FY 2012 Annual Report

Since my confirmation as Director of Operational Test and Evaluation (DOT&E) in 2009, I have implemented initiatives to improve the quality of test and evaluation (T&E) within the Department of Defense. I have emphasized early engagement of testers in the requirements process, improving system suitability by designing reliability into systems from the outset, and integrating developmental, operational, and live fire testing. Implementing these initiatives has revealed the need for an additional area of focus – the requirement to incorporate statistical rigor in planning, executing, and evaluating the results of testing.

There are significant opportunities to improve the efficiency and the outcomes of testing by increasing interactions between the testing and requirements communities. In particular, there should be early focus on the development of operationally relevant, technically feasible, and testable requirements. In this Introduction, I discuss the crucial role the T&E community can and should play as requirements are developed. Additionally, I describe DOT&E efforts to institutionalize the use of statistical rigor as part of determining requirements and in T&E. I also provide an update on the Department's efforts to implement reliability growth planning and improve the reliability and overall suitability of our weapon systems. And lastly, I describe challenges and new developments in the area of software T&E.

Last year, I added a new section to my Annual Report assessing systems under my oversight in 2010 – 2011 with regard to problem discovery during testing. My assessment fell into two categories: systems with significant issues observed in operational testing that should, in my view, have been discovered and resolved prior to the commencement of operational testing, and systems with significant issues observed during early testing that, if not corrected, could adversely affect my evaluation of those systems' effectiveness, suitability, and survivability during Initial Operational Test and Evaluation (IOT&E). This year, I am providing an update to the status of those systems identified last year, as well as my assessment of systems under my oversight in 2012 within those two categories.

THE ROLE OF T&E IN REQUIREMENTS

There is an inherent and necessary link between the requirements and the test communities. The requirements community must state our fighting force's needs in the form of concrete, discrete capabilities or *requirements*. The testing community must then assess a system that is developed and produced to meet those requirements to determine whether it provides the military capability being sought; that is, we evaluate the system's operational effectiveness and suitability when used by our forces in combat. In my opinion, the collaboration needed between the requirements and the test communities to discharge these responsibilities needs to be strengthened.

In my report last year, I discussed the Defense Acquisition Executive (DAE) independent assessment of concerns that the Department's developmental and operational test communities' approach to testing drives undue requirements, excessive cost, and added schedule into programs. The DAE assessment team "found no significant evidence that the testing community typically drives unplanned requirements, cost, or schedule into programs." However, they did note that there were four specific areas that needed attention:

> *"The need for closer coordination and cooperation among the requirements, acquisition, and testing communities; the need for well-defined testable requirements; the alignment of acquisition strategies and test plans; and the need to manage the tension between the communities."*

The lack of critically needed collaboration among the technical, test, and requirements communities is not new. The 1986 Packard Commission found that success in new programs depends on "an informed trade-off between user requirements, on one hand, and schedule and cost, on the other." It therefore recommended creation of a new body representing both military users and acquisition/technology experts. This ultimately led to the creation of the Joint Requirements Oversight Council (JROC), which includes the military operators as formal members but includes, as advisors only, the acquisition and test communities. In 1998, the National Research Council (NRC) identified the need for greater interaction between the test and the requirements communities; the NRC pointed out that operational test personnel should be included in the requirements

process in order to assist in establishing "verifiable, quantifiable, and meaningful operational requirements." And the National Defense Authorization Act for FY11 specifically named DOT&E as an advisor to the JROC. However, obstacles for close collaboration remain. I discuss below three specific areas where increased interactions could result in improved test outcomes, which should then result in systems with needed and useful combat capability being delivered to our forces more quickly.

Mission-Oriented Metrics

OT&E is defined in Title 10 United States Code as:

> *"The field test, under realistic combat conditions, of any item of (or key component of) weapons, equipment, or munitions for use in combat by typical military users; and the evaluation of the results of such tests."*

Weapon systems sit in the motor pool, at the pier, or on the runway. Individual systems do not have missions; it takes Soldiers, Sailors, Airmen, and Marines to make them work. Operational testing is about assessing mission accomplishment of the unit equipped with a system. To evaluate operational effectiveness we seek to answer the question, "can a unit equipped with the system accomplish the mission?" Operational effectiveness is defined in the Joint Capabilities Integration and Development System (JCIDS) manual as:

> *"Measure of the overall ability of a system to accomplish a mission when used by representative personnel in the environment planned or expected for operational employment of the system considering organization, doctrine, tactics, supportability, survivability, vulnerability, and threat."*

And the Defense Acquisition Guide emphasizes "the evaluation of operational effectiveness is linked to mission accomplishment." End-to-end testing with operational users across the intended operational envelope is essential to assessing the system's impact on mission accomplishment. Additionally, each system must be evaluated within the context of the system-of-systems within which it will operate.

