FY 2016 Annual Report

Introduction

I have served as the Director, Operational Test and Evaluation at the request of the President and Congress since September 2009. It has been an honor and a privilege to serve in this position for over seven years. During my confirmation, I pledged to assure that all of the Department's acquisition systems under my oversight undergo rigorous operational and live fire test and evaluation to determine whether they are operationally effective, suitable, and survivable. I also pledged to provide meaningful, credible test results on system performance to the Congress and civilian and military leaders so that they could make informed decisions regarding acquisition and employment of those systems. In my final annual report to Congress, I review the accomplishments of this office over my tenure, the challenges that the T&E community continues to face, and the consequences of repeatedly fielding equipment that cannot be counted on in combat – a trend that will continue unless rigorous independent operational testing is conducted early and adequately on all systems.

At the core of my pledge to ensure rigorous testing and credible results has been the use of scientific and statistical approaches to realistic operational test design and analysis starting at the beginning of a system's development. The test community has made enormous progress in increasing the use of scientific test design, increasing statistical rigor and improving the analytical capabilities of the Department of Defense (DOD) workforce. The National Research Council recommended the use of modern statistical techniques in defense test and evaluation in 1998, but these techniques were not fully embraced by the operational test community until I provided the direction and implementation guidance early in my tenure. The use of statistical test and analysis techniques is now standard procedure at all of the Operational Test Agencies (OTAs) and is similarly supported by the DOD's developmental test and evaluation office.

Implementation of rigorous test design and analysis provides defensible, factual information to support critical roles of this office. The topics below illustrate how my office has implemented rigorous test design, independent oversight, and objective analysis to support the DOD acqusition system:

- Data to support rapid fielding
- Opportunities for early problem discovery
- Rationales for not conducting testing
- Meaningful, testable requirements and test measures
- Rationales for test adequacy
- Efficient test plans that cover the operational envelope
- Characterization of performance across the operational envelope
- Optimum use of scarce resources
- Improved understanding of system usability
- Methodologies for cybersecurity testing and analysis
- Design for reliability
- Methodologies for combining data from multiple tests
- Rigorous validation of models and simulations
- Improved test resources for evolving threats

The remainder of this introduction summarizes some of the most critical impacts of this office over my tenure. Examples illustrate the value of our products to our primary customer, the soldiers, airmen, sailors, and marines who must ultimately use these systems to accomplish their missions.

IMPROVEMENTS IN TEST AND EVALUATION

The primary goal of operational testing is to understand how new and upgraded systems will perform under the stresses of realistic combat conditions, prior to the Full-Rate Production decision and fielding to combat units. Understanding the capabilities and limitations of systems before they are used in combat is important to commanders in the field and to the men and women who protect our country. Furthermore, the identification of problems permits corrective action before large quantities of a system are procured and minimizes expensive retrofitting of system modifications. Even for systems in which a few units (e.g., ships, satellites) will be acquired, operational testing is essential to find and fix problems, which often can only be found in operationally realistic test conditions, and characterize system performance across operational conditions before the warfighter has to use it in combat.

Rapid Fielding

One of my first priorities as Director was to support rapid fielding of new capabilities to meet urgent needs on the battlefields in Iraq and Afghanistan. My office relied on the use of all available data to provide information regarding performance of these systems. Since 2009, we have published more than 20 early fielding reports to Congress on critical combat systems such as countermeasures for helicopters, small form fit radios, air-to-ground munitions, and many naval systems including ship self-defense missiles, torpedo warning systems, and both variants of the Littoral Combat Ship (LCS). These reports identified performance problems that were either fixed before deployment or made known to the combatant commanders and joint forces that depended on them.

Early Problem Discovery

My office has advocated for earlier realistic testing and problem discovery so that acquisition decision makers can make timely decisions. The Undersecretary of Defense for Acquisition, Technology and Logistics' (USD(AT&L)) 2016 report on the defense acquisition system described $58 Billion in sunk costs over the last two decades on programs that were ultimately canceled. While this figure includes 22 major programs such as the Army's Future Combat System and Comanche Helicopter, it does not include other major programs developed outside the primary acquisition system such as the Airborne Laser and Air Force transformational satellites. To help avoid expensive programs continuing in development while not delivering military utility, my office now requires operational assessments (OAs) for all programs be conducted prior to the Milestone C production decision, when problem discoveries may highlight significant mission shortfalls and problems are cheaper to fix.

Early testing (both developmental test events and OAs) should inform the development process and enable the early identification of major problems. More than just providing an early opportunity for problem detection, an OA provides a chance to build knowledge on how the system will perform once placed in an operational environment. The use of Design of Experiments (DOE), even in early testing, allows efficient test designs that cover the operational envelope. Knowledge gained from OAs can help refine the resources necessary for the IOT&E, such as the most significant factors affecting operational performance, potentially reducing the scope for the IOT&E. In ideal cases, the use of sequential test design from early testing including OAs through IOT&E can provide even more efficient use of test budgets by combining information across test phases. While my office has successfully integrated information from OAs and IOT&Es, integrated developmental and operational testing is the exception and not the rule. One challenge in particular is having production-representative articles early enough to do realistic tetsing.

Rapid Realistic Testing Improves Design and Saves Lives: Mine Resistant Ambush Protected (MRAP)

Mine Resistant Ambush Protected (MRAP) vehicles are a family of vehicles designed to provide increased crew protection against battlefield threats, such as Improvised Explosive Devices (IEDs), mines, and small arms. Because of the urgent operational need for increased crew protection against battlefield threats in Iraq and Afghanistan, multiple MRAP vehicle configurations had to be procured, tested, and fielded on a highly accelerated basis.

DOT&E supported rapid, but operationally realistic testing. The MRAP Joint Program Office originally planned to conduct live fire testing against only Key Performance Parameter (KPP) threshold level of explosive underbelly and side attack threats. However, these KPP-level threats were smaller than known threats in the planned theaters of operation. Consequently, DOT&E required testing against larger explosive threats consistent with those documented in combat.

DOT&E worked with the Army and the Marine Corps to rapidly plan and conduct this testing, which revealed not only significant vulnerabilities against larger, more operationally realistic threats, but also revealed stark differences between the crew protection provided by the different MRAP variants as the threat sizes increased. Despite resistance from the Army, DOT&E immediately reported these newly discovered vulnerabilities and performance differences to the Department leadership and commanders in the field, leading the Program Office to develop, test, and implement design changes that could be retrofitted onto vehicles in theater as well as built into future production lines. The Army and the Marine Corps also considered these differences when selecting the MRAP variants they would retain in their enduring fleet. These timely reports resulted in equipment modifications and tactics changes that likely saved lives of American and Allied soldiers.

Conduct Operational Test Only when Systems are Ready

Having a clear understanding of the required testing provides a rationale for making decisions on when operational tests will or will not provide value to the community. While my office has been a strong supporter of OAs prior to Milestone C, operational testing should only be conducted when appropriate. In cases where systems are clearly not ready for rigorous, realistic testing, we have recommended against spending scarce resources to observe poor performance. Instead, DOT&E has advocated that those resources be reallocated to address capability shortfalls. In the case of the Remote Multi-Mission Vehicle (RMMV), my office recommended that the Navy cancel a planned OA because of well-documented reliability problems. We instead recommended that the Navy dedicate the resources allocated for the OA towards making improvements to the Increment 1 mine countermeasures (MCM) mission package. (See details in reliability section.)

My office also recommended the cancelation of the Army Integrated Air and Missile Defense (AIAMD) Limited User Test (LUT) in favor of a developmental test because of well-known problems with an immature system that was falling well short of performance requirements to demonstrate readiness for a Milestone C production decision. The LUT proceeded against our recommendation, but evaluated less than one-third of the effectiveness measures because of system immaturity and the lack of readiness of some AIAMD capabilities. As DOT&E predicted, the LUT was adequate to confirm poor effectiveness, poor suitability, and poor survivability. My office recommended that the Army fix all critical deficiencies and conduct another LUT to demonstrate the full range of capabilities identified in the May 2012 Test and Evaluation Master Plan (TEMP) under operationally realistic and system stressing conditions.

Early Problem Discovery:
CVN 78 USS *Gerald R. Ford*

CVN 78 is the lead ship in the Navy's newest class of aircraft carriers. USS *Gerald R. Ford* is scheduled to be delivered in 2017. The design incorporates several new systems including a new nuclear power plant, weapons elevators, radar, catapult, and arresting gear.

In the last two CVN 78 OAs, DOT&E examined the reliability of new systems onboard CVN 78 and noted that the poor or unknown reliability of the Electromagnetic Aircraft Launch System (EMALS), the Advanced Arresting Gear (AAG), the Dual Band Radar (DBR), and the Advanced Weapons Elevators (AWE) is the program's most significant risk to successful use in combat. These systems affect major areas of flight operations – launching aircraft, recovering aircraft, air traffic control, and ordnance movement. DOT&E noted that unless these reliability problems are resolved, which would likely require redesigning AAG and EMALS, they will significantly limit CVN 78's ability to conduct combat operations.

