

INTRODUCTION



FY 2013 Annual Report

The United States' Department of Defense (DoD) acquires some of the most complex systems known. Because of this complexity, they often require many years of development and testing; and if not tested properly, we run the very serious risk of delivering poorly performing equipment to the warfighter. Our Airmen, Sailors, Marines, and Soldiers rely on these systems to be effective, suitable, survivable, and lethal. Because in many respects their lives depend on weapons systems that work, it is essential that adequate testing is done to fully characterize those systems' capabilities and shortcomings across all of the relevant operational conditions in which the system is anticipated to be employed. Such characterization is needed in part so that well-informed acquisition and development decisions can be made, but also so the men and women in combat understand what these systems can and cannot do. As a nation, we cannot afford to field weapons systems that do not work, do not provide a clear improvement over existing systems, or are not militarily useful; nor can we afford to make these important fielding decisions without knowledge of the systems' operational effectiveness.

Time and again I have found that without adequate operational testing, we would not have understood the specific conditions in which a system is effective and suitable; my reporting continues to be focused on this characterization, since no system can provide perfect performance under all operational conditions or against all relevant threats. Provided the information gained from operational testing is used, characterization of performance as a function of operational conditions and threats enables developers to understand and fix problems quickly. Early testing (both developmental test events and operational assessments) can and should inform the development process and enable the early identification of major problems.

The requirement for adequate operational testing is part of the natural and healthy tension between the testing, acquisition, and requirements communities. This year, I have found several cases where the testing I determined to be adequate was beyond the narrow definitions in the requirements document(s) established by the Services and Joint Staff. I have observed two distinct limitations in requirements definitions:

- Requirements stated in terms of technical parameters that are not mission-oriented
- Requirements that are narrowly defined to specific conditions, when the Services will certainly employ the system in other conditions

I provided a specific example of the former case to the Vice Chairman of the Joint Chiefs of Staff. I found that the P-8A Multi-Mission Maritime Patrol Aircraft could be fully compliant with all Key Performance Parameter (KPP) and Key System Attribute (KSA) threshold requirements, and nonetheless possess significant shortfalls in mission effectiveness. The P-8 requirements define supporting system characteristics or attributes that are necessary, but not nearly sufficient, to ensure mission effectiveness. In an extreme case, the contractor could deliver an aircraft that meets all the KPPs but has no mission capability whatsoever. Such an airplane would only have to be designed to be reliable, equipped with self-protection features and radios, and capable of transporting weapons and sonobuoys across the specified distances, but would not actually have to have the ability to successfully find and sink threat submarines in an Anti-Submarine Warfare mission (its primary mission). The lack of KPPs/KSAs related directly to mission effectiveness will inevitably create a disconnect between the determination of operational effectiveness in test reports and the KPP and KSA compliance assessments that typically drive program reviews throughout development. The Department could therefore be making early acquisition decisions on the basis of standards that are useful, but do not capture the primary reason for procuring these systems: to provide a warfighting capability.

For the second case mentioned above, where requirements are too narrowly defined, I remain committed to conducting adequate testing in all the relevant operational conditions in which men and women in combat will employ the system. Requirements may be too narrowly defined because there is a common concern that failing to specify a certain, limited set of conditions could lead to an unwieldy or excessive test. The need to test neither too much nor too little is a key reason DOT&E is using Design of Experiments (DOE) methods to plan testing that efficiently spans the operational envelope. The DOE method is rooted in a structured and disciplined determination of the operational envelope. In some cases, a clear understanding of the operational envelope reveals the need to conduct testing in conditions not specified in the requirements documents, and such testing does indeed require funding for additional test events or test resources. Such investments are essential, and in my view, must be done to ensure prompt delivery of effective, suitable, and survivable warfighting

INTRODUCTION

capabilities. Test costs represent a small fraction of the cost of the program, and operational testing of conditions that are outside the scope of the requirements documents is usually the only venue by which system performance can be determined under those conditions. The Department cannot shrink from the need to conduct adequate testing.

As an important example of the above principle, I mention briefly the need for conducting testing of our Navy's current and future combat systems on destroyers, cruisers, and other "big-deck" surface ships. We currently use an unmanned, remotely controlled ship, called the Self-Defense Test Ship (SDTS), with the actual radars, weapons, and combat systems employed on some (not all) of these ships to examine the ability of these systems to protect against incoming anti-ship cruise missiles. The use of an unmanned, remotely controlled ship is essential, since conducting most engagements in the self-defense (close-in) region is not possible on manned ships due to safety considerations. Furthermore, modeling and simulation efforts, while useful, have not been able to reproduce the results of many of these tests. For the future radar and combat system now in development for the DDG 51 Flight III ships, we must conduct adequate testing under all relevant operational conditions. These conditions include examining end-to-end combat system performance against multiple simultaneous threat missiles within the self-defense zone of the ship, where manned testing is impossible. An SDTS is therefore essential for an adequate operational test. Previous testing has revealed for the combat systems of amphibious assault ships and carriers that without the use of an SDTS, critical problems in defending against certain threats would not have been found. Now, because of that test resource, many of those combat system problems have been corrected, and our Sailors are safer from harm. We cannot afford to not test future DDG combat systems and radars under stressing conditions in the self-defense zone, particularly since the DDGs themselves provide the defensive shield for the battlegroup. Our nation needs to pursue the testing and resources necessary to ensure system performance is understood in all regions of the operational envelope.

In the remainder of this Introduction, I briefly describe the other areas of focus for my office. These include:

- My continued emphasis on the need for statistical rigor in both the planning of operational tests and in the analysis of data from testing.
- My continued emphasis on the need to improve reliability of all weapons systems – here I include an assessment of new policies on reliability growth and tracking, as well as how the Department is progressing in improving reliability of weapons systems.
- My observations of software-intensive system development and testing, including the vulnerability of business systems.
- Other areas of interest, including cybersecurity testing and test protocols for personal protective equipment. My assessment of critical test resources is also a focus area, but discussed in a separate section of this report.
- An assessment of problem discovery during testing – this section of the report was added in 2011 based on concerns from Congress that significant problems in acquisition programs are being discovered during operational testing that arguably should have been discovered in developmental testing (page 13 in the DOT&E Activity and Oversight section).

CONTINUED EFFORTS TO ENSURE RIGOROUS, DEFENSIBLE, AND EFFICIENT TESTING

At my confirmation hearing in September 2009, I pledged to work to "assure that all systems undergo rigorous operational test and evaluation in order to determine whether they are operationally effective, suitable, and survivable." A rigorous operational test characterizes a system's end-to-end mission effectiveness across the operational envelope and quantifies the risk in such assessments. Statistical methods, including DOE, provide a defensible methodology for ensuring the adequacy of any test. These methods encapsulate the need to "do more without more," especially in light of a highly constrained fiscal environment. They provide a methodology for optimizing scarce test resources, ensuring that each test point provides the maximum information for my evaluation. They provide sound rationale for the level of testing prescribed, ensuring that we avoid either over-testing or under-testing weapons systems. Finally, they ensure we gather the data needed to provide men and women in combat confidence in evaluations of the performance of those weapons systems. In October 2010, I communicated to the Operational Test Agencies (OTAs) and Service T&E Executives my expectations regarding the use of DOE for developing rigorous, adequate, and defensible test programs and for evaluating their results.