In January 2010, I provided guidance to the Operational Test Agencies on the reporting of OT&E results reiterating that the appropriate environment for any operational evaluation includes the system being tested and all interrelated systems needed to accomplish an end-to-end mission in combat. I emphasized that the primary purpose of OT&E is to describe the operational effectiveness and suitability of the system being tested within that environment. A subsidiary purpose of OT&E, stated in DoDI 5000.02, is to determine if thresholds in the approved Capability Production Document (CPD) have been satisfied. The measures used for this purpose are appropriately referred to in the context of "performance" as in "key performance parameters (KPPs)," or "measures of performance." But these measures associated strictly with evaluating KPPs are not the full set necessary to evaluate operational effectiveness in combat.

Requirements are often stated in terms of technical parameters whose satisfaction is necessary, but not sufficient to determine a system's effectiveness, suitability, and survivability when used in combat. Ideally, KPPs should provide a measure of mission accomplishment, lend themselves to good test design, and encapsulate the reasons for procuring the system. However, DOT&E has seen many examples of KPPs that are not informative to an evaluation of mission accomplishment. For example, a previous ground combat vehicle had KPPs that only required it seat nine passengers, be transportable by a C-130, and have a specific radio system; these requirements could have been met by a passenger van. Another example was an amphibious ship with KPPs for the number of helicopter spots, the number of storage spaces, and the maximum speed of the ship; these requirements could have been measured with a stopwatch and a tape measure and could have been satisfied by a commercial ship with no capability to survive amphibious combat. While these technical performance requirements are important, they are not sufficient to determine whether the ground vehicle or ship can be used successfully in combat. In these cases, the test community encouraged the use of metrics for evaluation directly tied to mission success such as accomplishing geographic objectives while minimizing blue force losses or meeting an aircraft sortie generation rate and surviving likely attacks. If the test and requirements communities engage early, requirements can be stated in a manner that makes them directly relevant to mission success and therefore, both directly relevant to operational testing and much more capable than technically-oriented parameters of informing whether the sought-for combat capabilities have been achieved in the system to be produced.

Leveraging T&E Knowledge in Setting Requirements

Interactions between the requirements writers and the testers can also help identify alternatives to hard-to-test or impossible-to-test requirements. Requirements that cannot be verified in testing may as well not exist. The T&E community can help identify unrealistic, unaffordable, and un-testable requirements. Additionally, T&E knowledge of the current threat environment and test infrastructure can help the requirements community understand what resources will be needed to test a given requirement. We have seen Service requirements officers state they want demanding if not technically unachievable

requirements to drive vendors to deliver the best possible system performance; but, history has shown setting very high or unachievable requirements is particularly destructive to program success. For example, the Future Combat System program required high survivability ("tank-like") and tactical transportability (via C-130) that, together, were impossible to achieve. Additionally, reliability requirements for that system were much higher – nearly 10 times – that of our current systems, making achievement of those requirements both unrealistic and unaffordable. Clearly, we should not eliminate requirements simply because they are difficult to test. We must, however, carefully consider whether difficulty (or impossibility) of testing requirements implies the same for their achievement.

Testers have experience with the difficulty and cost associated with the testing needed to demonstrate whether certain metrics have been achieved. For example, consider a requirement for 99 percent reliability for completing a 6-hour mission. This is comparable to 600 hours between failures and would require at a minimum 1,800 hours of testing to verify. However, if the requirement were 95 percent reliability for completing the same 6-hour mission, the associated mean time between failures is only 120 hours and testing to that requirement could be accomplished in a minimum of 360 hours. If the testing revealed 40 hours between failures (instead of 120 or 600) that would indicate an 86 percent probability of completing a 6-hour mission. Would 95 percent or 86 percent be good enough? To answer that question, the rationale, or so-what factor, for the requirement should be fully explained. Accordingly, I intend to require that Test and Evaluation Master Plans (TEMPs) have an annex explaining the user's rationale for the requirements contained in the Capability Development Document. The requirements and their associated rationale should be revisited as often as needed as a program proceeds and knowledge is gained regarding the ability to achieve the program's currently stated requirements.

In addition to the value selected for a requirement, the manner in which a requirement is stated can also make testing expensive or impractical. For example, metrics stated as binomial probabilities (99 percent probability of detecting a target) are expensive to test because they require large sample sizes to gain statistical confidence in the results. Metrics that are physical, continuous, easily measured, and operationally meaningful can be used instead of such probabilities. For example, the "median miss distance" can be measured at high confidence with about a third the number of tests as the "probability of hit," and also provides more information from the resulting distribution of measurements (how close or far away) than a simple hit/miss answer. In many instances, the probabilities now often used to state requirements can be subsequently estimated using test data collected to evaluate continuous response metrics. Thus, wherever possible, I am requiring test plans that measure continuous performance variables as the basis for evaluating thresholds for requirements that have been written in terms of probabilities.