CVN 78 is intended to support high-intensity flight operations. The CVN 78 Design Reference Mission (DRM) specifies a 35-day wartime scenario. The DRM includes a 4-day surge with round-the-clock flight operations and 270 aircraft sorties per day. The DRM also includes 26 days of sustained operations with flight operations over a nominal 12 hours per day and 160 aircraft sorties per day.

Based on AAG reliability to recover aircraft, CVN 78 is unlikely to support high-intensity flight operations. AAG has a negligible probability (<0.0001 percent) of completing the 4-day surge and less than a 0.2 percent chance of completing a day of sustained operations without an operational mission failure.

EMALS has higher reliability than AAG, but its reliability to launch aircraft also is likely to limit flight operations. EMALS has less than a 7 percent chance of completing the 4-day surge and a 67 percent chance of completing a single day of sustained operations without a critical failure.

DBR's unknown reliability for air traffic control and ship self-defense is a risk to the IOT&E and for combat operations. The Program Office does not have a DBR reliability estimate based on test data. Because CVN 78 will be delivered soon and DBR hardware is already installed in the ship, it will be difficult to address any significant reliability issues should they arise.

Canceling the F-35 Joint Strike Fighter (JSF) Block 2B Operational Utility Evaluation

When asked in 2012 whether the Services supported the need for the Block 2B Operational Utility Evaluation (OUE), both the Air Force and the Navy stated that they would consider using the F-35 Block 2B aircraft in combat and hence required the testing planned for the Block 2B OUE.

In March 2014, I recommended not conducting the planned F-35 Block 2B OUE, scheduled for the summer of 2015 to evaluate the "initial warfighting capabilities" of the F-35A and F-35B aircraft. My recommendation was based on observations that the program was behind schedule in completing the Block 2B development, and the OUE would only delay the necessary progression to Block 3F development, which is needed to complete development and begin IOT&E. I predicted that the results of the OUE would confirm what we already knew – that the Block 2B F-35 would be of limited military utility. Also, there was substantial evidence that the aircraft would not be ready to support training of operational pilots and successful completion of a comprehensive operational evaluation. The USD(AT&L) and the JSF Program Executive Officer agreed with my recommendation, and the JSF Operational Test Team refocused their efforts from conducting the OUE to activities that would help the program progress toward completing Block 2B, and eventually Block 3F development.

Meaningful, Testable Requirements and Test Measures

My office has continually engaged with the requirements community in efforts to improve requirements and in doing so helped numerous programs refine their requirements early in the acquisition cycle, thereby saving time and resources from trying to achieve the unobtainable. We have pointed out unrealistic reliability requirements in programs like ground combat vehicles, tactical datalinks, and long-range air defense radars; these programs were able to establish the rationale for lower thresholds for providing desired mission performance.

The initial reliability requirement for the Joint Light Tactical Vehicle (JLTV) of 4,500 Mean Miles Between Operational Mission Failure (MMBOMF) was much larger than comparable systems such as the High Mobility Multi-purpose Wheeled Vehicle (HMMWV), and would have been very difficult to achieve. Based on feedback from my office and other stakeholders on what reliability is practically achievable and necessary to support mission objectives, user representatives reduced the requirement to 2,400 MMBOMF. This requirement has a clear, mission-based rationale and is verifiable within a reasonable operational test period.

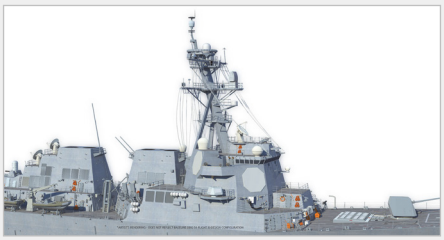
Early engagement also helps programs write requirements in such a manner that they are testable within a reasonable timeframe. We have encouraged the use of continuous metrics such as time, distance, and accuracy in place of binomial metrics such as probability of hit or probability of kill in order to reduce the testing required to confidently demonstrate compliance with requirements. Additionally, even in cases where requirements are not updated, the Service OTAs have now made it common practice to use continuous metrics to scope the operational test in addition to evaluating the required hit/kill-type requirements.

We continue to observe, that while necessary, Key Performance Parameters (KPPs) are not sufficient for testing military systems. KPPs often lack the context of the complex operational environment, including current threats. A few examples:

- P-8A Poseidon is a maritime patrol aircraft that will replace the P-3C Orion and conduct anti-submarine warfare (ASW) and other missions. However, the KPPs required only that the P-8A be reliable, be equipped with self-protection features and radios, and carry a requisite number of sonobuoys and torpedoes, but not actually demonstrate an ability to find and prosecute submarines. DOT&E, working with the Navy's OTA, focused the testing on examining quantitative mission-oriented measures, beyond the limited KPPs, in order to characterize the aircraft's ASW capabilities.
- *Virginia*-class submarine is a multi-mission nuclear attack submarine that is replacing the existing *Los Angeles*-class submarine. During the IOT&E, the submarine failed to meet two KPP thresholds. However, *Virginia's* performance was equivalent to or better than the legacy *Los Angeles*-class in all mission areas, leading my office to evaluate the *Virginia* as operationally effective and operationally suitable.
- Early Infantry Brigade Combat Team (EIBCT) systems were a collection of sensors the Army planned to use in infantry brigades to detect and provide warning of enemy activities. The KPPs for some of the sensors specified only that the systems produce images recognizable as human faces at specified distances—not an expected detection range or a probability of detection. DOT&E advocated and the Army agreed that the systems be tested under realistic combat conditions against a capable enemy threat, which revealed that enemy soldiers could easily spot the large antennas needed to transmit the images back to the operations centers. Additionally, many of the sensors were not useful to soldiers even though they met the KPPs. As a result, the Army canceled the portions of the program that were unnecessary.

As these examples clearly illustrate, operational context is necessary to fully evaluate systems, whether they meet their KPPs or not. My office continues to work with requirements organizations to ensure requirements are achievable, testable, and operationally meaningful, but some independent evaluation metrics will always be necessary, especially in the case of evolving threats.

Writing Measurable Requirements: Air and Missile Defense Radar (AMDR)



The Navy's new SPY-6 Air and Missile Defense Radar (AMDR) is intended to provide an improved Integrated Air and Missile Defense (IAMD) capability to the next flight of USS *Arleigh Burke* (DDG 51) class destroyers (i.e. DDG 51 Flight III). In 2012, DOT&E reviewed the Navy's draft Capability Development Document for AMDR. DOT&E's review noted that several of the program's requirements, including its IAMD Key Performance Parameter (KPP), were probabilistic in nature and would require an unachievable amount of operational testing. Verifying the IAMD KPP, for example, would have required hundreds of ballistic missile and anti-ship cruise missile surrogates. To improve the testability of the AMDR KPPs, DOT&E provided the Navy with alternative metrics using continuous variables like time and range for assessing the radar's capability. The Navy ultimately adopted metrics similar to those suggested by DOT&E, reducing required testing while maintaining the desired capability.

Defensible Rationales for Test Adequacy

Throughout my tenure I have emphasized that the statistical approaches of Design of Experiments (DOE) provide a defensible and efficient methodology for not only determining test adequacy but also ensuring that we obtain the maximum value from scarce test resources. DOE has proven to elicit maximum information from constrained resources, provided the ability to combine information across multiple independent test events, and produced defensible rationale for test adequacy and quantification of risk as a function of test size.

One clear advantage of statistical approaches to evaluating test adequacy is that they provide a means to quantify how much information can be derived from each test point. Clearly, the first time a projectile is fired at a helmet and does not penetrate we learn something new. The second, third, and fourth times, we learn about the robustness of that helmet and whether the first result was a fluke or a consistent trend. But if we fire 10 projectiles at 10 helmets, what is the value of firing the 11th projectile? As the test progresses, we are incrementally not learning as much as the first shot. Statistical methods provide a quantitative trade-space for identifying that point of diminishing returns and also the associated risks of making incorrect decisions based on limited test sizes. My office and the Service OTAs have found these methods invaluable when debating the cost/benefit of additional test points.

Efficient Test Plans that Cover the Operational Envelope

A critical aspect of operational testing is identifying how system capabilities are challenged when placed in operationally realistic conditions. However, today's modern systems are not only designed to contribute to multiple mission areas, but also work across a wide range of operational conditions. The constantly evolving threat further complicates the challenge of determining not only how much testing is enough, but also the conditions under which we need to test. My office has successfully used DOE to address how much testing is needed and also to select points that efficiently span the operational space to ensure that we have a complete picture of performance.

Statistically Rigorous Test Protocols: Enhanced Combat Helmet (ECH)

It is critical that we ensure that the protective equipment we provide to our soldiers meets the high quality that is demanded. After I was asked to assume oversight of personnel protective equipment, I directed that testing of these systems follow protocols that were comparable to existing statistically-based industry quality control methodologies. Employing a statistical approach allowed the Department to set quantifiable quality standards.

