The

ITEA Journal

December 2011
Volume 32, Number 4

Test as We Fight

Published quarterly
by the International
Test and Evaluation
Association

CONTENTS

DOD & INDUSTRY T&E EXECUTIVES

389 **Key Issues Causing Program Delays in Defense Acquisition** .... *J. Michael Gilmore, Ph.D.*

392 **Test As We Fight** .... *Steve Hutchison, Ph.D. and Luanne Overstreet*

397 **Myths and Truths of T&E** .... *Amy Markowich*

402 **Test as We Fight** .... *Dennis O'Donoghue*

ITEA BEST PAPER

411 **Intelligent Test and Evaluation Infrastructure** .... *Michael L. Curry, Nancy M. Sodano, and Noe B. Duarte*

TECHNICAL PAPERS

419 **The Art and Science of Orchestrating Free-Play in Operational Testing** .... *MAJ Cornelius Leonard Allen, Jr., MAJ Melvin T. Mitchell, and CW4 Kenneth L. Woods*

427 **JITC's DoD Interoperability Communications Exercise (DICE): Joint Interoperability Testing and Certification in a Realistic and Relevant Joint/Combined Tactical and Strategic Environment** .... *Richard Delgado, Jr., Alvin Slarve, and E. N. S. Christopher Brown*

431 **Distributed Integration and Testing: Empire Challenge** .... *MAJ Lee A. Brinker*

439 **Bringing Operational Interoperability Assessment to the Endeavors** .... *MAJ Robert R. Houston*

446 **Modeling to Support Test as We Fight** .... *Jamie L. Pilar, Windsor L. Jones, and Bruce Marihart*

452 **When Testing With Live Agents Is too Risky…Computer Modeling and Simulation May Be a Valuable and Cost-Effective Alternative** .... *Bill Rearick and David Godso*

457 **A Bioinformatics Approach to Finding Candidate Organisms for False Positive Testing of Biological Warfare Agent Detectors** .... *Charlie Holman, Ph.D. and Anika Dodds*

463 **Monte Carlo Techniques for Estimating Power in Aircraft T&E Tests** .... *Todd Remund and William Kitto*

471 **A Retrospective Analysis of Lessons Learned in Evaluating Advanced Military Technologies** .. .. *Craig Schlenoff, Brian A. Weiss, and Michelle Steves*

479 **Multi-Relationship Evaluation Design: Formalizing Evaluation-Design Input and Output Blueprint Elements for Testing Developing Intelligent Systems** .... *Brian A. Weiss and Linda C. Schmidt, Ph.D.*

489 **Reliability Program Scorecard and Reliability Growth Planning and Assessment Models** .... *Nancy Welliver and Marguerite Shepler*

DEPARTMENTS

377 President's Corner

378 Issue at a Glance

381 Guest Editorial: Restore Disciplined Testing & Test Oversight to DoD Acquisition ... . *Jack Krings, Tom P. Christie, and Pete Adolph*

385 Historical Perspectives: Military Wargaming: A Brief History of the Operating Environment .... *COL (R) Wilbur E. Gray, U.S. Army.*

ii Index to Advertisers

388 2012 ITEA Journal Themes

497 T&E News

499 Chapter Directory

510 Corporate Directory

511 Advertising Rates

On The Cover: 'Test as we fight' is a mantra that is underscored by the realities of modern combat and the testing methodologies required today. Modern combat relies now, more than ever, on sophisticated information and communication systems that must work as intended, the first time and every time. Often these systems are being demanded on the battlefield by the warfighter with an immediate need; seldom are there decades-long design, build, and test cycles. This changes some test methodologies, but does not excuse the need for testing–the need to test as we fight. The collage on the cover of our December issue was designed by Ed Holton, Defense Information Systems Agency, Indian Head, Maryland.

