan. RGA 7® software modeling tools have been validated for use on DoD programs. The RGA 7® modeling tools employ the Crow Extended model for reliability growth projections and the Crow Extended - Continuous Evaluation model that provides for iterative reliability growth plan adjustments once test data becomes available. For reliability growth planning, the ABC 10 program applied the Crow Extended reliability growth projection module.

Reliability Growth Strategy Methodology and Assumptions

As described in Section 1.0, the ABC 10 ship reliability growth strategy involves equipment-level and ship-level assessment processes designed to capitalize on lessons learned

from the legacy ABC 1 program; equipment/systems that possess demonstrated reliability performance; and equipment, integrated and ship-level reliability growth testing to achieve the ship-level MTBF requirement. The following sections provide details for the inputs and assumptions that were applied, the systems that were assessed and the accounting of their respective test hours, and the methodology and results for reliability growth at the equipment-level and ship-level.

Inputs and Assumptions

The Crow Extended model was used to construct the equipment-level and ship-level reliability growth curves previously described at an 80% confidence level. The supporting input values, assumptions and rationale are described below.

- Input Parameter:
  - Management Strategy = 0.75.
  - Assumption: The Shipbuilder and Government will implement fixes for 75% of the failure modes that have been identified in order to reduce the likelihood that the revised product design will fail due to those particular failure modes.
  - Rationale: Extensive equipment-level testing and prior demonstrated reliability of most systems resulted in a management strategy calculated at ship-level to be 0.75.

- Input Parameter:
  - Average Fix Effectiveness = 0.70.
  - Assumption: On average, corrective measures or fixes are effective 70% of the time. At this stage of the plan, the parameter represents an average value for all failure modes subject to corrective action.
  - Rational: Crow extended modeling recommends an initial overall value of 0.70.

Equipment-Level Reliability Growth

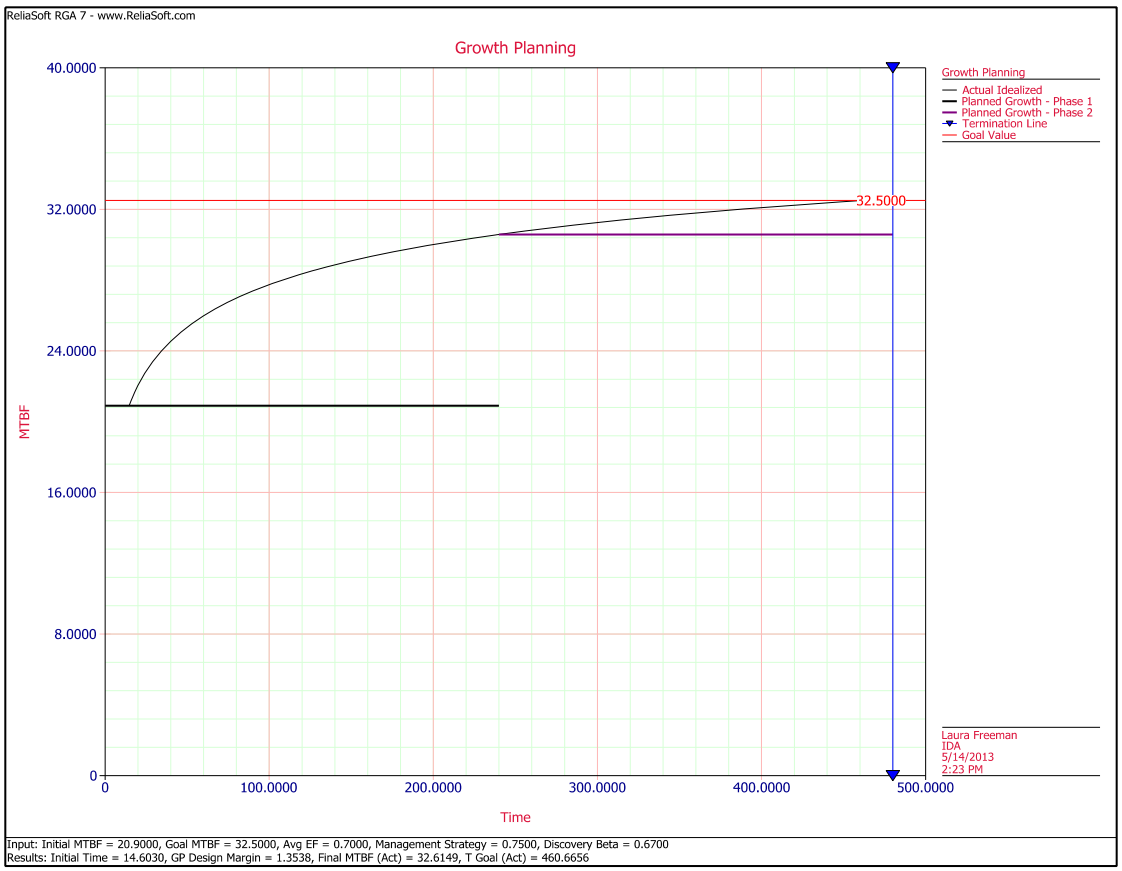
Reliability growth curves were constructed for each of the critical systems. The focus was to grow reliability on each of the sub-systems to a point where the full system level requirement can be achieved. The predicted values from column 4 of Table 2 were used as the growth goals for the equipment level growth curves. The individual reliability growth curves for the equipment level curves are in the reliability program plan.

Ship-Level Reliability Growth

The ship-level reliability growth model was developed based on the equipment-level reliability assessment. The strategy assumes 240 hours of ship-level test time required by the Shipbuilder in accordance with the contract and 240 hours of estimated reliability growth test

hours to be performed by the Government, and the input parameters described above as the inputs for the growth model

The initial MTBF was determined to be 20.9 hours based on the equipment-level assessment with a calculated management strategy of 0.75, which conservatively accounts for corrective actions/fixes expected to be in place after equipment-level testing and at entry into the Shipbuilder Ship-level test phase. The effective ship-level MTBF of 32.5 hours is reached within the 480 hour test period at a Growth Potential Design Margin (GPDM) of 1.35. Note that the GPDM value reflects the system's design maturity and required quality/reliability level as well as the program's level of aggressiveness. Figure 1 illustrates the reliability growth curve at the ship-level.



Software Algorithm Testing – Guidance

Summary

One of the three attributes for the Net-Ready KPP (NR-KPP) is that Information Technology (IT) must be able to support mission operations. For IT systems supporting operational mission threads, this means the mission threads must be executable within time periods that support the mission.

Each software system may be unique, but many computer software algorithm considerations are similar across the various systems. Software algorithms used for processing large amounts of data need to be efficient, incorporating industry best practices. This is especially important for fast searching, sorting, and merging of data files. Government testing, particularly during DT, may not look at actual data structure and algorithm coding within software modules. Instead, the software is considered a black box, with testing focused on input parameters, state variables, and results returned from the black box as well as the timeliness of receiving the outputs. The primary goal in looking at software algorithms during developmental testing is to ensure that industry best practices have been employed to ensure operational mission threads involving large data sets operate efficiently. Significant insights can be learned from focused testing in a DT controlled environment, even though the tester may not have direct access to the data structures or software code.

Algorithm performance testing should be considered during DT whenever large amounts of data are being manipulated, and the data processing time might be excessive to the point of potential mission impact.

Types of algorithms that may need performance testing

There are several types of algorithms that may need performance testing to try to ascertain whether the developer used industry best practices. Each of these categories of work needing to be performed can be categorized based on roughly how much longer the processing should take as the data set increases in size.

- Searching one or more large data sets to find data elements matching certain criteria, to include creation and execution of complex ad hoc data queries
- Sorting a large data set into a particular sorted order
- Merging two or more data sets, at least one of which is large, with resultant list possibly in some sorted order
- Optimization algorithms which seek to determine optimal routing of a delivery vehicle to visit multiple locations (for example, a optimizing a bomber route as it flies over or near multiple targets)

Industry best practices

The subject of combinatorial algorithms deals with the problems associated with performing fast computations on discrete data structures. Many types of algorithms can also be

found through simple internet searches, and Wikipedia will show the name of the algorithm and best case, average case, worst case, memory usage, and whether the algorithm is stable. http://en.wikipedia.org/wiki/Sorting_algorithm shows information for various sorting algorithms. Unless significant information is known about the data sets, industry best practices should generally use algorithms based on good average performance.

Big O notation characterizes functions such as the processing time according to their growth rates, usually providing an upper bound on the growth rate of the function. See http://en.wikipedia.org/wiki/Big_O_notation .

References

A New Approach for Delivering Information Technology Capabilities in the Department of Defense, Report to Congress, November 2010

CJCSI 6212.01F, 21 March 2012, Net Ready Key Performance Parameter (NR-KPP)

Examples

Software Algorithm Testing – Examples

Software Algorithm Testing – Examples

Software algorithms used for processing large amounts of data need to be efficient, incorporating industry best practices. This is especially important for fast searching, sorting, and merging of data files. The primary goal in looking at software algorithm performance during developmental testing is to ensure that industry best practices have been employed to ensure operational mission threads involving large data sets operate efficiently. Algorithm testing can also be used to compare the performance of proposed COTS solutions to aid in choosing the one that provides the best mix of capability and processing efficiencies. Significant insights can be learned from focused testing in a DT controlled environment, even though the tester may not have direct access to the data structures or software code. These insights should be sufficient to determine whether the developer has used industry best practices when writing data structures and algorithms. Because this type of testing does not focus on trying to manifest and detect software bugs, the process will be explained in more detail in this example. However, TEMP language can be very simple to insert.

Example TEMP language

Example 1 (generic): Algorithm performance testing will be executed during DT for those parts of mission thread execution involving the manipulation of large data sets supporting a major theater war level of scenario, where the response time may be excessive to the point of potential mission impact.

Example 2 (AOC-WS): Algorithm performance testing will be performed during DT for the Target List Merge Process that is used to create the Joint Integrated Prioritized Target List (JIPTL).

Example 3 (AOC-WS): Algorithm performance testing will be performed during DT for the auto-planning process used to determine aircraft routes to deliver weapons to multiple targets.

Types of algorithms that may need performance testing

There are several types of algorithms that may need performance testing to try to ascertain whether the developer used industry best practices. Each of these categories of work needing to be performed can be categorized based on roughly how much longer the processing should take as the data set increases in size.

- Searching one or more large data sets to find data elements matching certain criteria, to include creation and execution of complex ad hoc data queries
- Sorting a large data set into a particular sorted order
- Merging two or more data sets, at least one of which is large, with resultant list possibly in some sorted order

- Optimization algorithms which seek to determine optimal routing of a delivery vehicle to visit multiple locations (for example, a optimizing a bomber route as it flies over or near multiple targets)

Industry best practices

The subject of combinatorial algorithms deals with the problems associated with performing fast computations on discrete data structures. More information on this can be obtained through university-level course work on data structures and combinatorial algorithms. Many types of algorithms can also be found through simple internet searches, and Wikipedia will show the name of the algorithm and best case, average case, worst case, memory usage, and whether the algorithm is stable. http://en.wikipedia.org/wiki/Sorting_algorithm shows information for various sorting algorithms. Unless significant information is known about the data sets, industry best practices should generally use algorithms based on good average performance.

In this example, big O notation is used to classify algorithms by how they respond (e.g., in their processing time or working space requirements) to changes in input size (e.g. the number of data elements in a large data file). Big O notation characterizes functions such as the processing time according to their growth rates, usually providing an upper bound on the growth rate of the function. See http://en.wikipedia.org/wiki/Big_O_notation.

Performance is usually expressed in terms of the size of the data set. For example, if $\boldsymbol{n}$ represents the number of elements in a large data set, then the average performance of an algorithm operating on the elements of the data set would be expressed as being $O(\boldsymbol{n} \ln \boldsymbol{n})$ or $O(\boldsymbol{n}^2)$.

Algorithm performance testing – sorting example

When examining software data structures and algorithms in a black box environment, the goal is merely to determine whether the data structures and algorithms likely used in the software application belongs to a class exhibiting $O(\boldsymbol{n} \ln \boldsymbol{n})$ or $O(\boldsymbol{n}^2)$ average type behavior, for example.

Suppose the software function being tested is an algorithm that sorts a large data file into a particular sorted order. Good average performance for a sorting algorithm is $O(\boldsymbol{n} \ln \boldsymbol{n})$, whereas bad average performance would be $O(\boldsymbol{n}^2)$. Several industry standard sorting algorithms that exhibit “good” performance are Quick sort, Heap sort, and Merge sort. Sorting algorithms that do not exhibit good average performance would include Bubble sort, Insertion sort, and Selection sort, all of which exhibit $O(\boldsymbol{n}^2)$ average performance. Figure 1 illustrates the rate of growth, in time, based on $O(\boldsymbol{n} \ln \boldsymbol{n})$ or $O(\boldsymbol{n}^2)$ type performance, as the size of the data file increases.

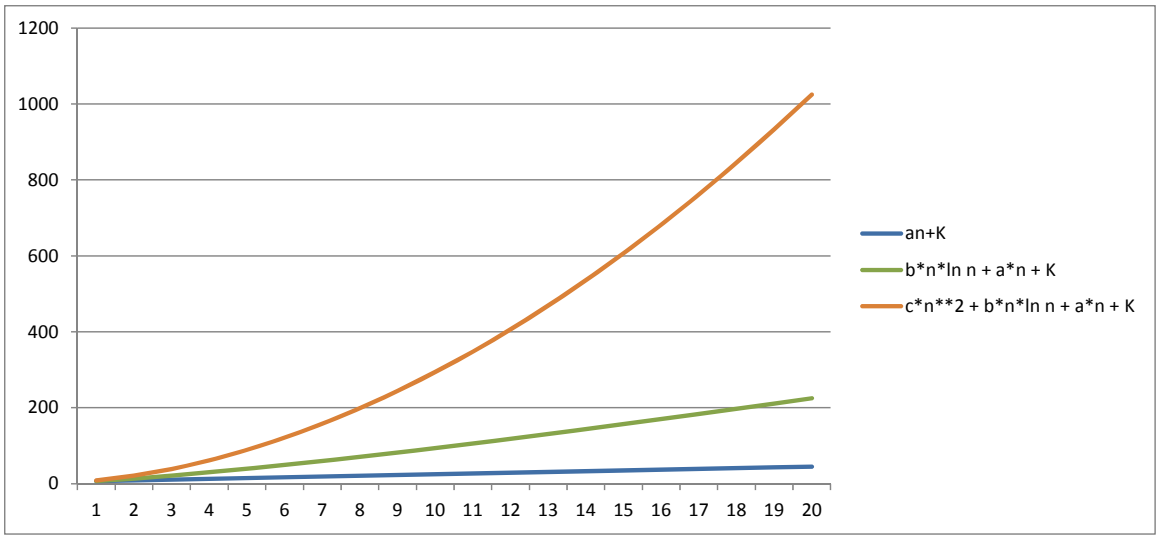


Figure 1. Performance Growth

In Figure 1, the bottom curve, expressed as "a*n + K", represents linear, or O(***n***) growth in time. The middle curve, expressed as "b*n ln n + a*n + K" represents O(***n*** ln ***n***) growth in time. The upper curve, expressed as "c*n^2 + b*n ln n + a*n + K" represents O(***n***^{2}) growth. This figure only serves to illustrate graphically how much faster an O(***n***^{2}) curve rises, compared to an O(***n*** ln ***n***) curve. Thus, a sorting algorithm that exhibits O(***n***^{2}) growth would become excessively slow as the file size increases.

Suppose during early DT testing, a process that sorts a list seems to take a relatively long time, to the point where the ability to accomplish a mission thread in a timely manner may be questionable. During early testing, the goal is not to determine whether the developer may have used Quick sort or Heap sort. Rather, it is simply to determine if the performance appears to be O(***n*** ln ***n***) or O(***n***^{2}), and whether more extensive testing or structured code walkthroughs may be required. For the example problem of sorting a large data set, four or five test runs sorting significantly different sized data sets could be completed with the response time plotted on the Y axis, against the data set size on the X axis. If the plotted data appears to be O(***n*** ln ***n***), additional testing of this kind may not be warranted. If the data suggests that the response time rises as O(***n***^{2}), then more thorough performance testing would be warranted, as well as a review of coding methods by individuals specifically trained in analysis of algorithms.