Those standards proved valuable following an engineering change proposal intended to increase manufacturing capacity for the ECH. The ECH failed the small arms component of the DOT&E-approved protocol. The helmet failed because of too many small arms penetrations, which demonstrated that the helmet did not provide the desired protection. The manufacturer ultimately decided it was necessary to use different ballistic shell laminate material to provide for an acceptable helmet against the small arms threat.

Designing an Efficient Test for a Multi-Mission Strike Fighter

The F-35 is a multi-role fighter aircraft being produced in three variants for the Air Force, Marine Corps, and Navy. The multi-dimensional operational space created by the mission types, aircraft variants, ground and air threats, and weapons loads is very complex, yet suited for the use of experimental design to efficiently ensure adequate coverage of the operational space for characterizing the performance of the F-35 in all mission areas. Additionally, experimental design enables a "matched pairs" construct for doing comparison testing between the F-35 and the legacy aircraft it is replacing.

The overarching test approach for the F-35 Block 3F IOT&E was to create detailed test designs for evaluating each of the core mission areas by defining appropriate, measurable response variables corresponding to operational effectiveness of each mission area. The test team divided the operational space – using DOE concepts – into factors that would affect the response variables, e.g., type of ground threat or number and types of red air threat, and varied those factors to ensure coverage of the operational space in which the F-35 may be used in combat. Also, the test team sought to maximize information collection by dividing the threat continuum into categories and then assigning coverage to the appropriate mission areas. The team also ensured that key capabilities would be assessed in at least one mission area. For example, finding, tracking, and engaging moving ground targets are enabled by the ground moving target indicator (GMTI) and ground moving target track (GMTT) functions of the radar, and are only covered in strike coordination and reconnaissance and close air support (CAS) missions. This allowed the test team to assess GMTI and GMTT capability without including moving ground targets in all of the mission areas.

The application of DOE to the test design process also supports the development of objective comparison tests. One of the purposes of operational testing is to provide realistic and objective assessments of how systems improve mission accomplishment compared to previous systems under realistic combat conditions. The F-35 requirements document states that the F-35 will replace legacy aircraft, including the A-10, in the CAS mission, so the test design includes a comparison test of the F-35A and the A-10 in this role.

Optimum Use of Scarce Resources

DOE and corresponding statistical analysis methods have supported extracting the maximum value from scarce test resources in a defendable manner. In cases where testing is expensive and there is pressure to reduce test sizes, DOE allows us to understand up front what information we are giving up. Additionally, these methods can assist in finding holes in our current knowledge and placing test points so that they provide the greatest information gain.

Improved Understanding of System Usability

A key aspect of operational testing is observing the quality of human-systems interactions and their impact on mission accomplishment. Operators are a critical component of military systems. Hardware and software alone cannot accomplish missions. Systems that are too complex for operators to use compromise mission success by inducing system failures and force the Services to invest in lengthy and expensive training programs to mitigate problems that arise because of poor interface design. DOT&E has provided guidance on the best practices of the use of surveys in operational test and evaluation

KC-130J Harvest Hercules Airborne Weapon Kit (HAWK)

The Navy is updating the Harvest HAWK that allows the KC-130J tanker/mobility aircraft to employ HELLFIRE and Griffin laser-guided missiles for close air support. Under an Urgent Operational Need Statement, Harvest HAWK has been deployed in theater since 2010 without a formal operational test. The updated Harvest HAWK includes a new sensor for targeting weapons and for laser designation and a new mission operator station. The Navy proposed a limited operational test with only a few end-to-end demonstrations of live munitions. My office proposed a more robust test design based on current tactics documents and munition capabilities. The Navy rejected that proposal, claiming that the system was adequately proven in combat and only limited testing was needed. The Navy provided the available combat data and our analysis showed that while the munitions generally perform well, there are significant gaps between where the system has been used in combat and the desired capabilities of the updated system. The combat data provided significant information on performance during the day, at one altitude, and against stationary targets. Very little information was available on different altitudes, at night, and against moving targets. The Navy is now working with my office to update the operational test design to collect the data that are necessary to fill those gaps.

Long Range Anti-Ship Missile (LRASM)

My office received a request from the Navy to reduce the number of free-flight test shots for the LRASM quick reaction assessment because of budget limitations. The Navy proposed reducing the number of weapons from the previously agreed upon 12 missiles to 6. The proposed reduction excluded important aspects of the operational engagements that looked at different target ranges and aspect angles, which I believe could affect the success rate and performance of the missile.

I was also concerned with having limited live testing to validate the modeling and simulation (M&S) tool. As it stands, the planned 12-shot free-flight program, provides limited opportunity to validate the M&S. Executing any less would not provide adequate information to detect differences between free-flight testing and the M&S. As a direct result, we would run the risk of mischaracterizing the performance of the weapon across the operational test space.

Through statistical analysis techniques, I determined the 12 missiles provided a minimally adequate test for assessing weapon performance and validating the M&S integral to this quick reaction capability. Therefore, I would not approve a test strategy with less than this minimum.

The Navy accepted this analysis and my decision.

Warfighter Information Network – Tactical (WIN-T) Usability Concerns

Original Interface (complex) New Interface (Simplified)

WIN-T is an Army communications system using both satellite and terrestrial datalinks. It allows soldiers to exchange information in tactical situations.

The initial testing of WIN-T focused on its technical performance. Testing revealed not only poor technical performance, but also problems with the complexity of the system. Even when the software and hardware were properly functioning, soldiers found the system difficult to operate. Usability has been a key concern as WIN-T has since been upgraded over the years.

Subsequent testing focused on improvements to the man/machine interface that soldiers use to operate the system on the battlefield. As depicted above, the original interface was complex and difficult to read. The interface had multiple sub-menus and when the system failed, it could take 40 minutes to an hour to restart it. The new interface is far simpler.

Testers used surveys to evaluate the difficulties that soldiers had when using the system. The Army initially constructed surveys that were complex, with nested questions and "Not Applicable" as a potential response. DOT&E encouraged the test and evaluation community to incorporate survey science into the testing, and worked with the Army to improve the surveys. The revised surveys are simpler, more meaningful, more likely to be completed reliably, and easier to interpret. Well-designed surveys allow operational evaluations to rigorously incorporate the soldiers' experience and are crucial for DOT&E evaluations and reporting to Congress.

to critically evaluate the usability of military systems as well as the workload, fatigue, and frustration that operators experience while employing the system. Surveys are often the only means to evaluate these issues; proper scientific survey design must be done to ensure that the data collected to evaluate the quality of human-system interactions are valid and reliable.

Methodologies for Cybersecurity Testing and Analysis

Improving our understanding of the cyber threat, including recognizing that cybersecurity applies to more than automated information systems, and improving the rigor of cyber testing rigor have been two of my office's more notable achievements. Most military systems, networks, and missions are susceptible to degradation as a result of cyber-attacks. DOT&E evaluates the cybersecurity posture of units equipped with systems and live DOD networks during operational testing and Combatant Command and Service exercises. Important efforts include our continued emphasis on identifying how cybersecurity affects operational missions, inclusion of cyber defenses in tests, improvement of Red Team skills, and analytical methodologies and measures. We have also advocated for overarching cyber assessments that focused on identifying cross-cutting problems for the Department to address. In 2014, I published comprehensive guidance to the OTAs, updating and reinforcing guidance we have been using since Congress directed DOT&E perform annual evaluations of Combatant Command and Service cybersecurity postures in 2002. The DOD acquisition process should deliver systems that provide secure and resilient cyber capabilities; therefore, operational testing must examine system performance in the presence of realistic cyber threats. My 2014 guidance specifies that operational testing should include a cooperative vulnerability and penetration assessment phase to identify system vulnerabilities followed by an adversarial assessment phase to exploit vulnerabilities and assess mission effects. My guidance encourages program managers to address cybersecurity vulnerabilities that are discovered during the cooperative vulnerability and penetration assessment, prior to conducting the adversarial assessment. Despite this, adversarial assessments often find exploitable mission-critical vulnerabilities that earlier technical testing could have mitigated.

My office continues to emphasize the need to assess the effects of a debilitating cyber-attack on the users of DOD systems so that we understand the impact to a unit's mission success. A demonstration of these mission effects is often not practicable during operational testing due to operational safety or security reasons. I have therefore advocated that tests use simulations, closed environments, cyber ranges, or other validated and operationally representative tools to demonstrate the mission effects resulting from realistic cyber-attacks. Representative cyber environments hosted at cyber ranges and labs provide one means to accomplish the above goals. Such cyber ranges and labs provide realistic network environments representative of warfighter systems, network defenses, and operators, and they can emulate adversary targets and offensive/defensive capabilities without concern for harmful effects to actual in-service systems/networks. For several years, I have proposed enhancements to existing facilities to create the DOD Enterprise Cyber Range Environment (DECRE), which is comprised of the National Cyber Range (NCR); the DOD Cybersecurity Range; the Joint Information Operations Range; and the Joint Staff J-6 Command, Control, Communications, and Computers Assessments Division. The need and use of these resources is beginning to outpace the existing DECRE capabilities. As an example, the NCR experienced a substantial increase in customers the last few years.