CONNECT with ITEA to LEARN, SHARE, and ADVANCE!
■ ITEA Headquarters: 4400 Fair Lakes Court, Suite 104, Fairfax, Virginia 22033-3899; Tel: (703) 631-6220; Fax: (703) 631-6221, E-mail: *itea@itea.org;* Web site: *http://www.itea.org.*
■ For over thirty years the International Test and Evaluation Association (ITEA), a not-for-profit education organization, has been advancing the exchange of technical, programmatic, and acquisition information among the test and evaluation community. ITEA members come together to learn and share with others from industry, government, and academia, who are involved with the development and application of the policies and techniques used to assess effectiveness, reliability, interoperability, and safety of existing, legacy, and future technology-based weapon and non-weapon systems and products throughout their lifecycle. ITEA members embody a broad and diverse set of knowledge, skills, and abilities that span the full spectrum of the test and evaluation profession. All of which is shared with others through the ITEA Journal–the industry's premier technical publication for the professional tester–and at ITEA's Annual Symposium, regional workshops, education courses, and local Chapter events. Join the thousands of ITEA members–your peers in the industry–in contributing to the ITEA Journal and participating at ITEA events so that you also can benefit from the opportunities to learn from others, share your knowledge, and help advance the T&E industry.
■ *The ITEA Journal* (ISSN 1054-0229) is published quarterly by the International Test and Evaluation Association at 4400 Lakes Court, Suite 104, Fairfax, Virginia 22033-3899. Single issue cover price for *The ITEA Journal* is $20. ITEA offers annual membership for individuals, full-time students, and any organization interested in advancing the test and evaluation profession. Annual membership dues are $50 for individuals ($70 if located outside of the U.S.) and $25 for full-time students. Annual Corporate Membership dues are $500 for small organizations (less than 50 employees), and $1,000 for large organizations (50 or more employees). Corporate Memberships include three (3) Individual Memberships for small organizations and five (5) Individual Memberships for large organizations. Annual dues include a one-year subscription to *The ITEA Journal*. The annual subscription rate for libraries and other organizations providing timely reference material to groups is $60. All overseas mail (air mail or AOA) requires an additional $20. *The ITEA Journal* serves its readers as a forum for the presentation and discussion of issues related to test and evaluation. All articles reflect the individual views of the authors and not official points of view adopted by ITEA or the organizations with which the authors are affiliated.

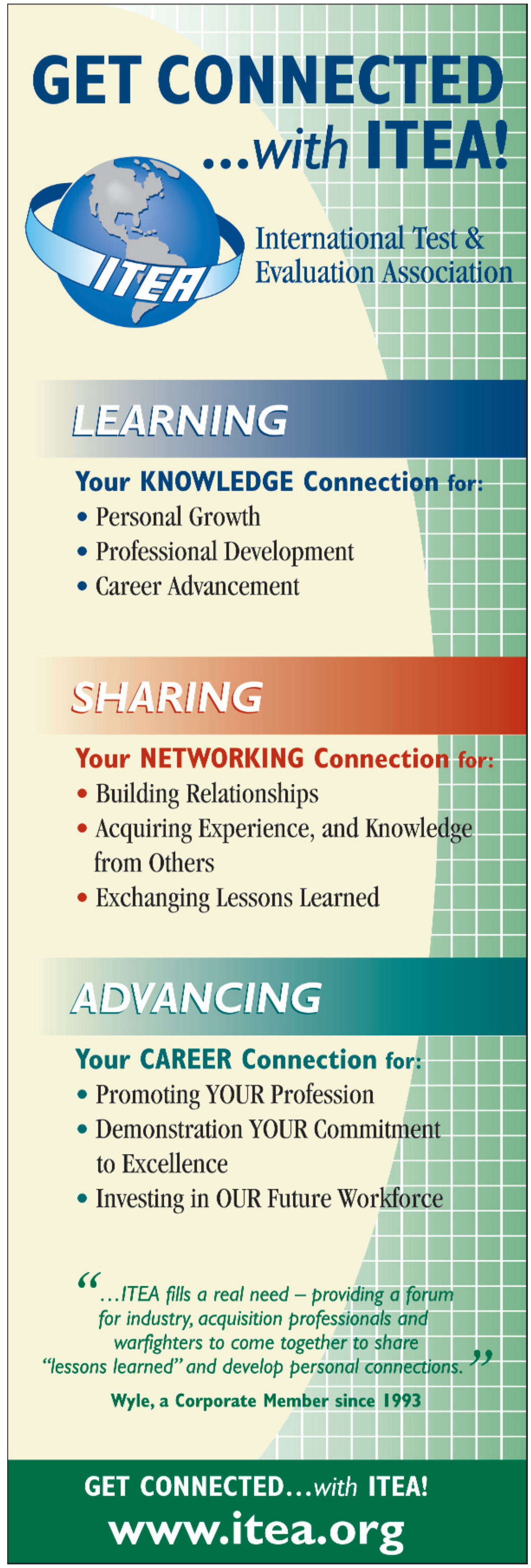
POSTMASTER: Send address changes to: ITEA, 4400 Fair Lakes Court, Suite 104, Fairfax, Virginia 22033-3899.

