FY 2014 Annual Report

The purpose of operational testing is to assure the Military Services field weapons that work in combat. This purpose has been codified in both USC Title 10 and in the Department of Defense's (DOD) 5000-series regulations for many years without substantive alteration. Operational testing is intended to occur under "realistic combat conditions" that include operational scenarios typical of a system's employment in combat, realistic threat forces, and employment of the systems under test by typical users (Soldiers) rather than by hand-picked or contractor crews.

Thorough operational testing should be conducted prior to a system's Full-Rate Production decision or deployment to combat in order to inform acquisition decision makers and operators in an objective way about how the system will perform in its combat missions. Under current law, the Director of Operational Test and Evaluation (DOT&E) is required to present his opinion on whether the operational testing conducted prior to the Beyond Low-Rate Initial Production decision is adequate or not. The Director must consider all the operational facets of a system's employment in combat when he determines what constitutes adequate operational testing, including the performance envelope the system must be able to achieve, the various operating conditions anticipated in a time of war, and the range of realistic operational threats.

In 2014, I investigated many examples of recent programs across all Services to identify common themes in operational testing. These themes illustrate the value that operational testing provides to the Defense community. Additionally, they highlight the continuing improvements we have made in the credibility and efficiency of OT&E during my tenure. A briefing covering these six themes and dozens of examples across all Services is posted on the DOT&E website.[1] These themes reveal a common conclusion: OT&E provides value to the Department by identifying key problems and clearly informing warfighters and the acquisition community about the capabilities our combat systems do and do not have. Furthermore, we are getting this information now more efficiently and cost effectively than ever by employing rigorous scientific methods in test planning, execution, and evaluation.

Identifying Problems

One of the primary purposes of operational testing is to identify critical problems that can be seen only when systems are examined under the stresses of realistic operational conditions, prior to the Full-Rate Production decision. This early identification permits corrective action to be taken before large quantities of a system are procured and avoids expensive retrofit of system modifications. For a recent example, operational testing of the Navy's Cooperative Engagement Capability (CEC) on the E-2D Hawkeye aircraft revealed several deficiencies. The CEC created many more dual tracks compared to the baseline CEC system, exhibited interoperability problems with the E-2D mission computer, and there was a degradation in CEC's ability to maintain consistent air tracks compared to the baseline E-2C version. As a result of these discoveries in operational testing, the Navy's acquisition executive decided to delay the Full-Rate Production decision until the root causes for these deficiencies could be found and fixed. The Navy is now implementing fixes to address these problems, and operational testing will be conducted to verify these fixes have corrected the problems. The value of such testing is abundantly clear if one considers the alternative: discovering these problems for the first time in combat, when it is too late to correct them.

Fixing, Not Testing, Delays Programs

Operational testing frequently reveals deficiencies in a system that require time and perhaps also training to correct. The acquisition executives who are responsible for programmatic decisions then have to weigh whether the problems discovered are of sufficient magnitude to warrant delays to the program while they are fixed (and re-tested). The assertion that testing causes programmatic delays misses the essential point: fixing the deficiencies causes delays, not the testing. Furthermore, taking the time to correct serious problems is exactly what we desire in a properly-functioning acquisition system; testing is the vehicle by which decision makers can be informed and make decisions that will ultimately benefit the Services and the Nation.

This year, my office updated a previous study that we conducted with USD(AT&L) in 2011 on the causes of program delays. This year's analysis examined case studies for 115 acquisition programs, which were selected because they had experienced

1. http://www.dote.osd.mil/pub/presentations/Value_of_OT_Final_Version_8.pdf

a delay of 6 months or more and had a Full-Rate Production decision after 2000. Delays on these programs ranged from 6 months up to 15 years, and in some cases, programs were cancelled after the delays (Figure 1 shows the distribution of these delays for these 115 programs). The reasons behind the delays are varied. In most cases, the delay is not due to a single reason; rather multiple reasons led to a delay (see Figure 2 and Table 1).

The study revealed that the least common reason for a delay was a problem associated with test conduct. As shown in Table 1, problems in test conduct occur in only 23 percent of the case studies, 26 of 115 cases. Furthermore, all programs that had problems in test conduct also had at least one other reason that contributed to the delay; test conduct, therefore, was never the sole reason for delaying a program. On the other hand, the most common reason that contributes to a delay is a performance problem discovered during developmental or operational testing that must be addressed before a program moves forward. A total of 87 of 115 cases examined (76 percent) discovered system performance problems during testing; 38 cases discovered problems in developmental testing only; 17 cases discovered problems in operational testing only; and 32 cases discovered problems in both developmental and operational testing.