Cybersecurity continues to evolve rapidly as both new threats and new defensive capabilities emerge and are fielded. Our ability to test and evaluate the DOD's cyber posture must keep pace with these advancements by accelerating development of appropriate tools and techniques. For example, Programmable Logic Controllers (PLCs) are ubiquitous in both fixed installations and deployable platforms, such as ships and aircraft. DOT&E has provided guidance on the necessity for caution in testing these components due to risk of platform damage caused by a PLC that is compromised, and has invested in the development of safe test and evaluation techniques for PLCs. Test agencies must continue to use all available tools and resources to assess PLCs and other industrial control systems used in DOD platforms. Other cybersecurity test challenges include:

- Systems with non-Internet Protocol data transmission (e.g., Military Standard 1553 data bus)
- Multiple Spectrum Cyber Threats (e.g., via non-computer based networks)
- Customized attacks
- End-to-end testing to include key subsystems, peripherals, and plug-ins
- Cloud computing

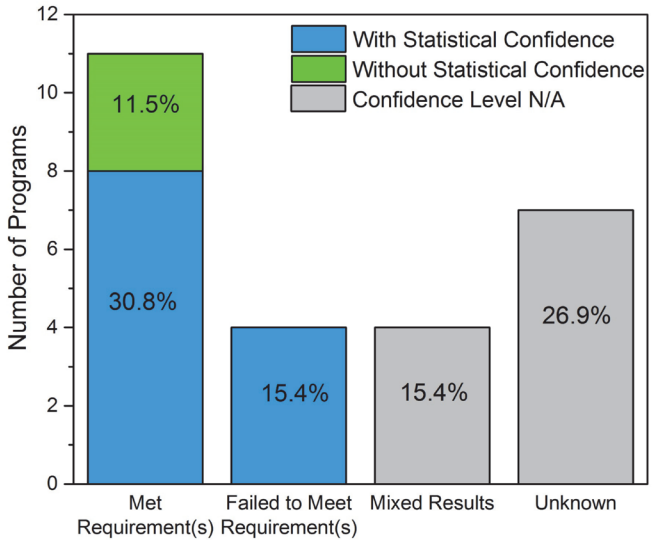
The Services' OTAs have established a cybersecurity technical exchange forum to discuss ongoing challenges and share solutions and lessons learned to improve overall cybersecurity operational test process. There were two meetings this year, which also included DOT&E participation. These interchanges are a good step forward for the operational test community to keep pace with the threat.

Design for Reliability

I similarly made improvement of system reliability a top priority – through initial design and early testing rather than discovering shortfalls at the end of development in operational testing. In my office's evaluation of oversight programs, we continue to see rising compliance with the policies set forth in the DODI 5000.02 and DOT&E guidance memos. The use of reliability growth curves as a tool to monitor progress of a system's reliability is now standard practice. The most successful programs are incorporating reliability growth into their contracts and have reliability thresholds as KPPs

However, change takes time and, despite the Department's continued efforts to emphasize the importance of reliability, defense systems continue to demonstrate poor reliability in operational testing. Only 11 of 26 systems (42 percent) that had a post-Milestone C operational test in FY16 met their reliability requirements. The remaining 15 systems either failed to meet their requirements (15 percent), met their requirements on some (but not all) parts of the overall system of systems (15 percent), or could not be assessed because of limited test data or the absence of a reliability requirement (27 percent).

Analysis of these recent operational tests indicates that one of the challenges in demonstrating whether a system meets its reliability requirement in operational testing is planning a long enough test. While tests are generally not scoped with respect to the reliability requirement, sufficient data should be captured throughout all testing phases to determine the reliability of the system as it compares to the requirements. The operational test scope for many systems is not long enough to demonstrate reliability requirements with statistical confidence. Over the past 3 years, 13 percent of requirements have planned test lengths shorter than the requirement itself. For systems with high reliability requirements, it is particularly important to intelligently use test data from all available sources. When system reliability is poor, even a short test might be adequate to prove the system did not meet its reliability requirement.



DISTRIBUTION OF RELIABILITY RESULTS FOR POST-MILESTONE C TESTING IN FY16 (UNKNOWN RESULTS INDICATE EITHER NOT ENOUGH DATA TO EVALUATE OR NO RELIABILITY REQUIREMENT)

Methodologies for Combining Data from Multiple Tests

While rigorous operational testing is paramount to this office's assessment of operational effectiveness, suitability, and survivability, it is not always possible or practical to obtain all of the information required for our assessments in an operational test. My office has supported the use of all information in operational evaluations in order to provide the best assessments available and use test resources in the most responsible fashion. In recent guidance updates, we have provided a pathway for using developmental test data in operational evaluations. We have enthusiastically advocated for considering all of the information available in reliability assessments.

Rigorous Validation of Modeling and Simulation (M&S)

Another focus area we are just beginning to influence is the rigorous validation of M&S that are to be used in the evaluation of a system's combat effectiveness and suitability. I expect the validation of M&S to include the same rigorous statistical and analytical principles that have become standard practice when designing live tests. All M&S, when used to support

Elements of a Successful Reliability Growth Program: Joint Light Tactical Vehicle (JLTV)

The JLTV is a partial replacement for the High Mobility Multi-purpose Wheeled Vehicle (HMMWV) fleet. The JLTV program presented a unique opportunity to understand the factors that contribute to a successful reliability outcome because three vendors competed during the Engineering and Manufacturing Development Phase. Each vendor implemented a reliability growth program and conducted extensive testing, but only one of the vendors met the program's reliability goals. Comparing the performance of the three vendors indicates that programs should:

- Review and approve failure definition scoring criteria early to improve vendors' understanding of government priorities.
- Encourage vendors to base initial reliability predictions on operationally representative test data, to include the system, test conditions, and approved failure scoring procedures.
- Allow adequate time and funding to grow system reliability.
- Address failure modes at all severity levels; non-aborting failures may degrade the system and cause system aborts. Addressing these failures early also reduces the maintenance and logistics burden and improves system availability. Ensure there will be enough testing to support a comparative evaluation of vendor reliability outcomes for competitive programs.

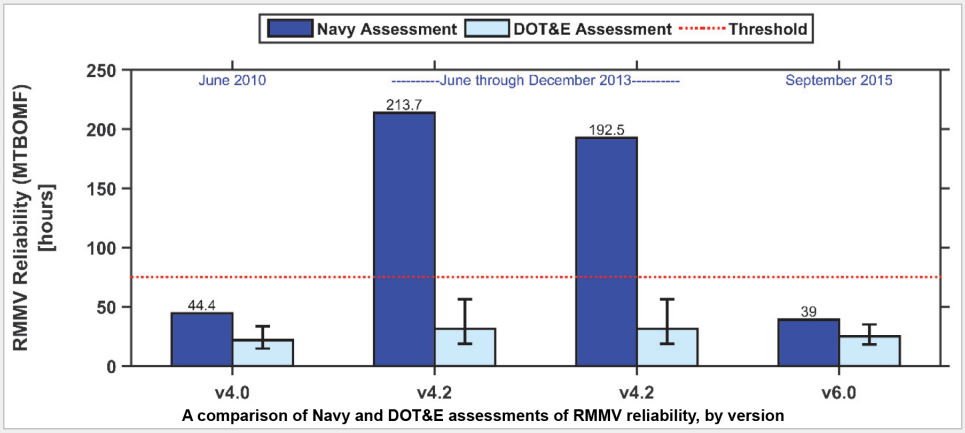
Statistically Based Reliability Analyses: Remote Multi-Mission Vehicle (RMMV)

The Remote Minehunting System (RMS) uses the RMMV, which is an unmanned, diesel-powered, semi-submersible vehicle, to tow a minehunting sonar (the AN/AQS-20 variable depth sensor).

From 2005 to 2009, the system exhibited reliability problems in nearly all periods of developmental and operational testing, twice failing to complete a planned IOT&E because of poor reliability, and ultimately experienced a Nunn-McCurdy breach. Following a Nunn-McCurdy review in 2010, USD(AT&L) directed the Navy to restructure the RMS program and fund and implement a three-phase RMMV reliability growth program.

Following combined developmental and integrated testing in 2013 (after the Navy concluded its reliability growth program), DOT&E assessed RMMV (v4.2) reliability as 31.3 hours Mean Time Between Operational Mission Failure (MTBOMF), less than half the Navy's requirement of 75 hours MTBOMF; further, DOT&E's statistical analysis of all test results indicated that reliability had not actually improved. Navy officials asserted that RMMV (v4.2) had demonstrated remarkable reliability improvements, testifying to Congress in 2013 that testing had shown reliability "substantially exceeding requirements" and in 2014 that the system "continues to test well." Throughout 2014, DOT&E detailed its analyses of RMMV v4.2 reliability in multiple memoranda to USD(AT&L) refuting the Navy's unsubstantiated claims that it had achieved reliability requirements and demonstrated readiness to restart low-rate initial production.