Evaluation Across the Operational Envelope

Another disconnect among the requirements, test, and operational communities is that often requirements are narrowly-focused and do not cover the operational envelope; a notional depiction is shown in Figure 1. To be adequate, the operational evaluation must report performance of the system across the operational envelope, not just at single conditions specified in the capabilities documents. There is a common concern that failing to specify a certain, limited set of conditions within requirements could lead to an unwieldy test. This is a key reason DOT&E is using Design of Experiments (DOE) to plan testing that efficiently spans the operational envelope. Requirements would be much more useful and meaningful if they identify multiple conditions in which the system is likely to be operated.

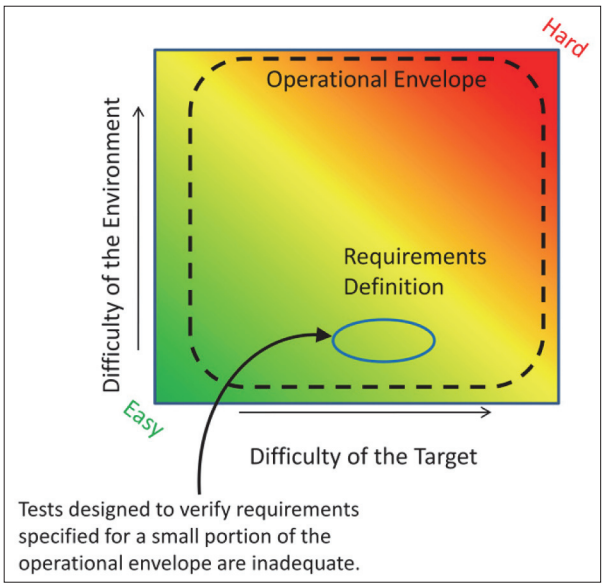

FIGURE 1. NOTIONAL TWO-DIMENSIONAL DIAGRAM OF A WEAPON SYSTEM'S OPERATIONAL ENVELOPE

I will continue to advocate for and require the use of DOE to plan and execute tests that span the operational envelope. One of the key tenets of a well-designed experiment is that all stakeholders must be engaged in the determination of the goals, metrics, operational envelope, and test risks. The requirements community is a key stakeholder that can provide valuable input regarding what the key factors (or conditions) are that will most influence mission performance and thus should be considered in operational test.

In summary, through early and continuous engagement between the testing and requirements communities, we can craft requirements that are technically feasible, mission-oriented, realistic, testable, and responsive to the limitations and opportunities revealed during system development.

INCREASING STATISTICAL RIGOR OF T&E

In support of all of my initiatives, I have advocated for increasing the statistical rigor employed in planning and executing T&E. To that end, my office has recently completed a roadmap to institutionalize Test Science and statistical rigor in T&E. The roadmap was a collaborative activity among DOT&E, Deputy Assistant Secretary of Defense for Developmental Test and Evaluation (DASD(DT&E)), the Service Operational Test Agencies (OTAs), and the Service T&E Organizations.

By increasing statistical rigor and using state-of-the-art test and analysis methodologies, we will ensure defensible and efficient T&E. The Test Science Roadmap accomplishes the following:

- Assesses the current state of analytic capabilities within each of the Services and Office of the Secretary of Defense (OSD)
- Develops qualification guidelines for personnel performing test design and analytic services for different kinds of T&E organizations
- Identifies the training, education, and other support that Services and agencies will need to attain the required test design and analytic capabilities
- Develops case studies of the implementation of scientific test design across test programs
- Provides guidance for the documentation of test design and statistical rigor in TEMPs, test plans, and reports
- Forms a permanent Advisory Board to continually identify and advocate for the use of methods to incorporate statistical rigor in all test planning and execution

We have made significant progress in the past two years across all areas of the roadmap, as discussed below.

Education & Training

DASD(DT&E) is leading the way in improving the educational materials needed by our T&E community, and I strongly support them in this initiative. In the past year, we have added courses and content on statistical methods for T&E to courses offered by the Defense Acquisition University. We have also made training widely available across DOT&E, DASD(DT&E), and all of the Services.

Case Studies & Best Practices

Case studies are an essential educational tool illustrating the application of statistical methods, including DOE to T&E. Over the past couple of years, my office has developed and published many case studies demonstrating the usefulness of applying DOE and statistical methods to T&E. Additionally, in the roadmap meetings, each of the Services shared case studies highlighting the application of DOE to solve their Service-specific problems. DOT&E has compiled these case studies as a resource for the T&E community (https://extranet.dote.osd.mil). They highlight challenges, areas for further research, and best practices.

Guidance & Policy

Policy that supports the use of scientific test techniques is essential to ensuring a continued commitment to Test Science in years to come. Both DASD(DT&E) and DOT&E have supported including more detailed language in DoDI 5000.02 on increasing statistical rigor of T&E. DOT&E also published a TEMP guidebook highlighting the important content for TEMPs and test plans. This guidance is available on the DOT&E public website (www.dote.osd.mil). DASD(DT&E) has also taken the lead on incorporating Test Science topics into other guidance documents including the T&E Management Guide and the Guide on Incorporating T&E into DoD Acquisition Contracts. All of these resources provide clear and consistent guidance to the T&E community on the importance of statistics in T&E. DOT&E insists that TEMPs and test plans submitted for approval include substantive documentation of the application of DOE to test planning, execution, and evaluation.