To perform more thorough performance testing, it is recommended that it be performed in a carefully controlled DT environment where "noise" contributions, such as resource contention from other sources, can be eliminated. One could start by establishing a data set of size ***n*** in a pseudo-random sorted order. This list could then be sorted, with the response time being recorded. This same list could be placed into a different pseudo-random sorted order, and then sorted again, with the time recorded. This process would be repeated several times, allowing one to compute the mean response time for sorting the lists of size ***n***. Even though the distribution of response times is not likely to be normally distributed, we use mean (instead of median) response

time since we are checking for the average performance characteristics of the algorithm used. Then, a similar process is followed for obtaining mean response times for sorting lists of size 2***n***, 3***n***, and 4***n***. The four mean response times for sorting lists of size ***n***, 2***n***, 3***n***, and 4***n*** would be used to solve a system of 4 equations, 4 unknowns for an equation of form

$$\text{Response Curve} = cX^2 + bX \ln X + aX + K$$

For the simple sorting of one large dataset example, the worst case behavior expected would be $O(n^2)$, so only coefficients "c", "b", "a", and "K" would be needed. If the estimated value for "c" is very near zero, then the algorithm is likely to be "good", since it would be exhibiting O(***n*** ln ***n***) performance. Conversely, if the estimated value for "c" is significantly greater than zero, then the sorting algorithm probably exhibits $O(\boldsymbol{n}^2)$ average behavior and is not "good" by industry best practices. "Bad" algorithms should be replaced with "good" algorithms early in development so that when the algorithms are used on large data sets during OT, significant performance degradation is not a problem.

Note we can also plot the worst case response times for sorting lists of size 2***n***, 3***n***, and 4***n***, and this may give additional insight into the worst case performance of the sorting algorithm used.

Algorithm performance testing – file merging example

Suppose the task is to merge two data files containing ***m*** and ***n*** records, respectively, and that duplicates must be removed. Suppose also that each list is sorted in some priority order, but not on a unique key field for each record. The merged list needs to be sorted in this same priority order.

For this type of problem, a "bad" algorithm would perform, on average, in O(***mn***) or possibly in O((***m***)(***m***+***n***)). It might not change the sorted order of either list, and instead would consider each element from the second list and walk through all elements of the first list to ensure no duplicates, and it would be inserted into the first list. This process would continue until all **n** records from the second list had been correctly processed into the first list, eliminating duplicates.

A "good" algorithm would perform in O((***m***+***n***) ln (***m***+***n***)), which is significantly better as ***m*** and ***n*** become large. Each list could be sorted in order by unique key, assuming the use of a good sorting algorithm. This would require, on average, O(***m*** ln ***m***) to sort the first list, and O(***n*** ln ***n***) to sort the second list. Then, it becomes very easy to make one linear pass through each list, inserting and removing duplicates, and this requires O(***m***+***n***) time. Finally, the merged list could be resorted into the desired priority order, requiring on average O((***m***+***n***) ln (***m***+***n***)), again assuming the use of a good sorting algorithm.

Similar to the approach for sorting a large data set, four or five test runs merging significantly different sized data sets could be completed with the response time plotted

on the Y axis, against the data set sizes (***m***+***n***) on the X axis. If the plotted data appears to be O((***m***+***n***) ln (***m***+***n)***)), additional testing of this kind may not be warranted. If the data suggests that the response time rises as O(***mn***) or worse, then more thorough performance testing, similar to that discussed for the sorting algorithm, would be warranted.

If the problem were merging three large data sets of size ***m***, ***n***, and ***p***, testing would consider O((***m***+***n***+***p***) ln (***m***+***n***+***p)***) type average behavior as good, with average performance worse than that being bad.

Algorithm performance in requirements or for product selection

For COTS solutions for which code development is desired to be minimized, algorithm performance can be used as one further attribute for product selection. Performance considerations could include algorithm characteristics such as best case, average case, worst case, memory usage, and stability.

Information Technology System Definition

Information Technology (IT) Systems are used in the automatic acquisition, storage, manipulation, management, movement, control, display, switching, interchange, or transmission or reception of DoD data of information regardless of classification or sensitivity.

Summary

Three metrics (whether specified as KPPs or KSAs) that cause testing issues for DoD IT systems are metrics specifying accuracy, timeliness, and data restoral. Although some aspects of data accuracy and timeliness may be assumed from the Net-Ready KPP (NR-KPP), this guidance provides separate examples to address specific accuracy, timeliness, and data restoral issues. Timeliness should be examined as part of early prototyping and discovery testing, thereby allowing for refinement of evaluation metrics between Milestone B and Milestone C. This prototyping and discovery testing should be described in the Milestone B TEMP.

CJCSI 6212.01F defines responsibilities and establishes policy and procedures to develop the NR KPP and NR KPP certification requirements for all IT and national security systems (NSS) that contain joint interfaces or joint information exchanges. The three NR KPP attributes are:

(1) IT must be able to support military operations.

(2) IT must be able to be entered and managed on the network.

(3) IT must effectively exchange information.

Normally, when JITC tests the third aspect of NR-KPP, they assume data transmission must be accurate in order to effectively exchange information, so accuracy issues would be cause to conclude the information exchange was not effective. A hypothetical NR-KPP example can be found in Appendix C of 6212.01F, so one is not included here.

Mission Assurance Category Requirements

Mission assurance category (MAC) requirements for data backup procedures and for disaster and recovery planning directly affect data restoral requirements and can be found in DoD Directive 8500.1 and DoD Instruction 8500.2.

- MAC I:
  - CODB-3 Data Backup Procedures

Data backup is accomplished by maintaining a redundant secondary system, not collocated, that can be activated without loss of data or disruption to the operation.

  - CODP-3 Disaster and Recovery Planning

A disaster plan exists that provides for the smooth transfer of all mission or business essential functions to an alternate site for the duration of an event with little or no loss of

operational continuity. (Disaster recovery procedures include business recovery plans, system contingency plans, facility disaster recovery plans, and plan acceptance.)

- MAC II:
  - CODB-2 Data Back-up Procedures

Data backup is performed daily, and recovery media are stored off-site at a location that affords protection of the data in accordance with its mission assurance category and confidentiality level.

  - CODP-2 Disaster and Recovery Planning

A disaster plan exists that provides for the resumption of mission or business essential functions within 24 hours activation. (Disaster recovery procedures include business recovery plans, system contingency plans, facility disaster recovery plans, and plan acceptance.)

Examples

Software Accuracy Evaluation – Example

Software Timeliness Evaluation – Example

Software Data Restoral Evaluation – Example

References

CJCSI 6212.01F, Net Ready Key Performance Parameter (NR KPP), 21 March 2012

DoDI 5000.02, 7 January 2015

Software Accuracy Evaluation – Examples

For software systems, the accuracy of data transmission or the accuracy of storing, maintaining, and retrieving data correctly to/from a database can be evaluated. Accuracy is also one aspect typically used as a criterion for interoperability testing by the Joint Interoperability Test Command.

Evaluation of Data Transmission Accuracy

Critical technical parameters (CTPs) should be used during DT to address engineering goals to identify, isolate, and fix data transmission channels that may not be working correctly. During OT, the accuracy KPP should measure a critical aspect of performance to ensure the operational mission can be accomplished.

Evaluation of Data Storing, Maintaining, or Retrieval Accuracy

When addressing storing, maintaining, and retrieving data correctly to/from a database, CTPs could be used to address individual aspects. If the system has built-in redundancy or accuracy correction methods to help address accuracy problems, then CTP testing could focus on each method separately. KPP testing during OT should account for the redundancy or correction methods provided users use them correctly, with the overall focus on a critical aspect of performance to ensure the operational mission can be accomplished.

An accuracy measure is particularly subject to data skewing during operational testing because users tend to avoid known failures and instead rely on methods that seem to work correctly. Data accuracy is routinely and incorrectly tested as

<number of errors> / <number of transmissions>

When measuring accuracy, the correct metric is

< number of elements with any error > / < number of elements >

An element is typically considered a data record, consisting of a number of data fields. Requirements are often ambiguous concerning data accuracy, and OTAs should seek clarification from the user representative so that the TEMP can be used to unambiguously build failure definition scoring criteria.

Hypothetical Example

Suppose our system transmits 100 data records, and each data record has 50 data fields. Suppose we observe the following: only 99 data records are received, and of those, 98 are totally correct (i.e. all 50 data fields correct in each of the 98 records). The one record received, but not totally correctly, has 5 data fields not correct. What is the point estimate of data accuracy, and how many data samples are counted? DOT&E interprets this as having 98 correct records, and 2 records not correct (1 not received, 1 containing errors). The point estimate would be 0.98, and there are 100 samples. The method of counting successes and failures should not be left ambiguous in the TEMP.

Software Accuracy Evaluation – Examples

Accuracy measures are particularly prone to skewing of samples during OT, since users tend to not repeat known errors. The following hypothetical example demonstrates this.

Hypothetical example of data skewing when testing accuracy:

Suppose the requirement is to return accurate track information to the user 95 percent of the time when the user clicks on a track displayed on the GCCS Common Operational Picture. Suppose the COP is displaying half ship tracks, half air tracks. Suppose if the user clicks on a ship track, the user receives an accurate data record, but whenever the user clicks on an air track, the user receives a record with incorrect data. Severe skewing would occur if the user were to click on an air track, note the error, and then click on one more air track to verify the error. Then the user might proceed to click on 85 ship tracks. While 85 successes out of 87 trials may meet 95 percent success rate with 80 percent level of confidence, the problem is that the data samples themselves are not independent, since the selection of tracks on which to click was not random and not representative of the population of tracks.

A key engineering goal of these KPPs is to identify, isolate, and fix the channels or software that are not working correctly. Accordingly, testers should also report any inaccuracies at the data field level. A report that details the errors found in each element will provide the PM with information needed to fix issues and will also be easily summarized with the correct metric.

Accuracy and the Net-Ready KPP

Both the first and third attributes of the Net-Ready KPP may require accuracy measures to help resolve the NR-KPP. Shown below are several accuracy KPPs, with a brief note about how they might be related to the NR-KPP. A separate note indicates if an ambiguity of how to measure data accuracy should be clarified.

Example 1

From Air Operations Center – Weapon System (AOC-WS): 99 percent of original content conveyed [assume correctly] to other divisions & process stations.

This KPP could be aligned under the third attribute of the NR-KPP, which requires the IT system to effectively exchange information. It is not clear whether the “content” is measured at the data field, or data record, level. This ambiguity should be resolved.

Example 2

From AOC-WS: Match air, space and information support resources to operations, Accuracy $\geq$ 95 percent (threshold).

This KPP could be aligned under the first attribute of the NR-KPP, which requires the IT system to be able to support military operations.

Example 3

From Global Combat Support System – Joint: Provide 95 percent accurate data from authoritative source.

This KPP could be aligned under the third attribute of the NR-KPP, which requires the IT system to effectively exchange information. It is not clear whether the data accuracy is measured at the data field, or data record, level. This ambiguity should be resolved. If not specified, DOT&E would assume at the data record level.

Example 4

From Global Combat Support System – Army: GCSS-Army must maintain an accurate funds available balance; allow verification of funds availability, and provide alerts for transactions that will exceed fund authorizations. Threshold: Based on a sampling, GCSS-Army achieves funds accuracy 95 percent of the time.

This KPP could be aligned under the first attribute of the NR-KPP, which requires the IT system to be able to support military operations.

Example 5

Joint Command and Control (JC2): Track to asset level visibility: Reports or queries will be delivered in less than 7 seconds from the time query is issued at 99.999 percent accuracy.

This KPP could be aligned under the third attribute of the NR-KPP, which requires the IT system to effectively exchange information. It is not clear whether the data accuracy is measured at the data field, or data record, level. This ambiguity should be resolved. If not specified, DOT&E would assume at the data record level. Even with no failures, 160,943 successful samples would be required to meet the accuracy requirement at the 80 percent level of confidence. DOT&E would recommend adjusting the requirement to a level that is affordable to test.

Mission Assurance Category Requirements

Mission assurance category (MAC) requirements for data backup procedures and for disaster and recovery planning directly affect data restoral requirements and can be found in DoD Directive 8500.1 and DoD Instruction 8500.2.

- MAC I:
  - Continuity of Operations – Data Backup (CODB)-3 Procedures

Data backup is accomplished by maintaining a redundant secondary system, not collocated, that can be activated without loss of data or disruption to the operation.

  - CODP-3 Disaster and Recovery Planning

A disaster plan exists that provides for the smooth transfer of all mission or business essential functions to an alternate site for the duration of an event with little or no loss of operational continuity. (Disaster recovery procedures include business recovery plans, system contingency plans, facility disaster recovery plans, and plan acceptance.)

- MAC II:
  - CODB-2 Data Back-up Procedures

Data backup is performed daily, and recovery media are stored off-site at a location that affords protection of the data in accordance with its mission assurance category and confidentiality level.

  - CODP-2 Disaster and Recovery Planning

A disaster plan exists that provides for the resumption of mission or business essential functions within 24 hours activation. (Disaster recovery procedures include business recovery plans, system contingency plans, facility disaster recovery plans, and plan acceptance.)

Example TEMP entry for MAC-I System:

Global Command and Control System – Joint (GCCS-J) is a command and control system rated as Mission Assurance Category I. The Joint Operations Planning and Execution System (JOPES) within GCCS-J has four primary, fully redundant strategic server enclaves (SSEs), with data also fully replicated across all four SSEs. The following criteria for JOPES have been summarized to capture the most relevant parts.

3.2 Evaluation Framework (for JOPES)

- System Availability: more than 99.7 percent.
- Disaster Recovery. Mean time to restore function (MTTRF) on any single system shall be within 24 hours. JOPES SSE database recovery backup must be within 12 hours.

Software Data Restoral Evaluation – Examples

- System ability to support mission essential JOPES activities (minimize in effect) following loss of one or more sites:
  - Capable of supporting users after loss of 50% of the sites for not less than 96 hours.
  - Capable of supporting users after loss of JOPES Network Support for not less than 4 hours.
- Strategic servers will have the capability to be mirrored, maintain data accuracy, and process data consistently.
  - Most current update available in a server to an authorized GCCS-J application user within 3 minutes.
  - JOPES SSE - Upload and network, to all available servers, a 150,000 Time Phased Force Deployment Decision (TPFDD) in an average of 8 hours.

Example TEMP entry for MAC-II System:

Global Combat Support System – Army (GCSS-A) is a tactical logistics data system rated as Mission Assurance Category II. GCSS-A has a primary server center and an alternate Continuity of Operations (COOP) center. Data is mirrored from the primary site to the alternate site at some specified interval of time which does not exceed four hours. The data restoral KPP for GCSS-A addresses both the disaster recovery time (24 hours threshold) as well as the mirroring frequency (not more than 4 hours).

3.2 Evaluation Framework (for GCSS-A)

(other information goes here)

KPP or KSA	Threshold	Objective
1.Continuity of Operations and System Restoration	GCSS-Army shall recover GCSS-Army critical capabilities within 24 hours (the MAC II requirement) of declaration of a disaster to a state not more than 4 hours prior (the data mirroring frequency) to disaster.	GCSS-Army shall recover GCSS-Army critical capabilities within 24 hours of declaration of a disaster to a state not more than 2 hours prior to disaster.

Software Evaluation – Guidance

This case study refers to the notional Information Technology ***Program X*** which is a world-wide web-based system accessing multiple databases. This case study is designed to illustrate the complexity of comprehensively specifying and measuring a responsiveness, or timeliness, KPP. The scope of Program X is limited to the large tan rectangle in the center of Figure 1. The red and blue circles represent data collection points for measurement of timeliness.