Index to Advertisers

DRS Training & Control Systems	inside front cover
Calculex	back cover
CSC	fmiii
NTSC International, LLC	fmiv
Wyle	379
JACOBS	380
L-3	380
EWA Government Systems, Inc.	409

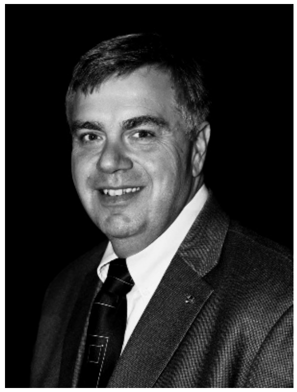

TAKE YOUR CAREER
TO NEW HEIGHTS
CSC
NAVY
CF-01
F-35
100
Find out just how far your talent and innovation can take you. We are a global leader because we keep the best on board. CSC rewards your talent with the resources, professional training and diverse opportunities to grow your career.
We are currently recruiting for immediate openings at
U.S. Naval Air Station Patuxent River:
• Software Engineers • Electrical Engineers • Field Engineers
For more information contact:
Joe Anderson • 301-737-5708 • jandersoniii@csc.com
EOE, M/F/D/V
www.csc.com/nps


The most experienced name in
HIGH SPEED
CAMERAS
Providing quality and reliability since 1958
Introduces the "NEW"
Memrecam HX-1...
- Highest Speeds
- Highest Resolutions
- Longest Record Times
- Highest Quality & Reliability
OF ANY HIGH SPEED CAMERA AVAILABLE ANYWHERE! PERIOD!
Memrecam HX-1
NEW!
nac
Image Technology
The visible difference.
Visit us on the web at
www.nacinc.com (800) 969-2711


GET CONNECTED
...with ITEA!
ITEA
International Test & Evaluation Association
LEARNING
Your KNOWLEDGE Connection for:
• Personal Growth
• Professional Development
• Career Advancement
SHARING
Your NETWORKING Connection for:
• Building Relationships
• Acquiring Experience, and Knowledge from Others
• Exchanging Lessons Learned
ADVANCING
Your CAREER Connection for:
• Promoting YOUR Profession
• Demonstration YOUR Commitment to Excellence
• Investing in OUR Future Workforce
"...ITEA fills a real need – providing a forum for industry, acquisition professionals and warfighters to come together to share "lessons learned" and develop personal connections."
Wyle, a Corporate Member since 1993
GET CONNECTED...with ITEA!
www.itea.org

ITEA Journal 2011; 32: 377

Let me begin my first President's Corner by saying that I am humbled and honored to be the President of the ITEA Board of Directors this year. For those who were at the Annual Symposium in Orlando, you heard that I consider my role as an enabler to help the local chapters as we lead ITEA and the Test and Evaluation (T&E) community through the challenges ahead. A recurring theme of the presentations at this year's Symposium was that we will be operating in a fiscally constrained environment for some time. This has real implications for both the T&E community in general and ITEA specifically. While we have fiscal constraints, we must make the maximum use of the resources we do have. For our community, that is our human capital.



Mark Brown, Ph.D.

This brings me to some questions for our membership. What do you see as the future for ITEA? What do you need ITEA to do for you and your organization? What are we doing right? What do we need to start doing to increase our value to you and your organization? And finally, is there anything we need to eliminate?