Advisory Board

Two different advisory groups have been formed in the past two years. The first is the Science of Test Research Consortium, funded by DOT&E and the Director of Test Resource Management Center; this academic consortium provides technical advice to the DoD on Test Science issues. The second is the Scientific Test and Analysis Techniques (STAT) Center of Excellence (COE). The STAT COE funded by DASD(DT&E) is charged with assisting program managers of major acquisition programs. Together, these two groups are working to operationalize Test Science in active programs.

Future Efforts to Institutionalize Statistical Rigor

Notwithstanding the significant progress that has been made in the past two years, there is still work to be done to utilize the full toolset the scientific community has available to support T&E. I have seen the Service OTAs modify their test design and planning techniques to incorporate DOE and take advantage of the efficiencies afforded by the use of its

methods. Further, I have observed an improvement to the quality of the TEMPs and test plans that are based on these methods. However, there are two areas requiring improvement as the Department's institutionalizes statistical rigor in testing:

- Execution of testing in accordance with the planned test design
- Analysis of test data using the advanced statistical methods commensurate with test designs developed using DOE

For the former, I have seen some cases where a test is well-designed, but the desired conditions of the test in the field are not the same as required by the original plan. This has the effect of limiting the conclusions that can be made from the subsequent data or, at worst, wasting time and resources. Since most of our tests are focused on characterizing the performance of the system across the actual conditions in which the operators will employ the system, it is crucial that the planned conditions are achieved during the test.

For the second area, I have not yet observed all of the OTAs employing the data analysis methods that would reap the benefits of the efficiencies afforded by DOE. In other words, although the OTAs use statistical rigor in their test planning, they are not always following up with the same rigor in their analysis of test data. The simplest case of this is where a test is designed to cover all or many of the important operational conditions, and is optimized to be extremely efficient in the number of test iterations in each condition, but the data analysis is limited to reporting a single average (mean) of the performance across all the test conditions. This result throws away all of the careful test design efficiencies afforded by the use of DOE. A more statistically rigorous analysis would enable all the available information to be extracted from the data, which is critical to evaluating the performance of systems across their full range of operational use. The more advanced statistical analysis also enables statements of system performance to be made with higher confidence in many cases, so that acquisition decisions can be based on certain knowledge rather than supposition.

I will work with the Service OTAs during the next year to rectify these remaining shortfalls in the application of DOE to test execution and analysis.

RELIABILITY ANALYSIS, PLANNING, TRACKING, AND REPORTING

Improving system reliability has been a DOT&E initiative since 2006; the Department has also recognized the significant adverse long-term life cycle cost impacts and reduced operational capability resulting from systems being unreliable. DOT&E initiatives have emphasized the need for reliability growth planning and assessment, establishment of reliability maturity goals and entrance criteria for each phase of testing and documenting the reliability test and evaluation strategy (TES) in the TEMP. Accordingly, the Under Secretary of Defense for Acquisition, Technology, and Logistics (USD(AT&L)) in 2011 released a Directive Type Memorandum (DTM 11-03) on Reliability, Analysis, Planning, Tracking, and Reporting; this DTM was continued into 2012 and will be incorporated into the updated DoDI 5000.02 "Operation of the Defense Acquisition System."

I am tracking the impact of the new directive on system reliability. The Office of the Deputy Assistant Secretary of Defense for Systems Engineering (DASD(SE)) is developing an implementation guide, which is in final staffing and should be available in early 2013. DOT&E has been an ardent advocate for the reliability concepts contained in the directive, and has institutionalized them in our priorities and policies. Figure 2 plots the outcomes of initial operational tests reported to Congress for systems tested between fiscal years 1997 to 2012. A total of 118 reports were included; each report includes an evaluation of operational effectiveness, operational suitability, and reliability.

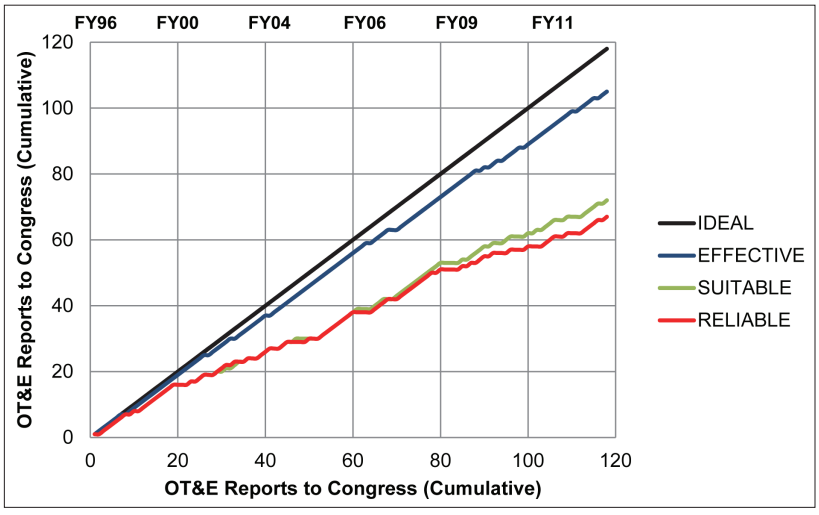

FIGURE 2. CURRENT TRENDS IN RELIABILITY

While evaluations of operational effectiveness and suitability are based on many factors, the evaluations displayed in this chart are based solely on whether the system met its required reliability threshold. As shown in Figure 2, reliability continues to lag; only 7/13 systems (54 percent) evaluated in 2012 met their reliability thresholds and overall between 1997 and 2012 only 67/118 systems (57 percent) were reliable.