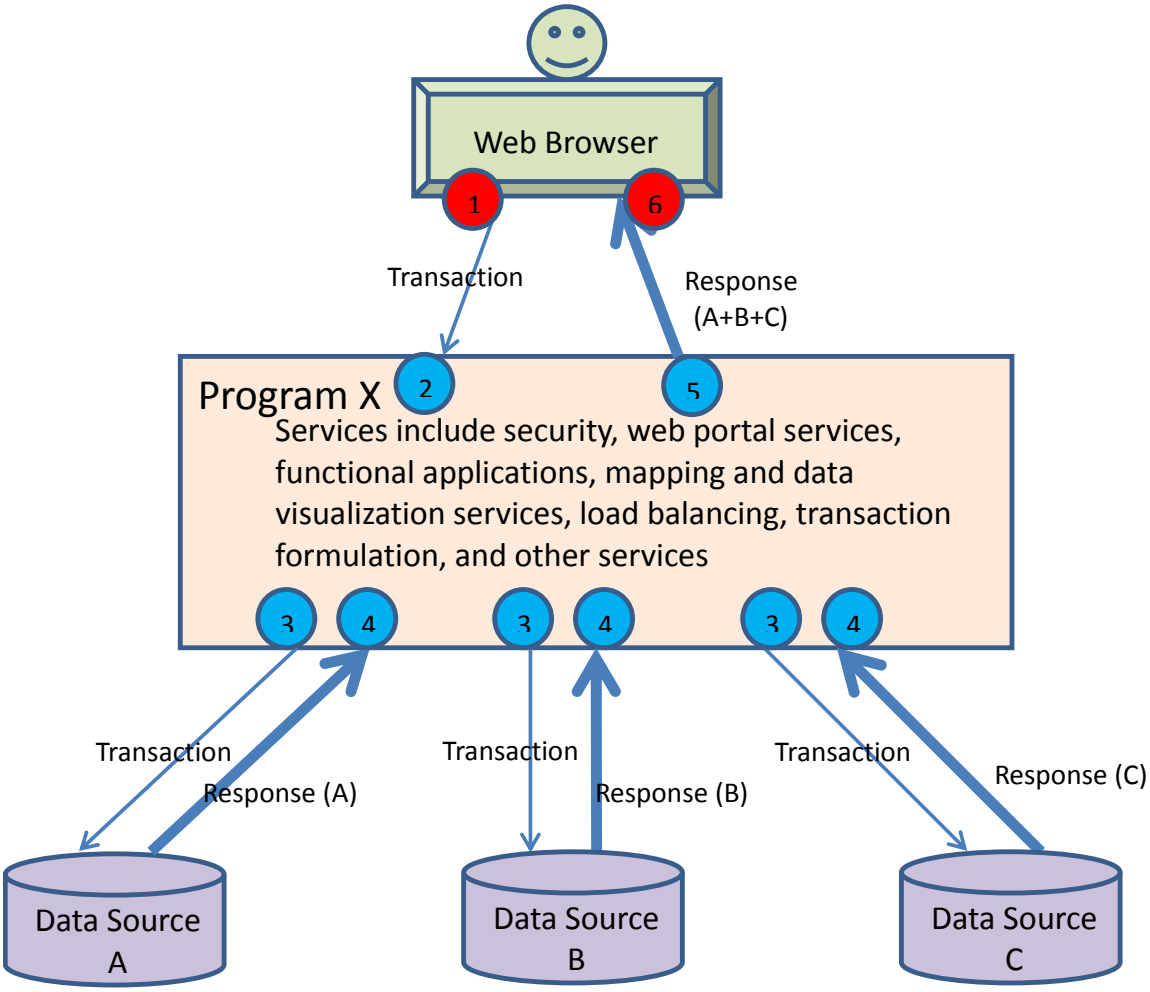


Figure 1: Program X Concept

Worldwide users access Program X services through their web browsers, accessing and sending query requests to the central Program X web site. Program X software forms the queries to access one or more underlying databases (not necessarily resident at the Program X site). Query information is returned to the Program X portal, which then forms the response to the user, finally sending the information to the user's web browser for display on web pages.

The responsiveness, or timeliness, KPP for Program X is shown is Table 1 below. Unfortunately common, this sort of KPP presents challenges for evaluation of Program X performance.

Table 1. Responsiveness KPP

1. KPP	2. Threshold	3. Objective
4. Responsiveness 5. (Asset Visibility)	6. single/multiple queries must be accomplished in less than 60 seconds, 95% of the time.	7. single/multiple queries must be accomplished in less than 30 seconds, 95% of the time.

There are several difficulties associated with a KPP like this:

- All queries, whether simple or complex are required to be completed in 60 seconds. As stated in Table 1, the KPP fails to describe the number of underlying databases that need to be accessed. The KPP also does not state how many simple and how many multiple queries might be expected in a day, week, or month. Both of these undefined factors will influence overall query timeliness.
- The KPP does not define the amount of data expected to be returned. It could range from zero or one record per query to well over 100,000 records.
- The KPP does not mention the possibility that some large queries that generate extremely large amounts of data could be satisfactorily processed during off-peak hours.
- The KPP does not define or accommodate the differing responsiveness of external databases that are beyond the influence of Program X. Other factors that could influence Program X responsiveness include the placement of external data servers, the location of users, network bandwidth, encryption, network reliability, packet retransmission, network loading, and information assurance threats.
- The KPP does not define how the system should perform if an external data source is temporarily inoperable or not responsive.
- The KPP can be evaluated by measuring the proportion of queries that meet the 60-second threshold. This method gives no credit for extremely fast queries and reduces our ability to understand how factors contribute, good or bad, to timeliness.
- Simple methods to measure responsiveness (time from red #1 to red #6, as shown in Figure 1) might be to use a stopwatch at the user terminal. This method may be reliable to within 1 second and inexpensive to use during testing, but is not good for helping a PM ensure the system remains responsive after fielding. Nor is it very useful for reconstructing network-wide symptoms and correlation of events among sites, as it only measures elapsed time and not absolute system start and stop times.

Refinement of Requirements

Early in the development of KPPs, the requirements community, program engineers, and the test community should draft and include more contextual information in the specification of the KPP. This contextual information will assist in Design of Experiments methods and become DOE factors for testing and early prototyping. Early prototyping could help characterize achievable performance levels and help shape the KPP.

The Milestone B TEMP should describe the early prototyping and DOE approach to characterize the key factors affecting the timeliness KPP. These factors and results should be used to adjust the KPP for the Milestone C TEMP.

Continuation of the Program X Case Study

Suppose that early testing revealed that three factors (the number of underlying databases needing to be queried, the location of the user (overseas or CONUS), and the number of records to be returned by the query) had a significant effect on query response time (RT). Suppose also that we learned the following information:

- Factor 1: When more than one database is queried, there is an increase in RT of 10 seconds per database queried.
- Factor 2: RT for queries from overseas users take roughly two times as long as queries submitted by CONUS users.
- Factor 3: RT increases 1 second for every 100 records returned.

Using these early test results, the KPP could be refined using a formula based on these three critical factors plus some constant K.

RT <= User Loc * [(10 * number of databases) + (Records Returned / 100) + K]

In this formula for the KPP, we could apply an overall multiplier of 2.0 for an overseas user, compared to 1.0 for a CONUS user. We could add 10 seconds per underlying database queried for the complexity factor, and one second per 100 records returned to address the third factor for the records returned. Then, the KPP requirement in the Milestone C TEMP could be expressed as 95 percent of the time meeting this formula.

Unresponsive external databases could be addressed through a requirements change by requiring the system to time-out after a period of time, and explicitly treating these responses as “no test” for purposes of meeting the timeliness KPP. Whether the system correctly timed out and responded accordingly to the user would be tested as a separate measure. The program manager could also implement a status board showing the up/down status of each underlying database to help address this problem (this was done for Program X). When considering overall mission accomplishment, too many instances of system timeout due to underlying database failures would negatively affect overall mission accomplishment, and thus they cannot simply be ignored. Other methods of addressing slow response time may be to include progress bars or the ability to spool the query or run it in batch mode. These considerations are all worked collaboratively between the user requirements representatives and the program engineers.

The next improvement would be to provide the OTA with historical data concerning the relative frequencies of various types of queries, and amounts of data expected to be returned. This would allow the OTA to construct a scenario for OT that would exhibit operationally realistic exercising of the system. For example, guidance on testing the KPP might state that simple queries are executed against Databases A, B, and C in a 20, 30, and 50 percent ratio, and that complex queries comprise 10 percent of the total queries and involve only two of the three databases (again at the summed ratio similar to the simple queries). Number of records returned could be expressed using a histogram, based on historical data. Network loading and contention could be based on historical data, if known.

Table 2 shows the number of data samples required to meet various pass/fail criteria, assuming an 80 percent level of confidence.

Table 2. Binomial Samples Needed

	Threshold Success Rates				
Failures	**80%**	**90%**	**95%**	**98%**	**99%**
0	8	16	32	80	161
1	14	29	59	149	299
2	21	42	85	213	427

When each data sample containing the response time data is reduced to a binary "pass/fail" data point, much information is lost. Simplistic methods of specifying performance requirements that reduce continuous data to binary pass/fail data may be acceptable for Milestone B TEMPs, but should be avoided in Milestone C TEMPs. For software systems operating in a network environment, response times should not be assumed to be normally distributed. Figure 2 shows a histogram for queries accessing a certain database that returned in 50 seconds or less. The tail of this data, not shown, would extend out to include two points just over 360 seconds (reflecting the timeout value). This data is not normally distributed. Early prototyping and engineering studies, combined with legacy data, should be used to better characterize expected timeliness data. This should allow specifying and testing response time requirements using continuous methods, thereby reducing sample sizes. Figure 3 shows a histogram for queries accessing a different database, and data has been binned in the histogram in groups of 10 seconds to better show that while the tail seems to get smaller and smaller, out at the "timeout" point, there can be a significant number of data samples (18 samples in this case). It is recommended that this aspect of system performance be considered for Critical Technical Parameter testing, and carefully addressed during operational testing if the frequency of timeouts affects overall mission accomplishment.

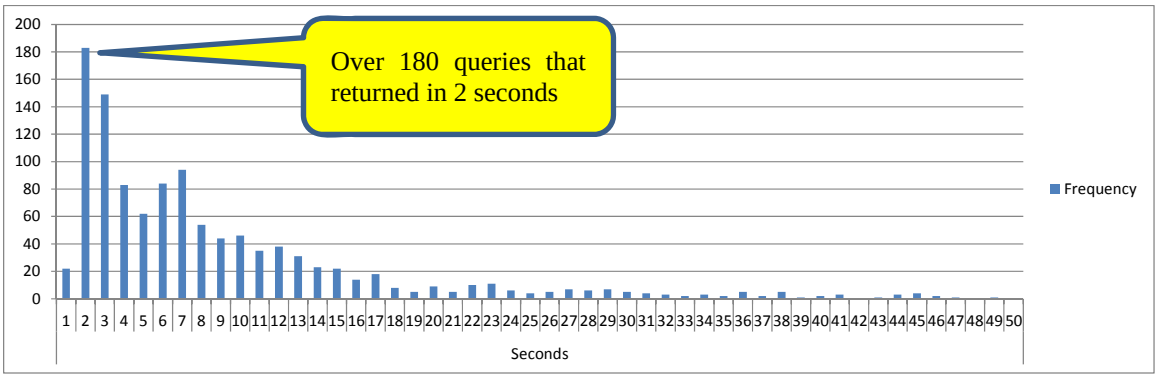


Figure 2. Data Histogram (Each bar shows number of queries returning in some number of seconds, as measured on the X axis)

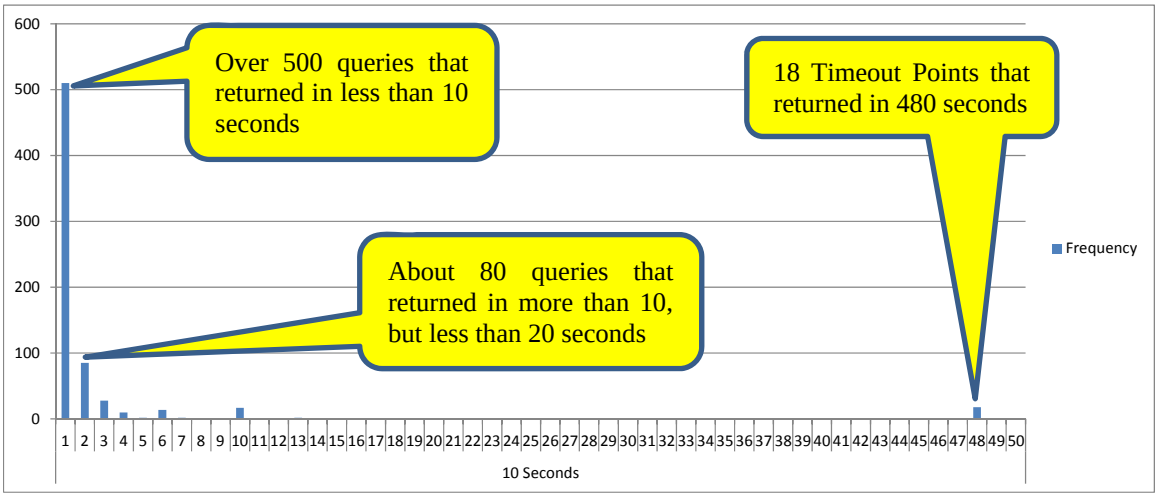


Figure 3. Data Histogram Showing Timeout (Each bar shows number of queries returning in some number of tens of seconds, as measured on the X axis)

The next suggested improvement concerns how to measure and report timeliness, not only during a few snapshots in time during OT, but also after fielding. If responsiveness is truly a KPP, then it is worth measuring and reporting on a monthly or quarterly basis, and should be accomplished by non-intrusive, automated means.

The Program X servers would be expected to be able to capture computer system time data and also the key factors affecting timeliness at the blue measurement points, but probably not at the red measurement points. System timeliness requirements are specified from an operational mission context which is what the user sees (meaning at the red points). The OTA can easily collect timeliness measurements at the blue points #2, #3, #4, and #5, but this does not represent the total waiting time experienced by the user, and hence cannot be used to fully answer the KPP requirement. Stopwatch methods tend to be limited to capturing relative elapsed time, and do not account for clock synchronization issues throughout the network. Thus, they are not very helpful for examining system performance across a network. They are also not conducive to continued performance monitoring post-fielding.

To help overcome the need to use stopwatches, there are commercially available methods for measuring web site performance. Two methods of gathering response time data are from Field Metrics and Synthetic Measurement. Field Metrics measure response time from real user traffic but have the advantage over stopwatch data in that they capture start and stop times using the system clock. This method relies on instrumentation of the pages, or toolbars to collect and log data. Field Metrics methods should be encouraged for Milestone C measurements that are truly of KPP importance, and these methods also allow continued monitoring of timeliness data post-fielding. Recording of user screens using the Defense Collaboration Services (DCS) collaboration tool is a field metrics method that can also be used to collect full round-trip response time during testing. However, use of DCS puts significant extra load on the system and cannot be used for monitoring system performance on an on-going basis. It can, however, be very useful for system debugging. Synthetic Measurement involves loading pages in one of a myriad of tools designed to collect metrics. Synthetic Measurement may be appropriate for early prototyping work when trying to identify the DOE factors, but it is important to collect the measurements over operationally realistic environments and not just in a lab. Finally, if system performance is critical for a network system, it is recommended to also test for overall system clock synchronization throughout the network being within a specified delta of Global Positioning Time.

Scientific Test and Analysis Techniques – Guidance

Background

The authors of the TEMP should employ scientific test and analysis techniques to develop a defensible analytical basis for the size and scope of the T&E program. Recent guidance in this regard has focused on Design of Experiments (DOE) for sizing test events, but there are other tools available in the scientific test planning toolbox. This guidance provides an overview of the scientific test and analysis techniques (STAT) that may be used in planning, conducting, and analyzing a test or experiment. Additionally, this guidance summarizes key content that should be summarized in the TEMP.

Guidance

Any program that applies scientific test principles should commence doing so early in the test planning process. The program should assemble a test and evaluation working integrated product team (T&E WIPT) of subject matter experts who can identify the primary quantitative mission-focused measures (in DOE terminology: response variables) of interest that will characterize the performance of the system in the context of a mission-focused evaluation. The T&E WIPT should identify environmental and operational factors that are expected to drive the performance of the system, as well as the levels of these factors (i.e., the various conditions or settings that the factors can take).

With these measures and factors in mind, the T&E WIPT should populate the Developmental and Operational Evaluation Frameworks and decide which of the following analysis techniques will best ensure adequate coverage of all important factors while evaluating the quantitative mission-focused measures through planned testing. The testing strategy should be iterative in nature throughout integrated test and evaluation to ensure an adequate Initial Operational Test and Evaluation (IOT&E). The testing strategy should accumulate evidence that the system performs across its operational envelope before and during IOT&E.

Elements of Scientific Test Design for the TEMP

A brief overview of the test design philosophy should be outlined in Section 3 of the TEMP. The information content may vary depending on the Milestone that the TEMP is supporting. Table 1 outlines information content that is appropriate for each milestone. Systems with legacy data will be expected to include more detail and have more robust test designs. Additionally, if previous test data will be used to augment operational testing, the methodology for using that data should be discussed. See the Bayesian Example for an example of how previous test data can be used to scope future testing. The details of each of the test designs should be provided in a supporting STAT appendix to the TEMP. Elements of any test design (regardless of statistical methodology) should include the following:

- The goal of the test (experiment).
- Quantitative mission-focused measures (also known as quantitative mission-oriented response variables) for effectiveness, suitability, and survivability. See Quantitative Mission-Focused Measures Guidance.

- Factors that affect those measures of effectiveness, suitability, and survivability. See Integrated Survivability Evaluation Guidance.
- A method for strategically varying factors across developmental, operational, and live fire testing with respect to responses of interest. See Integrated Testing Guidance.
- Statistical measures of merit (power and confidence) and corresponding effect sizes on the relevant quantitative mission-focused measures (response variables) (i.e., those for which doing so 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.

The execution of the test, including run plans/order, should be discussed in the Test Plan.

Commonly, the system under test (SUT) is a complex system with multiple missions and functionalities. The test design should reflect the complexity of the system. Often, multiple test designs will be necessary to fully characterize SUT mission performance. This might also require multiple experimental designs to capture all stages or aspects of mission execution.

