



## FY 2015 Annual Report

---

### Introduction

American national security is based on preparedness. By ensuring our armed forces' ability to deal with any extant challenge, we disincentivize threats to our interests and mitigate the effects of any attacks when perpetrated. To truly be prepared for the diverse body of threats facing the U.S., from aggressive nation-states to terrorists groups, in cyber and kinetic domains, and across land, sea, and air, weapons must be tested realistically in the environments in which they are to be used. This is the purpose of operational test and evaluation (OT&E). It is essential to assuring the men and women we send into combat can win.

In my tenure as the DOD's Director of Operational Test and Evaluation, I have made it my top priority to ensure that operational tests are adequate, particularly regarding the realism of the conditions under which the testing is conducted. In doing this, I consider all Service-defined operational conditions, including the system operational envelope, the intended mission(s), and the range of operationally realistic kinetic and cybersecurity threats. Conducting a rigorous and operationally realistic test capturing these key parameters is the only way to inform our forces what weapons systems actually can and cannot do.

I have also prioritized the objectivity and scientific rigor of operational tests. By leveraging scientific methodologies including Design of Experiments (DOE), survey design, and statistical analyses, DOT&E ensures defensible and efficient tests are conducted providing the critical information decision makers and warfighters require. Rigorous, scientifically-defensible analyses of the data ensure my reports tell the unvarnished truth. This introduction summarizes my office's continuing efforts to institutionalize these methods in the DOD test and evaluation (T&E) community.

Early stage testing can miss significant operationally relevant problems that are revealed during operational testing in realistic environments. In FY15, as in previous years, OT&E discovered problems missed during development and in previous testing. Finding and addressing these problems before production and deployment is critical, as the only other option is to discover them in combat, when the issues would endanger warfighter lives. In addition, identifying and fixing these problems once full-rate production is underway would be a far more expensive way to address deficiencies, as retrofits are rarely, if ever, cheaper than fixing the problems before full-rate production. Further details on problem discovery during OT&E are provided in a separate section (page 13). OT&E also highlights and exposes previously known problems, as many programs unfortunately choose to progress to operational testing with operationally significant unresolved problems identified in prior testing.

Also included in this introduction, I describe in more detail several focus areas of my office, including the following:

- My continued emphasis on the need to improve reliability of all weapon systems and my recent initiatives to include all relevant information in operational reliability assessments.
- The recently released updated DOT&E Test and Evaluation Master Plan (TEMP) Guidebook, which provides new guidance in my primary focus areas on what substance and level of detail, should be included in TEMPs.
- Recent improvements made in the area of cybersecurity and the need to continue to emphasize cybersecurity as a focus area for all DOD systems.
- Other topics of interest.

---

### RIGOROUS, DEFENSIBLE TESTING

In order to provide rigorous quantitative evaluations of combat performance, and to ensure that we fully utilize scarce test resources, I have advocated the use of scientific test design and statistical analysis techniques for several years. Since 2009, there have been substantial improvements in the use of these techniques within the Services, specifically at each of the Service Operational Test Agencies (OTAs). This improved capability has provided the Department with scientifically rigorous test results that identify what the systems the Services are acquiring can and cannot do in combat. These techniques have helped ensure adequate operational testing; providing sufficient information to characterize combat performance across the set of operational scenarios in which the Services themselves state the weapon systems will be used.

# FY15 INTRODUCTION

Both DOT&E and the Undersecretary of Defense (Acquisition, Technology and Logistics) (USD(AT&L)) updated OSD policy and guidance to promote the use of scientific approaches to test planning; in particular, the DOD Instruction 5000.02 now calls for universal employment of scientific approaches T&E. Specifically, the new instruction emphasizes that the test program should be designed to characterize combat mission capability across the operational environment using an appropriately selected set of factors and conditions.

Warfighters need to know under what conditions the system is effective and when it is not. This characterization is a key element of my guidance for OT&E. In OT&E, characterization ensures adequate information to determine how combat mission capability changes across the operational envelope. Under this concept, testers examine performance as a function of relevant operational conditions and threat types. This is in contrast to the historical approach where test results frequently have been averaged across the operational envelope. For example, a metric such as detection range was averaged across all conditions and compared to a single threshold requirement (or average historical performance). A simple average is not the best way to evaluate performance because it fails to identify differences in performance across the operational envelope, and consequently, it is not informative to the warfighter. Average performance across all conditions masks variances in performance across the operational envelope. An extreme example of this I have seen blended a 100 percent rating in one set of parameters with a 0 percent rating in another, saying the system was 50 percent effective across conditions. This statement is meaningless, and the conditions under which the system under test is ineffective need to be known by the users and developers of the system so that fixes or workarounds can be developed.

I have advocated for the use of scientific methods, including DOE, to ensure that this characterization is conducted as efficiently as possible. The methods that I advocate not only provide a rigorous and defensible coverage of the operational space, they also allow us to quantify the trade-space between the amount of testing and the precision needed to answer complex questions about system performance. They allow us to know, before conducting the test, which analyses we will be able to conduct with the data and therefore, what questions about system performance we will be able to answer. Finally, these methods equip decision makers with the analytical tools to decide how much testing is enough in the context of uncertainty and cost constraints.

The Deputy Assistant Secretary of Defense Developmental Test and Evaluation (DASD(DT&E)) has advocated the use of these methods through his Scientific Test and Analysis Techniques (STAT) T&E Center of Excellence (COE), which employs qualified statistics experts to aid acquisition program managers in applying advanced statistical techniques in developmental testing. The STAT T&E COE helps program managers plan and execute more efficient and effective tests beginning with early developmental testing. Initially 20 Acquisition Category I programs were partnered with the COE. To date, 36 programs have had dedicated COE support for development of test strategies, mentoring, or training. The COE is envisioned to eventually be funded by the Services' in order to expand in size and also provide support to program managers in smaller acquisition programs. I encourage all program offices to ensure that they have access to such a knowledge source.

As a community, we should always strive to improve our test methods. While I have seen improvements in several areas, continued improvement is possible. Important future focus areas include: statistical analytic techniques to examine test results, improving surveys in testing, validation of models and simulations, and using all the appropriate information to maximize the information available to decision makers and operators.