The statistical methods that I have made key to my assessment of test adequacy constitute well-established best practices in both industry and government at large. The pharmaceutical, automotive, agriculture, and chemical and process industries, where many of these techniques were originally developed, all use the same statistical methods for test design and analysis that I advocate. Furthermore, other government agencies such as the Food and Drug Administration, Census Bureau, the National Laboratories that ensure the integrity of our nation's nuclear stockpile, as well as the National Aeronautics and Space Administration, which also engage in the testing of large and/or complex systems (similar to the DoD), all rely on the use of these methods.

INTRODUCTION

There has been much progress in increasing the statistical rigor of test plans since 2009. Over the past several years, all of the OTAs have implemented DOE practices to varying degrees and have offered training to their staff on the statistical principles of DOE. Additionally, the Deputy Assistant Secretary of Defense for Developmental Test and Evaluation (DASD(DT&E)) endorses these methods and advocates them through his Scientific Test and Analysis Techniques (STAT) implementation plan. That office has also stood up a STAT Test and Evaluation Center of Excellence, which employs qualified statistics experts to aid acquisition program managers in applying advanced statistical techniques to the design of developmental tests and analysis of resulting data.

However, there is still variability in the application of these tools across the Services' T&E communities. To that end, my office has recently completed a "Roadmap" to institutionalize test science and statistical rigor in T&E (the published version can be found here: <http://www.dote.osd.mil/pub/reports/20130711TestScienceRoadmapReport.pdf>). Additionally, I continue to provide guidance on best practices on the employment of these methods in OT&E. This year, I provided two additional guidance memos that address misconceptions and highlight best practices for employing DOE in OT&E. Below, I provide a summary of this most recent guidance on the use of DOE in operational testing. I also discuss the major advances in the application of these tools to T&E in several key focus areas, highlighting resources available to the T&E community. Finally, I conclude with a look to the future and how we can further improve our capabilities to take advantage of state-of-the-art methodologies.

Working with the operational and developmental test communities, I will continue to employ advanced statistical methods, and continue to improve our acumen in this area, as it can only benefit the Department and ultimately, our men and women in combat, in the end.

2013 DOT&E Guidance Memos

In my review of Test and Evaluation Master Plans (TEMPs) and in discussions within the test community, I have learned that misunderstandings persist of what DOT&E advocates regarding the use of DOE when designing operational tests. In 2013, I provided two additional guidance memos; key points in those memos are highlighted below.

1. Clear Test Goals

The most essential element of any test design is clearly defined test goals. Operational testing should seek to characterize a system's end-to-end mission effectiveness across the operational envelope. Such characterization of performance informs the system operators, as well as strategic and tactical planners, of its capabilities and limitations in the various conditions that will be encountered during combat operations. The goal of operational testing is not solely to verify that a threshold requirement has been met in a single or static set of conditions. Using DOE enables test programs (including integrated testing, where appropriate) to determine the effect of factors on a comprehensive set of operational mission- and capability-focused quantitative response variables. The determination of whether requirements have been met is also a test goal, but is a subset of this larger and much more important goal.

2. Mission-Oriented Metrics

OT&E metrics must provide a measure of mission accomplishment (not technical performance for a single subsystem), be continuous rather than discrete so as to support good test design, and address the reasons for procuring the system. Good measures in OT&E often reflect the timely and accurate accomplishment of a combat mission.

3. Consideration of all Operational Factors and Strategic Control of them in the Test Plan

The users often employ the system in conditions that are different from those identified for system development and specification compliance. Operational testing must enable the evaluation of a system across the conditions under which it will actually be employed. By selecting test factors (the variables that define the test conditions across the operational envelope) and forcing purposeful control of those factors, we can ensure that the operational test covers those conditions, which the system will encounter once fielded. The test factors must be varied in a purposeful way, yet not overly constrain the operational realism of the test. This balance must be obtained while ensuring that the test will generate adequate information for my evaluation. Uncontrolled "free play" is not a defensible test methodology. Operational testing should consist of deliberate control of the conditions while still allowing operators and simulated opposing forces to react as they would in a true operational scenario. Factors should be varied in a way enabling diagnosis of the root cause of changes in performance across the operational envelope. Eliminating factors or specific conditions is usually the first tactic in reducing test costs, but this is a false economy. Time and money are saved by examining the operating envelope as early as possible and mitigating risks through rigorous testing across all phases of the acquisition life cycle.

4. Avoidance of Single-Hypothesis Tests

Single-hypothesis statistical tests and their corresponding statistical power calculations are generally inappropriate for designing operational tests because they do not provide the ability to characterize performance across the operational envelope. Nor do they provide insights on the placement of test points within the operational envelope.

5. Statistical Assessment of Test Designs

Statistical confidence and power continue to be essential tools in my assessment of test designs. When used correctly in the context of the goal of the test (which is to say, provided the test variables and factors have been well selected to address mission needs, as discussed above), these quantitative measures provide great insight into the adequacy of the test design. In an experimental design, power not only describes the risk in concluding a factor is not important when it really is, but also directly relates to the precision we will achieve in making quantitative estimates of system performance. The latter is key in my determination of test adequacy; without a measure of the expected precision we expect to obtain in the analysis of test data, we have no way of determining if the test will accurately characterize system performance across the operational envelope. A test that has low power to detect factor effects might fail to detect true system flaws; if that occurs, we have failed in our duty as testers.

It is also essential that we consider additional criteria in the evaluation of the statistical design. Other criteria that are important to consider are the prediction variance across the operational envelope and correlation between factors. I provided these criteria and others in a recent memorandum to the T&E community, the use of which will enable all of the Services to prepare good test designs.¹

Current Focus Areas

In an effort to institutionalize the use of scientific/statistical approaches to T&E, DOT&E has focused on several key areas including: developing the workforce, updating policy and guidance, developing case study best practices, and advancing state-of-the-art methodologies to address challenges unique to the T&E community. In June 2013, my Science Advisor published the DOT&E Test Science Roadmap Report, which captures the progress in each of these areas.

Workforce Development

The Test Science Roadmap Report indicates clearly that all of the OTAs could benefit by increasing the number of civilian employees with scientific, technology, engineering, and mathematics (STEM) backgrounds in their workforce. Additionally, the Commanders of each OTA would benefit from having a senior technical advisor who is well versed in the science of experimental design and data analysis and is responsible for ensuring technical rigor across the entire Command.

Education and training are essential in the development of our T&E workforce. At DOT&E, I ensure that my staff receives regular training on important topics such as experimental design, reliability growth and analysis, and survey design. I welcome members of the broader test community in these training opportunities, especially the OTAs. Additionally, there are many excellent training and education programs available to the T&E community (details can be found in the Roadmap Report).

Policy and Guidance Updates

Policy and guidance updates that are currently underway will support the institutionalization of a scientific approach to T&E. These updates include the Defense acquisition policy, the DoD Instruction 5000.02, and the Defense Acquisition Guidebook.

In addition to these broader policy documents, DOT&E has published a TEMP Guidebook, which provides an up-to-date resource for the T&E community. I continue to update the guidebook as new best practices and lessons learned are captured. The guidebook highlights the substantive content DOT&E is looking for in TEMPs. The TEMP Guidebook is available on the DOT&E public website (<http://www.dote.osd.mil/temp-guidebook>) and provides guidance on many test science topics, including:

- Design of Experiments
- Mission-oriented metrics
- Reliability growth
- Modeling and Simulation
- Information Assurance
- Software-intensive systems

¹ Memorandum dated July 23, 2013, "Best Practices for Assessing the Statistical Adequacy of Experimental Designs Used in Operational Test and Evaluation."