Table 1: Test Design Information Content for the TEMP

Milestone Supported	Information Content
A	Responsibilities of T&E WIPT for test design purposes The goal(s) to be addressed at each stage of testing Quantitative mission-focused measures for each goal/question An initial listing of factors for each quantitative mission-focused measure Language for the overall testing strategy, including: • Screening experiments to ensure important factors are considered in operational testing • Sequential experimentation Test designs to support resourcing for limited user tests (LUT), operational assessments (OA), and IOT&Es • While test designs may be very preliminary at MS A, it is essential that scientific test designs be used as a basis to estimate resources for a Request for Proposal near MS A. Therefore, special attention should be paid towards making sure adequate resources are allocated for long lead items (e.g., number of targets, weapons, specialize range capabilities, etc.).

B	Identify responsibilities of T&E WIPT for test design purposes The goal(s) to be addressed at each stage of testing Quantitative mission-focused measures for each goal/question Refined listing of factors and levels for each quantitative mission-focused measure Test designs to support resourcing for limited user tests (LUT), operational assessments (OA), and IOT&E • Test Designs should be updated from MS A to account for any new information. Language for the overall testing strategy, including: • Screening experiments to ensure important factors are considered in operational testing • Sequential experimentation
C	Identify responsibilities of T&E WIPT for test design purposes The goal(s) to be addressed at each stage of testing, focusing on IOT&E Quantitative mission-focused measure for each goal/question Refined listing of factors and levels, based on prior testing and the operational mission, for each quantitative mission-focused measure Details on how the factors and levels will be varied and controlled during each stage of testing Complete test designs to support resourcing for IOT&E Language for the overall testing strategy, including: • How previous knowledge is being used to inform IOT&E test planning. • Analysis plans to support power calculations

Scientific Test and Analysis Tools

There are many different scientific and statistical design and analysis tools that are appropriate to use in T&E. The most common tools and methods for T&E include:

- Design of Experiments (see DOE Guidance)
    a. DOE TEMP Body Example
    b. DOE Appendix D Artillery Example
    c. DOE Appendix D Precision Guided Weapon Example
    d. DOE Appendix D Software Intensive System Example
- Observational Analyses (see Observational Analysis Example)
- Survey Design and Analysis
- Reliability Test Planning
- Hypothesis Testing
- Statistical Methods for Validation of Modeling and Simulation
- Bayesian Analysis Methodologies (see Bayesian Guidance and Bayesian Example)

The STAT Appendix is an appropriate location to discuss any of these tools if they were

used to determine the required testing outlined in the TEMP.

Design of Experiments (DOE) requires that the tester has control over at least the important factors when executing a test. There are many types of experimental designs that allow for different test constraints. A list of common test designs and their applicability to operational testing is here. In addition to the content of Table 1, the specific test design strategy should be discussed if DOE is being used to plan the operational test. However, it is not always possible for a tester to control all test conditions. In these cases we can still analyze the adequacy of the test using observational analyses. It is still essential that the TEMP outline the key information in Table 1 as well as a minimum acceptable test size. If historical data is available, it should be used to approximate the test scope.

Surveys of operators and maintainers are essential aspects of operational testing. See Survey Design and Administration Guidance to read about what survey content belongs in the TEMP. The detailed surveys, administration plan, and other related information should be included in the Test Plan.

Simple hypothesis tests are rarely appropriate for scoping the operational test; however they do provide a methodology for assessing if the reliability testing is adequate. See the reliability test guidance for additional guidance.

Finally, Bayesian methodologies may be appropriate in cases where there is a lot of additional test data (operationally realistic developmental testing, operational assessments, etc.) that will be incorporated into the operational evaluation. In these cases, the TEMP should also discuss what information will be carried forward, the analysis methodology for doing so, and what testing must still be conducted in OT.

Guidance

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

Best Practices for Assessing the Statistical Adequacy of Experimental Designs Used in Operational Test and Evaluation, July 23, 2013

Guidance on the Use and Design of Surveys in OT&E, June 23, 2014

Discussion on the Use and Design of Surveys, 24 February, 2015

Discussion on Including Neutral Responses on Survey Questions, April 2, 2015

Survey Pre-Testing and Administration in Operational Test and Evaluation, 6 January 2017

Guidance on the Validation of Models and Simulation used in Operational Test and Live Fire Assessments, March 14, 2016

Clarifications on Guidance on the Validation of Models and Simulations used in Operational Test and Live Fire Assessments, 17 January 2017

References

Scientific Test and Analysis Techniques (STAT) in Test and Evaluation (T&E) Implementation Plan, DASD(DT&E), January 2012.

Montgomery, D. C. (2009), *Design and Analysis of Experiments*, John Wiley and Sons

Myers, R. H., and Montgomery, D. C. (2002), *Response Surface Methodology: Process and Product Optimization Using Designed Experiments*, John Wiley and Sons.

Common Test Designs

Design Type	Description and Applicability for Operational Testing
Full Factorial (2-level)	A design with two or more factors, each with two levels, where all possible factor combinations are tested at least once. Typically used in when the total number of factors and factor combinations is not too large (e.g., 3-5 factors). A full factorial design allows for the estimation of all main effects and interaction terms in the model. Full factorial designs tend to provide too much information (over powered) for large numbers of factors.
Fractional Factorial Design	A fractional factorial design consists of a strategically selected subset of runs from a full factorial design Useful when: Large number of factors and it is uneconomical to test every possible factor combination In screening experiments to identify the primary factors Typically, fractional factorial designs that allow for two-way interactions are adequate to characterize system performance Leverages sparsity of effects: most systems are dominated by some of the main effects and low order interactions
Full Factorial Design with center points	Center points add the ability to check for curvature across continuous factors Provide small increases to statistical power
Full Factorial (2-level) replicated	Replication can be used to increase statistical power and provide estimates of variation within a condition Often not possible in cost constrained operational tests In a constrained resource environment it is better to cover more of the operational space than to replicate (i.e., do not eliminate a factor for the sake of replication) A common middle ground is to only replicate a subset of the design (e.g., a center point)
General Factorial	Similar to a two-level factorial design a general factorial design has two or more factors, each with two or more levels, where all possible factor combinations are tested at least once. Only possible when the number of factors is not too large (e.g., 3-5 factors). Allows for the estimation of all main effects and interaction terms in the model. Less powerful as you add more levels to each factor For continuous factors, two-levels provides the highest power

Common Test Designs

Response Surface Designs	Response Surface Methodology is a collection of experimental designs Originally invented by the chemical industry to conduct sequential experimentation for process optimization Evolved to be a broad class of designs that characterize system performance Robust test design methodology fits second order models including quadratic effects for flexible performance characterization Types of Response Surface Designs: Central Composite Design, Face Centered Cube Design, Small Central Composite Design, Box-Behnken Designs, Optimal Designs
Optimal Design	Optimize the test points for a known analysis model and sample size Optimal designs are useful: Large number of factors Highly constrained design region (disallowed combinations of factors) Large number of categorical factors The optimal design fallacy Designs that are optimal under one criteria might be far from optimal under another criteria Optimal designs are similar to factorial designs and response surface designs for similar analysis models Always build in extra points to optimal designs to allow for incorrect model assumptions and statistical power
Combinatorial Designs	Highly efficient test designs that are commonly used in testing software Do not support cause-and-effect analysis like all of the above design types, rather they cover the space very efficiently to look for problems. Root-cause analysis must be conducted if problems are found

Bayesian Methods – Guidance

Background

Leveraging information from various sources to estimate reliability using Bayesian methods is becoming more common and has many advantages. Most notably, multiple sources of prior information (e.g., operationally relevant developmental testing or operational assessments) can be incorporated; complex systems (and their structures) can be analyzed without seriously increasing the computations; and uncertainty intervals are straightforward to calculate and interpret. These techniques require thought and understanding of both the system and the statistics. In addition to the actions taken to develop a classical test plan, a Bayesian approach needs to determine a prior distribution (what information to leverage) and establish an analysis framework (how to incorporate prior information).

Data Quality

Relevant prior information to be incorporated in OT analysis may include previous developmental or operational test data, engineering analyses, or information from modeling and simulation. The origin and quality of prior information should be indicated in the TEMP.

Any type of prior test data is reasonable to use in prior construction. However, the relevance of the prior information to the current data (i.e., DT data for OT analysis) should impact how heavily the prior information is weighted. The main point is not to introduce a bias into the current analysis if past test data do not support the results of the current test data.

Incorporating Prior Information

The prior data should always be used to aid in choosing which areas within the operational envelope to include in testing: knowing what combination of settings could be difficult or easy for the system, etc. Another approach to use the prior information is to shorten the test (or accepting a less powerful test) with the knowledge that the prior information will serve as the basis for the analysis. Assurance testing is a formal procedure to plan a test which combines information from various sources to reduce the amount of testing required to meet a requirement. See the references below for more information. For some systems, it might be reasonable to plan the OT to focus on the additional operational effects (i.e. the addition of an operator), because the prior information gives a good idea of the performance of the base system. Extremely informative priors which allow virtually no learning from OT data must be avoided.

Analysis Plan

The TEMP must indicate the intended analysis plan for the OT data when a Bayesian approach is proposed. That analysis plan should establish how the prior information is to be included in the operational evaluation.. For example, in a reliability analysis using Bayesian methodologies the operational test evaluation would specify the distribution that will be used to analyze the data (e.g., Exponential or Weibull), the prior distribution , and how it was derived from existing data.

General Guidelines for when Bayesian Methods are Appropriate

Bayesian Methods – Guidance

There are a few indications that Bayesian methods will be applicable and worth the additional effort:

- When relevant and defensible prior information is available. The Bayesian example illustrates a case when extensive information from DT is available at the subsystem level and the operational testing focuses on the integration of these subsystems with an operational user. Even when including prior information, the prior must have enough variability to allow the estimates to move away from what was previously seen if the data support such values.
- When assessing system or kill-chain reliability. Such analyses generally involve combining information from many areas or subsystems under a possibly complex system structure. Obtaining interval estimates for any system model is straightforward in a Bayesian analysis, but not in traditional analyses. Moreover, if any subsystem or component in a kill-chain or system analysis has zero failures, point and interval estimates are still attainable.
- To avoid unrealistic point estimates and to obtain interval estimates when measuring mean time between failure for short tests or highly reliable systems and zero failures are expected.

Bayesian Principles

Some basic, overarching principles to consider when planning a Bayesian analysis:

- Start with the properties of the parameter of interest: if a parameter needs to be positive, choose a distribution that is also non-negative.
- Decide on what prior information to use and how relevant it is to operational test evaluation.
- Allow for the analysis to change freely based on the data observed in operational testing.
- Check impact of your prior assumptions: explore prior predictions for bias and re-check in the analysis with a sensitivity study. A good model should be fairly robust to prior specifications.

Bayesian Example

References

Hamada, M.S., Wilson, A.G., Weaver, B.P., Griffiths, R.W., and Martz, H.F. (2014) "Bayesian Binomial Assurance Tests for System Reliability Using Component Data," *Journal of Quality Technology*, 46, 24-32.

Hamada, M.S., Wilson, A.G., Reese, C.S., Martz, H.F. (2008). *Bayesian Reliability*. New York: Springer.

Bayesian Methods – Example

System Description

A new mobile lab system is intended to analyze environmental samples for the presence of chemical, biological, and radiological material, and report the analytical results to directly support commander's force protection and force health surveillance decisions. Each subsystem (chemical, biological, and radiological) is comprised of a collection of components of various sensitivity, speed, and cost to run. Each set/system will be tailored to the specific operational user and their mission needs by incorporating specific capabilities from a common suite of Commercial-Off-the-Shelf (COTS) and Government-Off-the-Shelf (GOTS) analytical technologies and components. KPP performance requirement for each subsystem is to detect 85 percent of samples that come into the lab.

Prior Information

The subsystem components have completed multiple phases of testing to determine detection performance curves. Each phase increases the operational relevance of the testing: Phase 1 tested various targets at different concentrations in a pristine matrix on each component and Phase 2 tested various targets at different concentrations in operationally representative matrices such as soil, food, or swabs. There have been approximately 5500 tests on the subsystem to characterize the detection performance of the components for the target matrix combinations. Prior information to be incorporated in the analysis comes exclusively from prior test data that will be down-weighted for OT to take into account departures from operational realism (i.e. lab technicians versus soldier operators).

A logistic regression was used to analyze the Phase 2 data for with the factors target, matrix, and concentration. Dispersed priors are placed on each regression coefficient to obtain performance curves for each of the components and the target/matrix combinations:

$$\boldsymbol{logit(P_D) = \beta_1 * conc + \beta_{2,matrix} + \beta_{3,target}}$$

$$(\beta_1, \boldsymbol{\beta_2}, \boldsymbol{\beta_3}) \sim Multivariate\ Normal(\mathbf{0}, \mathbf{10^3}\boldsymbol{I})$$

Here, the evaluation explicitly forced a dependence on concentration (generating a curve) while leveraging all device runs to learn about each target/matrix combination[1]. The Figure 1 shows an example performance curves from the regression analysis.

Scoping the test

For the OT phase, each subsystem will be tested with various targets in various matrices by an operator according to sample processing and triage procedures. Most samples will be tested on multiple components within a subsystem, and then a final call will be made by the operator.

[1] MCMC techniques were used to generate posterior distributions for the regression coefficients. These posterior distributions can be used to calculate posterior distributions of the performance curves shown in the figure across concentration for any target/matrix combination from all four devices.

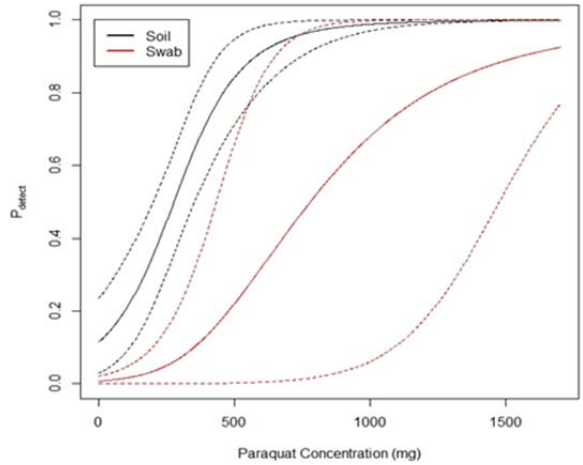


Figure 1. Probability of Detection for Paraquat Concentration in Soil and Swab

In the case of the chemical subsystem, the posterior from the Phase 2 analysis with an added degradation factor for moving from DT to OT serves as the basis for the assurance testing algorithm (see Hamada et.al. 2008).The OT plan needs to have 6 different concentration levels of 20 target/matrix combinations. The combinations are selected randomly from a list of threat representative agents of interest to the users as illustrated in Table 1.

Table 1: Target and matrix for OT

Target	Matrix
COI2 impure	Sand
GF	Soil
Sulfuric Acid	Pristine
VX	Sand
Methanol	Swab
GB-WGA	Air
COI3 pure	Vegetation
Lewisite	Sand
Sodium Cyanide	Water
GD	Soil
2-chlorovinylarsonic acid	Swab
Formaldehyde	Water
Paraquat	Vegetation
Octamethylpyrophosphoramide	Vegetation
Allyl Alcohol	Swab
Ammonia	Water
Thiodiglycol	Swab
Pinacolyl Methylphosphonic Acid	Sand
CVAOA	Soil
Methyl Bromide	Air

Determining Concentration Levels

Where information about threat representative or toxicity concentration levels is known, the OT concentration will be set at these levels. However, this information for some target/matrix combinations may not be known. The Phase 1 and Phase 2 analysis provides some insight into the range of values that each component, and each subsystem, can or might have difficulty detecting.

The lowest concentration of a given target/matrix combination will be set at the most sensitive device in the subsystem's $P_{detect} = 0.5$. This means that the lowest concentration level of any sample provided to the subsystem is set where the most sensitive of the components has a 50 percent chance of detecting. As shown in Figure 2 below, Component 2 of the biological subsystem can detect SEB in smaller concentrations than Component 1. The smallest of the concentrations in each matrix will be set where the performance curves for Component 2 cross 0.50. This would be a sample that the subsystem would have difficulty detecting, but controls the risk of setting all concentrations out of range for any component to detect.

Some concentration levels might be set so as to decrease the width of the performance curve to date. For instance, the analysis shown in Figure 1 suggests that 2 concentrations for Paraquat on a swab could be added between 500 and 1000 mg and at least one above 1000 mg on Component 1 of the chemical subsystem to add more information where the intervals are widest. By combining threat or toxicity level intelligence information and the Phase 1 and Phase 2 analysis, the 6 concentration levels for each agent/matrix combination can be set.