## **Statistical Analytic Techniques**

It is not sufficient to employ statistical methods only in the test design process; corresponding analysis methods should be employed in the evaluation of system performance, otherwise we risk missing important conclusions. Using statistical analysis methods instead of conventional approaches to data analysis, we have been able to learn more from tests without necessarily increasing their size and cost. In all of my reports, my staff uses rigorous statistical analysis methods to provide more information from operational tests than ever before. In the past few years, my staff has used these analysis techniques to identify areas of performance shortfalls. For example, in the operational test and of the Multi-Spectral Targeting System, which is intended to enable helicopters to target small-fast boats and employ HELLFIRE missiles, a logistic regression of the test results revealed a significant interaction between two factors that resulted in performance falling well below the required value in one of the scenarios, suggesting the need for a potential system algorithm improvement. In another example, the operational testing of the AN/TPQ-53 Counterfire radar, showed how performance degraded as a function of range and projectile elevation. This analysis was especially useful because in this case testers did not control all factors likely to affect performance in order to maintain operational realism. Regression techniques enabled DOT&E to determine causes of performance degradations across multiple operating modes, even with highly unbalanced data. Finally, we are using statistical analysis techniques to show statistically significant improvements between incrementally improved versions of systems. In the operational testing of the Acoustic Rapid Commercial Off-the-Shelf Insertion (A-RCI) sonar system an in-lab portion of testing was added to the traditional at-sea testing to evaluate operator detection capabilities across a range of environments and targets. Statistical analysis techniques (coupled with a robust experimental design) showed a statistically

# FY15 INTRODUCTION

significant improvement in the software build over the legacy build and allowed us to definitively claim that the improvement was universal across all operating conditions. It is important to note that if DOT&E had not pushed for these more rigorous analyses, all of these results would have been missed.

Rigorous methods should also be used for suitability analyses. In the past year, I have put a larger emphasis on the rigorous analysis of survey and reliability data. One notable example of this is the reliability assessment conducted for the Littoral Combat Ship (LCS). The LCS reliability requirement as stated would have been nearly impossible to test, requiring a core mission reliability of 0.80 for a 720-hour mission. Instead, my office focused on critical sub-systems that contributed to the core mission. Using Bayesian methodologies and series system models we were able to assess the core mission reliability defensibly, providing reasonable interval estimates of the reliability even in cases where the critical sub-systems had different usage rates and zero failures. This type of analysis also lays the groundwork for how different sources of information discussed below can be used to evaluate system reliability and performance.

Unfortunately, the implementation of rigorous statistical techniques is still far from widespread across all DOD T&E communities. Overall, statistical analysis methods such as logistic regression and analysis of variance, which supported the above discoveries, are underused. Until they are routinely employed in the analysis of T&E data, the OT&E community will miss opportunities to identify important performance results and truly understand system capability. Furthermore, we are not currently leveraging these methods in a sequential fashion to improve knowledge as we move from developmental testing to operational testing. Knowledge about the most important factors from developmental testing will improve our ability to clearly define an adequate operational test that avoids the unnecessary expenditure of resources.

## **Survey Design and Analysis**

In 2015, I issued additional guidance on the design and use of surveys in OT&E. Surveys provide valuable quantitative and qualitative information about the opinions of operators and maintainers as they employ and maintain weapon systems in an operationally realistic test environment. An objective measurement of these opinions is an essential element of my evaluation of operational effectiveness and suitability. However, I have noted that many of the surveys used in OT&E are of such poor quality they can actually hinder my ability to objectively evaluate the system. My office has worked closely with the Service OTAs to improve the quality of surveys used in operational testing.

Custom surveys, established surveys (e.g., NASA workload questionnaire), interviews, and focus groups all have important roles in OT&E. For example, focus groups are often essential venues to elicit operator opinions; however, focus groups should not be the sole source of operator opinion data. Focus groups can be affected by group dynamics and therefore should be used to obtain diagnostic information rather than quantitative information. To maximize the usefulness of focus groups, the test team should examine the survey responses immediately after administering them to look for trends. These initial results can then be used to help guide the focus group questioning which should occur after the written surveys but as soon as possible to ensure impressions are still fresh in the user's minds.

All of the OTAs are currently working on improving their own guidance on the use of surveys in OT&E. Once the scientific best practices I have advocated for are incorporated, I expect future evaluations to include better quality and usable survey results.

## **Validation of Modeling and Simulations**

Modeling and simulation (M&S) can and often does provide complementary information that is useful in my evaluations of operational effectiveness, suitability, and survivability. For example, there are cases in which not all of the important aspects of weapon system effectiveness or system survivability can be evaluated in an operationally realistic environment due to safety, cost, or other constraints. In these cases, M&S provides valuable information to my assessment. However, for M&S to be useful it must be rigorously validated to ensure that the simulations adequately represent the real-world performance under the conditions of its intended use (at a specific level of accuracy). A model that is validated under one set of operational conditions may not be valid under other sets of operational conditions.

Since my assessment of operational effectiveness includes the characterization of combat mission capability across the operational envelope, validation methods must ensure that M&S is valid across that operational envelope. We need to explore new scientific methods for validation that allow me to characterize where the M&S provides useful information to my assessments and where models do not represent the real-world conditions to a high enough level of accuracy. Historical methods of rolling up accuracy of the M&S across a variety of conditions do not provide this level of fidelity and must be improved upon using state-of-the-art scientific methods.

In my recent review of TEMP's that propose M&S as a key aspect of operational testing, I reviewed the selection of M&S points and the validation methods with the same scrutiny as the proposed live operational test points in order to ensure adequacy.

## Using All Information in Operational Evaluations

Operational testing occurs under realistic combat conditions, including operational scenarios typical of a system’s employment in combat, realistic threat forces, and employment of the systems under test by typical users rather than by hand-picked or contractor crews. History has shown us that emphasizing operational realism is essential in identifying critical system performance problems, many of which are only discoverable in an operationally realistic environment. However, operational testing is limited in that it typically spans a short period of time compared to the rest of the testing continuum. In many cases, it is beneficial to consider other test data in an operational evaluation. In doing so, we must account for the fact that these additional data were collected under less operationally realistic conditions.