We know that ITEA must continue to change in order to remain the preeminent T&E association and to grow in the future. ITEA was founded to address the education of the T&E workforce. I was asked at the Symposium to address what makes ITEA different from other associations. My response centered on our focus on the T&E workforce and training. This is a true differentiator. However, Mr. Duma, Principal Deputy to the Director, Operational Test and Evaluation, noted that it is the local Chapter that makes ITEA special. He is absolutely right.

We focus at the Chapter level on providing training opportunities, whether through Chapter meetings with keynote speakers or tutorials offered to our local community. All Chapters have scholarship dollars available for them to support the education of our future workforce. Some of our Chapters are better at using these funds than others, a challenge that the Board of Directors will be working on this year.

Several of our Chapters also are instrumental in providing the workshops that occur throughout the year. The Events Committee is continuously reviewing these to make sure the workshops remain relevant to our community and to make recommendations on new workshops for Board consideration.

ITEA headquarters has much the same focus on training and education of the T&E workforce. One of the areas we need to expand is on training across the full spectrum of T&E, making sure our workforce understands the processes and procedures needed to support developmental and operational testing. They also need to understand the policies that govern how T&E must be conducted. This includes contractor- and government-led testing. All too often we focus on the materiel solutions, and while these are important, our workforce's ability to plan and conduct tests, as well as analyze test results, is paramount. ITEA, by our very nature, is focused on the human capital involved in T&E. That is exactly where we need to focus in times of fiscal constraints.

In recognition of this need, the ITEA Education Committee recommended, and the Board of Directors approved, establishing a subcommittee focused on workforce development. One of the first activities of this subcommittee is to look at a certified T&E Professional curriculum. This certificate would result in a recognizable certification of the workforce, both government and contractor, similar to the Project Management Professional certification. Watch for more information on this in the near future.

One of the challenges we always face is volunteers. We are a volunteer organization. We only have six paid staff members to support our almost 2,000 members. ITEA exists for our members, but we only function with the support of our members. There are many ways you can get involved. Your local Chapter needs you. Attend your local Chapter meetings. Even better, bring a coworker, employee, or even your boss. Help recruit new members. Volunteer on a local committee or in a leadership position. Help your Chapter host a workshop or tutorial. Write an article for *The ITEA Journal*. All of these things help your local Chapter and ITEA grow.

I encourage you to really consider getting involved with your local Chapter. That is where we engage in addressing the real issues relevant to the local workforce. It is also where we engage the next generation of testers. The local Chapter is where "the rubber hits the road." So I ask again, "What can ITEA do to help you and your organization?" I also ask, "How can you help ITEA address that need?"

I look forward to working with you this year at all levels of the ITEA organization.

ITEA Journal 2011; 32: 378

Test as We Fight. Actual operations aren't merely realistic, they are real. 'Test as we fight' requires every cue that impinges on the senses to be considered, every constraint present, and every threat represented at the same level of fidelity as the friendly force. It entails operators who are trained as we fight and embodies integrated testing and its relationship to the operational environment. Outside of the Department of Defense the counterpart is testing in operationally representative environments, whether for air traffic control, power plant operations, harbormasters in busy commercial ports, or law enforcement.

The *Guest Editorial* is coauthored by three experienced Department of Defense test and evaluation (T&E) professionals, The Honorable John Krings, The Honorable Thomas P. Christie and Mr. Pete Adolph, attesting to the need for enforcing existing directives and guidance as a primary step in addressing shortcomings in the acquisition process. COL (R) Wilbur Gray describes the origin, history and use of military wargaming in *Historical Perspectives*.