Furthermore, when examining the length of the delay, no statistical evidence exists to support the claim that test conduct problems drive longer delays. Rather, the statistics support the assertion that performance problems discovered in testing significantly affect the length of the delay, not problems in conducting the test. For programs that discovered problems in operational testing, the length of the delay was more significant than for programs that discovered problems in developmental testing. This is not a surprising result, since problems seen for the first time in operational tests are frequently discovered late in the program's development, when designs are set and it is more difficult and time consuming to change them to correct problems.[2] Moreover, the statistical analysis revealed the largest drivers of delays are whether the program experienced manufacturing, software development, or integration problems and programmatic issues. A briefing with more details on this analysis is available on the DOT&E website.[3]

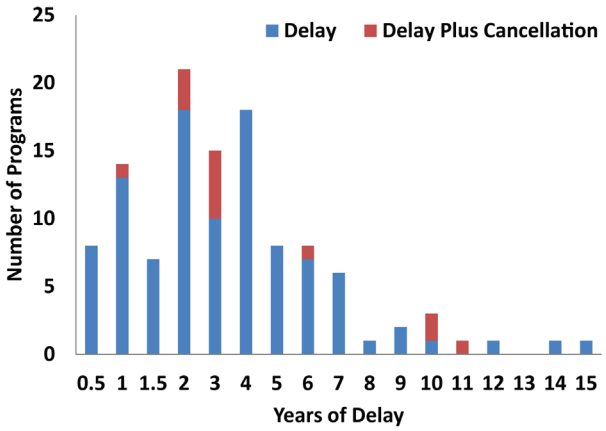


FIGURE 1. YEARS OF DELAY

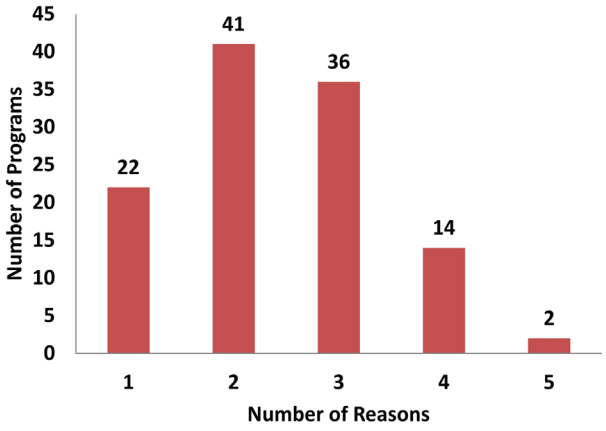


FIGURE 2. NUMBER OF REASONS FOR THE DELAY

TABLE 1. NUMBER OF PROGRAMS AFFECTED BY EACH REASON

Reason that Contributes to the Delay	Number of Programs Affected[1]
Problems conducting the test: problems with test resources, test instrumentation, or test execution that are typically beyond the control of the program manager	26
System problems identified during testing that must be addressed before the program can move forward: • During developmental testing only (38) • During operational testing only (17) • During developmental and operational testing (32)	87
Programmatic issues: funding, scheduling, or management problems	72
Manufacturing, software development, integration, or quality control problems	61
Critical Nunn-McCurdy breach[2]	34

1. The total number of programs affected is more than 115 because most programs had more than one reason for a delay.
2. A critical Nunn-McCurdy breach occurs when the program acquisition unit cost or the procurement unit cost increases by at least 25 percent over the current baseline estimate or by at least 50 percent over the original baseline estimate.

2. DOT&E employed a lognormal regression analysis that investigates the expected program delay duration as a function of each of the delay reasons listed in the table. The analysis revealed that delay duration is statistically significantly affected by the following factors: (small p-values, particularly those below 0.10, indicate that the factor significantly affects the delay duration). P-values were 0.08 for critical Nunn-McCurdy breaches; 0.08 for programmatic issues; 0.001 for manufacturing, software development, integration, or quality control problems; and 0.11 for problems discovered in operational testing. Problems discovered in developmental testing and problems in test conduct were not statistically significant factors, since their p-values were both greater than 0.30.
3. 2011 Study: http://www.dote.osd.mil/pub/presentations/20110830Program_delays_Nunn-McCurdy_final.pdf
   2014 Update: http://www.dote.osd.mil/pub/presentations/ProgramDelaysBriefing2014_8Aug_Final-77u.pdf

So, to reiterate, fixing problems discovered during testing causes program delays, not the testing itself.

In the remainder of this introduction, I describe in more detail several recent 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 the analysis of data from testing.
- My continued emphasis on the need to improve reliability of all weapon systems. I include an assessment of new policies on reliability growth and tracking, as well as how the Department is progressing in improving reliability of weapon systems.
- My new guidance on cybersecurity testing. Now and in the future, cybersecurity threats will arguably be some of the most dangerous threats our defense systems face. In 2014, I signed out guidance for testing the robustness of our combat systems' abilities to withstand cyber threats. In this introduction, I outline the highlights and the importance of this guidance, as well as recent cyber testing efforts.
- My emphasis on ensuring adequate test resources are available even when Department budgets are constrained.
- 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 development testing (page 13 in the DOT&E Activity and Oversight section).