In cases where other test data, especially that from operationally realistic developmental testing, operational assessments, and M&S, provide additional information we should use state-of-the-art analysis methods to include that information in our analyses. However, it is also essential that we avoid biasing the operationally realistic results in such analyses. Thoughtful application of statistical models, especially Bayesian models, has proven useful in this regard.

## IMPROVING SYSTEM RELIABILITY

Many defense systems continue to demonstrate poor reliability in operational testing. As shown in Figure 1, only 9 of 24 (38 percent) systems that had an Initial Operational Test and Evaluation (IOT&E) or Follow-on Operational Test and Evaluation (FOT&E) in FY15 met their reliability requirements. The remaining 15 systems either failed to meet their requirements (29 percent), met their requirements on some (but not all) platforms on which they were integrated (8 percent), or could not be assessed because of limited test data or the absence of a reliability requirement. In four instances where the system failed to meet its reliability requirement or did not have a reliability requirement, DOT&E assessed that the reliability demonstrated in testing was sufficient to support operational missions resulting in 13 of 24 (54 percent) programs being assessed as operationally reliable.

Various policies have been established to improve reliability performance. Most recently, the January 2015 update to the DOD 5000.02 codified the need for programs to employ best practices in reliability growth planning. The instruction requires program managers to formulate a comprehensive reliability and maintainability program that is part of the systems engineering process, assess the reliability growth required for the system to achieve its reliability threshold during IOT&E, and report the results of that assessment to the Milestone Decision Authority at Milestone C.

Since my office began monitoring reliability in 2005, programs have increasingly complied with these policies, but this has not yet translated to improved reliability performance. Common reasons why programs fail reliability requirements include lack of a design for reliability effort during the design phase; unrealistic requirements that are too large relative to comparable systems; lack of contractual and systems engineering support; insufficient developmental test time to identify and correct failure modes; absence of, or disagreement on, reliability scoring procedures; or failure to correct significant reliability problems discovered in developmental testing prior to operational testing.

Despite these shortfalls, there is some evidence that programs with a reliability Key Performance Parameter (KPP) are more likely to meet their reliability requirements. A 2014 National Academy of Sciences report commissioned by myself and Mr. Frank Kendall (USD(AT&L)) recommended programs develop a reliability KPP and ensure that all proposals explicitly designate funds for reliability improvement activities.<sup>1</sup> To follow-up on this recommendation, my office reviewed the requirements documents for programs that conducted an operational test in 2014. Of the 34 programs that had an IOT&E or FOT&E in FY14 and had a reliability requirement in their Capability Development Document (CDD), 8 had a reliability KPP and 26 did not. Seven of the eight programs (88 percent) with reliability KPPs achieved their reliability requirements while

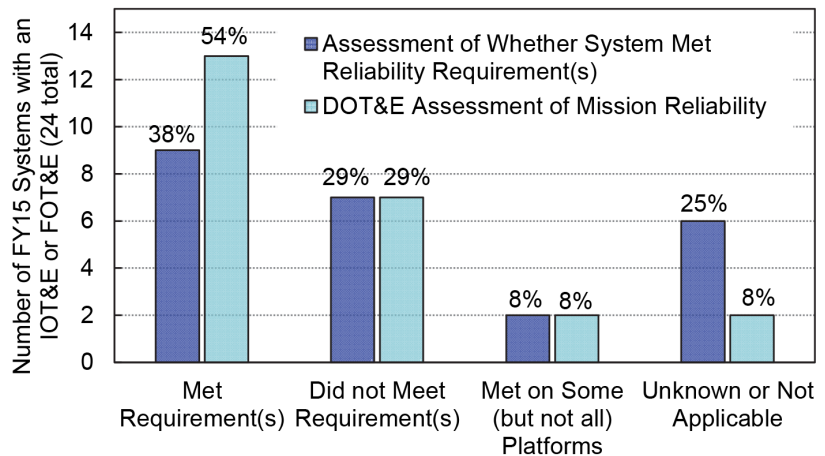


FIGURE 1. RELIABILITY ASSESSMENT FOR 24 SYSTEMS THAT HAD AN IOT&E OR FOT&E IN FY15

1. National Academy of Sciences, *Reliability Growth: Enhancing Defense System Reliability*, 2014.



# FY15 INTRODUCTION

only 11 of the 26 (42 percent) programs without reliability KPPs achieved their requirement. This initial result provides limited evidence that requiring reliability KPPs may be a good policy change for ensuring programs take the need for reliable systems seriously.

In the same annual review on reliability, my office noted that over a quarter (27 percent) of programs had operational test lengths that were shorter in duration than their reliability requirement. As part of my ongoing effort to ensure that testing is done as efficiently as possible, I have continually encouraged programs to intelligently use information from all phases of test, particularly when assessing reliability. Similar to the assessment of other system capabilities, it is important to understand the risks to both the government and the contractor when determining the appropriate length of a test. Overly simple rules of thumb such as testing for duration equal to three times the reliability requirement often lead to inconclusive assessments. In other cases, system reliability requirements can be so high that a test adequate for assessing effectiveness would only permit a limited assessment of reliability. This situation, in particular, benefits from the intelligent incorporation of developmental and early operational test data in the final reliability assessments. It is crucial to note that this does not mean simply adding developmental test data to operational test data. A rigorous statistical approach that accounts for the differences in test environments is necessary.

When a program intends to use developmental test data to support an operational assessment, it is crucial to involve the operational test community early in the data scoring process. Scoring conferences, used extensively by both the Air Force and the Army, provide a forum for stakeholders to discuss reliability, and I recommend that all programs use them. Even if a program does not intend to use developmental test data to supplement the operational assessment, including operational testers in scoring conferences for developmental tests provides the Program Office a better understanding of how issues observed in developmental testing are likely to effect the system's reliability assessment in subsequent operational testing. This helps program offices identify priority corrective actions. I have updated my guidance on reliability test planning in the recently updated DOT&E TEMP Guidebook to address my desire to incorporate all relevant information into operational reliability assessments.