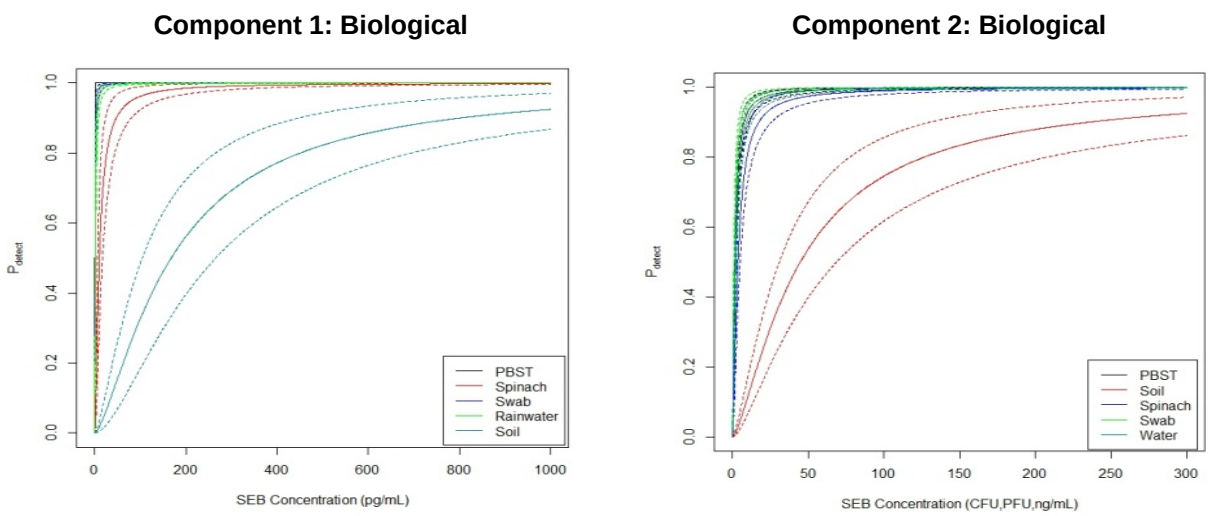


Figure 2. Probability of Detection for SEB Concentration for Two Components

Analysis Plan

To analyze the OT data, a logistic regression will again be used for each component of each subsystem with target, matrix, and concentration as factors. The Phase 2 posterior distributions will be used for the prior of the OT regression coefficients, with some additional variability.

$$\boldsymbol{logit(P_D) = \beta_1 * conc + \beta_{2,matrix} + \beta_{3,agent}}$$

$$(\beta_1, \boldsymbol{\beta_2}, \boldsymbol{\beta_3}) \sim Multivariate\ Normal(\boldsymbol{\mu_{Phase\ 2}}, \boldsymbol{c\Sigma_{Phase\ 2}})$$

The probability of a subsystem failing to detect is a function of how the components of the subsystem are structured. There are many types of system structures; some simple and commonly used are in series (all of the components must detect), in parallel (at least one component must detect), and k-of-n (at least k of the n components must detect). Here, a k-of-n system structure will be used based on the CONOPS. That is, the overall laboratory identification for a sample is made if at least 2 components in the subsystem detect a target in that sample. This will incorporate the component level information as well as account for the operator call.

STAT – Observational Example

D.1 Test Design Overview

The purpose of this appendix is to provide a framework for the Operational Test Activity's (OTA's) observational analysis of aircraft carrier flight deck operations. Testing will support an assessment of changes to Nimitz-class aircraft carrier flight decks that are designed to increase the number of aircraft sorties per a day from 120 to 135 in sustained operations. The test design is not based on a traditional approach in which specific factors and levels are controlled. Rather, the test design focuses on an observational analysis in which the test team collects and analyzes data from flight deck operations over which they exert limited control. In this case, the ship's crew will execute an operationally realistic air plan based on the Design Reference Mission (DRM) documented in the system's requirements. The air plan will include 6 days of sustained operations (12-hours per day). The OTA will collect data during flight operations to assess performance. The discussion below shows that sufficient data will be collected to analyze performance across relevant factors and to determine whether the number of sorties per day is increased.

The primary metric is Sortie Generation Rate (SGR), which measures the number of aircraft launched in a day. While there are factors that affect SGR, the OTA cannot control them. For example, the type of aircraft (e.g., helicopter versus F/A-18) and the mission assigned to the aircraft (e.g., tanker versus Combat Air Patrol) are likely to affect SGR. However, the missions assigned to aircraft and the order in which they are launched are highly constrained (e.g., the rescue helicopter will be launched first). The OTA cannot control this during a test. Furthermore, during flight operations, the flight deck crew makes numerous real-time decisions that affect SGR. For example, Nimitz-class carriers have four catapults for launching aircraft. During a typical launch cycle, multiple aircraft are launched and spare aircraft are available. If a problem occurs during a launch cycle, such as a catapult or an aircraft breaking, the flight deck crew will consider numerous factors. How long will the repair take? How critical is the aircraft to the mission? Where are they in the launch cycle? Based on this and other information, the flight deck crew may wait until the repair is completed, move the aircraft to a different catapult, use a spare aircraft, cancel the launch, or select another option. Artificially constraining the flight deck crew's options would not be realistic. Consequently, the test design does not control these factors and instead the test design is based on an operationally realistic air plan that the crew will execute during the test. The crew will be allowed to make real-time changes, as appropriate, following standard Navy procedures with the goal of achieving mission success for the scenario. The OTA will collect data on various aspects of flight deck operations; see Table D-1. The distributions of the times listed in the table and other metrics to be defined in the test plan will be analyzed to determine whether the flight deck changes improve flight deck operations and whether there are any bottlenecks in the process.

Table D-1. Primary Metrics for Observational Analysis

Metric	Definition
Sortie Generation Rate	Number of aircraft sorties launched in a flight day. Measured in a notional 12-hour day during sustained operations per the Design Reference Mission.
Recovery to Engine Shutdown Time	Measured from the time the aircraft is recovered until its engines are shut down.
Turnaround Time	Measured from engine shutdown until engine start and includes refueling, rearming, and maintenance time.
Engine Start to Taxi Time	Measured from engine start until the aircraft starts to taxi.
Taxi Time to Launch	Measured from start of aircraft taxi until aircraft launch.

Historically, Nimitz-class carriers are able to conduct 120 sorties per day in sustained operations. Over the planned 6 days of sustained operations, the test is expected to have a minimum of 720 sorties.

Overall, the flight deck design changes are expected to improve SGR performance from 120 to 135, and proposed testing has high power to detect SGR improvements of this magnitude. Figure D-1 shows the power of planned testing as a function of true SGR performance of the redesigned flight deck. As true SGR increases, the power to detect the improvement increases. At 95 percent confidence, power is 80 percent of detecting a difference in performance when the true SGR of the aircraft carrier is above 135 sorties per day in sustained operations.

The analysis should be able to detect relatively small differences in time on the order of a few minutes (e.g., a seven minute difference in turnaround time). To understand variability in the data, the OTA examined the results from the Navy's model of flight deck operations. This model was used to develop the changes to the Nimitz-class flight deck. As an example, Figure D-2 shows model results for turnaround time for aircraft that land, are refueled and rearmed, and are relaunched during the next launch cycle. Turnaround time is measured from engine shutdown to engine start and has a mean of 35.7 minutes, with a standard deviation of 7.0 minutes. Figure D-3 shows the power for an independent means test as a function of sample size, which can be used to compare test results to model predictions. As the number of turnaround times measured during the test increases, the power of the test increases. Measuring turnaround time in 35 events provides 90 percent power at 95 percent confidence for evaluating performance against expectations (effect size = standard deviation).

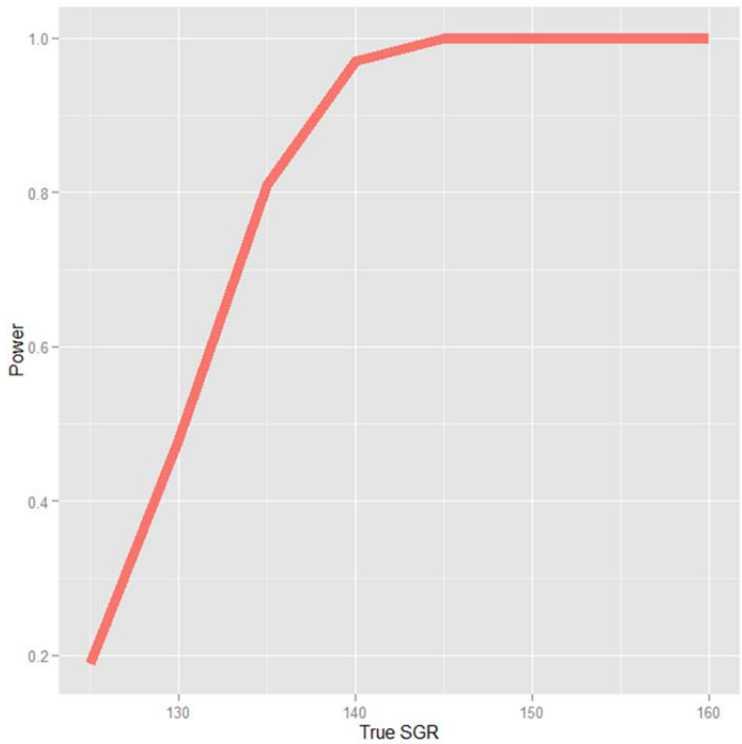


Figure D-1. Power estimate at 95 percent confidence for six days of sustained operations

As discussed above, planned testing is expected to have over 720 sorties based on historical Nimitz-class performance. Based on the planned airwing composition, these sorties will be distributed across six types of aircraft (e.g., E-2Cs, EA-18Gs, and F/A-18s) and six missions (e.g., Combat Air Patrol, Strike Support, and Interdiction). The proposed air plan, based on the Design Reference Mission, has a minimum of 35 sorties per aircraft type and minimum of 35 sorties per mission type. Consequently, the test should provide sufficient data to analyze results by aircraft type and mission.

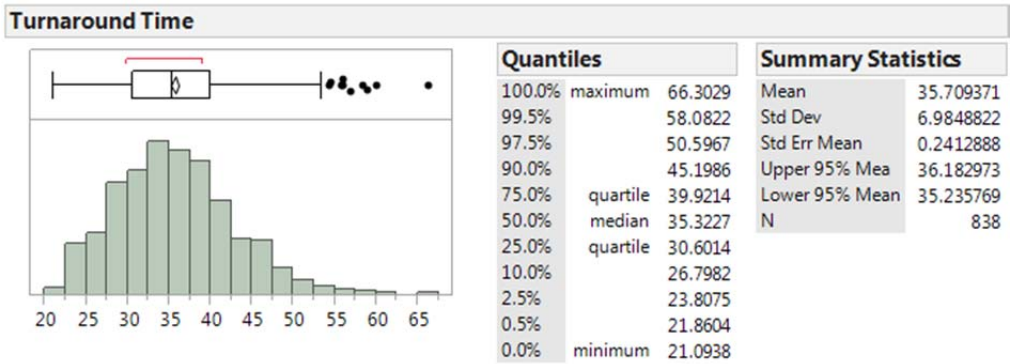


Figure D-2. Distribution of Turnaround Times for Aircraft that Land, are Refueled and Rearmed, and are Immediately Relaunched

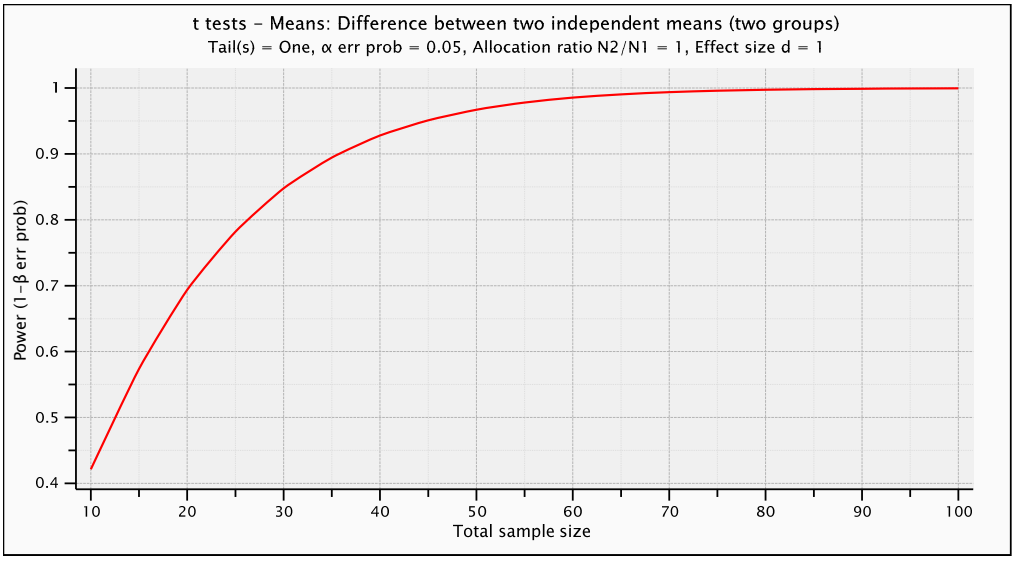


Figure D-3. Power for Independent Means Test

Overall, planned testing will provide sufficient power to determine whether flight deck changes improve SGR performance to 135 sorties per day and to determine whether there are any major deviations from expected turnaround times and other metrics.

Survey Design and Administration – Guidance

Summary

Surveys are systematic measures of people's thoughts, feelings, and opinions. They are important tools in operational test and evaluation because they help determine if military systems are suitable for use by military personnel. Surveys should be used to supplement objective human and system performance data or in instances where objective measurement is not feasible. Surveys should be developed in order to address specific questions related to the effectiveness and suitability of systems, such as the effect of operator workload on detection time, rather than as problem discovery tools. Interviews and focus groups are more effective methods for identifying unforeseen problems. Surveys that yield insights into system effectiveness and suitability require careful planning. Not only must the operational test community design valid and reliable survey instruments, but they must successfully integrate surveys into their test design, administering them systematically across test conditions to a representative group of users. This guidance identifies the types of information that the operational test community should consider when planning surveys into operational tests as well as the specific information needed in TEMPs. For information on the type of information required for Test Plans please reference the DOT&E guidance memo published in January 2017 entitled Survey Pre-testing and Administration.

Survey Design & Administration Plans in TEMPs

TEMPs should specify which measures will be addressed by surveys, the purpose of those measures, the type of user that each survey will be administered to, and whether survey data will be collected using an empirically-vetted or custom-made instrument. An example of how to present this information is displayed in Table 1.

Table 1. Presenting Survey Design & Administration Information in TEMPs

Measure	User Type	Evaluation Plan
Workload during target detection tasks*	Radar operators	Compare workload ratings to time to detect first target under high and low density battlefield conditions

**Empirically-vetted instrument*

Such information is useful for understanding the types of users that operational test agencies (OTAs) need to secure prior to operational testing, the number of unique scales[1] that must be developed as part of the test plan, and if those scales require pre-testing.

Pre-testing is a deliberate review of the survey instrument to ensure that the answers provided by respondents are similar to the data that the OTA intends to collect during operational

[1] Scales measure a specific psychological attribute like mental demand, physical demand, learnability, or usability. Surveys can be comprised of one or more scales.

testing. Empirically-vetted instruments are existing scales that have undergone rigorous testing to establish that they are reliable and valid measures of specific concepts like usability and workload, and consequently, do not require pre-testing prior to use in operational tests. Custom-made instruments, by contrast, require pre-testing to ensure that they actually measure the concept that they are designed to measure prior to use. In addition to the information in Table 1, the TEMP should identify when custom-made instruments will be developed, how they will be pre-tested, when revisions from the pre-test will be completed and how they will be reported, including required staffing timelines. Any necessary resources, such as outside experts or representative users, should also be documented in the TEMP.

Several methods have been developed for pre-testing survey instruments. For example, Chapter XI of Army Research Institute's *Questionnaire Construction Manual* describes a set of guidelines for pre-testing survey instruments. These guidelines are useful for identifying (1) questions that are difficult for respondents to understand and interpret, (2) issues with formatting, (3) instructions that may make the survey difficult for respondents to navigate or compromise their interpretation of particular questions, and (4) the amount of time required for respondents to complete the survey. OTAs should refer to the *Questionnaire Construction Manual* or another reputable source to decide on an appropriate method for pre-testing custom-made instruments. A pre-testing example is available in Appendix A of the guidance memo entitled Survey Pre-testing and Administration published by DOT&E in January 2017.

References

Guidance on the Use and Design of Surveys in Operational Test and Evaluation (OT&E), DOT&E, 23 June 2014

Discussion on the Use and Design of Surveys, DOT&E, 24 February 2015

Discussion on Including Neutral Responses on Survey Questions, DOT&E, 2 April 2015

Survey Pre-Testing and Administration in Operational Test and Evaluation, DOT&E 6 January 2017

Test Funding – Guidance

Guidance

For reporting T&E funding requirements in the TEMP, use the Resource/Cost Element Example to define resource and cost elements. The example is consistent with cost elements in test resource plans, detailed test plants, and budgetary TE-1 reporting forms.

T&E Funding Elements

Include all funding elements that apply to the T&E strategy.