RIGOROUS, DEFENSIBLE, EFFICIENT TESTING

Since my appointment as Director, I have required thorough operational tests that provide adequate information to characterize system performance across a variety of operational conditions. This information is essential to my evaluation of system operational effectiveness, suitability, and survivability. I have advocated the use of scientific methodologies, including experimental design or design of experiments (DOE) to ensure that this characterization is done as efficiently as possible. The methodologies that I have advocated for not only provide a rigorous and defensible coverage of the operational space, they also allow us to quantify the trade-space between the amount of testing and the precision needed to answer the complex questions about system performance. They allow us to know, before conducting the test, which analyses we will be able to conduct with the data and therefore, what questions about system performance we will be able to answer. Finally, they equip decision makers with the analytical tools to decide how much testing is enough in the context of uncertainty.

There has been much progress in increasing the statistical rigor of test plans since 2009. Over the past several years, all of the Service Operational Test Agencies (OTAs) have implemented DOE practices to varying degrees and have offered training to their staffs on the statistical principles of DOE. Additionally, the Deputy Assistant Secretary of Defense Developmental Test and Evaluation (DASD(DT&E)) endorses these methods and advocates them through his Scientific Test and Analysis Techniques (STAT) T&E Implementation Plan. That office has also overseen the conduct of the STAT in T&E Center of Excellence (COE), which employs qualified statistics experts to aid acquisition program managers in applying advanced statistical techniques in developmental testing. However, these steps are not enough. In the DOD, we acquire some of the world's most complex systems, but our test and analysis capabilities lag behind the state of the practice, let alone the state of the art for statistical techniques. The DOD Test and Evaluation community should be setting the standard for test and evaluation, not struggling to apply methods that have been applied for decades in other test organizations.

Moreover, it is not sufficient to only employ statistical methods in the test design process; the corresponding analysis methods should be employed in the evaluation of system performance, else we risk missing important conclusions. One example of the benefits of statistical analysis methodologies was revealed during the operational test of a Navy helicopter program, the Multi-spectral Targeting System, which is intended to enable helicopters to target fast,small-boat threats and employ HELLFIRE missiles at safe-standoff distances. The testing conducted thoroughly examined performance under the variety of operational and tactical conditions that a crew might expect to encounter, including a variety of threat types and operating profiles, as well as engagements in different ocean and daylight conditions. A simple analysis of the results combining all of the data together into a single average (a particularly limiting but unfortunately common analysis technique) suggested the system was meeting requirements. However, only when the more complex and rigorous statistical analysis was employed did the testers discover that the system was significantly failing requirements in a subset of the operational conditions. The unique set of conditions in which performance was poor revealed a weakness in the system, which can now be addressed by system developers. It is important to note that if DOT&E had not pushed for this rigorous analysis, this result would have been missed completely.

While there has been a lot of progress, much work remains. The implementation of these techniques is still far from widespread across all DOD T&E communities. Overall, statistical analysis methods such as regression and analysis of

variance, which supported the above discovery, are underused. Until they are routinely employed in the analysis of T&E data, many situations such as the Multi-spectral Targeting System example are likely to be missed. Furthermore, we are currently not leveraging these methods in a sequential fashion to improve knowledge as we move from developmental testing to operational testing. Sequential learning is at the heart of the experimental method, which all testing is based on, and we need to employ such approaches in DOD T&E. Doing so will aid in improving our defense systems by enabling early problem discovery, supporting integrated testing, and improving our ability to clearly define an adequate operational test that avoids the unnecessary expenditure of resources.

DOT&E Efforts to Institutionalize Test Science

Institutionalizing scientific approaches to testing in the DOD T&E community requires a change from business as usual. My office has worked to provide the motivation, resources, education and training, and overall support to the T&E community to make this change possible. DOT&E has worked to institutionalize test science principles by:

- Updating policy and guidance to reflect scientific best practices
- Developing educational and training resources to advance the current workforce
- Developing case studies and examples to illustrate the value of using such techniques in test and evaluation
- Forming advisory groups and a research consortium to provide the T&E workforce with advice, help solve challenging problems, and develop the future workforce

A 2013 report summarized the efforts to institutionalize scientific methods in the DOD T&E community and discussed each of these focus areas. This year, we have continued to advance each of those areas and provide support for the T&E community; our efforts are described below.

Policy and Guidance

Both the developmental and operational testing portions of the Interim DOD Instruction 5000.02, "Operation of the Defense Acquisition System," now call for a scientific approach to testing. The operational test section calls for documenting each of the elements of the test design, as well as basing all test resources on statistically-derived test design and quantification of risk. I have authored several guidance memos on the use of statistical methods in OT&E, and the new 5000.02 guidance codifies the driving principles of those guidance memos in DOD policy.