---

## TEMP GUIDEBOOK 3.0

Throughout my tenure, I have always strived to provide clear guidance on my expectations. This year my office updated the DOT&E TEMP Guidebook to complement the January 2015 version of DOD Instruction 5000.02. While the updates also included formatting updates, strict or immediate adherence to the new TEMP format is not required as my evaluation of TEMP adequacy is based on the TEMP's content, not the format. The TEMP Guidebook 3.0 follows the updated DOD 5000.02 TEMP organization; there are bold blue font callouts with links to DOT&E guidance and examples. The callouts have been placed throughout TEMP Guidebook 3.0 at locations where DOT&E and other applicable policies apply. The combination of guidance and examples is intended to highlight areas of emphasis to me, and provide clear examples how my guidance should be interpreted.

There are several key content areas that my office revised in this third iteration of the TEMP Guidebook based on lessons learned over the past several years. The primary areas where substantive updates were made were the creation of an operational evaluation framework, methods for combining information from multiple phases of testing, reliability test planning, and cybersecurity.

I have also expanded my guidance on the use of developmental test data for operational test evaluation. In the current fiscal climate, it is important we test enough to provide the warfighter with valuable information on system capability without testing too much. I have taken every opportunity to use all information available to me to ensure we provide valuable information as efficiently as possible. The Integrated Testing section and the Bayesian guidance section capture best practices for leveraging all available information while still ensuring operational assessments reflect performance in the operational environment.

There is a new section on reliability test planning, which is distinctly different from the reliability growth section. This new section provides clear guidance on my expectations for planning reliability tests as well as what information I expect to be in a reliability growth program.

Additionally, TEMP Guidebook 3.0 contains expanded guidance and examples for implementation of the DOT&E memorandum, "Procedures for Operational Test and Evaluation of Cybersecurity in Acquisition Programs" dated August 1, 2014. These examples are based on lessons learned from cybersecurity test successes and challenges in the past year of implementing the 2014 DOT&E cybersecurity procedures memorandum.

# FY15 INTRODUCTION

## CYBERSECURITY

DOT&E observed improvements in several cybersecurity areas within the DOD this past year; however, operational missions and systems remain vulnerable to cyber-attack. Observed improvements during training exercises include enhanced protection of network elements, greater challenges for cyber opposing forces attempting to access networks, and growing awareness by DOD leadership that cyber-attacks can degrade key systems and critical missions. In some networks, vulnerabilities routinely available elsewhere were mitigated by timely upgrades and software patches. Operational tests of isolated systems experienced much less success in preventing and detecting cyber intrusions highlighting the importance of cyber defense-in-depth. A layered approach to stop primary attack vectors, such as phishing, proved effective at defending some networks. Application whitelisting, where network defenders allow only “known good” applications to operate on a network, also hindered the cyber opposing force from expanding its foothold in the network. However, these improvements were insufficient to ensure that networks and systems can continue to support DOD missions in the presence of a cyber adversary.

In FY15 operational tests and exercise assessments, cyber opposing forces frequently attained a position to deliver cyber effects that could degrade operational missions, often significantly. Unfortunately, exercise and test control authorities seldom permitted aggressive cyber-attacks to affect systems and networks, or allowed non-cyber forces to exploit compromised information in their operations. These restrictions limit insights on both the scope and duration of associated mission effects and preclude the opportunity for training in representative cyber-contested conditions. Acquisition programs, Combatant Commands, Services, and cyber defenders need realistic operational tests and training events that include cyber-attacks and mission effects representative of those expected from advanced capability cyber adversaries.

The demand on DOD-certified Red Teams, which are the core of the cyber opposing forces teams, has more than doubled in the past three years. In the same timeframe, the Cyber Mission Force and private sector have hired away members of Red Teams, resulting in staffing shortfalls during a time with increasing demand. To reduce administrative overhead and increase the realism in portraying cyber threats, DOT&E worked with U.S. Pacific Command, U.S. Northern Command, U.S. Strategic Command, and U.S. Cyber Command to establish permissions for continuous Red Team operations on selected DOD networks and systems. DOT&E also helped Red Teams access advanced cyber capabilities so that they can better emulate advanced capability cyber threats. However, these efforts alone will not offset the Red Team staffing and capability shortfalls, which the DOD must address to retain the ability to assess DOD systems and train Service members against realistic cyber threats.

## ADDITIONAL TOPICS OF INTEREST

In this section, I provide details on specific test resources and test venues that have had significant action on my part this year. For more details on the Multi-Stage Supersonic Target (MSST), self-defense test ship (SDTS), Radar Signal Emitters (RSE), Warrior Injury Assessment Manikin (WIAMan), and Fifth-Generation Aerial Target (5GAT), see the Resources section of this Annual Report (page 397).

### DOT&E Staffing

The FY08 National Defense Authorization Act (NDAA) expressed concern about the adequacy of DOT&E staffing and directed a manpower study be conducted. As a result of that study, the Secretary of Defense authorized 22 additional government billets for DOT&E, increasing civilian authorizations from 54 to 76. Subsequently, in FY10, the DOD evaluated contractor support Department-wide and authorized in-sourcing of inherently government functions while directing a reduction in the levels of contractor support for headquarters organizations. As a result, DOT&E in-sourced 17 inherently government positions and reduced contractor support by a total of 47 (from 72 in 2008 to 25 in 2015 and beyond). Multiple OSD efficiency reviews further reduced DOT&E civilian authorizations from 93 to 67 by FY20.