Test Articles. Assets directly supporting T&E:

- Test Assets to be expended in test (as in LFT&E)
- Joint Assets (other platforms participating in the operational test)
- Targets (Actual or surrogates)
- Threats (Actual or surrogates, jammers, opposing forces, air defense systems)
- Weapons, ammunition, pyrotechnics, chaff, flares
- Other assets that participate in T&E (support aircraft, captive carry weapons, real-time casualty assessment instrumentation)

Test Resources Categories. Itemize only those test facilities that are used in T&E. Test facilities might include:

- Costs to operate on an Open Air Range (OAR), test range, training facility, at sea, or any facility where T&E is conducted
- Digital Modeling and Simulation (DMS) Facility (or Digital Models and Computer Simulations)
- Measurement Facility (MF)
- System Integration Laboratory (SIL)
- Hardware in the Loop (HITL) Facility
- Installed System Test Facility (ISTF)
- Distributed Live, Virtual, and Constructive (LVC) environment

Other Test Resources. Other test costs not previously mentioned and itemized

- Evaluation (evaluators, JITC participants, DISA participants assessment of IA (Cybersecurity)
- Support Contractor (if not already costed above)
- TDY and Travel
- Other

- Computer and office supplies
- Transportation of test assets, equipment, and personnel to/from the test site
- Instrumentation (if not already costed above)

Funding Elements that should not be Included

- Costs paid to the developing contractor to develop and produce the system under test.
- Military and Government personnel costs.
- Operations and Support costs (spare parts, fuel, training, or other logistical services that will be provided for the system under test upon fielding)

T&E Funding Sources

T&E funding is provided by the program office of the system under test, by the Developmental or Operational Test Activity, by Joint organizations, or by Service-managed accounts. Included in Service-managed accounts are flying hour programs, joint or Service support assets, weapons, targets, ammunition, training ranges, exercises, or anything else that contributes to T&E but is not funded by the test activity or the program office.

Examples (pdf files)

Resource/Cost Element Example

Test Funding Aircraft

Test Funding Space Observation Radar

Test Funding Clean

Downloadable Excel Spreadsheet Files (These will take a few moments to download)

Test Funding Aircraft

Test Funding Space Observation Radar

Test Funding Clean

Test Funding – Combat System Example

Figure 4.X. Test Resource/Cost Elements Summary

T&E Resource Cost Element	Funding Source	FY13	FY14	FY15	FY16	FY17	FY18	FY19	FY20	FY21	FY22	FY23	FY24
Developmental Testing													
DT&E Evaluation	PM	10	5	5	10	20	34	44	33	35	12	10	5
Developmental Test Center/Range	PM		5	5	10	1600	2143	3400	2200	2150	1458	1390	95
Contractor Facility	PM	10	15	30	60								
Threats	PM								20	20			
Ammunition	Service Acct								*	*			
Information Assurance	PM			5	10		13	3			30		
Modeling and Simulation	PM		15	30	60	135							
System Integration Lab (SIL)	PM			5	10	490	540	3					
Hardware in the Loop (HWIL)	PM			20	40	55	670	50	42	50			
Targets	PM								60				
Operational Testing													
OT&E Evaluation	OTA						5	10	30	20	50	30	10
Joint Assets	SOCOM								10		35	10	
OTA and Support Contractor	PM						5	40	970	380	3452	970	
TDY and Travel	PM								50	50	175	50	
Instrumentation	OTA							20	75		260	75	
Test Range/LUT/IOT&E Test Site	PM							30	650	50	2275	650	
Operational Forces	Service Acct								100		350	100	
Threats and Targets	PM								20		70	20	
Ammunition	Service Acct								95		333	95	
Live Fire Testing													
LFT&E Test Assets	PM			1000									
LFT&E Test Range	PM					10	90	280	380	480			
LFT&E Evaluation	PM					10	10	20	20	20			
DT&E Total		20	40	100	200	2300	3400	3500	2400	2300	1500	1400	100
OT&E Total							10	100	2000	500	7000	2000	10
LFT&E Total				1000		20	100	300	400	500			

* Ammunition is provided by the Army – no cost or funding data available

Aircraft T&E Funding Profile

T&E Series Profile		FY13	FY14	FY15	FY16	FY17	FY18	Notes and Caveats
DT&E								First OT&E period is OA concurrent with DT&E, no dedicated test assets
OT&E								
T&E Resource Estimate Profile		**FY13**	**FY14**	**FY15**	**FY16**	**FY17**	**FY18**	**Notes and Caveats**
T&E Articles	**($K)**							
1x - Aircraft	$2,010			$402	$1,608			Munitions are long lead items; expendables all in place at start of IOT&E.
Munitions - Expendable	$2,516		$2,516					
Chaff and Flare - Expendables	$25			$25				
Moving Target	$630			$630				
Static Targets	$7			$7				
Threat Representation	$265			$265				
Test Support - KC-10/135	TBD				$0			
Test Resource Categories		**FY13**	**FY14**	**FY15**	**FY16**	**FY17**	**FY18**	**Notes and Caveats**
M&S - EW/IRCM Patterns	$20	$4	$8	$8				Includes M&S for Integrated Survivability Analysis
M&S -WPN Performance Models	$20	$4	$8	$8				
M&S -LFT&E	$250	$50	$100	$100				
MF - McKinley Lab	$228				$228			
Range - Melrose	TBD			TBD	TBD			
Range - WSMR	$1,200			$300	$900			
Range -China Lake	$60			$15	$45			
Aircraft Instrumentation	$54		$54					
COMSEC	$12			$12				
SATCOM Time	$3			$1	$2			
Other Test Resource		**FY13**	**FY14**	**FY15**	**FY16**	**FY17**	**FY18**	**Notes and Caveats**
Test Support Aircraft (TRANSCOM) C-17	$500			$250	$250			
Deployed Location	$231			$58	$173			
Test Facility	$100			$25	$75			
Support Contractor	TBD			TBD	TBD			
TDY and Travel	$172			$86	$86			
Other	TBD							
T&E Funding Profile	**($K)**	**FY13**	**FY14**	**FY15**	**FY16**	**FY17**	**FY18**	**Notes and Caveats**
DT&E		Not included						Does not include other TBD totals
OT&E	**$8,303**	**$58**	**$2,686**	**$2,192**	**$3,367**			

Scope/Resource/Funding Profile

T&E Series Profile					FY12	FY13	FY14	FY15	FY16	FY17	FY18	FY19	FY20
DT&E		Contractor	Gov										
OT&E		EOA	OA	OT&E									
Test Resource					**T&E Funding Profile**								
T&E Articles		**AFOTEC**	**Program**	**Other**	**FY12**	**FY13**	**FY14**	**FY15**	**FY16**	**FY17**	**FY18**	**FY19**	**FY20**
System Under Test	(sub) component, cabinet, computer strings / Radar Prototype		X										
	Ascension Island Radar, and SOC		X										
	Aus Radar, Ascension Island Radar, and SOC		X	X									
Joint Support Assets	JSpOC			X									
	DODIN			X									
	NASIC			X									
	6 radars, 3 Electro-Optical sensors, TBD Laser Ranges	2 radars, Laser R	2 radars, Laser R	X									
	Airborne and on-orbit in-band E/M Receivers		X										
On-Orbit Targets	GPS receivers, calsats, cubesats, spheres, non-symmetrical objects		cubesat and calsat	X									
Threats	Cyber, EMI, other TBD	X	X										
Test Resource Categories		**AFOTEC**	**Program**	**Other**	**FY12**	**FY13**	**FY14**	**FY15**	**FY16**	**FY17**	**FY18**	**FY19**	**FY20**
Digital Modeling and Simulation	SSA Performance Assessment Tool		X										
Measurement Facility	Developmental Contractor Test Facility		X										
System Integration Laboratory	CONUS Prototype Facility		X										
Other Test Resources		**AFOTEC**	**Program**	**Other**	**FY12**	**FY13**	**FY14**	**FY15**	**FY16**	**FY17**	**FY18**	**FY19**	**FY20**
TDY and Travel to Ascension Island, Eglin (SOC), JSpOC, NASIC, Huntsville, and Aus (FOC)		X	X										
Fly and/or ship (sea) test equipment to Ascension Island and Aus (FOC)		X	X										
Personnel Resource		**AFOTEC**	**Program**	**Other**	**FY12**	**FY13**	**FY14**	**FY15**	**FY16**	**FY17**	**FY18**	**FY19**	**FY20**
Program Office T&E Support			X		4	8	12	7	9	7	5	4	TBD
46 Test Squadron (DT)			X		3.5	4.5	9	9	5	5			
AFOTEC		X			3	3	3	3	3	5	1	TBD	TBD
Joint Interoperability Test Command			X		2	2	2	2	3	3		TBD	TBD
92nd Information Operations Squadron (cooperative cyber assessment)		X								3			TBD
177th Information Aggressor Squadron (adversarial cyber assessment)		X								3			TBD
AFSPC/A9 Analysts (Evaluation, Data reduction)		X							2	2			TBD
MIT/LL and MITRE Analysts (Evaluation, Data reduction)			X							3			TBD
Estimated T&E Funding Levels *				**($K)**	**FY12**	**FY13**	**FY14**	**FY15**	**FY16**	**FY17**	**FY18**	**FY19**	**FY20**
			DT&E	20,059	2182	2764	2712	3000	3469	3412	2520		
			OT&E	434	29	32	32	32	64	225	20	TBD	TBD

Assumptions, Caveats, Notes:

Approx. $100k to build and launch cubesat;

Sensors below fill a particular testing need:

USSTRATCOM Dedicated SSA sensors:

o FPS-85 UHF phased array at Eglin AFB.

o Ground Based-Electro-Optical Deep Space Surveillance (GEODDS) 4 40-inch telescopes.

USSTRATCOM Collateral (not dedicated to SSA) sensors:

o Cavalier AFS, North Dakota, PARCS UHF phased array radar

Other agency sensors:

o Haystack, X-band dish radar

o Haystack Auxiliary, Ku-band dish radar

LASER Range are procured through a Central Bureau operated by NASA at the Goddard Space Flight Center

*Note that Funding levels do not reflect the "Other" Test Resource that are provided at no cost to the Program

T&E Funding Profile

T&E Series Profile		FY13	FY14	FY15	FY16	FY17	FY18	Notes and Caveats
DT&E								
LFT&E								
OT&E								

OT&E Resource Estimate Profile		FY13	FY14	FY15	FY16	FY17	FY18	Notes and Caveats
T&E Articles	($K)							
System Under Test								
Joint Assets (other platforms participating in the operational test)								
Targets (Actual and surrogates)								
Threats (Actual and surrogates, EW systems, opposing forces, air defense								
Ammunition, Pyrotechnic, chaff, flares								
Other Assets								

Test Resource Categories		FY13	FY14	FY15	FY16	FY17	FY18	Notes and Caveats
Open Air Range								Includes M&S for Integrated Survivability Analysis
Training Facility								
System Integration Laboratory								
Hardware in the Loop								
Installed System Test Facility								
Measurement Facility								
Digital Modeling and Simulation								
LFT&E								

Other Test Resource		FY13	FY14	FY15	FY16	FY17	FY18	Notes and Caveats
Evaluation								
Test Organizations								
Instrumentation								
Other								

T&E Funding Profile	($K)	FY13	FY14	FY15	FY16	FY17	FY18	Notes and Caveats
Totals	**$0**	**$0**	**$0**	**$0**	**$0**	**$0**	**$0**	

Test Instrumentation – Guidance

Summary

In the conduct of operational testing, instrumentation is vital to identify with clarity what happens during test events. However, instrumentation data alone is generally not sufficient to explain why events unfold as they do and thus requires other sources of information, including interviews with operators and commanders. In general, instrumentation data is helpful in characterizing the environment and assessing Measures of Performance, but makes up only a portion of the data needed to assess Measures of Effectiveness.

When preparing a TEMP, specify in detail what instrumentation will be used to collect data on the system under test, and precisely what the instrumentation data will be used for in the evaluation. Factors and levels that are crucial to the evaluation should be identified in the Design of Experiments methodology. When possible, both DT and OT events should use common instrumentation to facilitate interpretation of the instrumentation outputs. The instrumented data should be collected carefully during the event to ensure that harvesting does not interrupt the operational context.

In addition to specifying the system performance instrumentation, the TEMP should delineate the real-time casualty assessment (RTCA) instrumentation to be used in OT events. This should include the description of the RTCA systems to be used and their quantities in both the Red and Blue forces.

Best Practices

An example of instrumentation used in support of operational testing is the Instrumented Field Data Collector[1] (IFDC) used in the Force XXI Battle Command Brigade and Below (FBCB2) and Early Infantry Brigade Combat Team (E-IBCT) assessments. The instrumentation system was physically attached to the test vehicles to capture and record all of the electronic message traffic that passed through the FBCB2, and was crucial to understanding the volume of message traffic flow between combat units, and the degree of situational awareness subordinate units had as a result of the presence of the digitization equipment. However, the presence of the IFDC was not sufficient to disclose everything necessary about the FBCB2 during the OT. Other sources of information, such as interviews with unit leaders and system operators, were also needed to assess the impact of improved situational awareness during operations.

Time/position/velocity/acceleration sensors are commonly used in developmental and operational testing.

References

Reporting of Operational Test and Evaluation (OT&E) Results, DOT&E January 6, 2010

Examples

1 IFDCs monitored digital message traffic and provided data on message completion rates.

Test Instrumentation – Examples

Example 1

3.4.2.4 Test Instrumentation. The Instrumented Field Data Collector[1] (IFDC) will be used in the Force XXI Battle Command Brigade and Below (FBCB2) and Early Infantry Brigade Combat Team (E-IBCT) assessments. The instrumentation system is physically attached to the test vehicles to capture and record all of the electronic message traffic that passes through the FBCB2, and is crucial to understanding the volume of message traffic flow between combat units, and the degree of situational awareness subordinate units have as a result of the presence of FBCB2. Other sources of information, such as interviews with unit leaders and system operators, will be used to assess the impact of FBCB2 on unit situational awareness.

Example 2

3.4.2.4 Test Instrumentation. On-board instrumentation for the Dakota attack helicopter FOT&E will record aircraft state data (Roll, pitch, yaw, warnings, position, speed, etc), video, and transmit video to the ground-based test control center. By design, the Dakota routinely records and stores mission video, fault detections, aircraft state data, and maintenance data. The OTA will coordinate with the Dakota PM to develop vendor software for the extraction and interpretation of recorded data. Video from the primary EO/IR sensor will be transmitted by the Air-to-Air-to-Ground (AAG) system to the test team to assist in coordination, control, and direction of each test event.

The Dakota uses the Tactical Engagement Simulation System (TESS) for force-on-force testing and training. This hardware and software uses a laser-based scoring system to portray realistic force exchange. TESS will be integrated with a ground-based RTCA system for all Red and Blue forces to adjudicate engagements between Dakota and ground forces.

1 IFDCs monitored digital message traffic and provided data on message completion rates.

Test Limitations – Guidance

Guidance

Ideally, the test and evaluation strategy would have no limitations that could degrade or prevent resolution of the critical operational issues (COIs) or formulation of conclusions concerning system effectiveness, suitability, or survivability. In those instances when test limitations cannot be avoided, the TEMP should enumerate them. For each limitation, the TEMP should explain the problem(s) in enough detail to describe specifically how the limitation will affect the evaluation and the conclusions that can be drawn from the test.

A program might have test limitations that affect DT, LFT&E, and/or OT. Each limitation should be addressed in the appropriate TEMP sections 3.2.4 DT Test Limitations, 3.5.4 OT Test Limitations, or 3.6.3 LFT&E Test Limitations, as appropriate. Cybersecurity test limitations should be addressed in the appropriate DT/OT Test Limitation section (if integrated with DT or OT), or in Appendix E, Cybersecurity.

Rarely should a TEMP that anticipates a critical limitation for planned test events be submitted to DOT&E for approval. The TEMP should explain plans, if any, to mitigate limitations.

Definition

Generally, test limitations are constraints that cause differences between the test environment and the expected operational environment (combat or peacetime, as appropriate), which in turn could cause the test results to differ from the results in the expected operational environment. A test might also have limitations if it is impossible to establish ground truth or evaluate results with certainty. The test might be limited in scope because there are inadequate resources to test in all of the relevant operational environments, e.g., extreme cold or hot weather. Other limitations might include altered procedures because of safety concerns, constrained test infrastructure, lack of threat surrogates, inadequate target realism, or the immaturity of the system or any subsystems.

3.2.6 Test Limitations

Aerial targets will not fully represent the full spectrum of threat anti-ship cruise missiles (ASCM) in terms of speed, altitude profile, maneuverability, radar cross section, size and shape, infra-red (IR) signature, countermeasures, counter-countermeasures, radar emissions, and survivability (in the event of warhead-configured Sea Sharks). In those areas where the target fidelity differs substantively from the most prevalent ASCM threat, the Sea Shark and its supporting NCS may not be stressed to a comparable extent as they would be by the actual threat, thereby bringing into question the relevance of the operational test results when using the lower fidelity target. The areas in question are the target speed and the target altitude profile.