To further understand the reliability trends in Figure 2, I surveyed 52 programs for which I approved TEMPs or TESs in FY11 following up on the survey I did in FY10 for all oversight programs. The TEMPs approved in FY11 continue the positive trends I am seeing for all TEMPs approved after June 2008 (when the Department began initiatives to improve reliability). These trends include programs:

- Having an approved System Engineering Plan
- Incorporating reliability as an element of the test strategy
- Having a reliability growth strategy and documenting it in the TEMP
- Incorporating reliability and availability requirements

Unfortunately, the programs reviewed in FY11 did not show improvement in establishing reliability-based milestone or operational test entrance and exit criteria. However, I believe the recent emphasis on reliability has had some demonstrable positive impacts. Having reliability growth curves alone did not correlate with attainment of reliability requirements, but programs with comprehensive reliability plans were more likely to meet their reliability requirements. A larger fraction of programs that establish growth curves with intermediate goals; anchor milestone or entrance/exit criteria to reliability performance; use metrics to ensure reliability growth is on track; predict changes caused by the implementation of corrective actions; and calculate reliability growth potential met their operational test reliability entrance and exit criteria compared to programs that do not follow these practices.

Examining the TEMP survey trends by Service shows that higher percentages of Army and Air Force programs: have added a reliability growth strategy since June 2008; have reliability growth curves; and are calculating the reliability growth potential. Army and Navy programs show increasing improvement in ensuring there is time in the schedule to implement and verify corrective actions and document the reliability test strategy. Army programs are most likely to: use reliability growth curves and intermediate reliability goals; put systems into the hands of representative users before Milestone C; and document reliability changes caused by implementation of corrective actions. Figure 3 shows the fraction of systems meeting reliability thresholds at IOT&E for programs on DOT&E oversight between 1997 and 2012 (the same programs depicted in Figure 2 now broken out by Service).

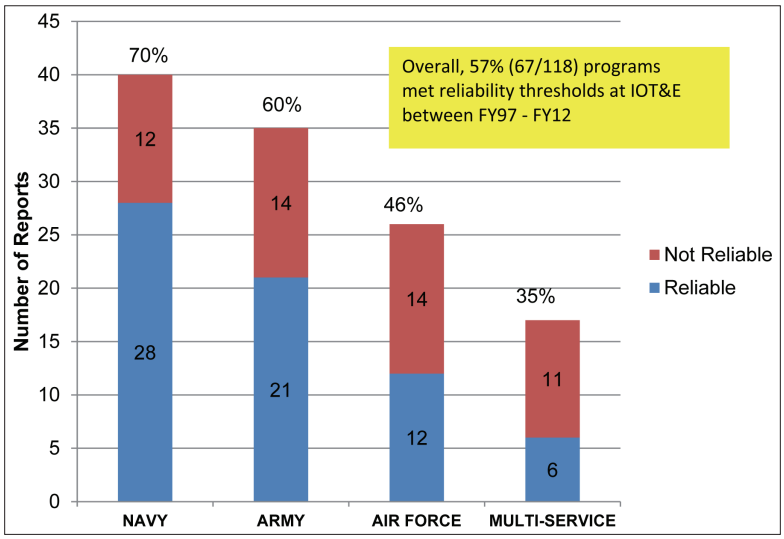

FIGURE 3. FRACTION OF PROGRAMS MEETING RELIABILITY THRESHOLDS AT IOT&E, BY SERVICE (FY97-FY12)

I am not yet seeing more systems actually meet their reliability requirements than in past years, but I believe the recent emphasis on reliability planning has had some demonstrable positive impacts. While the majority of programs now have and are documenting their reliability growth strategy in the TEMP, they are not fully incorporating the design for reliability tenets described in the ANSI/GEIA-STD-0009 Reliability Program Standard for Systems Design, Development, and Manufacturing. In particular, programs are failing to "get on" their planned reliability growth curve at the beginning. I have seen evidence that programs with a procedure for calculating reliability growth potential (a calculation that places emphases on initial reliability, which in turn requires that the system be designed for reliability) have a much greater likelihood of meeting reliability based entry criteria for operational test phases.