INTRODUCTION

Case Studies, Best Practices, and Lessons Learned

In recent years, DOT&E, the Service OTAs, as well as the broader T&E community have captured many case studies that highlight best practices and lessons learned. These case studies are available in the Test Science Roadmap Report. Additionally, many of the best practices are captured in my most recent guidance memos. Best practices I advocate include:

- Provide clear justification for all designs – every design requires the quantification of acceptable risks and a determination of what changes in performance (effect size) need to be captured by the test design. These elements need to be clearly described and justified by the operational context.
- Use existing system and developmental test data. Operational test designs have the greatest chance of succeeding if they leverage all existing data on the system and its intended employment.
- Use continuous metrics where possible, since they provide the maximum information from a given test size; furthermore, they enable at least a 50 percent (and likely greater) reduction in test size over comparable pass/fail metrics for similar test goals.
- Ensure that power calculations are consistent with test goals and avoid single hypothesis tests. Additionally, use power curves to show trade-offs in resources and risk.
- Include all relevant factors (cast as continuous where possible) in design; mitigate risks by leveraging data and information from developmental testing.
- Do not limit test goals to verifying requirements under limited set of conditions; focus on characterizing performance across the operational space.
- Use statistical measures of merit to evaluate the trade-space in the test design.

Test Science Research Consortium

In conjunction with the Department's Test Resource Management Center, DOT&E continues to fund a multi-year research consortium to address the unique needs of the T&E community. This consortium funds several graduate-level research projects on advanced statistical techniques. By doing so, it not only enables these projects to be focused on topics of benefit to the Department's T&E needs, but also creates a pool of professionals with strong technical skills who can contribute to solving the many problems the Department confronts in improving its ability to acquire and field complex weapons systems.

Scientific Test and Analysis Techniques Test and Evaluation Center of Excellence (STAT T&E COE)

The STAT T&E COE, stood up by DASD(DT&E), provides direct T&E support to the program offices of Major Defense Acquisition Programs (MDAPs). The STAT experts are assigned to the program's T&E leads and work directly with the larger teams to assist by injecting more statistical rigor into defensible test planning, design, execution, and assessment processes. In 2013, the COE supported a total of 25 major programs, as well as various ad hoc requests. STAT experts have created and delivered multiple two-day STAT courses for various program test teams. These courses educate and inform testers and program office personnel on the value and implementation of a rigorous test methodology.

Looking to the Future

While significant progress has been made in recent years, there is still work to be done in ensuring that the scientific community's full toolset is available to support T&E. All programs need to employ best practices identified over the past several years. In addition to implementing these best practices, I have noted further areas for improvement that I will emphasize in the upcoming year. These specific areas for improvement include:

- Conducting data analysis commensurate with DOE design. Although most in the T&E community are now using statistical rigor to develop test designs, they are not always following up with the same rigor in the analysis of the data. The worst case of this occurs when a test is designed to cover the important operational conditions efficiently through DOE techniques, yet the data analysis is limited to reporting a single average (mean) across the test conditions. A more comprehensive statistical analysis is needed to fully realize the efficiencies and increased information provided by a rigorous experimental design. We must employ standard statistical tools, such as regression analysis techniques, that utilize all of the factors that affect system performance (meaning the "recordable variables" that were not controlled in the test design, as well as the factors that were). Additionally, we must improve our capabilities to verify these empirical statistical models to ensure they accurately reflect the data.
- Employing advanced methods. Many tests are complicated by data that require more than the "standard" or "simple" analysis methods. In these cases, we should embrace the opportunity to employ advanced methods. I plan to continue efforts to employ these advanced statistical tools where appropriate, and will continue to encourage the use of and train the community on these methods. Some examples include--
 - Bayesian approaches (especially in a reliability context) allow us to leverage information from multiple phases of test while ensuring the results still reflect the operational reliability.

INTRODUCTION

- Censored data analysis allows us to incorporate information from continuous measures in cases where traditional pass/fail metrics would have been the only option.
- Generalized linear models and mixed models allow flexible analysis methodologies that truly reflect the character of the data.
- Improving the use of surveys in OT&E. Surveys provide essential information for the evaluation of systems. However, I have observed that their use in OT&E often does not reflect best practices of the survey community. The result is data that have limited utility in my evaluations. In the upcoming year, I will provide guidance on the appropriate use of surveys in OT&E.

RELIABILITY ANALYSIS, PLANNING, TRACKING, AND REPORTING

I, and other Department leaders, have placed emphasis on improving the reliability of DoD systems via several reliability improvement initiatives, and I continue to emphasize the importance of reliability in my assessments of operational suitability.² Test results from the last few decades indicate that the DoD has not yet realized significant statistical improvements in the reliability of many systems. However, there is evidence that those systems that implemented a comprehensive reliability growth program are more likely to meet their development goals.

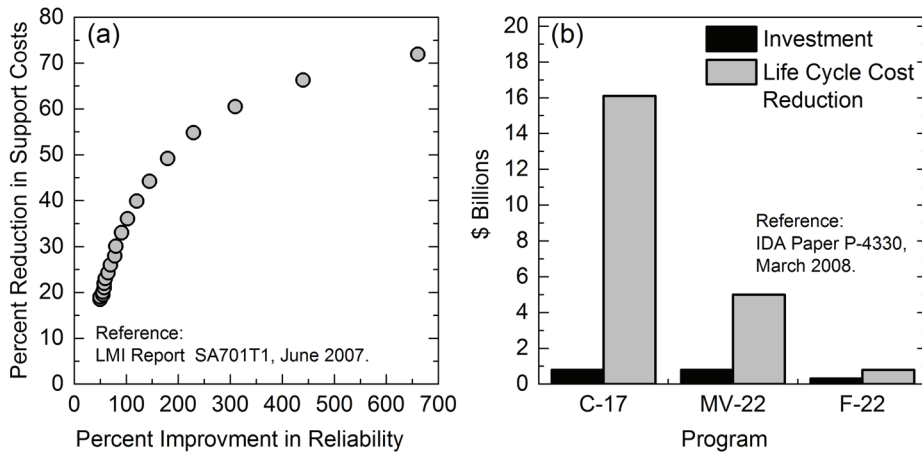


FIGURE 1:

(A) REDUCTION IN SUPPORT COSTS DUE TO RELIABILITY IMPROVEMENTS FOR A VEHICLE ELECTRONICS SYSTEM
 (B) LIFE CYCLE COST REDUCTION DUE TO INVESTMENT IN RELIABILITY FOR SELECT PROGRAMS

While always important, it is especially important in the current fiscal climate that system reliability is emphasized early in the acquisition process. Reliable systems cost less overall (because they require less maintenance and fewer spare parts), are more likely to be available when called upon, and enable a longer system lifespan. Reliability is more effectively and efficiently designed-in early (design for reliability) vice being tested-in late. While more upfront effort is required to build reliable systems, the future savings potential is

too great to ignore. The Department has recognized these potential cost savings. Figures 1a and 1b are examples from two studies that illustrate how investments in reliability lead to reduced life cycle costs. Programs that invest in reliability improvements early in their life cycle, such as the C-17 in Figure 1b, are expected to get the greatest return on investment and concomitant reduction in life cycle costs.