Planned mitigation efforts include:

- NCS and Sea Shark modeling and simulation will explore Sea Shark missile performance and in-flight support against all expected threat/target speed/altitude profiles. This will be followed by validation of the M&S simulation with developmental test results and pre-shot predictions for operational testing.
- Development and procurement of an upgraded threat target that can match the speed/altitude profile of the most challenging threats.

Background for Maritime Air Defense Example

This example is for the hypothetical Sea Shark missile (ship-launched, anti-air, semi-active radar homing missile, supported by the hypothetical Neptune Combat System (NCS)). Critical operational issues (COIs) for Sea Shark and its supporting combat systems include:

- Area Air Defense Capability (Can Sea Shark, supported by the NCS, provide air defense for other ships within the Aircraft Carrier Strike Group?)
- Own Ship Air Defense Capability (Can Sea Shark, supported by the NCS, provide own ship defense against air threats while also conducting Area Defense?)
- Availability (Can Sea Shark, after a representative shipboard storage time in the vertical launch cell, provide the required launch availability?)
- Reliability (Can Sea Shark, after a representative shipboard storage time in the vertical launch cell, provide the required in-flight reliability?)

3.6.3 Test Limitations

LFT&E will not confirm or demonstrate through ballistic testing the actual vulnerability of the wiring or avionics subsystems of the Dakota aircraft. LFT&E and combat data have shown that ballistic damage to wiring or avionics can result in loss of mission critical systems such as: EO/IR sights/displays, communications, and weapons systems. In mitigation, the effects of avionics and wiring failures will be tested through fault insertion in the Avionics Integration Laboratory. Those results will then be incorporated into the system-wide M&S vulnerability assessment.

Background for Maritime Air Defense Example

This example is for the hypothetical Sea Shark missile (ship-launched, anti-air, semi-active radar homing missile, supported by the hypothetical Neptune Combat System (NCS)). Critical operational issues (COIs) for Sea Shark and its supporting combat systems include:

- Area Air Defense Capability – Can Sea Shark, supported by the NCS, provide air defense for other ships within the Aircraft Carrier Strike Group?
- Own Ship Air Defense Capability – Can Sea Shark, supported by the NCS, provide own ship defense against air threats while also conducting Area Defense?
- Availability – Can Sea Shark, after a representative shipboard storage time in the vertical launch cell, provide the required launch availability?
- Reliability – Can Sea Shark, after a representative shipboard storage time in the vertical launch cell, provide the required in-flight reliability?

3.5.4 Test Limitations

Quantities of Sea Shark missiles will be limited, possibly precluding re-engagement of surviving simulated threats. In some scenarios, threats/surrogates might survive initial engagement, thus requiring deployment of a second Sea Shark. The test plan does not provide enough Sea Shark missiles to support a second launch. This is a departure from operational realism. At most, the test unit will conduct a simulated Sea Shark missile launch against surviving surrogates.

Planned mitigation includes:

- Once the M&S is validated with the initial IOT&E results, conduct simulation using the available Office of Naval Intelligence digital models for the threats and simulated Sea Shark missile re-engagement of surviving threats. This would provide an early prediction of how Sea Shark and the NCS could respond against surviving Anti-Ship Cruise Missiles (ASCM) threats.
- Follow-on OT&E (FOT&E) will be scheduled at the earliest opportunity when production Sea Sharks are available to support OT addressing re-engagement of simulated threats that survive initial engagement.

Current test range target launch and control capability will limit the number of simultaneous targets in flight and thus, the size of simulated ASCM raids. Sea Shark is required to defend against multiple simultaneous threats, but the test range is unable to launch and track multiple simultaneous threat systems.

Mitigation efforts include the following:

- Once the Sea Shark and NCS M&S capability is validated by initial IOT&E results, simulated engagements will be conducted against threat large ASCM raids to predict results for interim fleet tactics development.
- The Navy will upgrade the test range facilities to support multiple simultaneous engagements prior to the first FOT&E.

Missiles will not have representative shipboard magazine storage times by the time of operational testing. Missiles must be fielded and in representative storage magazines for one year before steady-state availability and reliability levels will be known.

Mitigation efforts include the following:

- The reliability growth curve will estimate system reliability after fielding. The growth curve will be adjusted as needed based on results of IOT&E and accelerated life testing of guidance, fuze, and propulsion components.
- Availability and reliability of Sea Shark missiles with representative magazine storage times will be evaluated during the first FOT&E.

3.5.4.1 Cybersecurity Test Limitations

Both the CVPA and AA will be conducted in-port, as the testing will necessarily decertify the platform. Ship's crew will be executing mission threads using simulation data sources to support mission effects data collection during the AA.

If crew safety or equipment damage concerns preclude the evaluation of any systems (e.g., industrial control systems such as PLCs) while onboard the ship, independent laboratory testing of these systems will be performed. This data will be included in the CVPA report and cyber exploitations based on the findings will be white-carded in the AA.

Test Plan Approval Matrix – Example

Test	Document	Delivery Date	DT&E	DOT&E
LFT&E				
	Armor Coupon Detailed Test Plan	30 days before test		X
	BH&T OTA TP	60 days before test		XX
	BH&T Detailed Test Plan	30 days before test		X
	Controlled Damage Experiment Detailed Test Plan	30 days before test		XX
	FUSL Pre-Test Predictions	15 days before test		X
	FUSL OTA TP	60 days before test		XX
	FUSL Detailed Test Plan	30 days before test		XX
	M&S Accreditation Report including V&V Report(s)	Before start of FUSL test		X
	M&S Comparison Report	90 days after final FUSL test event		X
Developmental Testing				
	Component Qualification Test Plans	60 days before each test	X	X
	Weapons Performance Test Plan	60 days before test	X	X
	Sensor Performance Test Plan	60 days before test	X	X
Operational Testing				
	Operational Assessment Test Plan	60 days before test	X	XX
	IOT&E Test Plan	60 days before test	X	XX
	FOT&E Test Plan	60 days before test	X	XX

X – Denotes Review XX – Denotes Review and Approval

BH&T	Ballistic Hull and Turret
OTA TP	Operational Test Agency Test Plan
FUSL	Full-up system-level
M&S	Modeling and Simulation
V&V	Verification and Validation

Test Planning Documents – Guidance

Summary

For all operational tests, live fire tests, and all other tests that support DOT&E evaluations, the TEMP should include a matrix that identifies which test planning documents will be submitted for DOT&E approval and which will be submitted for information and review only. The lead OTA shall brief the DOT&E on T&E concepts for the Operational Test Plan as early as possible but no fewer than 180 days prior to start of any such testing. The lead OTA shall deliver the Operational Test Plan to DOT&E for approval no fewer than 60 days before the start. Use of developmental test data for an operational assessment or evaluation should be coordinated with the lead OTA and DOT&E prior to the start of testing, and, when feasible, shall receive prior approval. The DOT&E shall require approval of LFT&E strategies, LFT&E plans, and survivability test plans for covered systems.

References

LFT&E Statute: 10 USC 2366

Timeliness of Operational Test and Evaluation (OT&E) Plans, DOT&E, 24 June 2011

Defense Acquisition Guidebook, Chapter 9

Examples

Document Approval Matrix Example

Test Resources – Guidance

Guidance

The program manager, in coordination with all T&E stakeholders, must identify and plan for all T&E resources needed to adequately support DT&E, OT&E, and LFT&E. The first step is to develop data requirements in the Developmental and Operational Evaluation Frameworks. From those data requirements, develop the resources needed at each stage of the program to generate the required data. Part IV of the TEMP should flow directly from the analyses and identify the test resources to conduct the tests described in Part III of the TEMP. At each TEMP update, the resource estimates should be updated based on information learned in previous test phases and other programmatic changes. (Reference, DoDI 5000.02)

Best Practices

DOT&E will be particularly interested in the size of the test unit and threat force, the number of test articles, other operational force test support (personnel and equipment) (including provisions for baseline systems where appropriate to the evaluation strategy), test location and duration, OT-related modeling and simulation, ammunition, munitions, targets, and OT-related instrumentation (particularly instrumentation that requires separate developmental efforts).

Scientific Test Analyses and Techniques (STAT) should generate statistical measures of merit (power and confidence) on the relevant response variables (mission focused metrics). These statistical measures are important to understand how much testing is enough. DOT&E supports neither too little nor too much testing. The STAT should form the basis for the appropriate scope of testing.

Programs should follow the TEMP format and itemize the test resources of interest. Section 4.2.1 addresses Test Articles; Section 4.2.2 addresses Test Sites and so on. Programs may create separate tables to address each section of Part IV, or may consolidate all test resources into a single table as shown in the examples.

Examples

Test Resources Example

Test Resources Aircraft Example Spreadsheet

Test Resources Space Observation Radar Example Spreadsheet

Test Resources Clean Spreadsheet

Test Resources – Example

One or more tables similar to this should appear in Part IV of the TEMP and should provide a concise summary of the required test resources. The resources in this table should be consistent with the narrative in Part I, the schedule in Part II, and the T&E strategy in Part III. The tables should address DT, IT, OT, and LFT&E as described in other portions of the TEMP.

Operational Test Events						
Test Event	**Date (Qtr/FY)**	**Test Articles**	**Test Sites**	**Funding" ($000)**	**Threat Representation Test Targets/Ammo**	**Operating Forces (OPFOR) (Personnel and Vehicles)**
Single Vehicle Directional Stability DT/OT	1Q/09	1 MCVP (EMD vehicle)	CamPen	Provided in Part IV	None	17 Marines with approach march load
Multi-Vehicle Directional Stability DT/OT	2Q/09	2 MCVP (EMD vehicles)	CamPen	Provided in Part IV	None	2 Reinforced Rifle Squad
Land Gunnery DT/OT	3Q/09-4Q/09	2 MCVP (EMD vehicles)	29P	Provided in Part IV	600 MK268 APFSDS-T; 600 MK264 MPLD-T/MK266 HEI-T LINK; 600 MK239 TP-T; 4000 7.62mm; 20 2.5D & 3D targets (BMP, BMD, BTR, BRDM)	None
Hot Weather DT/OT	4Q/09	2 MCVP (EMD vehicles)	29P	Provided in Part IV	2500 MK239 TP-T; 7200 7.62mm; 20 2.5D & 3D threat targets (BMP, BMD, BTR, BRDM, T72); 5 2.5D friendly targets (LAV, Bradley)	2 Reinforced Rifle Squad
MS C OA	2Q/11	3 MCVP & 1 MCVC (EMD vehicles)	CamPen, 29P	Provided in Part IV	600 MK268 APFSDS-T; 600 MK264 MPLD-T/MK266 HEI-T LINK; 4200 MK239 TP-T; 15000 7.62MM; 20 2.5D & 3D threat targets (BMP, BMD, BTR, BRDM, T72): 5 2.5D friendly targets (LAV, Bradley)	1 Reinforced Rifle Platoon, 1 Battalion Staff, 1 AAV Section w/crews, 1 M1A1 Section w/crews, 2 LAV Sections w/crews (1 section designated as OpFor), MAGTF Afloat Node, 1 Amphibious Ship (LPD), 1 LCAC, 1 81mm Mortar Section, 1 60mm Mortar Section Engineer Squad w/designated attachments, 1 Inf Co FST, FoF OpFor (2-4 LAV Sections and 1-2 Platoons of dismount infantry)
PABM DT/OT	1Q/12	2 MCVP (EMD vehicles)	29P	Provided in Part IV	700 rds MK239 TP-T; 2100 rds PABM; 4000 rds 7.62MM; 20 2.5D & 3D threat targets (BMP, BMD, BTR, BRDM, T72); 2 BTRs; 1 BRDM; 60 3D ballistic plywood mannequin	None
Regimental COC DT/OT	3Q/12-4Q/12	1 MCVP & 1 MCVC (EMD vehicles)	29P	Provided in Part IV	None	1 Regimental Staff
HW (Hot Wx) OA	3Q/12-4Q/12	3 MCVP & 1 MCVC (EMD vehicles)	29P	Provided in Part IV	2500 MK239 TP-T; 7200 7.62mm; 20 2.5D & 3D threat targets (BMP, BMD, BTR, BRDM, T72); 5 2.5D friendly targets (LAV, Bradley)	1 Reinforced Rifle Platoon, 1 Regimental Staff with COC, 1 Battalion Staff, 1 AAV Section w/crews, 1 M1A1 Section w/crews, 2 LAV Sections w/crews (1 section designated as OpFox), 20 threat representative targets (BMP, BMD, BTR, BRDM)
CW (Cold Wx) OA	2Q/13	3 MCVP & 1 MCVC (EMD vehicles)	CRTC, Valdez AK	Provided in Part IV	1000 Mk239 TP-T 3000 7.62mm	1 Infantry Platoon (reinf), 1 Bn Staff (Composition TBD), 20 Data Collectors, 1 DC Chief, 8 Control Cell, live fire and maneuver ranges, 1 Amphibious Ship (LPD)
IOT&E	4Q/14-2Q/15	12 MCVP 2 MCVC (LRIP Vehicles)	CLNC, CamPen, 29P	Provided in Part IV	7800 rds 30mm (AP and HE); 7000 rds 7.62mm; 5000 rds 40mm; 2500 rds 50cal Threat Rep EW & targets 8000 rds 30mm; 4000 rds 7.62mm; 5000 rds 30mm; 2500 rds 7.62mm; 5000 rds 40mm; 2500 rds 50cal 100 2.5D & 3D threat targets (BMP, BMD, BTR, BRDM, T72); 2 BTRs; 1 BRDM; 60 3D ballistic plywood mannequin	14 AAVP7A1, 1 reinforced rifle company(-), 1 AAVC7A1, Bn/Reg HQ staff, 4 M1A1 tanks, 10 LAVs (6 LAV-25, 2 LAV-AT, 1 LAV-L, 1 LAV-C) 8 Javeline Msl Sys, Mortar/Arty FDCs, 8 weapons vehicles (4 Mk19, 4 M2, 50 cal), GSR, 2 AH-1Ws, 1 UH-1N w/C&C or Airborne Relay, 2 AV-8s (20 flight hours), 2 F-18s, live fire test range, USN – 10 steaming days LSD/LPD (Flag configured), 2 LCACs, 2 RACs; Exercise control group personnel at MAGCC 29 Palms and CamPen CSSG Maint. Detachment; 1 CAX BLT exercise and 1 RLT size exercise

Threat Representation – Guidance

Guidance

Threat systems, tactics, and overall capabilities must be adequately represented in operational testing to yield credible, valid results of a system's performance in a realistic operational environment. Information and guidance for characterizing threat systems, tactics, and overall capabilities is provided by the Defense Intelligence Agency (DIA), the Service intelligence production centers, and other intelligence agency reporting. To obtain the additional threat system intelligence that is necessary for test planning, but which is beyond the level of detail captured in the System Threat Assessment Reports (STARS), test planners should consult related intelligence documentation such as Integrated Technical Evaluation and Analysis of Multiple Sources (ITEAMS) reports and Joint Country Operating Force Assessments (JCOFA). To obtain information on the missions and targets of greatest interest to the system under test, and for operational context, planners should consult system employment documents such as field manuals, concepts of employment, analyses of alternatives, and operational mission summary/mission profile documentation.

Emphasis should be placed on adequate representation of threats, threat attributes, and threat environments that are most relevant to the evaluation of the system under test, including evaluation of system lethality and survivability.

The TEMP should illustrate that threats will be adequately represented in testing by including plans to:

- Section 1.3.4. System Threat Assessment: Identify the threats and threat attributes of most interest to the evaluation of the system under test. Review intelligence community assessments and reports to determine the threats the system is likely to face in the operational timeframe(s) and theaters of interest. Perform a preliminary appraisal of threats and threat attributes that are likely to have the greatest impacts on operational effectiveness. Consultation with technical and tactical subject matter experts may be required. (Example)
- Section 1.3.6. Special Test or Certification Requirements: The threat assessment may reveal that critical threats, targets, or threat attributes are not available to support operational or live fire testing. The TEMP should describe the need for development of special threat or target systems and any activities necessary to validate these systems for use in testing. (Example)
- Section 3.5. Operational Evaluation Approach: Summarize the operational test events, key threat simulators and/or simulation(s) and targets to be employed, and the type of representative personnel who will operate and maintain the system. (Example)
- Section 3.5.4. Operational Test Limitations: Identify projected critical/severe or major test limitations stemming from inadequate threat representation, and plans to mitigate those limitations. (Example)

- Section 4.2.5 and Section 4.2.6. Threat and Target Resources: Identify the necessary quantity (numbers of troops, attack aircraft, surface-to-air missiles, torpedoes, tanks, etc.) of threat systems or threat surrogates necessary for all test events. Specify responsibilities, timeframe and resources required to complete validation of threat surrogates. Include threat targets for LFT&E lethality testing and threat munitions for vulnerability testing. (Example)

Each Service is responsible to conduct technical and operational comparisons (validation) between the actual threat attributes and the attributes of planned threat systems (actual or surrogate) for operational or live fire testing. Validation activities should be planned, budgeted, and scheduled to complete well in advance of operational or live fire testing.