The Navy subsequently upgraded the RMMV v4.2 to make it compatible with the Littoral Combat Ship's (LCS) communications and launch, handling, and recovery systems and commenced ship-based testing of the so-called RMMV v6.0. This version of the system continued to experience reliability problems. In an August 2015 memorandum, DOT&E advised USD(AT&L) that the reliability of the RMS and its RMMV v6.0 was so poor that it posed a significant risk to the planned operational test of the *Independence*-variant LCS and the Increment 1 mine countermeasures (MCM) mission package and to the Navy's plan to field and sustain a viable LCS-based minehunting and mine clearance capability prior to FY20. Test data continued to refute the Navy's assertion that vehicle reliability had improved and statistical measures employed by DOT&E showed "no confidence or statistical evidence of growth in reliability over time" between RMMV v4.0, v4.2, and v6.0.



A comparison of Navy and DOT&E assessments of RMMV reliability, by version

In October 2015, the Navy delayed operational testing of the *Independence*-variant LCS equipped with the first increment of the MCM mission package pending the outcome of an independent program review, including an evaluation of potential alternatives to the RMS. The Navy chartered the review in response to an August 21, 2015, letter from Senators John McCain and Jack Reed, Chairman and Ranking Member of the Senate Committee on Armed Forces expressing concerns about the readiness to enter operational testing given the significant reliability problems observed during testing in 2015. In early 2016, following the completion of the independent review, among other actions, the Navy canceled the RMS program, halted further RMMV procurement, abandoned plans to conduct operational testing of individual MCM mission package increments, and delayed the start of LCS MCM mission package IOT&E until at least FY20. After canceling the RMS program, the Navy also announced its intention to evaluate alternatives to the RMS.

Ironically, the Navy's mine warfare resource sponsor identified a multi-function unmanned surface vessel (USV) as a "game changer" and potential RMMV replacement in 2012. In the years that followed, however, Navy officials touted RMMV reliability improvements that never materialized, reported inflated reliability estimates based on incorrect analysis, and funded additional RMMV development. The Navy did not use robust statistical analysis to assess RMMV performance objectively nor did it prioritize development of a multi-function USV capable of integrating with the RMS's towed sonar. These choices have left the Navy without a viable means of towing improved sonars when the contractor delivers initial production units next year and could delay realistic testing and fielding of the system until FY20. By accepting objective analysis of RMMV performance and committing to the USV sooner, the Navy could have avoided this unfortunate position and saved millions in RMMV development costs.

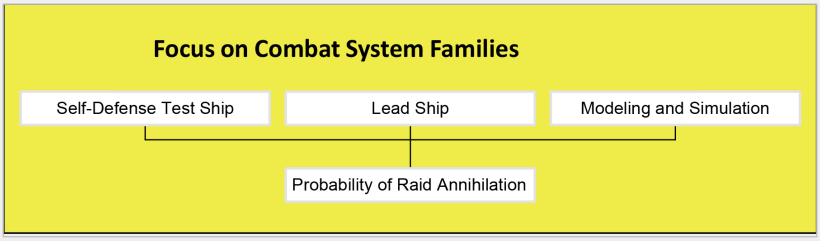
Despite DOT&E's reporting, USD(AT&L) published in its annual Developmental Test and Evaluation (DT&E) reports in March 2015 and March 2016 that RMMV v6.0 "improves vehicle performance and reliability," and that RMMV v4.2 "demonstrated sufficient reliability growth to satisfy Nunn-McCurdy requirements," citing a debunked, inflated reliability estimate of 75.3 hours MTBOMF. Such assurances from USD(AT&L) and the Navy misled their audience as to the seriousness of the problems the RMS program faced in delivering a necessary capability to the warfighter.

operational tests and evaluations, should not be accredited until a rigorous comparison of live data to the model's predictions is done. Testers should focus on the validation of the full system or environment being emulated.

Scientific Test and Analysis Techniques Center of Excellence

The Deputy Assistant Secretary of Defense for Developmental Test & Evaluation (DASD DT&E) / Director, Test Resource Management Center (TRMC) and my office continue to work collaboratively to advance the use of scientific approaches to test and evaluation. In 2011, DASD DT&E signed the Scientific Test and Analysis Techniques (STAT) Implementation Plan, which endorses these methods and created the STAT Center of Excellence (COE). The STAT COE provides program managers with the scientific and statistical expertise to plan efficient tests that ensure that programs obtain valuable information from the test program. Since 2012 when the STAT COE was formed, I have noted that programs who engage with the STAT COE early have better structured test programs that will provide valuable information. The STAT COE has provided these programs with direct access to experts in test science methods, which would otherwise have been unavailable. However, the COE's success has been hampered by unclear funding commitments. The COE must have the ability to provide independent assessments to programs (independent of the program office). Furthermore, the COE needs additional funding to aid program managers in smaller acquisition programs. Smaller programs with limited budgets do not have access to strong statistical help in their test programs and cannot afford to hire a full-time PhD-level statistician to aid their developmental test program; having access to these capabilities in the STAT COE on an as-needed basis is one means to enable these programs to plan and execute more statistically robust developmental tests. Finally, the STAT COE has also developed excellent best practices and case studies for the T&E community.

Enterprise Strategy – Testing Naval Air Defense



In 1996, the Navy defined the self-defense capability against anti-ship cruise missiles (ASCMs) that all new ship classes were required to have. This probabilistic self-defense requirement is known as the probability of raid annihilation (PRA) requirement. The PRA requirement states that a ship must defeat a raid of ASCMs, arriving within a short time window, such that no ASCMs hit the ship, and specifies with what probability of success this must be achieved. With assistance from DOT&E, the Navy developed a strategy for assessing this requirement with end-to-end testing of integrated combat systems for all new ship classes (e.g., USS *San Antonio* class, USS *America* class, USS *Zumwalt* class.). The combat systems on U.S. Navy ships are composed of many systems, which are developed by separate program offices. Before this new "enterprise" strategy, no one program office was responsible for developing the overall test program. One goal of the strategy was to consolidate all testing requirements from all sources, developmental or operational testing, for individual systems or for the overall ship, and truly create an integrated test program.

Among other things, this new enterprise strategy intended to address testing the ship-class PRA requirement and to provide for a more efficient use of test resources for conducting anti-air warfare ship self-defense testing. By addressing multiple ship class and combat system element requirements in an integrated test strategy, the Navy was able to reduce the total amount of testing required. Before using the enterprise strategy, each ship class and individual system would develop its own test program. With the enterprise strategy, a test program for the family of combat systems is developed. This allows testing to focus on the overall end-to-end mission of ship self-defense and eliminates duplicative testing. As an example, USS *San Antonio* and USS *America* are both amphibious ships that operate in similar environments against similar threats. The equipment on the *San Antonio* is a subset of the equipment on the *Americ*a.

This enterprise strategy was successfully applied to the USS *San Antonio* class. For the USS *America* class, the enterprise approach permitted testing to focus on the added components (SPS-49 radar and Evolved SeaSparrow Missile (ESSM) integration) and on incremental upgrades to the other systems. As with the USS *San Antonio* assessment, the USS *America* assessment is satisfying the ship's PRA requirements, requirements for the Block 2 Rolling Airframe Missile (RAM Blk 2), and for the Mark 2 Ship Self-Defense System (SSDS MK 2). Prior to the enterprise strategy, the Navy pursued individual test programs for each system that would have required many tests, each very similar in nature, be executed. Before adopting the enterprise approach, the Navy estimated they would spend $1.1 Billion on ship self-defense testing against cruise missiles between FY05 and FY15. The enterprise strategy reduced those costs by $240 Million and continues to provide a means to optimize the use of scarce and expensive resources.

Additionally savings related to the enterprise strategy are the results of a common modeling and simulation (M&S) paradigm for assessing the PRA requirement and some other combat system requirements. In the case of RAM Blk 2 and USS *America*, both programs needed end-to-end representations of the ship's combat system to test requirements. In this example, the M&S suite developed to assess the ship's PRA requirement is also being used to assess the missile probability of kill requirement. By using the same M&S paradigm, the live testing needed to support the verification, validation, and accreditation is also reduced. A similar approach will be applied to the next flight of the USS *America* class (i.e. LHA 8) and its combat system elements (SSDS MK 2, the Block 2 ESSM, and the Enterprise Air Surveillance Radar) and to other new ship programs (e.g., USS *Arleigh Burke* Flight III) and their combat system elements (e.g., SPY-6 Air and Missile Defense Radar).

Science of Test Research Consortium

As we work to apply more rigorous approaches to the test and evaluation of defense systems, challenges inevitably arise that demand new approaches. In collaboration with TRMC since 2011, my office continues to fund the Science of Test Research Consortium. The consortium pulls together experts in experimental design, statistical analyses, reliability, and M&S from Naval Post Graduate School, the Air Force Institute of Technology, and six additional universities. The Science of Test Research Consortium supports both the development of new techniques as well as a link between academia and the T&E community and a pipeline of graduates who could enter the T&E workforce. As advances occur in statistics, the research consortium keeps the T&E community aware of those changes. Additionally, they are working to focus research efforts on the unique challenges of operational test and evaluation that require new statistical methods. The consortium is essential for ensuring we remain well-informed of new techniques and improvements to existing techniques.