Evidence of Continuing Reliability Problems

Despite the implementation of the previously cited policies intended to encourage development of more reliable systems, the fraction of DoD systems assessed as reliable during operational testing has not improved. From FY97 to FY13, 56 percent (75 of 135) of the systems that conducted an operational test met or exceeded their reliability threshold requirements as compared to nearly 64 percent between FY85 and FY96. Figure 2 shows performance by Service.

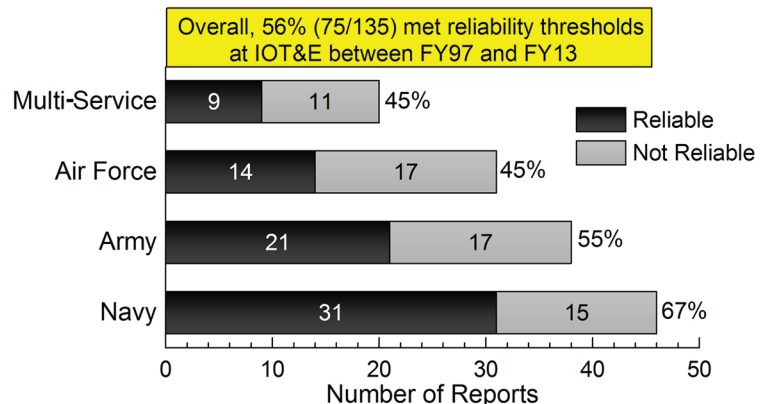


FIGURE 2: FRACTION OF DOT&E OVERSIGHT PROGRAMS MEETING RELIABILITY THRESHOLDS AT IOT&E BY SERVICE (FY97-FY13)

² e.g., Reliability Growth section of the DOT&E TEMP Guidebook version 2.1; USD(AT&L) policies including July 21, 2008, "Reliability, Availability, and Maintainability Policy" and March 21, 2011, Directive-Type Memorandum (DTM) 11-003 – "Reliability Analysis, Planning, Tracking, and Reporting."

INTRODUCTION

Figure 3 shows the relationship between previous and current policies for reliability as ratios of achieved reliability to threshold requirements between FY85 and FY13. The yellow highlight with two-year lag is the period of the prescriptive policy described in MIL-STD-785B; the green highlight also with two-year lag is the period of non-prescriptive, commercial best-practices; and the red is the current policy with emphasis on design for reliability and reliability test planning and growth. All data points greater than or equal to 1 indicate the system demonstrated reliability at or above its threshold requirement. Data points below 1 indicate the system failed to demonstrate its reliability threshold in operational testing. A linear fit to the data suggests there has

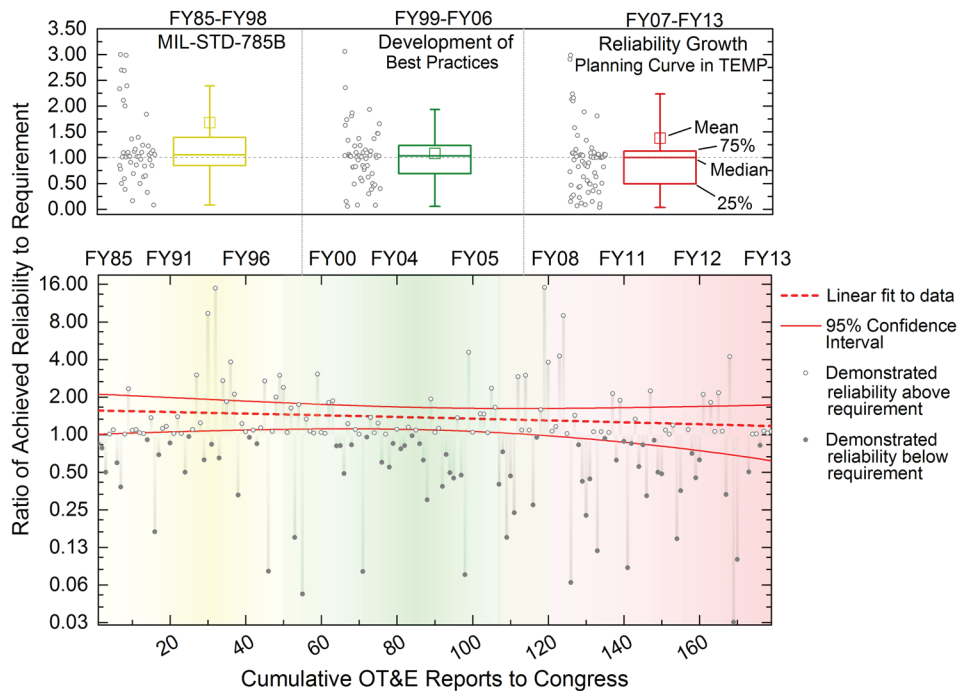


FIGURE 3: RELIABILITY TRENDS VERSUS POLICY PERIODS FOR YEARS FY85 TO FY13

been no improvement in program reliability over time. The boxplots in Figure 3 show that the three groupings have similar median values, but a larger fraction of data in the first grouping (FY85 to FY98) is concentrated at somewhat higher values compared to the latter two groupings. Although the plots suggest a decreasing trend in reliability, the trend is not statistically significant. Nonetheless, the data are conclusive that the reliability of DoD systems has not significantly improved over time.

The Department has acknowledged this poor track record of meeting system reliability requirements in March 2011 when USD(AT&L) issued a Directive Type Memorandum (DTM 11-003) on “Reliability, Analysis, Planning, Tracking, and Reporting.” The DTM 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. To instantiate reliability reporting in support of Defense Acquisition Executive Summary (DAES) reviews, DOT&E has worked with DoD Systems Engineering to implement a systematic process of tracking MDAP reliability status. MDAPs in system-level developmental testing with a documented reliability growth curve in the Systems Engineering Plan and TEMP will be required to report reliability data on a quarterly basis. The data will be used to inform the DAES selection process, review MDAP reliability performance-to-plan, and support reliability growth planning for future programs. At the direction of Acquisition Resource and Analysis, MDAPs that meet the criteria for reporting will submit their reliability data starting in FY14.

Evidence of Some Success

To better understand these trends, I have conducted a survey of programs under DOT&E oversight in each of the past five years to determine the extent to which reliability-focused policy guidance is being implemented and to assess whether it is leading to improved reliability. The most recent survey focused on 90 programs that submitted either a Test and Evaluation Strategy (TES) or TEMP to DOT&E, and/or had an operational test in FY12.

The survey results indicate, not surprisingly, that systems with a comprehensive reliability growth program are more likely to reach reliability goals compared to those that do not employ a growth program. In particular, the results show the importance of establishing and meeting operational test Reliability, Availability, and Maintainability (RAM) entrance criteria before proceeding to operational test. While many programs did not establish or meet operational test RAM entrance criteria, those that did were far more likely to demonstrate reliability at or above the required value during operational test. There is also evidence that having intermediate goals linked to the reliability growth curve improves the chance of meeting RAM entrance criteria.

INTRODUCTION

The survey results indicate that programs are increasingly incorporating reliability-focused policy guidance. In FY12:

- 92 percent of programs had a reliability growth strategy, with 90 percent documenting it in the TEMP.
- Likewise, 78 percent of programs incorporated reliability growth curves into the TEMP.
- 59 percent of programs used a reliability growth curve to develop intermediate goals.
- 87 percent of programs used reliability metrics to ensure that growth was on track to achieve requirements.
- 49 percent of programs had a process for calculating growth potential.