DOT&E monitors the validation and approves – through the test plan – the use of all threats and threat surrogates for operational and live fire testing.

References

Defense Acquisition Guidebook, Chapter 9

Guidance on Threat Representation in Operational Test and Evaluation of Space Systems, 4 March 2016

Threat Representation – Operational Evaluation Approach Example

<u>Example 1 – Sea Shark</u>

3.5 Operational Evaluation Approach

The IOT&E for the Sea Shark missile will employ a new threat surrogate that represents the latest anticipated anti-ship cruise missile (ASCM) threat in altitude, speed, radar cross-section, maneuverability, and radar emission capabilities. The Program Manager will fund the development of 10 surrogate threat systems and the associated verification/validation studies. The Operational Test Activity will accredit the surrogates for use in IOT&E. In addition to developing a high fidelity threat surrogate for IOT&E, the Navy will develop the capability to launch multiple simultaneous threat surrogates to support the first FOT&E.

<u>Example 2 – Dakota Helicopter</u>

3.5 Operational Evaluation Approach

The IOT&E for the Dakota Helicopter will feature force-on-force missions which employs RTCA instrumentation to enforce the use of appropriate tactics by blue and red forces. The performance of Dakota-equipped Air Weapons Teams (AWT) will be compared to the performance of Legacy-equipped AWTs in the performance of reconnaissance and attack helicopter missions. The test will be conducted in a joint integrated operational environment to include indirect fires, and J-STARS against an appropriate validated threat.

<u>Example 3 – Generic Air-to-Air Missile (GAAM)</u>

3.5 Operational Evaluation Approach

The IOT&E for the GAAM will emphasize employment of the Modern Stealthy Fighter Target (MSFT) utilizing threat representative electronic attack waveforms against the GAAM. GAAM performance will be compared with the legacy Earlier Generic Air-to-Air Missile (EGAAM). GAAM performance will be compared with EGAAM in the areas of target range, high and low altitude, and electronic attack. GAAM will be launched off all relevant fighter aircraft. Modeling and simulation, validated by flight test, will supplement the limited number of flight tests used in IOT&E in order to develop a Probability of Target Kill (P_{tk}).

Threat Representation – Special Test or Certification Requirements Example

Example 1 – Sea Shark

1.3.6 Special Test or Certification Requirements

Anti-Ship Cruise Missiles (ASCMs) are the primary threat to Naval Surface Ships. Critical attributes of ASCMs include speed, altitude profile, maneuverability, radar cross section, size and shape, infra-red (IR) signature, passive homing capability, countermeasures, and radar emissions. In planning for IOT&E, the ship-launched Sea Shark missile must intercept several ASCM threats, including the most prevalent ASCM, which has a cruise speed of 1.5 Mach and, upon achieving radar lock on its ship target, accelerates to 2.0 Mach, and maintain that speed while homing on the target until ship impact. The threat also has the ability to descend from a 50-foot cruise altitude to 25 feet.

The available aerial threat surrogate has a relatively constant speed of 1.2 Mach and can be flown no lower than 50 feet. Accordingly, the adequacy of the IOT&E for the Sea Shark missile will hinge on the development of a new threat surrogate that more closely matches the anticipated threat in altitude, speed, and radar emissions. The altitude and speed capabilities will demonstrate Sea Shark's kinematic capability to intercept the threat. Radar emission capability will allow the electronic support capability of Sea Shark's combat system to detect and identify the threat during the engagement time-line. The evaluation will also leverage missile flight test results from developmental testing to validate an end-to-end simulation model of threat and Sea Shark engagements. In addition to developing a high fidelity threat surrogate for IOT&E, the Navy will develop the capability to launch multiple simultaneous threat surrogates to support the first FOT&E.

Example 2 - Dakota

1.3.6 Special Test or Certification Requirements

A simulator/stimulator for Band IV infrared Man-Portable Air Defense Systems (MANPADs) is needed to participate with other Real Time Casualty Assessment (RTCA) instrumentation during IOT&E. This simulator/stimulator will have the visual signature of an actual MANPADS, will require the gunner to employ appropriate target tracking within range before simulated launch, will emit appropriate missile launch signatures, will adjudicate the engagement outcome, and will transmit the engagement outcome to the RTCA instrumentation integrator. As a battlefield entity, the simulator/stimulator will be vulnerable to engagement from the Dakota helicopter and if adjudicated as killed by Dakota weapons systems, will be inactivated until appropriately restored.

<u>Example 3 – F-35 Joint Strike Fighter</u>

1.3.6 Special Test or Certification Requirements

Early certification and release of software loads and capability to OT&E is necessary to enable early assessment of system capabilities by OT pilots and maintenance personnel, to afford the opportunity for operationally representative training prior to OT periods, to facilitate test data collection planning, and to reduce risk by maximizing the effectiveness of integrated test. Coordination in this regard must include software safe-for-flight certification for employment by OT pilots in pre-fleet-release mission systems software loads on OT&E aircraft and supporting systems.

<u>Example 4 – Generic Air-to-Air Missile (GAAM)</u>

1.3.6 Special Test or Certification Requirements

The Modern Stealthy Fighter Target (MSFT) is required to represent modern low signature fighters. The MSFT is required to achieve the radar and infrared signatures of low signature fighters as described in the System Threat Assessment Report dated XXXX. Additionally, the MSFT will be required to carry full Digital Radio Frequency Memory (DRFM) electronic attack capability and emulate modern threat Active Electronic Steering Array (AESA) radars. MSFT will be able to fly throughout the entire threat envelope, including high-g maneuvers. MSFT must also be able to carry internally all necessary range instrumentation, including lethality assessment hardware.

Threat Representation – System Threat Assessment Example

Example 1 – Sea Shark

1.3.4 System Threat Assessment

The System Threat Assessment Report (STAR) contains the Defense Intelligence Agency-validated threat to the Sea Shark weapon systems and was validated in 2013. This threat assessment also considered fleet procedures for air defense at sea.

Threats of most interest for evaluation during operational testing of the Sea Shark are:

- Anti-Ship Cruise Missiles
- Infrared or laser-guided rockets and munitions
- Airborne fighters and bombers
- GPS jammers
- Cyber security exploitation

Example 2 - Dakota

1.3.4 System Threat Assessment

The Dakota Threat Assessment Report (STAR) prepared by the Intelligence Division, U.S. Army Aviation and Missile Command, contains the Defense Intelligence Agency-validated threat to Dakota. The Dakota STAR was validated in April of 2010. This threat assessment also considered Analysis of Alternatives, the Dakota Operational Mission Summary/Mission Profile, and FM 3-04.126, Attack Reconnaissance Helicopter Operations.

Threats of most interest for evaluation during operational and live fire testing of the Dakota helicopter are:

- Man-Portable Air Defense Systems
- Laser-guided munitions
- Laser Designators
- Ballistic weapons including rifles, machine guns, rocket propelled grenades, and tank rounds
- Forward echelon mobile radar air defense systems
- GPS jammers
- Cyber security exploitation

Targets of most interest to operational and live fire testing of the Dakota helicopter are:

- Ground forces (infantry, artillery, armor, command and control headquarters)

- Armored vehicles (tank and armored personnel carriers)
- Wheeled vehicles
- Fast boat formations at sea
- Unmanned aircraft

<u>Example 3 – F-35 Joint Strike Fighter (JSF)</u>

1.3.4 System Threat Assessment

The System Threat Assessment Report (STAR) for the F-35 Joint Strike Fighter contains the Defense Intelligence Agency-validated threat to the JSF weapon systems and was validated in 2013.

Threats of highest interest for evaluation during operational and live fire testing of the JSF are:

- Fighter Aircraft
- Radar-guided and infrared-guided air-to-air missiles s
- Mobile and fixed site radar surface-to-air missile (SAM) systems
- Radar, communications, and GPS jammers
- Cyber security exploitation

Targets of most interest to operational and live fire testing of the F-35 JSF are:

- Mobile and fixed site SAM systems
- Fighter aircraft
- Armored vehicles (tank and armored personnel carriers)
- Bunkers
- Buildings

<u>Example 4 – AC-130J</u>

1.3.4 System Threat Assessment

The AC-130J System Threat Assessment Report (STAR) contains the Defense Intelligence Agency-validated threat to the AC-130J weapon systems and was validated in 2013. This threat assessment also considered the Integrated Technical Evaluation and Analysis of Multiple Sources (ITEAMS) and the Joint Country Operating Force Assessment (JCOFA).

Threats of most interest for evaluation during operational and live fire testing of the AC-130J are:

- Air-to-Air infrared or laser-guided munitions

- Mobile and fixed site radar air defense systems
- GPS jammers
- Cyber security exploitation

<u>Example 5 – Generic Air-to-Air Missile (GAAM)</u>

1.3.4 System Threat Assessment

The GAAM System Threat Assessment Report (STAR) contains the Defense Intelligence Agency-validated threat and was published in XXXX.

Threats of most interest for evaluation during operational and live fire testing of the GAAM are:

- Modern stealthy fighters
- Modern DRFM electronic attack
- GPS jammers
- Infrared flares.

Threat Representation – Threat Resources Example

Example 1 – Dakota Helicopter

4.2.5 Threat and Target Systems for Testing

Threat Nomenclature	Test				Source
	DT	LUT	IOT	FOT&E	
MANPADs	1	3	6	6	PM ITTS/TSMO/ YPG
Red APC	2	2	4	4	PM ITTS/TMO/YPG
Red Tank (T72 or later model) [1]	4	5	5	5	PM ITTS/TMO/YPG
Red Truck (2.5T variant)	1	2	4	4	PM ITTS/TMO/YPG
Mobile Ground Radar		1	2	2	PM ITTS/TMO/TSMO
C3 (van)		--	1	1	PM ITTS/TMO
Militarized Civ Vehicles (Mix truck/SUVs/sedans)		6	10	10	YPG
HMMWV or Trucks	2	2	6	6	FORSCOM/YPG
IFV (M2/3)	1	2	5	5	FORSCOM
Blue Tank (M1)	2		5	5	PM ITTS/TMO/YPG
Fast Attack Craft (CG-41 small boat or equivalent)[1]	1		5	5	PM ITTS
Fast Inshore Attack Craft (High Speed Maneuverable Surface Target, HSMST, or equivalent)[1]	1		5	5	PM ITTS

[1] Identified threat for live fire test in the Live Fire Strategy

TEMP Guide 3.0 Index

Page	DOT&E Topic
28	Baseline Evaluation – Guidance
30	CONOPS – Guidance
31	CONOPS – Example
32	Cybersecurity OT&E – Guidance
37	Cybersecurity – TEMP Body Example
42	Cybersecurity – Appendix E Command and Control System Example
49	Cybersecurity – Appendix E Shipboard Example
54	Cybersecurity – Appendix E Tactical Aircraft Example
59	Defense Business Systems – Guidance
61	Defense Business Systems – Examples
66	Design of Experiments – Guidance
69	Design of Experiments – TEMP Body Example
71	Design of Experiments – Artillery Howitzer Example
79	Design of Experiments – Precision Guided Weapon Example
89	Design of Experiments – Example for Software-Intensive System
94	End-to-End Testing – Guidance
96	End to End Testing – Examples
98	Force Protection and Personnel Casualties - Guidance
99	Integrated Survivability Assessment – Guidance
101	Integrated Testing – Guidance
103	IOT&E Entrance Criteria – Guidance
104	IOT&E Entrance Criteria – Examples
106	LFT&E Strategy - Guidance
110	Critical LFT&E Issues – Aircraft Example
113	Critical LFT&E Issues – Ground Combat System Example
116	Critical LFT&E Issues – Ground Tactical System Example
118	LFT&E Threat/Target Matrix - Examples
119	M&S for Test and Evaluation - Guidance
123	Mission Focused Evaluation – Guidance
124	Mission Focused Evaluation – Examples
125	Mission Focused Measures – Guidance
128	Operational Evaluation Framework – Guidance
131	OEF - Aircraft Example
132	OEF - Space Surveillance Radar Example
134	OEF - Clean Example
135	Operational Testing of Software-Intensive Systems - Guidance

Page	DOT&E Topic
137	OT of Software-Intensive Systems – Example
139	Production Representative Test Articles – Guidance
140	Production Representative Test Articles – Examples
141	Realistic Operational Conditions - Guidance
142	Realistic Operational Conditions - Examples
144	Reliability Growth – Guidance
149	Reliability Growth – Example
151	Software Reliability Tracking – Example
153	Reliability Test Planning – Guidance
157	Requirements Rationale – Guidance
159	Ship Reliability Growth – Guidance
161	Ship Reliability Growth – Mature Ship Example
163	Ship Reliability Growth – New Ship Example
168	Software Algorithm Testing – Guidance
170	Software Algorithm Testing – Examples
175	Software Evaluation – Guidance
177	Software Accuracy Evaluation – Examples
180	Software Data Restoral Evaluation – Examples
182	Software Timeliness Evaluation – Case Study
188	STAT (Scientific Test and Analysis Techniques) – Guidance
195	STAT Bayesian Methods – Guidance
197	STAT Bayesian Methods – Example
201	STAT – Observational Example
205	Survey Design and Administration
207	Test Funding – Guidance
209	Test Funding – Example
213	Test Instrumentation – Guidance
214	Test Instrumentation – Examples
215	Test Limitations – Guidance
216	Test Limitations – DT Examples
217	Test Limitations – LFT&E Examples
218	Test Limitations – OT Examples
220	Test Plan Approval Matrix – Example
221	Test Planning Documents – Guidance
222	Test Resources – Guidance
223	Test Resources – Example
224	Threat Representation – Guidance
226	Threat Representation – Operational Evaluation Approach Example
227	Threat Representation – Special Test or Certification Requirements Example
229	Threat Representation – System Threat Assessment Example
232	Threat Representation – Threat Resources Example

TEMP Guide 3.0 – Navigation Guidance

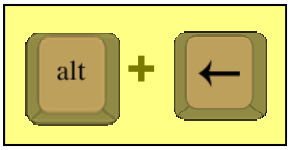
Guidance

The pdf version of TEMP Guide 3.0 uses standard Adobe navigation tools for pdf files. The most useful feature of Adobe navigation is the Alt + Left Arrow key. This hotkey combination returns the viewer to the most recent page or view. Successive uses of Alt + Left Arrow return to prior pages that were viewed in the current session.

To return to the Previous Page/View

Left-click

On this icon



OR

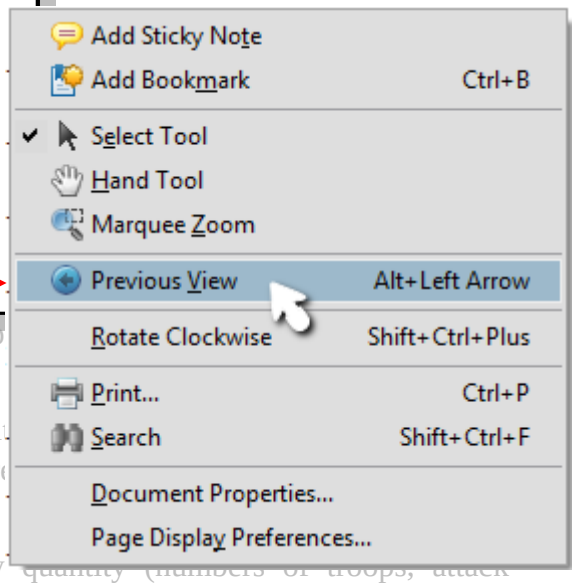
Press Alt + Left Arrow

On your Keyboard

OR

Right-click your mouse and select

"Previous View" in this popup menu



adequately represented to assist in
rational environment. The goal for
the system under test (SUT), based
e intelligence threat assessments.
presentation of threats that are most
Threat systems serve as targets for
SUT survivability.

be adequately represented in testing

profile, maneuverability, size and shap
proposed threats for IOT&E

- Section 3.6.3: Identify projected cri
stemming from inadequate threat repr
limitations (Internal Link)
- Section 4.1.4: Identify the necessary quantity (numbers of troops, attack aircraft, surfa -air missiles, torpedoes, tanks, etc.) of threat systems necessary f st events (External Internet Link)

Identificati escription of certain th in the Service or DIA threat assessments m early conclusion (that should d as early as Milestone
presentative su
e IOT&E.

Internal Links will send you to the appropriate page in the pdf version of TEMP Guide 3.0

External Links look just like Internal Links, but will send you to the appropriate site on the Internet