Figure 4 shows the distribution of root failure causes for the 51 programs that did not meet their reliability thresholds between 1997 and 2012. The root causes include: 1) inadequate systems management (failures traceable to incorrect interpretation or implementation of requirements, processes, or procedures); implementation of "bad" requirements (missing, inadequate, ambiguous, or conflicting); or failure to provide the resources required to design and build a robust system; 2) inadequate design margins (failures resulting from lack of design robustness to the stresses and loads in usage environment); 3) inadequate software (failures of a system to perform its intended function due to software issues); 4) induced failures (failures resulting from externally applied stresses such as operator or maintainer interfaces); 5) part quality (random failures); and 6) manufacturing anomalies.

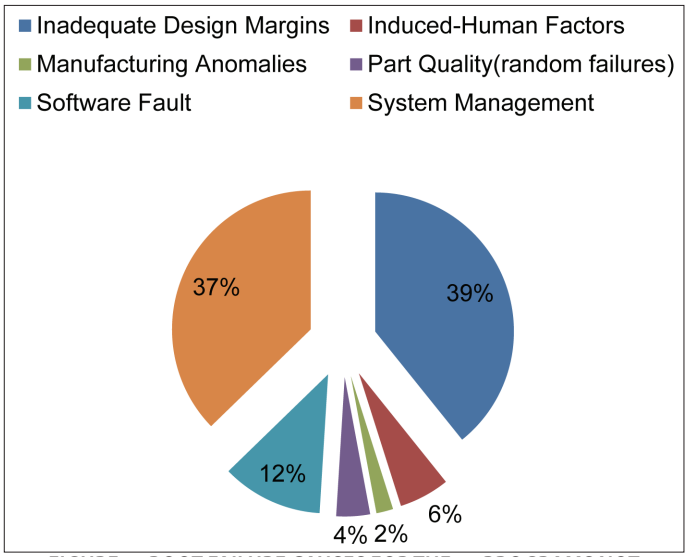

FIGURE 4. ROOT FAILURE CAUSES FOR THE 51 PROGRAMS NOT MEETING RELIABILITY THRESHOLDS BETWEEN FY97 AND FY12

Inadequate design margins and system management combine to account for 76 percent of the root causes for reliability failures in these data. Clearly, inadequate attention to reliability during engineering design, and inadequate management focus on best practices for reliability design and growth testing have been and continue to remain a concern – improvements in these areas, particularly using a Design For Reliability strategy, would help programs get on their planned reliability growth curve and have a greater likelihood of meeting their ultimate reliability goals. Additionally, software reliability design and growth testing are of concern. The 12 percent of systems that failed due to software root causes in Figure 4 are mostly software intensive systems like APG-79 Active Electronically Scanned Array (AESA) Radar (software immaturity causes excessive and inexplicable radar hang-ups; the built-in test function is not automated to isolate software failures); F-15 Mission Planning System (suitability is poor due to software instability, high frequency of system crashes); and Large Aircraft Infrared Countermeasures (LAIRCM) Phase II (software design bugs caused 19 critical failures; bugs were traced to software coding errors).

SOFTWARE TESTING

I continue to see software issues in programs of all types. Most commonly, programs do not create adequate ability to track software reliability and test software patches. Software requirements are poorly stated and in some cases wrongly tested. There are also unique needs for the special class of programs, business systems, which are being developed by the Services to meet the 2014 and 2017 Congressional deadlines for auditability.

Software Reliability

Software reliability is broadly similar to reliability for any weapon system with subtle distinctions in failure definitions, defect tracking, and the speed of the test-fix-test cycle. The overall effect of these distinctions has led me to conclude that new policy is needed that will the mandate the use of some software test automation for most programs that utilize software.

Failure Definition and Defect Tracking: Software is nearly always multi-functional. Software use is not well represented by failure-per-hour metrics. Except in cases where the same operation is performed repeatedly (for example spacecraft during planetary cruise), programs should simply track counts of defects. Defects should always be categorized by severity in accordance with Institute for Electrical and Electronics Engineers standards. Programs should track open and closure rates of the defects in each category. For multi-functional systems, it is helpful to track defects against distinct capabilities as well.

Test-Fix-Test: The test-fix-test cycle for software is faster and less visible than for other systems types. For many software issues, there is no meaningful distinction between maintenance and follow-on development. Given the speed

of software development, the inability to oversee software in detail, and the fact that one must develop code to fix code, the line between fixing defects and adding features is nearly always blurred. Given the pace at which new security patches and product updates and changes in the computing environment occur, there is also essentially no such thing as a stable software system. For all of these reasons, I have concluded that operational testing of software must include a demonstration of the program's ability to perform robust and repeatable testing in support of software maintenance.