In 2014, I issued a new guidance memo on the design and use of surveys in OT&E. Surveys provide valuable quantitative and qualitative information about the thoughts and feelings of operators and maintainers as they employ weapon systems in an operationally realistic test environment. An objective measurement of these thoughts is an essential element of my evaluation of operational effectiveness and suitability. However, I have noted that many of the surveys used in operational T&E are of such poor quality they can actually hinder my ability to objectively evaluate the system. For this reason, I issued a policy memorandum, based on best practices from the academic community, that provided concrete guidance for the use of surveys in OT&E. The key elements included:

- Use surveys only when appropriate; do not ask operators about system performance attributes that are more appropriately measured by the testers (e.g., accuracy/timeliness of the system).
- Use the right survey and leverage established surveys when appropriate.
- Employ academically-established best practices for writing and administering surveys.

I strive to ensure that my guidance is widely available to the T&E community. The DOT&E website is a convenient source for all DOT&E guidance memos.[4] Additionally, the website has a copy of the current Test and Evaluation Master Plan (TEMP) Guidebook, which is an important resource for DOT&E guidance on TEMPs.[5] This guidebook provides references to all guidance and policy memoranda, describes in plain language what I am looking for in my review of TEMPs, and provides several examples taken from various TEMPs that meet my expectations. The TEMP format was revised with the new 5000.02. New content was added, including a requirements rationale and overview of the concept of operations. My staff is currently working on the development of the next version of the TEMP Guidebook, which I expect will be available in 2015.

Education and Training

The use of statistical methods in DOD T&E has been limited by inadequate access to education and training opportunities by our T&E practitioners. Many great training opportunities exist, and I encourage DOD leadership to make a commitment to improving the access of our T&E professionals to education and training. Select members of the workforce need to have graduate degrees in fields related to test science (statistics, applied mathematics, operations research, etc.). Additionally, all members of the T&E workforce, including the system engineers who develop and test these systems prior to formal

4. http://www.dote.osd.mil/guidance.html
5. http://www.dote.osd.mil/temp-guidebook/index.html

developmental and operational testing, should have a base level of training in experimental design and statistical analysis methods. A combination of both longer-term education and short-term training is necessary for our test organizations to truly improve the rigor of all testing. At DOT&E, we have developed custom training for our action officers on DOE, reliability, survey design, and statistical analyses. Additionally, we have developed advanced training materials for our analysts on power analysis and statistical analysis methods. I am always happy to share these resources with the rest of the T&E community and welcome the participation of OTAs and other organizations in our DOT&E training.

In addition to providing training, DOT&E is committed to developing an online knowledge center for the DOD T&E community. We have initiated development of a web-based interface to this knowledge center, which will include training material, resources for best practices, tutorials, and web-based tools related to DOE and statistical analysis. This website is being built in collaboration with many organizations across the DOD T&E community. An initial version of the knowledge center is scheduled for completion in 2015.

Case Studies

Case studies provide valuable insight on the application of statistical methods, including DOE in operational testing. Over the past several years, DOT&E has developed many case studies illustrating the application of DOE and statistical analysis in T&E. Many of these case studies are summarized in previous DOT&E publications (e.g., Test Science Roadmap published in 2013), but new case studies continue to be developed.[6] As these new case studies have become available, we have shared them with the OTA leadership.

Testing of the complex systems the DOD T&E community encounters often requires non-standard applications of the tools and methods currently available in the statistics and test literature. DOT&E has used case studies to illustrate how advanced methods can be used to improve test outcomes and sculpt existing methods to meet the needs of the T&E community. One example of this occurs in the translation between probability-based metrics and more informative continuous metrics, such as time, distance, etc. Binary or probability-based requirements such as probability-of-detection or probability-of-hit, provide operationally meaningful and easy-to-interpret test outcomes. However, they are information-poor metrics that are extremely expensive to test. Having a continuous metric can reduce test sizes by 50 percent or more and provide more information in the analysis, but it can be unclear how to convert between the probability metric and the corresponding continuous metric. DOT&E has developed several case studies illustrating this translation for different test outcomes. We are now using statistical cumulative density functions and censored data analyses, resulting in much more efficient tests that require smaller sample sizes than would be required to accurately measure the related binary metric.

Advisory Groups

Engagement with the academic community and leading experts in the field of T&E is essential to advancement of these rigorous statistical techniques in DOD T&E. In 2014, DOT&E renewed funding for the Test Science Research Consortium in partnership with the Department's Test Resource Management Center. This multi-year research consortium is tasked with addressing the unique needs of the T&E community. This consortium funds several graduate-level research projects on advanced statistical techniques, enabling these projects to focus on topics of benefit to the Department's T&E needs and preparing a pipeline of students with strong technical skills to learn about to the Department and the T&E community. This research consortium has already produced several new members of the T&E community with advanced degrees in statistics and related fields.