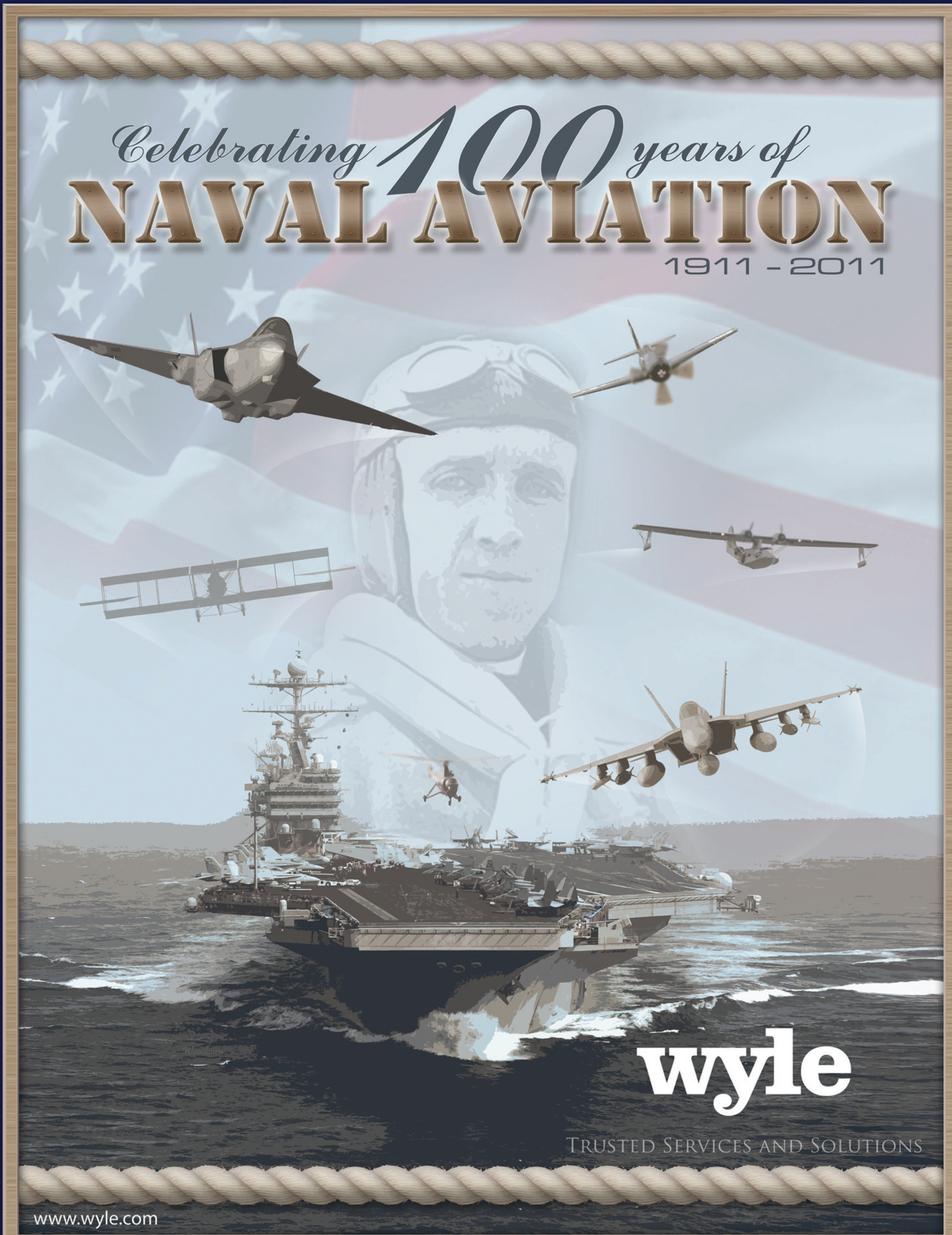
December sees a new special section featuring perspectives from three of the Department of Defense and one industry T&E Executive. The Honorable Dr. J. Michael Gilmore, Director of Operational Test and Evaluation, summarizes two recent studies that examined acquisition program delays, concluding that none of the reasons cited included T&E. Dr. Steven Hutchison, Defense Information Systems Agency (DISA) T&E Executive, describes the need for designing a test to determine if the product does what the user expects it to do, and the DISA lessons captured from the information technology sector to help craft an improved approach to testing. Ms. Amy Markowich, Senior Executive for Department of the Navy Test and Evaluation, "busts" some of the myths of T&E to dispel commonly held false beliefs and reveal kernels of truth. Finally, Mr. Dennis O'Donoghue, Vice President of Boeing Test & Evaluation, explores test as we fight as a cultural shift in test philosophy necessitated by complex system-of-systems testing.