In support of robust and repeatable within-program testing, I have begun enforcing the following test automation policies, which will be contained in the next version of the DoDI 5000.02:

- *At Milestone A, program managers shall identify an approach to software test automation, including when key test automation software components or services will be acquired and how those decisions will be made. The test automation approach shall be updated in the Milestone B and Milestone C TEMPs as appropriate.*
- *Program managers shall demonstrate system sustainment maturity at IOT&E. Sustainment maturity shall include routine T&E to support routine technology upgrades. For Information Systems, Defense Business Systems, and software components of Weapons Systems, program managers shall demonstrate mature test automation to include an end-to-end trace of test information from requirements to test scripts to defect tracing*

This year, I recommended the following programs demonstrate this test-fix-test cycle: Next Generation Enterprise Network (NGEN), Integrated Strategic Planning and Analysis Network (ISPAN), Defense Enterprise Accounting and Management System (DEAMS), EProcurement, and Global Combat Support System – Army (GCSS-Army). Because the development of automation tools can be time consuming given the complexity of many of these programs, I anticipate that most programs will take several years to create an automated test-fix-test approach to satisfy these recommendations. Currently, very few acquisition programs have mature test automation solutions for regression testing that can be demonstrated at IOT&E and even fewer can create the environments and conditions to validate their regression testing processes. Without substantial help from a central resource, it is likely that most programs will have this deficiency assessed during IOT&E.

The need for test automation will create demand for the corresponding expertise in program offices. Program managers need a resource in the form of a center of excellence to help meet that demand, and DOT&E is taking the initial steps to establish such a center. The center of excellence would work with vendors and government providers to promote the use of various test automation solutions under the construct of "Test as a Service (TaaS)." A center of excellence will:

- Centralize knowledge of the various automation approaches
- Assist programs in applying test automation
- Create "in-house" test automation expertise

A center of excellence TaaS capability may lessen the tendency of program offices and vendors to use a "stove-piped" approach to test automation, may reduce duplicative resources (technological and human), should increase programs' use of existing capabilities, and should improve the consistency and adequacy in the types of testing accomplished.

Testers do not have questions about system maturity that are distinct from the questions the systems managers should have. System managers should always know how well the system is functioning. If testers have reasonable questions about system performance that the system managers cannot answer with the data they are already gathering, then the system management probably is not as mature as it should be. Examples of performance parameters that should be routinely and continuously reported to the system management include defect tracking, helpdesk use, system productivity/utilization, schedule of upcoming changes (commercial releases, changes in interfacing system, etc.), staff turnover rate, and training and documentation adequacy.

Software Requirements

Software requirements typically involve KPPs for system response time, data loss and restoration, and data transmission accuracy. DOT&E has seen many examples of metrics that incorrectly capture this information. For example, a KPP might specify 95 percent accuracy for information retrieval; but if a random 5 percent of your data is garbled every time you use the system, the utility of that system is very much in question. Some programs include requirements for data loss in event of an outage or other emergency that requires a system restore from backups, and these are almost always expressed as percentages. The data loss requirements should be expressed in time, not percentages. In every case, the system is expected to lose 100 percent of the data that has been entered following the most recent backup interval before the outage. I have seen that testers are dutifully reporting the amount of data loss, but that is not meaningful. Rather, testers should always perform a demonstration test that verifies the ability of the system to backup and restore data on a schedule consistent with the operational need for the system to be available for use.

Finally, some programs have percentage KPPs for data accuracy. These KPPs reference a variety of technical circumstances: transmission across interfaces, retrieval from databases, account balances, and so on. These are often treated as global metrics but they should be treated as percentages that apply to some relevant set of channels. For example, in general, once an interface is correct it is always correct. It is much less important to know that 95 percent of the data transmitted across all interfaces is correct than it is to know which 5 percent of the interfaces are transmitting incorrectly. The metric should not simply be the global number of errors per the number of transmissions. The mission need is for the data elements with errors to be limited. Therefore, the metrics should be looking at counts of element types containing errors. Global metrics also contribute less to finding and fixing problems than would differentiated metrics.

Vulnerability of Business Systems

The HOUSE ARMED SERVICES COMMITTEE PANEL ON DEFENSE FINANCIAL MANAGEMENT AND AUDITABILITY REFORM FINDINGS AND RECOMMENDATIONS (January 24, 2012), Recommendation 4.9 directed DOT&E and others to identify and address shortfalls in workforce levels and corresponding skill sets for Enterprise Resource Programs (ERPs). A clear shortfall in the testing of these systems is in identification of financial vulnerabilities. I have accordingly begun directing that the financial vulnerabilities of ERPs be probed in a manner analogous to Information Assurance, and anticipate that such testing will draw, at least initially, on the existing commercial services that provide such testing. The programs to which this applies are:

- Air Force Integrated Personnel and Pay System (AF-IPPS)
- Defense Agency Initiative (DAI)
- Defense Enterprise Accounting and Management System – Increment 1 (DEAMS – Increment 1)
- Defense Enterprise Accounting and Management System – Air Force (DEAMS – AF)
- EProcurement
- Future Pay and Personnel Management Solution (FPPS – Navy) Pre-MAIS
- General Fund Enterprise Business System (GFEBS)
- Global Combat Support System – Army (GCSS – Army)
- Integrated Personnel and Pay System – Army (Army IPPS)
- Navy Enterprise Resource Planning (ERP)

In support of this initiative, the DCMO has initiated a study of commercial providers of financial Red Team test services. In general, commercial vendors of these services focus on protect and detect capabilities (both system and people). They work with their clients to identify likely targets for fraud or theft within the system; they may attempt (within established rules of engagement) to circumvent controls and processes; and they assess the audit processes that are in place to catch fraud or theft. In addition, together with the Deputy Chief Management Officer, DASD(DT&E), and DASD(SE), we will ensure that developmental and operational testing helps fulfill the Federal Information System Controls Audit Manual requirements.