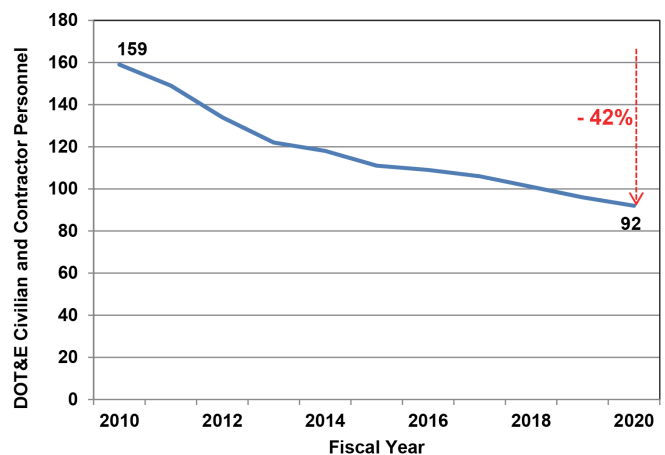


FIGURE 2. DOT&E CIVILIAN AND CONTRACTOR STAFF PROJECTION BETWEEN 2010 – 2020

Between 2010 and 2020, DOT&E civilian and contractor personnel will shrink by 42 percent, and DOT&E anticipates further reductions in budgets and/or manpower authorizations. It is noteworthy that DOT&E, unlike other headquarters staffs, did

# FY15 INTRODUCTION

not receive any additional manpower or funding to support the missions of Operation Iraqi Freedom (OIF) and Operation Enduring Freedom (OEF). Because headquarters staff reductions Department-wide are intended to reduce those staffs that grew larger to support OEF and OIF, the impact to DOT&E staffing is especially significant. To preserve its Title 10 responsibilities, it is likely that DOT&E will have to terminate some non-core, non-Title 10 activities.

## **Multi-Stage Supersonic Target (MSST)**

The Navy's MSST program was intended to provide a threat representative surrogate for a specific class of Anti-Ship Cruise Missiles (ASCMs). Unfortunately, the MSST program, originally intended to cost \$297 Million, ballooned to \$962 Million and was nearly five years behind schedule. Moreover, recent analysis by the Navy's intelligence community indicated the target, if completed, would likely have been a poor surrogate for the threats it was intended to emulate. For these reasons, the Navy directed that the program be terminated.

I agree with the Navy's decision to terminate the MSST program. I also strongly recommended to the Navy that it not pursue a segmented, highly artificial test approach as a substitute for the MSST that the Navy estimated would have cost more than \$700 Million to implement. The artificialities of the alternative proposed by the Navy would have hopelessly confounded the interpretation of any results obtained from its use, making it unwise, unwarranted, and a waste of resources. Nevertheless, without a threat representative surrogate for the threats the MSST was intended to emulate, I will not be able to assess the ability of Navy surface combatants to defend against such threats.

## **Aegis Self-Defense Test Ship (SDTS)**

The Navy's Aegis cruisers and destroyers are charged with defending our Carrier Strike and Amphibious Ready Groups against ASCM attacks. Without such a defense, the self-defense systems on our carriers and amphibious ships may be overwhelmed. It is thus critical that our Aegis ships be able to defend themselves against ASCM attacks so they can survive and complete their air-defense missions. These facts are reflected in the self-defense requirements for all new ship classes and combat system elements to include the Navy's new flight of DDG 51 destroyers (DDG 51 Flight III), the Air and Missile Defense Radar (AMDR) that is to be installed on DDG 51 Flight III, the upgraded Aegis Weapon System planned for DDG 51 Flight III, and the Block 2 upgrade to the Evolved SeaSparrow Missile (ESSM Block 2).

Operationally realistic testing of DDG 51 Flight III, AMDR, the Aegis Weapons System, and ESSM Block 2 requires demonstrating the ship's combat system's ability to defeat raids of ASCMs including a particularly menacing and proliferating set of threats--supersonic ASCMs flying directly at the ship (stream raids). Navy sea-range safety restrictions do not permit ASCM surrogates to be flown directly at crewed ships; even with a cross-range aim-point, the surrogate threats cannot fly within the ranges necessary to test the ship's self-defense combat system. Amphibious ship classes and aircraft carriers have used a crewless SDTS in combination with live firings and M&S to evaluate their self-defense systems. However, the Aegis combat system has never been installed on a test ship. For nearly three years, my office has engaged the Navy regarding the need for an AMDR- and Aegis-equipped SDTS. In doing so, my office has detailed numerous problems found on other Navy surface combatants only as a direct result of testing on a SDTS. Without those tests, critical failure modes would not have been found and could not have been corrected.

In 2015, OSD Cost Analysis Performance Assessment (CAPE) studied various options for acquiring an Aegis- and AMDR-equipped SDTS. The CAPE study, which was based on Navy cost data, showed that an appropriately-equipped SDTS could be acquired for \$320 Million. DOT&E has raised this issue to the Secretary and Deputy Secretary for resolution in the FY17 program and budget review. Meanwhile, DOT&E continues to work with the Navy to develop an integrated test plan for live firings using crewed ships, the SDTS (if available), and M&S.

## **Radar Signal Emitters (RSE)**

In order to improve realism of electronic warfare threats at open air ranges, DOT&E is collaborating with the Test Resource Management Center (TRMC) and Army Threat Systems Management Office (TSMO), to procure a fleet of mobile, programmable radar signal emulators (RSEs) designed to replicate a wide variety of modern, ground-based threat air defense radars. These test assets are essential for creating operationally realistic, multi-layered air defense scenarios for open-air testing of many new systems that are required to operate in an Anti-Access Air Denial (A2AD) environment. These systems include the Joint Strike Fighter (JSF), F-22, B-2, Long-Range Strike Bomber, and the Next Generation Jammer for the EA-18G, as well as others. The first two RSEs are schedule to be delivered to the Nevada Test and Training Range (NTTR) for testing and integration in FY16. A total of 16 systems are under contract and scheduled to be delivered and integrated at Air Force and Navy open-air test ranges.

Now that the JSF Program Office has decided to discontinue the Lockheed Martin Verification Simulation, a high-fidelity manned simulation that had been central to JSF's operational test plans, the ability of open-air testing to replicate more

# FY15 INTRODUCTION

realistic and stressing operational environments is paramount. Having the RSEs integrated on the test ranges and available for the JSF IOT&E is essential.

Significant progress was made this year on the development, production, and planning for testing and range integration of the first two RSEs. Each RSE is capable of high-fidelity emulation of the output power, signal parameters, and performance of long-range surface-to-air missile radars, and is mounted on its own highway-certified and range-road-capable trailer with integral cooling for all weather operability. Once delivered to NTTR, these systems will each be paired with a tow vehicle that incorporates a generator for powering the RSE, communications equipment for connecting to range networks, and an operator control cabin. The RSEs are rapidly reprogrammable and capable of emulating the signals of a wide variety of radars found in modern air defense environments. They employ active electronically-steered array radar technology with high-powered, high-efficiency transmit and receive modules.