Finally, the STAT T&E COE has for three years provided direct T&E support to 24 program offices. The COE has provided these programs with direct access to experts in test science methods, which would otherwise have been unavailable. I have observed much benefit and value from the COE's engagement with programs. However, the COE's success has been hampered, in part, by unclear funding commitments in the out-years. Furthermore, the analysts are often constrained to only answering specific, and sometimes narrowly-defined questions, as opposed to providing independent thought on the full content of a program's development test strategy. In the case of the self-defense testing for the Air and Missile Defense Radar and DDG 51 Flight III Destroyer, the COE analysts were constrained to constructing a test for a limited set of conditions, particularly excluding the self-defense region near the ship where the complex interactions between multiple combat system components (missiles, radars, and control systems) are not known. Although the test design provided was robust for the limited question asked of the COE, it egregiously missed the most crucial region of the battlespace, and gave the false impression that results from such a test design were adequate to fully characterize performance of the combat system and that a self-defense test ship was unneeded to examine performance in the self-defense region. I will continue to advocate that programs have access to a STAT COE and make use of these excellent capabilities; however, the COE must

6. http://www.dote.osd.mil/pub/reports/20130711TestScienceRoadmapReport.pdf and http://www.dote.osd.mil/pub/reports/20130711Appdxes2theTestScienceRoadmapReport.pdf

have the ability to provide independent assessments to programs. Furthermore, the COE needs to be appropriately funded to be successful and needs to expand in size to aid program managers in smaller acquisition programs (Acquisition Category II and III). 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.

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. There is evidence that those systems that implement and enforce a comprehensive reliability growth program are more likely to meet their reliability goals; however, test results from the last few years indicate the DOD has not yet realized statistically-significant improvements in the reliability of many systems.

The Department has acknowledged this poor track record of meeting system reliability requirements. In 2011, the Under Secretary of Defense for Acquisition, Technology and Logistics (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 Initial Operational Test and Evaluation (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's Systems Engineering office in USD(AT&L) to implement a systematic process of tracking the reliability status of a Major Defense Acquisition Program (MDAP). Beginning with FY14, MDAPs in system-level developmental testing with a documented reliability growth curve in the Systems Engineering Plan or TEMP were required to report reliability data on a quarterly basis. At present, 18 programs are actively reporting reliability data via this process, making a system's progress relative to its expectations (seen through its reliability growth curve) a visible factor for the DAES process to consider. While the number of systems reporting these data to DAES is increasing, the information is not yet being used to trigger programmatic reviews or decision-making.

Current Reliability Trends

To better understand ongoing trends in reliability, my office has conducted a survey of programs under DOT&E oversight in each of the past six 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 either submitted a TEMP to DOT&E and/or had an operational test in FY13.

The survey results indicate that programs are increasingly incorporating reliability-focused policy guidance. Since FY13:

- 89 percent of programs had a reliability growth strategy, with 92 percent documenting it in the TEMP.
- Likewise, 83 percent of programs incorporated reliability growth curves into the TEMP.
- 88 percent of programs had interim reliability metrics prior to the system requirement.
- While only 28 percent of programs with FY13 TEMPs included a discussion of producer and consumer risk for passing the reliability threshold in IOT&E, this represents significant progress because only one program had done this in the past.

Differences have been observed in the implementation of policy across the different services. The Army has been a leader at implementing reliability policy and the use of reliability growth planning curves, and while the Air Force has caught up considerably, many Navy programs have yet to adopt these methods. This includes the use of reliability growth curves, the use of intermediate goals based on demonstrating reliability thresholds at operational test events, and discussing producer and consumer risk (statistical power and confidence) in the TEMP.

Despite these improvements in policy implementation, we have not observed a similarly improving trend in reliability outcomes at operational test events. Reliability growth curves are excellent planning tools, but programs will not achieve their reliability goals if they treat reliability growth as a "paper policy." Good reliability planning must be backed up by sound implementation and enforcement.

The survey results indicate that two essential elements of this implementation are 1) including the reliability program in contracting documents and 2) having reliability-based test entrance criteria. Programs that implemented at least one of these actions were statistically more likely to meet their reliability requirement in operational testing. This is to be expected, as inclusion of the reliability growth program in contracting documents provides the government with additional leverage to

ensure that contractors deliver reliable systems. The survey results revealed that the Army and Air Force have been more likely than the Navy to include the reliability growth plan in contracting documents and meet entrance criteria based on reliability, availability, and maintainability for operational test events. Unfortunately, the survey also revealed that it is not common practice for any Service to implement these steps.

While following this guidance may lead to improvements in reliability in the future, at present, many programs still fail to reach reliability goals. The reasons programs fail to reach these goals are numerous, but include overly-ambitious requirements, unrealistic assumptions about a program's capability for reliability growth, lack of a design for reliability effort prior to Milestone B, and/or failure to employ a comprehensive reliability growth program. For example, the reliability thresholds for some programs were unachievably high or disconnected from what was really needed for the mission. In some cases, a program's reliability growth goal, though documented in a TEMP or SEP, was not supported by contractual obligations or funding. A larger fraction of surveyed programs met their reliability thresholds after fielding during Follow-On Operational Test and Evaluation (FOT&E) (57 percent) rather than before fielding during IOT&E (44 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. Furthermore, until we as a Department demonstrate commitment to enforcing these principles by including them in contracting documents and enforcing test entrance criteria, programs will have little incentive to actively pursue and fund system changes that lead to improve reliability.