OTHER AREAS OF INTEREST

Electronic Warfare Test Infrastructure

In February 2012, I identified shortfalls in electronic warfare test resources that prevent adequate developmental and operational testing of many systems, including, but not limited to, the Joint Strike Fighter. I am working to address these shortfalls in government anechoic chambers, open-air ranges, and the Joint Strike Fighter electronic warfare programming laboratory. My staff participated in a "tiger team" assigned by the USD(AT&L) to review the issue, which concurred with my conclusions and recommended additional enhancements.

Cyber Testing

Implementation of the February 2011 Chairman of the Joint Chiefs Execute Order (EXORD), which directed increased cyber-adversary realism for training events, has been modest. During FY12, most of the exercise assessments and tests involved operations largely against low- and mid-level cyber threats and on networks that were only moderately stressed in terms of loading or network degradation. In the cases where the adversary team portrayed higher-level threats, exercise training audiences frequently misinterpreted these portrayals as maintenance issues, poor system performance, or anomalies. This indicates that the Department has not yet developed sufficiently advanced cyber defensive tactics to counter advanced adversary tactics and to consistently operate in degraded cyber environments. Following publication

of the FY11 Annual Report, I provided a separate and classified amplification of findings, which resulted in a series of meetings with the Deputy Secretary of Defense on the topic of how these operational and training shortfalls might be resolved. A number of actions resulting from these discussions are in progress, including the consolidation and enhancement of training support capabilities, additional guidance on meeting the intent and requirements of the EXORD, and improving the way the Department ensures that critical shortfalls are resolved. Additionally, the lessons garnered from operational network assessments are being applied to the acquisition and testing of information systems to ensure that subsequent systems procurement does not contain cyber shortfalls already discovered and documented by the Department. I also remain closely engaged with U.S. Cyber Command and other key stakeholders to ensure priority is given to the necessary investments supporting improved Red Team availability, capability, and accessibility.

Testing Protocols for Personal Protective Equipment

I continue to exercise oversight over the testing of personal protective equipment. The National Academy of Sciences' Committee to Review the Testing of Body Armor published its final report in May of 2012 and I and the Services are pursuing the report's recommendations. Congressional interest in the testing of the Army's Advanced Combat Helmet (ACH) resulted in the Department's Inspector General initiating a technical assessment of the ACH. In response to this Congressional interest in the ACH, we have also asked the NRC to conduct an independent review of the helmet testing protocols. This is a direct follow-up to the NRC's independent review of hard body armor testing, which included a review of test protocols. One of the objectives of the review is to examine the rigor of statistical metrics. My staff will leverage the knowledge of some of the nation's leading statisticians to improve and advance the use of statistical techniques in test. I will also conduct a comprehensive technical assessment of the ACH to characterize its ballistic performance more comprehensively than is possible with existing data. The results of these assessments will provide the basis for any changes to the current helmet test protocols that might be appropriate.

Warrior Injury Assessment Manikin (WIAMan)

I am sponsoring a five-year research and development program to increase the Department's understanding of the cause and nature of injuries incurred in underbody blast combat events and to develop appropriate instrumentation to assess such injuries in testing. This program, known as the Warrior Injury Assessment Manikin, utilizes expertise across multiple commands and disciplines within the Army to generate a medical research plan from which data will, at pre-determined times, be transitioned to the materiel and T&E communities. These data will feed the design of a biofidelic prototype Anthropomorphic Test Device designed to capture occupant loading from the vertical direction, reflecting the primary load axis to which occupants are exposed in an under-vehicle blast event.

Environment and Renewable Energy Effects on Test Ranges

The Department's ranges are experiencing encroachment from infrastructure associated with the electrical energy production and transmission industry. This encroachment can affect test operations as well as systems under test through a variety of means. These include physical obstructions, electromagnetic interference, and thermal effects. The sources of such encroachment include wind turbines, solar power towers, photovoltaic panels, and high voltage bulk power transmission lines. I will continue to cooperate with the Department's Siting Clearing House and the Services to identify potential encroachment of our ranges resulting from renewable energy infrastructure and work to mitigate the impact of such encroachment.

CONCLUSION

Since my first report to you in 2009, we have made significant progress increasing the scientific and statistical rigor of OT&E; we have engaged early with the requirements community to develop realistic, feasible, and testable requirements; we have focused attention on reliability management, design, and growth testing; and we continue to support rapid fielding through flexible and early operational test events. I submit this report, as required by law, summarizing the operational and live fire T&E activities of the Department of Defense during FY12.

J. Michael Gilmore
Director