In the 2011 ITEA Technology Review Best Paper, Mr. Michael Curry, Mr. Noe Duarte and Ms. Nancy Sodano present a T&E infrastructure for data-driven decision-making and development of an integrated, end-to-end T&E strategy.

The contributed articles open with a need for creating a more relevant operational environment by crafting realism and enabling free-play in operational testing, by MAJ Cornelius Allen, Jr and his coauthors. Three articles from the Joint Interoperability Test Command (JITC) directly speak to test as we fight. Mr. Richard Delgado, Jr *et al.* explain JITC automated test tools and procedures and their value in providing rigor to the testing process and raising the confidence level of the test results. MAJ Lee Brinker discusses the importance of distributed integration and testing in realizing test as we fight. MAJ Robert Houston reviews multi-national testing for assessing operational interoperability of various command, control, communications, and computer systems.

Ms. Jamie Pilar and co-workers introduce modeling and simulation tools that are flexible and network intensive to fully investigate the value-added of systems during evaluation. Mr. Bill Rearick demonstrates how computer modeling and simulation has the ability to elevate testing and training to a whole new level of realism and fidelity. Dr. Charlie Holman and Ms. Anika Dodds illustrate the use of bioinformatics tools for improving the accuracy of biological warfare agent detection systems. Mr. Todd Remund and Mr. William Kitto apply Monte Carlo techniques for achieving statistical rigor in aircraft T&E. Mr. Craig Schlenoff *et al.* review lessons learned in evaluating military systems, citing decisions made during the evaluation design stage as possibly the most critical. Mr. Brian Weiss and Prof. Linda Schmidt describe a methodology for automating evaluation design and test plan generation for complex systems. The issue closes with an article by Ms. Nancy Welliver and Ms. Marguerite Shepler that describes two methods for meeting system reliability requirements and reducing life cycle support costs for major weapon systems.

Finally, the Publications Committee salutes one of its own, Dr. Mark Brown, newly elected President of ITEA. Congratulations Mark. Once your tenure is complete we will welcome you back into the fold.

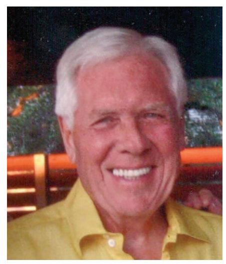

Celebrating 100 years of
NAVAL AVIATION
1911 - 2011
wyle
TRUSTED SERVICES AND SOLUTIONS
www.wyle.com

ITEA Journal 2011; 32: 381–384

Restore Disciplined Testing & Test Oversight to DoD Acquisition

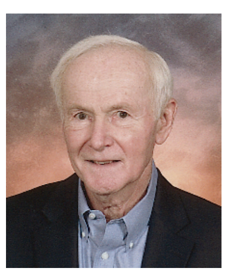
Jack Krings
Austin, Texas
Tom P. Christie
Vienna, Virginia
Pete Adolph
Albuquerque, New Mexico

Most people would agree that there are serious problems with today's Department of Defense (DoD) acquisition process. The shortcomings have been addressed in numerous studies, but they all boil down to a lack of discipline in implementing and enforcing existing directives and guidance throughout the product definition, technology development, system development, and test process. A lack of robust testing during early development is one, but by no means the only, major shortcoming for most current developmental programs. There is ample evidence to show that the management of developmental programs, with a few notable exceptions, has deteriorated markedly in the last 2 decades. Numerous indicators exist, including significant increases in developmental timelines; cost overruns unprecedented in magnitude and frequency, often leading to Nunn-McCurdy breaches; and dramatic increases in suitability failure rates. Additionally, because of the lack of discipline in adhering to an event-based strategy, large production increments are increasingly funded prior to initial operational test and evaluation (IOT&E), or worse, adequate developmental test and evaluation (DT&E).