It is also essential that we collect enough information to adequately test system reliability. The survey results showed IOT&Es and FOT&Es often are not adequately sized to assess the system's reliability requirement with statistical confidence and power. For many programs, such testing is not achievable based on concerns such as cost and schedule. In other cases, the requirements were either not testable or not operationally meaningful. In these cases, as always, my assessment of system reliability was based on how the systems' demonstrated reliability would impact the warfighters' ability to complete their mission. Despite the fact the survey revealed many operational tests are not statistically adequate to assess requirements, in most of these cases, DOT&E had sufficient data to assess system reliability performance. When system reliability is substantially below the requirement, it is possible to determine with statistical confidence the system did not meet its requirement with substantially less testing than would otherwise be required. In other cases, other sources of data can be used. This overarching result demands that we must think about reliability testing differently. The next version of my TEMP Guidance will include discussion on how the TEMP should be used to specify which data sources will be used in assessing system reliability at IOT&E, as well as the fidelity these sources must achieve to be included in this assessment. This will assist programs in adequately scoping IOT&E and FOT&E test lengths, helping them to allocate their T&E resources more efficiently.

National Academy of Sciences (NAS) Reliability Study Results

Recently, the National Academy of Sciences (NAS) published a report on reliability and reliability growth for defense systems; this report was the result of a study commissioned by myself and Mr. Frank Kendall, the USD(AT&L).[7] In this report, NAS offered recommendations for improving the reliability of U.S. defense systems. The recommendations advocated for many of the same principles that I support, including:

- Implementing reliability growth programs that include failure definitions and scoring criteria as well as a structure for reporting reliability performance over the course of the acquisition process.
- Using modern design-for-reliability techniques supported by physics of failure-based methods.
- Planned test lengths that are statistically defensible.

NAS also suggested that these plans be updated periodically throughout the life of the program, including at major design reviews and program milestones.

The NAS study addresses the need for appropriate reliability requirements. NAS recognizes, as I have for years, the need for technically-justified, testable, mission-relevant requirements. These requirements must also balance acquisition costs and lifetime sustainment costs. Systems that push the limits of technical feasibility will be more expensive to acquire initially, but may reduce lifecycle costs. However, reliability requirements that greatly exceed current capabilities may be unachievable and drive acquisition costs unnecessarily. As systems evolve, the requirements may need to be updated as the system engineering becomes more fully understood, but all changes in these requirements should be considered in the context of the mission impact of the change.

The NAS report also points to the importance of reliability-focused contracting. Making reliability a Key Performance Parameter on all new systems and ensuring all proposals include explicit language designating funds and describing the

7. http://www.nap.edu/catalog/18987/reliability-growth-enhancing-defense-system-reliability

design for reliability activities (including early system reliability testing) will provide the DOD with leverage to ensure delivered systems are reliable. As mentioned above, my survey of acquisition programs has found that including the reliability growth plan in the contracting documents does indeed make systems more likely to meet their reliability threshold. NAS also recommends the use of rigorous reliability-based entrance criteria prior to operational testing, stating,

> *"Near the end of developmental testing, the USD(AT&L) should mandate the use of a full-system, operationally-relevant developmental test during which the reliability performance of the system will equal or exceed the required levels. If such performance is not achieved, justification should be required to support promotion of the system to operational testing."*

I have also found that making sure systems meet their entrance criteria prior to entering their IOT&E makes them much more likely to perform well in the operational test.

Recent Lessons on Reliability Requirements

The first step in ensuring reliable systems is to ensure that requirements are appropriate. Sound reliability requirements are grounded in the operational relevance of the missions the system will support. They also ensure technical feasibility based on existing systems and engineering limitations, balance acquisition costs and sustainment costs, and are testable. Two recent examples have illustrated the tendency to develop reliability thresholds for programs that are unachievably high and/or disconnected from what is needed for the mission.

Three-Dimensional Expeditionary Long-Range Radar (3DELRR)

Three-Dimensional Expeditionary Long-Range Radar (3DELRR) will be the principal Air Force long-range, ground-based sensor for tracking aircraft and other aerial targets. It replaces the aging TPS-75 radar, which is incapable of detecting some current and emerging threats and has become difficult and expensive to maintain.