With close cooperation of the Air Force NTTR range personnel, the integration and implementation of the RSEs for the JSF IOT&E was defined. Several test events are currently being planned for initial check out. Operational testing of the RSEs is expected to begin by the end of 2016.

Additionally, we are now working closely with the Navy range personnel (Point Mugu Sea Test Range) to implement enhancements at that range necessary to incorporate the RSEs. The Navy will eventually take ownership of 5 RSEs and the Air Force the other 11 for the purposes of operations and maintenance. However, the mobility of the systems is such that any or all of the RSEs would be available for any test program that requires them, and they are readily transportable by air (C-17 or C-130) or over the road to a variety of test ranges.

## **Warrior Injury Assessment Manikin (WIAMan)**

There have been over 23,000 casualties from underbody blast (UBB) events due to improvised explosive devices (IEDs) in the Iraq and Afghanistan conflicts; furthermore, the UBB threat has been an effective enemy tactic over the past decade and a half, and it is likely to remain so. The need to protect our Service members from this threat in the future is clearly reflected in the force protection requirements developed by the Services for their ongoing combat and tactical wheeled vehicle programs. The Army has spent over \$2 Billion to retrofit existing vehicles with UBB protection. New vehicles such as the Joint Light Tactical Vehicle, the Amphibious Combat Vehicle, and the Mobile Protected Firepower Light Tank are being procured with requirements to protect occupants against UBB threats. However, the Department remains without an adequate test device and scientifically-defensible injury criteria to effectively evaluate the protection provided by our combat and tactical wheeled vehicles.

The Department's inability to assess injuries due to UBB events was made clear during the early (2007 – 2009) LFT&E of the Mine-Resistant Ambush Protected (MRAP) vehicles, when the Army could not evaluate differences in the degree of force protection provided to occupants by the different MRAP variants due to non-biofidelic instrumentation and poor injury assessment capability. The DOT&E MRAP assessment, published in 2010, highlighted these test resource deficiencies. Despite these shortcomings, the same ineffective instrumentation and injury criteria used in those tests remain in use today. As part of a retrospective review of MRAP procurement and performance, the DOD directed a status review of UBB M&S to determine if an enhanced UBB M&S capability could have identified the MRAP performance differences prior to the publication of the DOT&E report. The review identified 10 major gaps in the Department's capability to accurately model the effects of UBB; the top three gaps were all associated with the shortcomings in test instrumentation and criteria to assess human injury in the UBB environment. This study highlighted that the current T&E techniques used to address occupant injuries in UBB LFT&E (using automotive crash test dummies and injury criteria designed and developed for forces and accelerations in the horizontal plane to address frontal impact-induced injuries) are not appropriate to assess the effects of the vertical forces and accelerations imparted from a combat UBB event. To address these gaps, I submitted an issue paper in 2010 that ultimately provided \$88 Million for five years of funding for an Army-led research and development program to increase the Department's understanding of the cause and nature of injuries incurred in UBB combat events, and to develop appropriate instrumentation to assess such injuries in testing. This project is known as the Warrior Injury Assessment Manikin, or WIAMan.

In 2013, the Army created a dedicated office (the WIAMan Engineering Office (WEO)) under the Army Research, Development, and Engineering Command (RDECOM) to lead its execution of the program. However, in early 2015 the office of the Assistant Secretary of the Army for Acquisition, Logistics, and Technology determined the WIAMan project would become an Acquisition Category II program of record under the Program Executive Office for Simulation, Training, and Instrumentation (PEO STRI). Army PEO STRI and RDECOM are developing a Test Capabilities Requirements Document based on the previous five years of research by the WEO, which I intend to approve upon its completion. Finally, PEO STRI worked with the WEO to develop and validate a formal Program Office Estimate for full funding of the program.



# FY15 INTRODUCTION

Unfortunately, the Army elected not to program any funding for the WIAMan project after its initial five years of funding was to end in FY16, despite knowing the project would not be completed by then. This delay was, in part, due to the Army's early mismanagement of the biomechanics testing, which necessitated restructuring the project in its third year. This restructuring resulted in cost overruns and schedule delays that the Department has not accounted for in its allocation of resources to WIAMan. The Assistant Secretary of Defense (Health Affairs) has committed Science and Technology funding to the program post-Milestone B to ensure critical injury biomechanics research is completed, but this commitment has not been matched by a similar commitment from the Army to program for the anthropomorphic test device (ATD) production and procurement.

Some within the Army question whether the DOD still needs a combat-specific injury assessment capability for UBB test events; however, it is entirely appropriate for the DOD, and in particular for the Army, to accord the same high priority to testing and verifying the protection provided to Soldiers by their combat vehicles that the commercial automotive industry accords to testing and verifying the protection provided to the U.S. public by their automobiles. For example, the U.S. automotive industry has developed ATDs tailored to the multiple axes of impact that occur in civilian car crashes. This includes, but is not limited to, ATDs to assess injuries from frontal impacts, rear impacts, and side impacts. There is no single ATD that is acceptable for all automotive impact conditions, even for the relatively slow impacts of a car crash and none of these automotive ATDs are acceptable for impact conditions observed in combat. The Army's lack of a commitment to completing this project required me to submit an issue paper this year for additional funding of \$98 Million through FY21 that would enable the continuation of development of defensible injury criteria, predictive modeling and simulations, and two generations of prototype ATDs.

## **Fifth-Generation Aerial Target (5GAT)**

DOT&E investigated the need for an aerial target to adequately represent the characteristics of Fifth Generation threat aircraft in light of the emergence of threat aircraft like Russia's PAK-FA and China's J-20. The Fifth-Generation Target study effort began in 2006 and examined the design and fabrication of a dedicated 5GAT that would be used in the evaluation of U.S. weapon systems effectiveness. The study team, comprised of Air Force and Navy experts, retired Skunk Works engineers, and industry, completed a preliminary design review for a government-owned design. DOT&E and the TRMC have invested over \$11 Million to mature the 5GAT government-owned design. Further investment is required to complete the prototype. DOT&E submitted an issue paper this year for \$27 Million to complete final design, tooling, and prototyping efforts. The prototyping effort will provide cost-informed, alternative design and manufacturing approaches for future air vehicle acquisition programs. These data can also be used to assist with future weapon system development decisions, T&E infrastructure planning/investment, and could support future analysis of alternative activities.