Despite these policy implementation improvements, many programs still fail to reach reliability goals. In other words, the policy has not yet proven effective at changing the trends displayed in Figure 3. The reasons programs fail to reach reliability goals include inadequate requirements, unrealistic assumptions, lack of a design for reliability effort prior to Milestone B, and failure to employ a comprehensive reliability growth process. For example, the reliability thresholds for some programs were unachievably high or disconnected from what was really needed for the mission. Other unrealistic assumptions include choosing an initial reliability value for their reliability growth curve that was significantly higher than comparable systems have been able to achieve, or choosing an optimistic initial value for the growth curve without an adequate design-for-reliability effort (which should occur prior to the growth program) to achieve that initial value. In some cases, a program's reliability growth goal, while documented in a TEMP or Systems Engineering Plan, was not supported by contractual obligations or funding. As a result, a larger fraction of surveyed programs met their reliability thresholds after fielding during FOT&E (72 percent) rather than before fielding during IOT&E (50 percent). I conclude from this study that although we are in a period of new policy that emphasizes good reliability growth principles, without a consistent implementation of those principles, the reliability trend will remain flat.

Recommendations for the Future

In the future, programs need to do a better job incorporating a robust design and reliability growth program from the beginning that includes the design for reliability tenets described in the ANSI/GEIA-STD-0009, "Reliability Program Standard for Systems Design, Development, and Manufacturing." Programs that follow this practice are more likely to be reliable.

There should be a greater emphasis on ensuring that reliability requirements are achievable, and reliability expectations during each phase of development are supported by realistic assumptions that are linked with systems engineering activities. I recommend that all programs establish operational test entrance criteria and ensure these criteria are met prior to proceeding to the next test phase. Examples of effective RAM entrance criteria include (1) demonstrating in the last developmental test event prior to the IOT&E a reliability point estimate that is consistent with the reliability growth curve, and (2) for automated information systems and software-intensive sensor and weapons systems, ensuring that there are no open Category 1 or 2 deficiency reports prior to operational test. I also reemphasize USD(AT&L) policy described in DTM 11-003, "Reliability Analysis, Planning, Tracking, and Reporting" that reliability growth curves/programs should be constructed with a series of intermediate goals, with time allowed in the program schedule for test-fix-test activities to support achieving those goals. System reliability should be tracked through system-level T&E events until the reliability threshold is achieved.

Second, when sufficient evidence exists to determine that a program's demonstrated reliability is significantly below the growth curve, I recommend that the program develop a path forward to address shortfalls and brief their corrective action plan to the acquisition executive. Such efforts might include a reexamination of the requirements and updates to the assumptions made in the growth curve, and may reveal the need for the program to perform lower level re-design work to get back on course. This will help encourage sound development processes, including the use of design-for-reliability efforts, and allow the growth curve to be a much more useful tool for decision makers.

Based on findings from surveys, reliability trend analysis, and other lessons learned, I continue to update and refine the reliability growth guidance section of DOT&E's TEMP Guidebook. The latest edition, updated July 12, 2013, provides specific reliability growth planning guidance for different types of systems, including hardware-only systems; hybrid systems containing a combination of software, hardware, and human interfaces; and software-intensive systems. The Guidebook also provides an overview of the key systems engineering and design activities that constitute a comprehensive reliability growth program and requires the TEMP to include a description of these activities for each of the three system types, with emphasis on the latter two. For hybrid systems (e.g., weapons systems composed of both hardware and software, such as radars), the TEMP requires plans for categorizing hardware failures versus software failures, for tracking software failures, and for regression testing software failure fixes. Software-intensive systems, starting in the design phase, should describe a plan to track software reliability to include defined entrance and exit criteria for system reliability at critical decision points. Finally, the latest Guidebook illustrates how to use operating characteristic curves to quantify allowable test risks (consumer's and producer's risk) and develop the reliability growth goal.

INTRODUCTION

TESTING OF SOFTWARE-INTENSIVE SYSTEMS

Over the last several decades, the Department's reliance on and procurement of software-intensive systems has significantly increased. These Major Automated Information Systems (MAIS) provide key capabilities to the Department, including financial and budgetary management functions, command and control, medical records management, and logistics and inventory management. Furthermore, nearly every military system is based upon software to provide functionality and capability. Because of the importance of the issue, and because many capability shortfalls are directly related to software failures and poor software maintenance capabilities, I have increased my involvement in testing these systems.

I note four areas are of interest in testing of software-intensive systems. First, I continue to observe that many MAIS programs do not create adequate software maintenance capabilities early enough to support deployment. Second, software requirements continue to be poorly stated. Third, as a new area of interest within the last several years, I am focusing on testing the financial vulnerabilities of systems that have direct accounting or logistics functions. Finally, as the Department begins to examine how its test processes can and should be adjusted to accommodate the Agile software development model, I provide three distinct models of how Agile concepts can be applied to operational testing.

Software Maintenance

Current Department acquisition practices categorize software maintenance as a sustainment activity – something that begins after software is deployed. This is problematic as it sets our programs up for failure. Disciplined software maintenance (by which I mean configuration control, defect tracking and prioritization, maintenance of a high fidelity test environment, and automated testing within that environment) must begin as soon as there is any software to maintain. Software that is deployed in the absence of a robust maintenance capability typically has poor operational results, and the reliability of such software can grow steadily worse with each new upgrade or patch to the software.

Illustrative examples of late development of software maintenance capabilities include the DoD Automated Biometric Identification System (ABIS), Defense Enterprise Accounting and Management System (DEAMS), and Navy Enterprise Resource Planning (Navy ERP).

- **DoD ABIS.** A key action item for the program manager from the stakeholder meeting following the fourth failed deployment attempt of ABIS 1.2 (see “Problem Discovery Affecting OT&E” in the Activity and Oversight section of this report) was to determine and document what functionality ABIS 1.0 provides to its users. How DoD ABIS can be developed and maintained without comprehensive knowledge of the capability it currently provides is a key question.
- **DEAMS.** DEAMS had 2 operational assessments in 2 years, each identifying 200+ defects. DEAMS appears to be improving after the program manager implemented improved configuration control and defect tracking, as well as rudimentary regression testing.
- **Navy ERP.** The Navy ERP system demonstrated significant reliability shortfalls due to software maintenance in early testing. After developing an improved software maintenance capability, the program is now operationally effective and operationally suitable. The program has a functioning software configuration control board and defect management process that is expeditiously correcting new deficiencies, particularly high-severity ones. The regression testing process is efficient, being almost entirely automated. Between the 2008 IOT&E of Release 1.0 and the 2013 FOT&E of Release 1.1 (which is to say, the five years following initial deployment), the Program Office instituted disciplined software management practices. It probably would not have taken so long to reach the full deployment decision if the software had been better managed early on. For example, during the Release 1.1 IOT&E in 2010, the discovery rate for new system defects was 125 per month with a backlog of nearly 500 defects remaining at the conclusion of testing. After the 2010 IOT&E, the Program Office improved the defect management process, which included reviewing outstanding defects more frequently and increasing the emphasis on maintaining accurate status on all defects. Navy ERP is now the Department's second successfully deployed ERP system.

To promote earlier attention to software maintenance, I have begun enforcing the following test automation policy, which was put into effect recently in the interim (November 26, 2013) Defense acquisition policy, the DoD Instruction 5000.02:

For software in any system, the evaluation of operational suitability will include a demonstrated capability to maintain the software. Program managers must sustain an operationally realistic maintenance test environment in which software patches can be developed and upgrades of all kinds (developed or commercial) can be tested.