While working with the Air Force to develop plans to test 3DELRR, DOT&E observed the 720-hour Mean Time Between Critical Failure (MTBCF) appeared to be unnecessarily high and disconnected from the related availability requirement, set at 0.947. When asked, the Air Force's rationale for this requirement was initially presented as "the system should be operational for 30 days, 24 hours per day, failure free." The initial Service position was that establishing an operational availability (A_O) of 0.947 with an associated 720-hour MTBCF would ensure the capability to sustain operations for 30 days in austere locations with minimum external support. However, the probability of completing the 30-day mission with zero critical failures is about 0.37, assuming a system MTBCF of 720 hours. Achieving mission reliability values higher than 0.37 would require very-high MTBCF values; requiring a 90-percent probability of completing a 30-day mission without failure would require an MTBCF of over 6,800 hours. A lower MTBCF of 300 hours would provide availability near 0.90, with each 100 hours of reliability adding just a fraction to the overall A_O. Based on these observations, DOT&E recommended the Service review the reliability requirement to determine if an MTBCF of 720 hours was truly needed to achieve reasonable reliability. Additionally, DOT&E recommended that once the reliability requirement was validated, the Service should implement a design for a reliability program and implement a reliability growth program.

After multiple discussions and review of the logistics supply concept, as well as the concept of operations for completing missions, the Air Force recognized that a 720-hour MTBCF was not, in fact, required. After further review, the Service set the requirement that the system will achieve an MTBCF of 495 hours by the end of government-conducted developmental T&E, along with an A_O of 0.947. Furthermore, the Air Force designed an acceptable reliability growth program that adheres to best practices, and DOT&E approved the program's TEMP for Milestone B.

Ground/Air Task Oriented Radar (G/ATOR)

The Marine Corps' Ground/Air Task Oriented Radar (G/ATOR) is a phased array, multi-role radar that is being designed to be used initially in the air surveillance and air defense roles, with follow-on capabilities that will be developed to support the ground weapon locating/counter-targeting and air traffic control mission areas. During a recent operational assessment period, G/ATOR met its A_O Key Performance Parameter; however, Key System Attributes reflecting system reliability were well below thresholds. DOT&E issued an operational assessment report, as well as a separate memorandum discussing the program's proposed reliability growth plans in the related draft TEMP, that again noted several problems with the system's reliability and growth-planning assumptions. As a result of G/ATOR not meeting planned reliability growth milestones, and with no clear means to grow the reliability to that required and maintain program timelines, the Navy stood up a "Blue Ribbon Panel" made of Department experts and stakeholders (including DOT&E) to assess the program's ability to achieve current reliability threshold requirements. The panel's findings included:

- The Mean Time Between Operational Mission Failure reliability threshold requirement is disconnected from the mission (not operationally relevant).

- There is no clearly defined G/ATOR system definition for government- and contractor-furnished equipment.
- The rationale for excluding certain government-furnished equipment from reliability calculations is ambiguous.
- There is no closed loop in the failure reporting, analysis, and corrective action system; specifically, the program's Failure Review Board determines which failures are valid but does not formally adjudicate the effectiveness of corrective actions.
- Reliability growth planning models used optimistic planning factors and growth curves, and were based on Mean Time Between Operational Mission Failure/Mean Time Between Failure initial values that were not previously realized during testing.
- Definitions for failures and operating time during previous developmental test are not consistent.

The findings were recently briefed to the Milestone Decision Authority for G/ATOR. Recommendations to the above findings are currently under review.

Both the 3DELLR and G/ATOR programs reveal the need to carefully consider the reliability requirements in relation to what is essential for completing a mission within the Services' own concepts of operations. Acting on the recommendations of the NAS study, as well as those others and I have stated, will ensure programs not only are successful in achieving their reliability goals, but also that the goals are realistic and achievable.

CYBERSECURITY OPERATIONAL TESTING AND ASSESSMENTS DURING EXERCISES

Cyber adversaries have become as serious a threat to U.S. military forces as the air, land, sea, and undersea threats represented in operational testing for decades. Any electronic data exchange, however brief, provides an opportunity for a determined and skilled cyber adversary to monitor, interrupt, or damage information and combat systems. 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 a realistic cyber threat. My assessment of operational effectiveness, suitability, and survivability is determined in part by the results of this crucial testing.

During 2014, cybersecurity testing of more than 40 systems showed improvements must occur to assure secure and resilient cyber capabilities. One important conclusion from my 2014 review of DOD programs was that operational testing still finds exploitable cyber vulnerabilities that earlier technical testing could have mitigated. These vulnerabilities commonly include unnecessary network services or system functions, as well as misconfigured, unpatched, or outdated software, and weak passwords. Developmental testing over the course of the program, including the process to grant a system the authority to operate on DOD networks, could have found most of these vulnerabilities; yet, such vulnerabilities are still found as late as during the IOT&E. My review of these systems also identified the need to increase the participation of network defenders and assessment of mission effects during threat-representative, adversarial assessments.

In August 2014, I published updated policy and procedures for cybersecurity assessments in operational T&E; the new guidance specifies that operational testing should include a cooperative vulnerability assessment phase to identify system vulnerabilities followed by an adversarial assessment phase to exploit vulnerabilities and assess mission effects.[8] The adversarial assessment phase includes system users and network defenders to detect the adversarial actions, react to those actions, and restore the system to full/degraded operations after a cyber-attack. My office continues to emphasize the need to assess the effects of a debilitating cyber-attack on the users of these systems so that we understand the impact to a unit's mission success. A demonstration of these mission effects are 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, my office has 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 in FY14, and the Test Resource Management Center, which oversees the NCR, has initiated studies to examine new capabilities to further expedite the planning, execution, and sanitization of NCR events.