## **Network Integration Evaluation (NIE)**

In FY15, the Army executed two Network Integration Evaluations (NIEs) at Fort Bliss, Texas, and White Sands Missile Range, New Mexico. NIE 15.1 was conducted in October and November 2014, and NIE 15.2 was conducted in April and May 2015. The purpose of the NIEs is to provide a venue for operational testing of Army acquisition programs, with a particular focus on the integrated testing of tactical mission command networks. During NIE 15.1, the Army executed an FOT&E for Warfighter Information Network – Tactical (WIN-T) Increment 2. During NIE 15.2, the Army conducted an FOT&E for the Distributed Common Ground System – Army (DCGS-A) and a Limited User Test for the Mid-Tier Networking Radio (MNVR). Individual articles on these programs are provided elsewhere in this Annual Report. Beginning in FY16, the Army will devote one NIE a year to operational testing and another annual event to experimentation and force development. The latter event is to be called an Army Warfighting Assessment; the first of these was conducted in October 2015.

The Army Test and Evaluation Command's Operational Test Command and the Brigade Modernization Command, continue to develop realistic, well-designed operational scenarios for use during NIEs. The Army should continue to improve its instrumentation and data collection procedures to support operational testing, including refining its method for the conduct of interviews, focus groups, and surveys with the units employing the systems under test. The Army continues to improve threat operations during NIEs, particularly with respect to threat information operations, such as electronic warfare and computer network operations. NIEs should incorporate a large, challenging regular force threat that includes a sizeable armored force and significant indirect fire capabilities.

Network components, both mission command systems and elements of the transport layer, remain excessively complex to use. The current capability of an integrated network to enhance mission command is diminished due to pervasive task complexity. It is challenging to achieve and maintain user proficiency. While networked communications at lower tactical levels may create enhanced operational capability, the use of these networking waveforms brings negative attributes, which need to be fully evaluated and understood. The challenge of integrating network components into tracked combat vehicles

remains unresolved. Due to vehicle space and power constraints, the Army has yet to successfully integrate desired network capabilities into Abrams tanks and Bradley infantry fighting vehicles. It is not clear how the desired tactical network will be incorporated into heavy brigades. The WIN-T FOT&E conducted during NIE 15.1 revealed significant problems with the integration of WIN-T into Stryker vehicles. Integration of the tactical network into an Infantry Brigade Combat Team has not been evaluated at NIEs due to the lack of a light infantry unit assigned to the NIE test unit. Integration of the network into the light forces will be challenging given the limited number of vehicles in the Infantry Brigade Combat Team. The intended tactical network places a greater demand upon the available electromagnetic spectrum than has been the case with non-networked communications. An integrated tactical network introduces new vulnerabilities to threat countermeasures, such as threat computer network attacks, and the ability of a threat to covertly track friendly operations. The Army has yet to integrate radios into its rotary-winged aircraft, which are capable of operating in the same network as ground forces at the company level and below. Units remain overly dependent upon civilian Field Service Representatives to establish and maintain the integrated network. This dependency corresponds directly to the excessive complexity of use of network components.

## **Ballistic Missile Defense**

The Ballistic Missile Defense System (BMDS) is a system of sensors and weapons that have not yet demonstrated an integrated functionality for efficient and effective defense. Currently, the BMDS relies on man-in-the-loop processes to integrate across the blue force instantiations for mission execution coordination within each Combatant Command because the Command and Control, Battle Management, and Communications (C2BMC) element does not provide engagement management capability to the BMDS. The Missile Defense Agency (MDA) should continue C2BMC development efforts to provide an engagement management capability to the BMDS.

In its ongoing efforts to demonstrate BMD theater defense, the MDA conducted several system- and weapon-level flight and ground tests in FY/CY15 using Aegis Ballistic Missile Defense (Aegis BMD), Terminal High-Altitude Area Defense (THAAD), and Patriot. However, the MDA still needs to prioritize development and funding for a BMDS simulation-based performance assessment capability including M&S validation, verification, and accreditation and the ability to produce high-fidelity and statistically-significant BMDS-level performance assessments. Aegis BMD has demonstrated the capability to intercept short- and medium-range ballistic missiles with Standard Missile-3 (SM-3) Block IB interceptors, but the reliability of that interceptor needs to be improved. A key component of the MDA's efforts to improve SM-3 Block IB reliability is the redesign of that interceptor's third-stage rocket motor aft nozzle system, which must be sufficiently ground and flight tested to prove its efficacy. DOT&E recommends that a flight test of the THAAD system against an intermediate-range target should occur as soon as possible. The first THAAD flight test against an intermediate-range ballistic missile (the expected threat class for defense of Guam where THAAD is currently deployed) was scheduled for 2015, but was delayed because of problems with other BMDS test events.

At the BMD strategic defense level, the MDA did not conduct a Ground-based Midcourse Defense (GMD) interceptor flight test in FY/CY15. To improve and demonstrate the capability of the GMD and the reliability and availability of the operational Ground-Based Interceptors (GBIs), the MDA should continue diligently extending the principles and recommendations contained in the Independent Expert Panel assessment report on the GBI fleet to all components of the BMDS instantiation for Homeland Defense and should continue with their plans to retest the Capability Enhancement-I Exo-atmospheric Kill Vehicle in 4QFY17 to accomplish the test objectives from the failed Flight Test GBI-07 (FTG-07) mission. In addition, DOT&E recommends that the MDA should also determine additional sensor capability requirements for a robust Defense of Hawaii capability.