- (1) IOT&E or a prior test event will include an end-to-end demonstration of regression test, preferably automated, in the maintenance test environment from requirements to test scripts to defect tracing.*
- (2) IOT&E or a prior test event will include a demonstration of processes used to update the maintenance test environment so as to replicate deficiencies first found in the operational environment.*

INTRODUCTION

I have also worked in the last year to help programs make the transition to the use of automation for regression testing. My staff has initiated a Test Automation Center of Excellence (TACE), which is now helping to automate the third of their target list of seven highly similar MAIS programs. In the last year, by working closely with the Defense Logistics Agency (DLA) sustainment staff and support contractors (for the Department's first successfully deployed ERP, the DLA's Enterprise Business System), the TACE has trained 38 DLA staff in the use of automation; 6 in the development of automation; and transitioned 12 validated scripts to operational use. These scripts (and associated setup) take 18 human-at-keyboard minutes on average to execute as compared to 142 minutes on average for the corresponding manual scripts. Five scripts were executed in November 2013 as part of normal operations, including two that were developed by the DLA staff. DLA has made substantial progress in one year (and I expect another year will be needed to make DLA fully self-sufficient) at a direct cost of \$500,000, as opposed to the \$11.5 Million over 5 years originally quoted to DLA by a leading market analysis group.

The Services have begun making efforts to include planning for software regression testing and automation. Seventeen of the 63 unclassified TEMPs, TESSs, or Operational Test Plans that I signed out between December 1, 2012, and December 1, 2013, included detailed discussion of software regression testing methods and/or test automation.

Finally, the importance of these software testing efforts is amplified by the push to deploy the Joint Information Environment (JIE). The JIE is envisioned to be a shared and upgraded information technology infrastructure that will, amongst other things, consolidate existing net-centric systems into a reduced number of data centers and operations centers using a common computing model for virtualization. This means that for each existing net-centric system there should at some point be two copies: the current system and the new virtualized, JIE version of the system. No existing system should be shut off until the JIE version is shown to perform at least as well, and that testing should be automated. That automated validation would then ideally be reused for subsequent regression testing.

Software Requirements

In most cases, it will be possible to develop software that automatically provides performance metrics. If operational testers cannot answer reasonable questions about software system performance from data that the system owners are already gathering, then the system owners also, clearly, do not fully understand how well their system is performing. This is operationally important for the same reason as software maintenance: the software will change over time. In order to maintain and improve system performance, parameters that are key to the capability should ideally be automatically measured and monitored by the Program Office vice being checked manually during operational tests. The bias and presumption in operational software testing should be toward independent review of automatically gathered performance metrics. Interactions between testers and users often provide helpful insights; however, human execution of highly repetitive, precise operations is an unnecessary expense and a missed opportunity. In the latter case, operational testing should verify that automated performance metrics exist and that the Program Office is organized to utilize those metrics in its ongoing software maintenance.

I would not want nor expect a Program Office to optimize software around a performance metric that was not relevant to mission accomplishment. Unfortunately, software KPPs and their associated measures are often uninformative with respect to mission accomplishment. The measures can be seen to carry a bias and presumption toward testing that consists of human review of compliance checklists. Human review is open-loop. Program Office use of automated metrics is closed-loop, which will be better. The F-22A program, Theater Medical Information Program – Joint (TMIP-J), and Air Operations Center – Weapon System (AOC-WS) programs provide examples of open-loop and closed-loop review processes.

- **F-22A.** The Net-Ready KPP in the F-22A TEMP (January 2013) is geared toward paperwork compliance instead of mission-relevant, automated performance measures. The KPP is: “Key Interface Profiles will be satisfied to the requirements of the specific joint integrated architecture products and information assurance accreditation.” This KPP is stated in terms of documents and accreditation, and was translated in the TEMP into various measures of compliance (for example, one measure requires all “policy enforcement controls designated as enterprise level or critical in the joint integrated architecture”). In the future, I will require that TEMPs and test plans evaluate this KPP using mission-oriented measures collected using monitoring of the operational network. In particular, the KPP should be evaluated using continuous observation of measures, including time to detect protocol deviations and error tolerance levels.
- **TMIP-J.** The TMIP-J Increment 2 TEMP (May 2013) has a Critical Operational Issue (COI) for Supportability which translates into nine different surveys and subject matter expert evaluations. The COI “Are TMIP-J Increment 2 features, training plans, characteristics, processes, procedures, and resources adequate to sustain its intended operations?” is clearly mission-critical; the TMIP-J operators certainly need to know if and when the system becomes inadequate. However, the COI would better lend itself to appropriate automation and use by the Program Office if it were phrased or interpreted as: “Does TMIP-J Increment 2 provide reporting on its features, training, characteristics, processes, procedures, and

INTRODUCTION

resources sufficient to determine that it is fulfilling its intended operations?” As in the previous example, the COI should be understood in terms of continuous monitoring rather than occasional compliance-checking via surveys.

- **AOC-WS.** Conversely, the AOC-WS TEMP (October 2013) has a good measure for its Data Accuracy capability: “Percent of missions flown linked to Air Operations Directive tactical tasks.” This measure indicates that all targets must be “linked” to their desired effects. The linkage requires the AOC-WS machine-assisted capability to maintain a connection to the planned operational assessment results throughout the development of all AOC products. The connection links actions to effects and traces effects to the expected data sources. This measure of accuracy can be achieved through automation, and it will help AOC commanders evolve tasking orders during engagements by ensuring that the software can always trace planned actions to desired effects and then trace observed effects back to their associated actions, which must then be repeated or updated in subsequent tasking orders. It is important to the mission that this metric be satisfied, and it can assist in software maintenance by automatically identifying mission areas where the linkage is not working properly.

With few exceptions, software KPPs should support ongoing software management by requiring automated measurement and reporting (for system managers) of help desk use, interface throughput, system productivity/utilization, training adequacy, reliability metrics, and other (less generic) mission critical performance parameters. Such reports would also answer most software OT&E questions. To promote improved requirements, I have begun enforcing the following polices, which were put into effect recently in the interim (November 26, 2013) Defense acquisition policy, the DoD Instruction 5000.02:

Beginning at Milestone A, every TEMP will include an annex containing the Component’s rationale for the requirements in the draft Capability Development Document (CDD) or equivalent requirements document.

Program managers for software acquisitions will provide plans at Milestone B indicating how system logs and system status records will interface with operational command and control. At IOT&E or a prior test event, program managers for software acquisitions will demonstrate performance monitoring of operational metrics to manage and operate each system capability (or the whole system, as appropriate).

Financial Vulnerabilities

I have 13 accounting or logistics systems on oversight, and all will be required to undergo operational testing geared to their unique vulnerabilities.³ These systems are typically being acquired so as to achieve full auditability by 2017 in accordance with the National Defense Authorization Act (NDAA) for FY10. They will homogenize the sometimes obscure or conflicting legacy accounting practices within the Department, but in the process they may also expose the Department to new or expanded vulnerabilities to theft, fraud, or nation state manipulation. Losses due to fraud in the commercial sector are estimated at 5 percent of revenues each year.⁴ Common fraud controls – such as those required by the Government Accountability Office Federal Information System Controls Audit Manual – should result in significant reductions in both the amount lost and the undetected time span of fraudulent activities. The Defense Intelligence Agency has not yet evaluated the potential threat to U.S. supply lines and/or U.S. markets through manipulation of the Department’s accounting and logistics systems, and there is currently no guidance for mitigating these risks.