Science of Test Workshop

This past year my office, in collaboration with NASA and the Institute for Defense Analyses, supported the inaugural Test Science Workshop, which was designed to build a community around statistical approaches to test and evaluation in defense and aerospace. The workshop brought together practitioners, analysts, technical leadership, and statistical academics for a 3-day exchange of information, with opportunities to attend world-renowned short courses, share common challenges, and learn new skill sets from a variety of tutorials.

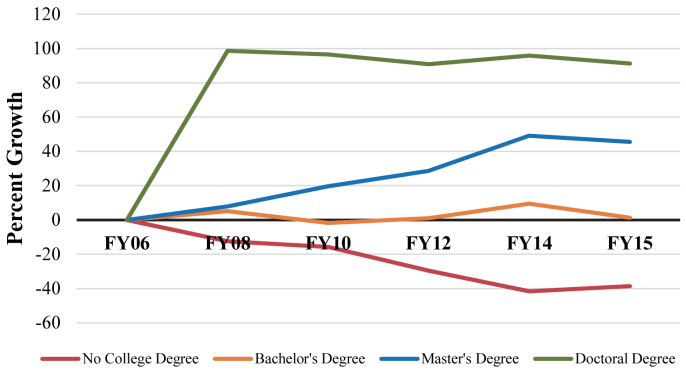
The Workshop promoted the exchange of ideas between practitioners in the T&E community with academic experts in the research consortium. Over 200 analysts from across the federal government and military Services benefited from training sessions, technical sessions, and case studies showcasing best practices. The feedback from participants was overwhelmingly positive, reinforcing that the event was much needed in the DOD and NASA analytical communities. The high response rate and enthusiastic comments indicated a clear desire to attend such events in the future.

Workforce

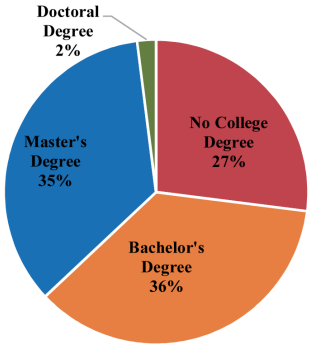
Rigorous and operationally realistic testing requires a skilled workforce capable of understanding the systems under test and applying scientific, statistical and analytical techniques to evaluate those systems. It is critical that personnel in the Operational Test Agencies (OTAs) have strong scientific and analytical backgrounds. In 2012, DOT&E conducted a workforce study and recommended that each OTA (1) increase the number of civilian employees with scientific, technology, engineering, and mathematics (STEM) backgrounds, (2) acquire at least one subject matter expert with an advanced degree in statistics, operations research, or systems engineering, and (3) continue to recruit military officers with operational, fleet experience.

Currently, the OTA workforce consists of roughly half civilian (51 percent) and half military (49 percent) personnel. While the overall size of the workforce has declined since 2006, the proportion of civilian personnel with advanced degrees has grown by 136 percent. The number of civilian personnel with master's and doctoral degrees increased by 45 percent and 91 percent, respectively. Currently, 2 percent of civilian personnel hold doctoral degrees, 35 percent hold master's degrees, 36 percent hold bachelor's degrees, and 27 percent do not possess a college degree. These trends are similar for each OTA and indicate that overall, OTA civilian personnel are more educated today than they were a decade ago.

Only 56 percent of civilian personnel in the OTA workforce currently hold a degree in a STEM field. However, this number includes all OTA civilian personnel, including those who do not directly engage in operational testing, such as administrators and security personnel. The proportion of civilian personnel with a degree in a STEM field increases to 72 percent when



Left: Growth in the number of civilian personnel with different degree types from FY06-FY15



Right: Proportion of civilian personnel with different degree in FY15

EDUCATION DISTRIBUTION OF CIVILIAN PERSONNEL IN THE OPERATIONAL TEST AGENCIES, FY06-FY15

these individuals are excluded, closely mirroring the proportion reported in 2012 (75 percent). Since 2012 all OTAs have acquired at least one expert with a background in statistics, operations research, or systems engineering.

The OTAs are making steady progress toward achieving the recommendations that DOT&E outlined in the 2012. The two most notable improvements since 2012 are they have all acquired expertise in statistics, operations research, or systems engineering and overall there has been an increase in the number of personnel with master's degrees.

All of the OTAs have also made significant investments in improving their capabilities for implementing rigorous statistical methods. They have updated their internal guidance and procedures to reflect DOT&E guidance. Additionally, they have all invested in training on experimental design and survey design enabling the existing workforce to better use these methods in developing and analyzing operational tests.

As military systems grow in complexity and capability, however, the need for personnel with advanced analytical capabilities, who understand scientific test design and statistics techniques, will become increasingly important and OTA hiring processes will need to continue to emphasize STEM fields.

VALUE OF INDEPENDENCE

In 1983, Congress directed OSD to create the DOT&E office, and the Director was given specific authorities in title 10 U.S. Code. The Congressional concerns that led to the establishment of this office were many, but included: poor performance of weapon systems, inaccurate reports from the Services, shortcuts in testing because of budget pressure, and a lack of realistic combat conditions and threats in testing. The unique independence of this office, free from conflicts of interest or pressure from Service senior leadership allows us to:

- Illuminate problems to DOD and Congressional Leadership to inform their decisions before production or deployment
- Tell the unvarnished truth
- Ensure operational tests are adequately designed and executed

As Director, OT&E, I do not make acquisition decisions but inform those who make them about weapon system performance under combat conditions. My staff is composed of over one-third active duty military officers from all Services in addition to civilians with advanced engineering and science degrees. Our mission is to inform acquisition officials about how weapons will work in combat, including live fire survivability and lethality, before the systems are deployed.

The independence of this office allows us to require adequate and realistic operational testing and to advocate for resources to improve our T&E capabilities. I have observed that some of the most important capabilities or tests that we have prescribed have been met with substantial resistance from the Services, sometimes requiring adjudication by the Deputy Secretary of Defense; I describe the most important of these decisions below (the T&E Resources section of this report provides details of FY16 focus areas). In light of the remarkable resistance from the Services to prioritize adequate testing and test assets in their acquisition programs, it is even more apparent that the independence of this office is critical to the success of finding problems before systems are used in combat.

Improved Test Resources for Electronic Warfare

An alarming trend I have seen during my tenure is that our threats are increasing their capabilities faster than our test infrastructure. Through the yearly budget review process, I have advocated for resources to improve test range infrastructure to support rigorous testing of modern combat systems. Most notably, in 2012, I convinced the Department to invest nearly $500 Million in the Electronic Warfare Infrastructure Improvement Program (EWIIP) to upgrade open-air test ranges, anechoic chambers, and reprogramming laboratories in order to understand performance of the F-35 Joint Strike Fighter (JSF) and other advanced air platforms against near-peer threat integrated air defense systems. The open-air test and training ranges owned and operated by both the Air Force and Navy are lacking advanced threat systems that are being used in combat by our adversaries today, are proliferating, or are undergoing significant upgrades; yet both Services strongly resisted incorporating these modern threats that we proposed until directed to do so by the Deputy Secretary.

REPROGRAMMABLE GROUND-BASED RADAR SIGNAL EMULATOR FOR USE IN OPEN-AIR TESTING OF ADVANCED AIR PLATFORMS, INCLUDING THE JOINT STRIKE FIGHTER

Moreover, an important part of the JSF mission systems is the mission data file, which contains the settings that the JSF sensors use to identify signals detected from the threat's integrated air defense systems. The United States Reprogramming Laboratory (USRL) is responsible for building the mission data file. The USRL is also a recipient of resources DOT&E argued for with the EWIIP program. Unfortunately, even though funding for upgrades was provided in 2014, preventable but now insurmountable delays configuring the USRL will delay its ability to support JSF combat capabilities until at least mid-2018.

In 2016, my office again requested funding for infrastructure to support testing and training of additional advanced air warfare systems such as the Next Generation Jammer. This funding is intended to enable the test ranges and the models and simulations (that must be validated with test data) to assess the performance of U.S. systems against the key challenges of near peer threat air defense networks of the 2020s.

Fifth-Generation Aerial Target (5GAT)

In 2006, DOT&E sponsored a study on the design of a dedicated Fifth Generation threat aircraft to adequately represent characteristics of threat aircraft being deployed by our adversaries. Since then, DOT&E and TRMC have invested over $11 Million to mature the government-owned design. The Department provided funding to complete the final design, tooling, fabrication, and flight tests. The prototyping effort will provide cost-informed alternative design and manufacturing approaches for future aerial target acquisition programs. These data can also be used to assist with future weapon system development decisions as well as T&E planning and investment, and will support future T&E analysis of alternative activities.