8. http://www.dote.osd.mil/pub/policies/2014/8-1-14_Procs_for_OTE_of_Cybersec_in_Acq_Progs(7994).pdf

Also in 2014, my office conducted 16 cybersecurity assessments in conjunction with Combatant Command and Service exercises. A notable improvement over previous years was the increased participation of higher-echelon computer network defense service providers and local defenders, resulting in a more comprehensive assessment of cyber defensive postures. Despite the improved defenses, my office found that at least one assessed mission during each exercise was at high risk to cyber-attack from beginner to intermediate cyber adversaries. I have placed emphasis on helping Combatant Commands and Services mitigate and reduce those persistent cybersecurity vulnerabilities observed from assessment to assessment. My continuing focus is on finding problems, providing information and assistance to understand and fix problems, and following up to verify cybersecurity status and ability to conduct operations in a contested cyberspace environment. At the request of several Combatant Commands, I have implemented more frequent operational site assessments during day-to-day operations on live networks to provide feedback on specific areas of interest such as status of patching or defense against specific attacks (e.g., phishing) and cybersecurity implications of physical security. Additional continuing efforts include working with the intelligence community to improve cyber threat realism, and to develop a persistent cyber opposition force with the capability to operate across several Combatant Commands.

TEST RESOURCES

Adequate funding of test resources remains a crucial aspect to fielding weapons that work. My office continues to monitor DOD and Service-level strategic plans, investment programs, and resource management decisions to ensure the Department maintains the capabilities necessary for adequate and realistic operational tests. I have continued to emphasize the need for these resources despite the constrained fiscal environment. There are some who argue that in a constrained fiscal environment, particularly in the face of sequestration, all testing should be cut commensurate with cuts in program budgets. That is, if the Department's budgets are reduced by 20 percent, then testing should also be reduced by 20 percent. Yet, we are fielding the weapons that are developed to satisfy 100 percent of their concepts of operation against 100 percent of the actual threat. In particular, what constitutes adequate operational testing under realistic combat conditions is determined not by fiscal constraints, but by our war plans and the threats we face—the enemy (always) gets a vote. It would therefore, be a false economy and a disservice to the men and women we send into combat to make arbitrary budget-driven reductions to either developmental or operational testing.

The T&E Resources section of this Annual Report details the projects and issues on which I am most concerned or focused. Of particular note this year is that I remain concerned about the substantial year-after-year staffing reductions taken by the Army T&E Executive and his office within the Office of the Deputy Under Secretary of the Army, as well as reduction in staff levels in both the Army Operational Test Command and the Army Evaluation Center. These reduced staff levels will cause delays to developmental and operational testing, the inability to conduct simultaneous operational test events, and longer timelines for the release of test reports. Furthermore, the Commander, Operational Test and Evaluation Force continues to try to enhance his workforce by growing in-house technical talent and hiring in personnel with advanced degrees.

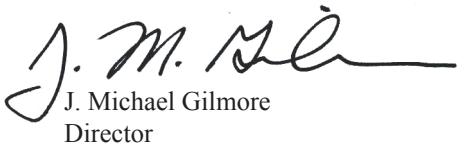
As the Department moves forward in considering important test resource infrastructure and investments in the face of constrained budgets, I will continue to advocate for the need for the most crucial test assets. These include:

- The need for an Aegis-capable self-defense test ship 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. (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 Evolved SeaSparrow Missile Block 1) and ship classes (LPD-17, LHA-6, Littoral Combat Ship, LSD 41/49, DDG 1000, and CVN-78), the Navy has not made a similar investment in a self-defense test ship equipped with an Aegis Combat System, Air and Missile Defense Radar, and Evolved SeaSparrow Missile Block 2 for adequate operational testing of the DDG 51 Flight III Destroyer self-defense capabilities.)
- The DECRE, as discussed above.
- The electronic warfare infrastructure, which the Department is funding and for which some slow progress is being realized.
- The electronic warfare assets for anti-ship cruise missile seeker emulation and the jamming simulators for the assessment of Army communications networks.

Other resource needs that I consider crucial for the Department to pursue are detailed in the T&E Resources section of this report (page 339).

SUMMARY

Since my first report to you in 2009, we have made progress increasing the scientific and statistical rigor of operational T&E; there is much work to be done, however, since the Department's test design and analysis capabilities lag behind the state of the practice. Additionally, we have focused attention on reliability design and growth testing, and in improving cybersecurity operational testing. Operational testing continues to be essential to characterize system effectiveness in combat so 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 T&E activities of the DOD during Fiscal Year 2014.



J. Michael Gilmore
Director