This year, the Navy ERP program conducted the first fraud vulnerability test. The test identified 1,799 user accounts that had multiple segregated roles (and who could therefore potentially commit fraud without assistance). The Navy ERP Program Office was not aware if any of those user accounts had in fact been used fraudulently. Accordingly, subsequent financial vulnerability scans and assessments will include forensic accounting activities so as to provide immediate information on the extent to which identified vulnerabilities have been exploited. The Navy ERP test was also similar to a “Blue Team” Information Assurance vulnerability scan (as opposed to a “Red Team” penetration test). The second fraud vulnerability test (for DEAMS) will complete in early 2014. DEAMS data from the last year have been provided to forensic accountants for analysis. A certified and accredited Red Team paired with trained accountants will conduct the penetration test. If the Red Team is able to penetrate the system cyber defenses, then the accountants will assess the potential operational effects that they will be able to cause. These assessments will occur in four threat scenarios that include insider threat and nation state threat scenarios.

³ Air Force Integrated Personnel and Pay System (AF-IPPS); Defense Agency Initiative (DAI); Defense Enterprise Accounting and Management System (DEAMS); Expeditionary Combat Support System (ECSS); EProcurement; Future Pay and Personnel Management Solution (FPPS); General Fund Enterprise Business System (GFEB); Global Combat Support System – Joint (GCSS-J); Global Combat Support System – Marine Corps (GCSS-MC); Global Combat Support System – Army (GCSS-Army); Integrated Personnel and Pay System – Army (IPPS-A); Logistics Modernization Program (LMP); Navy Enterprise Resource Planning (ERP)

⁴ According to the 2012 Report to the Nations of the Association of Certified Fraud Examiners, <http://www.acfe.com/rtrn-highlights.aspx>

INTRODUCTION

Finally, I have directed my staff to develop a new Enterprise Cyber Range Environment (ECRE) to mimic the software stack supporting U.S. Transportation Command. This ECRE will enable observation of the undetected duration and magnitude of the operational effects of nation state cyber attacks that might be launched to disrupt U.S. supply lines.

Agile Operational Testing of Software

This year, I have approved three operational assessments that provide three distinct models of Agile operational testing.

- For the Integrated Electronic Health Record (iEHR) program, I established that the responsible OTA, the Army Test and Evaluation Command (ATEC), would observe all tests (including developmental testing) and send me a report or synopsis. An ATEC tester is now embedded with the iEHR program.
- For DEAMS, I approved a two-stage test. The first stage took less than one month from execution to reporting. In the first test phase, my staff interviewed DEAMS managers and users following deployment of the new (Release 2) software to existing users. The interviews were sufficient to determine that the DEAMS software management had improved, that deploying Release 2 did not disrupt operations, and that I could support the decision to deploy Release 2 to new users. The second test phase will provide me with data to evaluate the Release 2 capabilities. In this model of Agile OT&E, a rapid check on the gross features of an initial software deployment to existing users is followed by a risk-appropriate level of test of the system within a new group of users and the existing users.
- For the Integrated Personnel and Pay System – Army (IPPS-A) Increment 1, I have approved an operational test concept that will largely utilize data gathered organically by IPPS-A. The program manager and ATEC were able to implement an inexpensive email dialogue and survey process. This process will continuously track for all IPPS-A users whether their Soldier Record Brief (SRB) data are correct and, if not, what data are incorrect, and, later, whether the user has been able to successfully use the instructions for correcting their data. The survey process will also assess the usability of the IPPS-A system. Once the data have been corrected in the legacy systems (which remain the authoritative data sources in Increment 1), the final automated user survey will ask the user to review their SRB and verify whether the corrections are now displayed in their SRB. As discussed in the Software Requirements section above, this process will provide IPPS-A system owners with valuable ongoing self-monitoring information relevant to the system’s customer service needs, and it also predominantly meets operational test needs for Increment 1.

With these working models of Agile operational testing in hand, I expect to be able to craft appropriate test approaches for subsequent Agile acquisitions.

OTHER AREAS OF INTEREST

Electronic Warfare Test Infrastructure

In February 2012, I identified significant shortfalls in the test resources required to test mission systems electronic warfare capabilities under operationally realistic conditions. The Department programmed for an Electronic Warfare Infrastructure Improvement Program starting in FY13 to add both closed-loop and open-loop emitter resources for testing on the open-air ranges, to make at least one government anechoic chamber capable of providing a representative threat environment for electronic warfare testing, and to upgrade the electronic warfare programming laboratory that will produce threat data files. These test capabilities are essential to many programs, including F-35 Joint Strike Fighter (JSF), F-22 Increment 3.2 A/B, B-2 Defensive Management System, Long-Range Strike Bomber, Next Generation Jammer for the EA-18G, Integrated Defensive Electronic Countermeasures upgrades, as well as several other programs. However, progress in selecting sources and beginning development of the test resources has been slower than needed to assure these resources are available in time for the JSF Block 3 IOT&E in 2018. Without these resources, the JSF IOT&E of Block 3 capability will not be adequate to determine the system’s effectiveness in existing operationally realistic threat environments.

Aegis-Capable Self-Defense Test Ship (SDTS)

As mentioned above, the test community currently relies on an unmanned, remotely controlled ship, called the SDTS, with the actual radars, weapons, and combat systems employed on some (not all) of the Navy’s currently deployed ships to examine the ability of these systems to protect against incoming anti-ship cruise missiles. Navy range safety restrictions prohibit close-in testing on manned ships because the targets and debris from successful intercepts will pose an unacceptable risk to the ship and personnel at the ranges where these self-defense engagements take place. The importance of this testing and the need for such a test resource is underscored by the recent incident in November 2013, where two Sailors were injured when an aerial target struck the USS *Chancellorsville* (CG-62) during what was considered to be a low-risk test of its combat

INTRODUCTION

system. The Navy employs a high-fidelity modeling and simulation capability that relies heavily on data collected from testing with the SDTS, as well as data from manned ship testing, so that a full assessment of ship self-defense capabilities of non-Aegis ships can be completely and affordably conducted. While the Navy recognizes the capability as integral to the test programs for certain weapons systems (the Ship Self-Defense System, Rolling Airframe Missile Block 2, and the Evolved Sea-Sparrow Missile Block 1) and ship classes (LPD-17, LHA-6, Littoral Combat Ship, DDG 100, and CVN-78), the Navy has not made a similar investment in an Aegis-capable SDTS for adequate operational testing of the DDG 51 Flight III Destroyer (with Aegis Advanced Capability Build “Next” Combat System and Air and Missile Defense Radar (AMDR)) capabilities. The current SDTS lacks the appropriate sensors and other combat system elements to test these capabilities.

I continue to strongly advocate for the development of an Aegis-capable SDTS to test ship self-defense systems’ performance in the final seconds of the close-in battle and to acquire sufficient data to accredit ship self-defense modeling and simulation test beds. Other methods that are being examined and desired in lieu of an STDS, in my estimation, are wholly inadequate to fully examine the complex, close-in battlespace where multiple components of the combat system must work simultaneously to orchestrate shooting down multiple incoming highly-capable anti-ship cruise missiles, all within an engagement timeline of tens of seconds. The estimated cost for development and acquisition of an SDTS capability over the Future Years Defense Program is approximately \$284 Million. Of that, \$228 Million would be recouped after the test program completes by installing the hardware in a future DDG 51 Flight III hull. I have disapproved the Milestone B AMDR TEMP because, contrary to its predecessor AMDR TES, the TEMP did not provide for the resources needed to equip an SDTS. Similarly, I will disapprove the DDG 51 Flight III TEMP if it omits the resources needed to equip an SDTS.