Self-Defense Test Ship

In 2013, the Navy sadly re-learned in the accident aboard the USS *Chancellorsville* (CG 62) where a target drone impacted the ship, that the only safe way to test the complex close-in self-defense capabilities of a ship is to mount those capabilities on a remotely controlled, unmanned self-defense test ship (SDTS). And this was not the first time such an accident occurred. In 1983, a sailor was killed onboard USS *Antrim* (FFG 20) during a test. The safety risks associated with testing short-range, self-defense systems are significant and increasing with the increasing capabilities of modern anti-ship cruise missiles. Hence, it is necessary to have test assets such as the unmanned SDTS to conduct such testing.

The SDTS has been integral in the past in testing weapons systems and ship classes. Without it, significant limitations in the Navy's ability to defend surface combatants would not be understood. Furthermore, efforts to overcome these limitations could not be tested. Unfortunately, the Navy has been reluctant to extend the same investment to developing an SDTS equipped with an Aegis Combat System, Air and Missile Defense Radar (AMDR), and Enhanced SeaSparrow Missile (ESSM) Block 2 for adequate operational testing of the DDG 51 Flight III destroyer self-defense capabilities. The current SDTS lacks the appropriate sensors and other combat system elements to test these capabilities.

In 2014, the Navy published a study that claimed an Aegis-equipped SDTS was not necessary for operational testing; however, DOT&E refuted these claims, which use flawed justifications There is no short cut. Safety considerations preclude testing against realistic threats onboard manned ships. It has been demonstrated on numerous occasions that data from less stressing manned ship testing, where targets must be fired at large crossing angles and turned away from the ship at significant ranges, cannot be extrapolated to stressing, realistic threat encounters. Modeling and simulation (M&S) cannot replace live testing because without the SDTS there are no data to ensure that the M&S accurately portray live results.

In December 2014, the Deputy Secretary of Defense commissioned a study by the Director of Cost Assessment and Program Evaluation (CAPE) to provide options to deliver an at-sea test platform adequate for self-defense operational testing of the DDG 51 Flight III, the AMDR, and the ESSM Block 2 programs. CAPE provided three affordable alternatives and the Deputy Secretary directed the Navy to procure long-lead items to begin procurement of an Aegis-equipped SDTS. The Deputy Secretary further directed the Navy to work with DOT&E to develop an integrated test strategy for the DDG 51 Flight III, AMDR, Aegis Modernization, and ESSM Block 2 programs, and to document that strategy in a draft Test and Evaluation Master Plan (TEMP) to be submitted by July 2016.

Despite the clear need for an Aegis-equipped SDTS and the unambiguous direction of the Deputy Secretary, the Navy has, as of the signing of this report, not yet provided an integrated test strategy for these crucial programs; and although the Navy provided funding for the long-lead AMDR components, the Navy did not program funding in the Future Years Defense Plan to complete all other activities (including procuring Aegis Combat System equipment and targets) necessary to modify the SDTS and support adequate operational testing of the DDG 51 Flight III's self-defense capabilities in FY23 as planned. In November 2016, the Deputy Secretary again directed the Navy to fully fund those activities.

Full Ship Shock Trial (FSST) for CVN 78 and DDG 1000

In hostile areas, ships commonly face the threat of underwater shocks created by non-contact detonations of torpedoes, mines, or near miss air delivered weapons. These threats do not require precise targeting or the ship to sink because the shock from

a nearby miss can defeat critical mission capabilities by knocking motors and generators off-line and breaking equipment not adequately shock-mounted. Consequently, DOT&E requires shock trials for ships to test them for survivability against these widely prevalent threat types. The shock trial subjects combat-equipped ships to as operationally realistic an underwater shock load as possible while avoiding potential for crew injury and catastrophic damage. These trials are required before the first deployment of any ship class to allow for design improvements to the ship to make it more survivable in combat. Identifying these problems early in the construction of the class allows design changes to be more economically incorporated into follow-on ships. The early execution is especially critical, as each shock trial results in hundreds of findings of shock deficiencies that require correction and would not appear in M&S.

Unfortunately, the Navy, despite admitting in its technical warrants that "shock trials do have value and a return on investment," recommended in 2013 that the ship acquisition program forgo the use of shock trials as part of LFT&E or to meet Navy shock-hardening requirements. The Navy further attempted to delay shock trials on CVN 78 and DDG 1000 to later ships in the class, citing program schedule, cost, or operational availability above any scientific rationale. If the shock trial is delayed to later ships, it will occur after many years of operational deployment, exposing these ships to unnecessary risk from undiscovered and uncorrected vulnerabilities. After the Senate Armed Services Committee Chairman and Ranking Member expressed concern with this plan and urged restoration of the shock trial to the lead ship in the CVN 78 class, the Deputy Secretary directed the Navy to conduct shock trials on CVN 78 prior to first deployment, and on DDG 1000 or 1001 prior to the deployment of any ship of that class.

Warrior Injury Assessment Mannequin (WIAMan)

Commercial automotive crash test dummies were designed to assess injuries from the forces most commonly seen in civilian car accidents – sharp accelerations parallel to the ground as the car is rapidly (over milliseconds) pushed from the back, front, or side. In 2009, and repeatedly since, evaluations of combat injury data and the Department's underbody blast M&S capabilities have revealed these dummies, used only out of necessity, are wholly inadequate for predicting injuries in the direction that military vehicles and their occupants were being pushed in the field – upwards and over orders of magnitude shorter time frames resulting in completely different shock impacts. The fundamentally different nature of this impact and its effects on warfighters in vehicles exposed to an under-vehicle Improvised Explosive Device (IED), required initiating a new effort to increase DOD's previously poor understanding of the cause and nature of injuries incurred in underbody blast events, and as well as designing a military-specific anthropomorphic test device (ATD) to use in live fire test events replicating IED events.

The Department's shortcomings in this domain were a cause for concern for the Secretary of Defense in 2010. The DOT&E vulnerability assessment of the Mine Resistant Ambush Protected (MRAP) family of vehicles revealed that combat injuries, and not test data, proved that some MRAP variants provided significantly less protection than others. Upon receiving this news, Secretary Gates directed a review of the Department's underbody blast M&S capability gaps, and the top three gaps were all related to the ability to predict injuries to vehicle occupants after under-vehicle explosions. The subsequent directive to address these gaps came from senior OSD leadership, and, with initial funding from DOT&E, the Army began this project known as the Warrior Injury Assessment Manikin (WIAMan.)

Unfortunately, Army leadership continues to question the need for this capability, which threatens the successful execution of the WIAMan project, even though these threats are likely to persist into the future. The Army requirements community recognizes this threat, as demonstrated by the fact that all of their current and future ground platforms have some form of underbody protection requirement. Despite these survivability requirements for future ground combat vehicles, Army leadership continues to renew resistance to almost every aspect of the WIAMan project, from its requirements to its cost, and some claim, despite overwhelming evidence to the contrary, that the Department's current injury assessment capability is good enough. The Army Research Laboratory did not agree that the Department's current capability was adequate, and created the WIAMan Engineering Office (WEO) in 2012 to oversee the scientific research and ATD development to advance the state of the science. The WEO has led 5 years of successful research on injury assessment criteria by a consortium of university and government laboratories and the production of a prototype mannequin. Subsequently, in 2015 the Army decided that WIAMan should become an Acquisition Category II acquisition program of record similar to a combat weapon system with a formal program manager, but the Army did not provide any additional funding to establish this acquisition program office. All of the bureaucratic minutiae associated with a establishing a major program of record to build 40 articles costing less than $1 Million each has had a significantly negative impact on cost and schedule, with no demonstrable benefits. The personnel and resources required to stand up a program office whose only function is to support contracting is a questionable use of funding on a resource-constrained program. The Army should remove the WIAMan project from its acquisition system (thereby eliminating unnecessary bureaucratic overhead) and allow the WEO to develop a build-to-print prototype concept ATD; once its performance has been assessed as adequate by the WEO, the Army should solicit bids from industry to build the new ATD. A separate (unfunded) program office should not be required for this approach. As

the project is currently unfunded in its entirety past FY18, DOT&E remains concerned that the Army does not intend to ultimately complete this project.

The development and fielding of the WIAMan ATD will bring the Department on par with the civilian automotive world in its ability to accurately assess injuries from traumatic events. Despite the 2011 OSD and Army approval of a well-documented project scope driven by combat injuries, Army leadership is now requiring yet another round of justification on the injuries selected for inclusion in the WIAMan ATD, and Army acquisition leadership is expressing unease with incorporating these ATDs into live fire testing up to, and including, the Advanced Multi-Purpose Vehicle. In the view of DOT&E, 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.

MYTHS ABOUT OPERATIONAL TESTING

Over the course of more than 25 years in public service, I have found it lamentable that the acquisition bureaucracy in the DOD routinely promulgates unfortunate falsehoods. I have seen and heard many inaccurate claims of what DOT&E does and does not do, and inaccurate claims about system performance that are subsequently recanted or proven wrong by this office. These falsehoods can have deleterious impacts on programs. When a program manager makes false assertions regarding the impact of operational testing on programs, there is always a risk that people in leadership positions, who have little detailed knowledge of the program, will nonetheless believe the program manager and unwisely attempt to curtail operational testing – despite the fact that operational testing requires a small fraction of the overall program's cost and schedule and all too frequently identifies significant problems with performance for the first time.