## **Combat Data**

Combat operations over the past 14 years have resulted in a large number of rotary-wing aircraft hit by enemy fire resulting in aircraft losses and personnel casualties (fatalities and injuries). In 2009, Congress directed the DOD to conduct a study on rotorcraft survivability with the specific intent of identifying key technologies that could help reduce rotary-wing losses and fatalities. However, since non-hostile and non-combat mishaps accounted for more than 80 percent of the losses and 70 percent of the fatalities, conclusions from the 2009 study were concentrated towards preventing mishaps rather than surviving direct combat engagements. Since then, DOT&E has continued to analyze combat damage to rotary-wing, fixed-wing, and unmanned aircraft to provide insight on the threats (including small arms, Man-Portable Air Defense Systems, and rocket-propelled grenades), aircraft components and systems, and operational conditions that led to the loss or damage of aircraft and personnel casualties. Additionally, analyses of combat-damaged aircraft have been compared to live fire testing to determine if any changes need to be made in how live fire test programs are conducted.

This year, analyses of combat data have been conducted for aircrew currently engaged in combat operations. Forward deployed aircrews have found the data extremely valuable and have modified their tactics, techniques, and procedures based on these data. Because combat data are perishable if not collected immediately, I made a recommendation to institutionalize an Air Combat Damage Reporting (ACDR) process across the Department. Institutionalizing the ACDR will improve the Department's ability to understand and rapidly respond to changes in enemy tactics and weapons resulting in fewer losses of personnel and aircraft through mitigation actions. The value of the data provided by ACDR is lessened due to a lack of structured and enforced combat damage data collection in theater. Integrating ACDR into operational scenario planning and subsequent manning to deploy combat damage assessment teams as an integrated element of the aviation fighting force will fill the capability gap. It will enable the timely collection of perishable combat damage data to support the Department's rapid and long-term response to current and evolving threats.

## **Full Ship Shock Trial (FSST)**

In combat, even momentary interruptions of critical systems can be catastrophic when those systems are crucial to defending against incoming threats. This is why the Navy has historically required mission-essential systems to remain functional before, during, and after shock. The Navy's shock qualification specification states that a momentary malfunction is acceptable only if it is automatically self-correcting and only if no consequent derangement, mal-operation, or compromise of mission essential capability is caused by the momentary malfunction. The FSST will provide critical information regarding a ship's ability to survive and continue to conduct combat operations after absorbing hits from enemy weapons. Understanding these vulnerabilities is essential. Discoveries made by conducting the FSST on the first-of-class ships will enable timely modification of future ships of the class to assure their survivability.

At the direction of the Deputy Secretary of Defense, the Navy is planning to conduct an FSST on CVN 78 before her first deployment—to do otherwise would have put CVN 78 at risk in combat operations. Historically, FSSTs for each ship class have identified previously unknown mission-critical failures that the Navy had to address to ensure ships would be survivable in combat. We can expect that CVN 78's FSST results will have significant and substantial implications on future carriers in the *Gerald R. Ford* class and any subsequent new class of carriers.

Shock trials are routinely conducted on first-of-class ships, recently including PGH 1, LCC 19, DD 963, CV 59, LHA 1, FFG 7, DDG 993, LSD 41, MCM 1, LHD 1, and MHC 1. However, on occasion, various circumstances have caused some shock trials not to be conducted on the first-of-class, with the primary reason being to ensure testing is conducted on the most representative ship of the class. For example, FSSTs will not be conducted on the first-of-class LCSs because numerous significant design changes are being incorporated in later ships. Nonetheless, the preference is to perform the FSST on the first-of-class, so as to identify and mitigate mission-critical failures as soon as possible.

Some have argued component-level testing and M&S are sufficient to identify and correct shock-related problems on fully-integrated ships. However, the mission-critical failures occurring during every FSST, which are conducted at less than the design-level of shock, discredit this theory. For CVN 78, the FSST is particularly important given the large number of critical systems that have undemonstrated shock survivability. These systems include the Advanced Arresting Gear (AAG), Electromagnetic Aircraft Launching System (EMALS), Dual Band Radar (DBR), the 13.8 kilovolt Electrical Generation and Distribution Systems, the Advanced Weapons Elevator (AWE), a new reactor plant design, and a new island design and location with a unique shock environment.

It is noteworthy that the conduct of an FSST on CVN 78 prior to her first deployment had been a part of the program of record since 2004; therefore, the Navy has had ample time to plan for this event. Nonetheless, a number of claims have been and are being made regarding the potential delay in CVN 78's deployment caused by conducting the FSST prior to the ship's first deployment. These claims span months to years; however, only the former is consistent with the Navy's conduct of the FSST on CVN 71, USS *Theodore Roosevelt*. Commissioned in October 1986, CVN 71 was underway most of January and February 1987 conducting crew and flight operations as part of shakedown. From March to July 1987, CVN 71 underwent a post-shakedown availability. The month of August was used to prepare for the FSST, which was conducted during the period spanning August 31, 1987, to September 21, 1987. Upon completing the FSST, CVN 71 returned to Norfolk Naval Station for a two-week period to remove specialized trial equipment and to complete repairs to systems essential to flight operations. After completing those mission-critical repairs, CVN 71 returned to sea to conduct fleet carrier qualifications. From November 1987 to January 1988, the ship underwent a restricted availability to complete all post-FSST and other repairs. CVN 71 was then underway for most of the remainder of 1988, conducting independent steaming exercises and other activities, departing on its first deployment on December 30, 1988. The effect of conducting the FSST on CVN 71's


# FY15 INTRODUCTION

availability for operations following the shock trial was two weeks to conduct mission-critical repairs, and the total time required to prepare for, conduct, and recover fully from the FSST was about five months, including the restricted availability. Currently, FSSTs for LCS 5 and 6 are planned for 2016. The inevitable lessons we will learn from these tests will have significant implications for LCS combat operations, as well as for the construction of the future frigate, which may be based on one of the LCS designs.

Despite the benefits of expedited FSST, the Navy intends to delay FSST from DDG 1000 to DDG 1002—a decision that I do not support. Conducting FSST on DDG 1000 is critical to finding and correcting failures in mission critical capabilities prior to her first deployment. I submitted an issue paper this year to restore the funding for this test.

---

I submit this report, as required by law, summarizing the operational and live fire test and evaluation activities of the Department of Defense during Fiscal Year 2015.



J. Michael Gilmore  
Director