Cybersecurity Testing

DOT&E continues to focus cybersecurity testing for all systems subject to information systems certifications and exposure to information networks. A review of the existing cybersecurity T&E procedures is underway in anticipation of the coming updates to the processes by which the Department certifies and accredits systems to operate on DoD networks (a shift from the DoD Information Assurance Certification and Accreditation Process to the National Institute of Standards and Technology “Risk Management Framework” in use by other federal agencies). A review of testing over the past several years continues to indicate the need to discover and resolve information system security vulnerabilities as early as possible in program development. The majority of system vulnerabilities discovered in operational testing over the last two years could and probably should have been identified and resolved prior to these tests. These challenges are also discussed in the Information Assurance and Interoperability Assessment section of this report.

Testing of Personal Protective Equipment

I continue to exercise oversight over personal protective equipment. The Services and the U.S. Special Operations Command (USSOCOM) continue to implement rigorous, statistically-principled testing protocols approved by DOT&E for hard body armor inserts and military combat helmets. In partnership with the Services and USSOCOM, I am developing a soft armor vest testing protocol that will standardize testing of soft armor vests and require them to meet rigorous statistical measures of performance. In its final report, the National Academy of Sciences’ Committee to Review the Testing of Body Armor supported the use of statistically-based protocols that allow decision makers to explicitly address the necessary and unavoidable risk trade-offs that must be faced in body armor testing.

As a result of Congressional interest, the Department’s Inspector General completed a Technical Assessment of the Advanced Combat Helmet (ACH) in May 2013. The assessment found that the DOT&E test protocol for the ACH adopts a statistically principled approach and represents an improvement from the legacy test protocol with regard to increased sample size. In response to a recommendation in this assessment, I will conduct characterization testing of new combat helmet designs that are being procured: specifically, the lightweight ACH, the Enhanced Combat Helmet, and the Soldier Protective System Integrated Head Protection System. Based on these data, I will determine whether the relevant test protocols should be updated to be more consistent with the products’ demonstrated performance. Additionally, we developed a specific statistical procedure that provides increased confidence that combat helmets meet minimum performance standards for all helmet sizes and test environments. I asked the National Research Council to conduct an independent review of the helmet testing protocols. Their report is anticipated to be released in FY14 and I will act on its findings.

As noted by the National Research Council of the National Academy of Sciences in their final report on the Testing of Body Armor and in my report to Congress on the Live Fire Test and Evaluation of the Enhanced Combat Helmet, medically validated injury criteria for behind-armor and behind-helmet blunt trauma do not exist. This is a serious limitation for the T&E of all personal protective equipment. Body armor and helmets made from modern materials deform rapidly during a bullet or fragment impact. The blunt force of the impact to the torso or of the impact of the deforming helmet shell on the

INTRODUCTION

head might cause injury or death even if the threat does not penetrate. The current acceptance criteria for helmets are based on the ability to withstand penetration and on acceptable levels of deformation in the event a bullet impacts but does not penetrate. The requirements for the latter were not established using medical data nor were they informed by how much deformation would be acceptable to prevent serious injury from bullet impact. Therefore, using Joint Live Fire funds, I have funded an effort to establish injury risk criteria for one type of injury due to behind-helmet blunt trauma.

My office is also monitoring a multi-year Army program to investigate behind-helmet blunt trauma, determine injury mechanisms and risks, and develop an injury criterion that can be used for helmet testing. The results of such testing have the potential of changing the way we evaluate helmets, and the protocols for testing these helmets may need to change. My office is also overseeing and participating in an Army effort to improve helmet test mount headforms by developing multiple-sized headforms to replace the single-sized headform currently used to test all helmet sizes (a recognized limitation to the current test method). Finally, I have provided funding to help characterize new potential ballistic clay formulations for use in the testing of personal protective equipment. The Army is pursuing a ballistic clay formulation with a more consistent dynamic response; these efforts have the potential to reduce the variability in the clay's response to an impact, thereby providing a better measure of the true performance of the tested equipment. I continue to work with the Services and USSOCOM to incorporate improved test procedures as they are developed and to update personal protective equipment test standards based on the results of these studies.

Warrior Injury Assessment Manikin (WIAMan)

In 2010, I brought to the Department's attention the lack of validated medical criteria and adequate instrumentation by which to assess occupant injuries in underbody blast Live Fire tests conducted against ground combat and tactical wheeled vehicles. This is a serious limitation to the T&E of all ground combat and tactical wheeled vehicles. In 2011, the Deputy Secretary of Defense directed the Army, with OSD oversight, to execute a project to conduct medical research to develop underbody blast-specific injury criteria, as well as an anthropomorphic test device (ATD) designed specifically for the underbody blast environment.

The WIAMan project made significant progress in 2013 after I directed a major restructuring to address delays in medical research planning and execution. The WIAMan Project Office now resides at the U.S. Army Research Laboratory, and under this new management has begun to execute medical research, as well as ATD development. The university research performers on the WIAMan project are some of the premier injury biomechanics researchers in the country and provide the project with the requisite experience and laboratory capabilities. The first phase of medical research is well underway, and the results from that research, as well as from anthropometric studies, are informing the concept for the initial ATD prototype. The project has also provided insights into the shortcomings of the current ATDs used in Live Fire Test and Evaluation. By using a unique, purpose-built test device that is able to expose ATDs and other test subjects to a controlled, blast-driven, vertical accelerative load environment, the research revealed the lack of biofidelity of the currently-used ATD when compared to the human response. These results further reinforce the need to continue this important work. To this end, I have provided Joint Live Fire funds to support the Army's efforts on this project and will continue to work with the Army to update underbody blast test standards and procedures to incorporate the results of this project.


Fifth-Generation Aerial Target

With the advent of fifth-generation aerial threats, to include low observability, low probability of intercept sensors, and embedded electronic attack, the feasibility of completing operationally realistic testing will decline significantly without developing adequate test capabilities that will assure U.S. air superiority in future conflicts. Over the past seven years, my staff has developed an alternative, low-cost fifth-generation aircraft design that will enable end-to-end testing to evaluate U.S. weapons systems effectiveness, from post-launch acquisition to end-game fusing, against fifth-generation fighter threats in Anti-Access/Area Denial missions. The Department, in partnership with the Canadian government, is considering funding a three-year, \$80 Million critical design, prototyping, and flight test effort that could provide an essential developmental and operational T&E capability.

INTRODUCTION

CONCLUSION

Since my first report to you in 2009, we have made progress increasing the scientific and statistical rigor of operational test and evaluation; there is much work to be done, however, to improve and consistently apply these techniques. Additionally, we have focused attention on reliability management, design and growth testing, and the improvement in testing software-intensive systems. Operational testing continues to be essential to characterize system effectiveness in combat so that well-informed acquisition and development decisions can be made, and men and women in combat understand what their equipment and weapons systems can and cannot do. 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 2013.


J. Michael Gilmore
Director

INTRODUCTION