Constrained defense budgets have existed throughout my tenure, which has resulted in questions about the value of operational testing. It has also been asserted that testing is a major cause of delays in defense programs and adds uncontrolled costs. A primary purpose of operational testing, and a key value of such testing, is to identify critical problems that can be seen only when systems are examined under the stresses of realistic combat conditions, prior to the Full-Rate Production decision and fielding to combat units. This identification permits corrective action to be taken before large quantities of a system are procured and avoids expensive retrofit of system modifications. The assertion that testing causes delays misses the essential point: fixing the deficiencies causes delays, not the testing. Furthermore, taking the time to correct serious performance problems is exactly what we desire in a properly-functioning acquisition system. We are not engaged in bureaucratic game play here; testing is not a game to be won. What we do is very serious. And yes, we need to highlight the performance problems that need to be fixed so that they can be fixed.

In response to the cost of operational testing, it is relevant to consider these costs relative to the acquisition costs of the systems themselves. Numerous studies have identified that the marginal cost of operational testing is small, in general less than 1 percent of a program's overall acquisition cost. This small relative cost stands in stark contrast with the potential savings from problems identified that can be corrected before full-rate production and the likely result that the system will work when called upon in combat.

While there has been concern over the cost of operational testing throughout my tenure, I have had the opportunity to observe firsthand how necessary an independent, objective operational test is to our acquisition system. Independent, operational testing not only provides objective information for

Inaccurate claims about Operational Testing

The USD(AT&L) requests yearly assessments from program managers concerning the challenges they face; these assessments are routinely shared with the defense community without critical factual review. In a recent assessment, a program manager expressed concern regarding the negative impacts of operational testing. The program manager asserted that three releases of a major automated information system had taken an average of 12 to 18 months to complete operational testing, and that:

> … the testing community has taken almost as long to operationally test the software as the program office took to develop it in the first place. Over time, this has contributed to the cost and schedule overruns … [and] delays in delivering important capabilities to users.

The program manager went on to say that this type of operational testing issue is "systemic to defense acquisition." These are classic examples of falsehoods routinely promoted by the acquisition community to divert attention away from the real issues of problems discovered in testing that must be fixed. In this case, the operational testing revealed the system was neither operationally effective nor survivable.

The claim that the operational tests took almost as long as development is refuted by the calendar: from the beginning of this program in 2006 to the end of the Multi-Service Operational Test and Evaluation in 2015 (9 years), it took a total of five months to conduct three operational tests, less than 5 percent of the program's duration. System design and development activities required the majority of the 9-year period. The claim that operational testing delayed delivering capabilities to users is also false, not only because operational testing did not contribute to delays, but also because DOT&E is not responsible for fielding decisions. In fact, limited fielding was authorized in 2006-2007 based on an urgent operational need.

the Congress and Defense leadership, but also provides critical information to programs on improving systems so warfighters are properly equipped.

Programs clearly have an incentive to denounce testing as unfair when it reveals performance problems. Cost and schedule overruns, especially those that are the direct result of poor program management, reflect poorly on program managers and program executive officers. However, by engaging in bureaucratic games, rationalizing problems, and minimizing testing, the result is a great disservice for the people for whom we work – men and women in combat whose lives depend on the systems we field to them. There's a terrible fear that exists that a negative DOT&E report will kill a program; however, it is much more likely that performance problems reported by DOT&E lead to a greater allocation of resources and time to fix them.

Bureaucratic process is no substitute for thought and common sense. Programs often complain that DOT&E requires testing beyond threshold requirements, or even threshold KPPs. As I discussed earlier, if programs were tested solely to their KPPs, we often would not be able to evaluate whether systems can accomplish their primary missions. While we must always pay attention to requirements documents, we also have to interact with the operators. We have to pay attention to the concepts of operation, to the war plans, to the intelligence information on the latest threats, and all of those things will tell us how to do an operational test under the circumstances the system will actually be used in combat and enable us to characterize the performance of systems across their operational envelope – not just at one key parameter. For example, I have heard program managers claim there are no requirements for cybersecurity, and therefore cybersecurity should not be tested. This is an extreme example of not using common sense but hiding behind ambiguous language in DOD directives.

Exaggerated Costs of Testing

DOT&E approved a TEMP in 2012 for a program with multiple software releases planned. Separate OT&E periods were planned for selected releases depending on the capabilities introduced. Operational testing was not required for versions without meaningful mission capability enhancements. In 2014, the Service restructured this program and approved critical KPP capabilities to be delivered with one of the versions that was not originally planned to have operational testing – the Service changes were a result of development of previous releases taking much longer than predicted. Successive rounds of developmental testing revealed repeated instability and inadequate performance. After the restructure, DOT&E required the Program Office to update their TEMP to reflect the new reality. In response, the program reported to USD(AT&L) that operational test requirements would add 3 months and $9 Million additional cost and schedule. This was contrary to the Service's Operational Test Agency (OTA) estimate that the testing would take approximately 30 days and cost approximately $300,000. The delays identified by the program manger were the result of unrealistic assumptions about development and integration time periods – not because of operational testing.

Inaccurate Claims Regarding Cybersecurity Test and Evaluation

Earlier this year, the USD(AT&L) requested Program Executive Officers (PEOs) provide him assessments of the challenges they confront in their jobs; these assessments were published in the Defense Acquisition University (DAU) online magazine without critical factual review. One PEO wrote that cyber testing and the ability to achieve a survivable rating from DOT&E was nearly impossible, adding that test criteria are not well defined. The PEO went on to say that threat portrayal exceeds the capabilities of a Blue Force Team (i.e., nation-state threat going against a brigade-level formation) and focuses on insider threats of unreasonable proportions. It was especially unfortunate for this to be published widely without comment because it could inevitably undermine the efforts the operational test community has taken to find and fix the significant cybersecurity issues present in most of our acquisition programs.

While the Joint Staff is making progress formalizing cybersecurity within the survivability KPP, Secretary Carter clearly stated his common-sense requirement that all the Department's weapon systems must undergo cybersecurity assessments. And consistent with DOT&E's statutory authority, we have published specific procedures and metrics to be used to conduct cybersecurity test and evaluation for over a decade.

We have routinely seen that DOD Red Teams need to use only novice skills to successfully attack our systems. Nonetheless, the intelligence community states that virtually all major defense acquisition programs will face advanced, nation-state cyber threats. Our assessments report results for both types of threats separately.

The intelligence community also consistently describes insider threats as the primary cybersecurity threat to acquisition programs. Bradley Manning and Edward Snowden are two insiders we know; we clearly do not know about all potential insider threats. Hence it would be grossly irresponsible for OT&E to not assess insider threats, which are obviously real.

LOOKING TOWARD THE FUTURE

As a community we have made immense progress in the past seven years. The need for rigorous and defensible approaches to test and evaluation is not going away. As our systems become even more complex, and autonomous, continuous and integrated testing will be necessary. We will need to continue to evolve our application of state-of-the-art methodologies to confront these new challenges. We will continue to need to update range resources.

Over the past seven years, we have put the framework in place, establishing the research consortium, science of test workshop in partnership with NASA, developing guidance including the TEMP Guidebook and others. However, this office as well as the Service test organizations, need to keep moving the trajectory forward so that we continue to provide valuable information to decision makers.

The operational test community should continue to provide independent, fact-based information to senior leaders and decision makers. The Service operational test organizations, like my office, are organized to be independent from the acquisition leadership. This is so that the facts, the unvarnished truth, can be reported to senior leadership without undue influence. However, in order for real change to take place in the acquisition system and to minimize future acquisition failures, leadership must actually make itself aware of the information provided by independent assessments of systems, critically question all the information they have, and use it to make sound decisions. I have provided numerous examples in this introduction where plenty of facts about systems are available; I have provided numerous methods and techniques to obtain the facts in an effective and efficient manner depending on the program involved. But unless leaders in the department display the intellectual curiosity to create a demand signal for accurate information about their programs, and the moral courage to act faithfully on that information once it's generated, acquisition reform cannot occur. Only when leaders have the authority and confidence to say "No," when the facts reveal that a course deviation is essential to a program, change will occur. The willingness and ability to say "No" to high-risk schedules, optimistic cost estimates, and optimistic claims of technical readiness and to support those decisions within and outside the Department using cogent arguments based on the facts are essential. Leadership that does this sends a strong message by directly challenging the powerful incentives that can otherwise lead to the adoption of unachievable requirements embodied in high-risk programs that fail. While there is constant criticism of DOT&E and the Services' independent activities and pressure to constrain our independence, continued strong support by the Congress and successive Administrations of these pockets of independent and objective expertise and evaluation remains, in my view, essential.

I cannot emphasize enough the need for early, adequate, realistic, and rigorous independent operational testing on all systems to ensure what is being developed will, in fact, provide our Service men and women the capabilities they need in combat. This is especially true during this period of tight budget controls as there are not sufficient resources to correct significant problems once systems are fielded.

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 2016.

J. Michael Gilmore
Director