Jack Krings

Tom P. Christie



Pete Adolph

Although the developmental programs of the 1970s, 1980s, and early 1990s were not without problems, there are numerous examples of well-managed programs as compared with recent ones. For example, highly successful aircraft development programs include the F-15, F-16, F-18, A-10, and F-117. Many factors have contributed to the recent degradation. They include poor requirements definition, inadequate attention to technology readiness for critical subsystems, changes made as a result of the implementation of some acquisition reform initiatives (e.g., the elimination of numerous specifications and other guidance documents), and the decimation of the government technical workforce. In effect, in the last decade or two, DoD and industry threw out the acquisition playbook, and then proceeded to get rid of the majority of their most experienced players.

Some changes implemented in the last 15 years were aimed at eliminating what was perceived by some as bureaucratic impediments to the acquisition process that resulted, in their view, from excessive government oversight. There have certainly been some instances of overly stifling government oversight contributing to acquisition problems. However, even in these cases, that oversight was clearly not the cause of major program slippages and cost growth. In the late 1990s, the government T&E community became the primary target for cutbacks—one person observed that the government strategy changed from oversight to out-of-sight. Many Service test activities reverted to what was described as a rent-a-range mode of operation, and the developmental test oversight organization in the Office of the Secretary of Defense (OSD) was disestablished. With the advent of total system performance responsibility (TSPR), combined with workforce reductions, government T&E involvement and/or oversight was often eliminated.

Numerous studies have shown that a lack of attention to reliability during development is the most significant sustainment cost driver.

As a result of actions taken in the last 2 decades, with some exceptions, neither the Service T&E communities nor the Office of the Undersecretary for Acquisition, Technology and Logistics in the OSD (OSD/AT&L) have been performing effective developmental test oversight. Many of the problems that occur as a normal part of the developmental process could have been detected and remedied much earlier had Service and OSD leadership been informed and acted in a timely manner. As a consequence, with few exceptions, programs were permitted to advance into production and even deployment with the expectation (often unrealized) that procurement funds would be used to correct the problems. Furthermore, delay of reliability, availability, and maintainability (RAM) fixes resulted in expensive retrofits and increased life cycle costs. Numerous studies have shown that a lack of attention to reliability during development is the most significant sustainment cost driver. Continued reliance on industry, without adequate Service and OSD oversight of developmental testing, will only prolong many of the systemic shortcomings of the acquisition process that contribute to cost overruns, schedule delays, poor performance and reliability, and ultimately failure to meet the needs of the warfighter.

Our focus in this discussion is on the role of T&E, and what needs to be done in the test arena to remedy some of the current problems. While not a cure-all, test process improvements are essential and will go a long way towards correcting numerous problems that currently exist. Government DT&E involvement from the outset of an acquisition program provides opportunities to shape the developmental test program through participation in program planning. When the program enters the test execution phase, technical insights from experienced test personnel combined with objective evaluation of developmental progress are critical to making more informed decisions during acquisition program reviews.

An examination of problems encountered by numerous major programs illustrates the need for objective, independent evaluation of risk, performance, and cost throughout the development process. For example, a strong, independent DT&E voice may have delayed the F-22's entry into low rate initial production (LRIP) until critical developmental issues had been resolved, thus avoiding the expensive and time-consuming retrofits required to correct deficiencies discovered after that entry. A competent developmental test organization would have insisted that the V-22 complete the full flight control system development and flying qualities flight envelope expansion before entering IOT&E. These tests are an indispensible element of a disciplined developmental test effort on any aircraft. In the original Terminal High Altitude Aerial Defense (THAAD) program, design stress/qualification and system integration testing were waived. The THAAD program was characterized as "a rush to failure" by an independent senior oversight group. OSD belatedly shut the program down, until a more disciplined approach to development was taken. A few of many other recent programs with serious test process/reporting deficiencies include the Space Based Infrared System (SBIRS) Increment 1, the Airborne Laser Program, the Advanced Seal Delivery System, and the Expeditionary Fighting Vehicle.

When an acquisition program runs into trouble and there are developmental delays, there is often a tendency for program managers to "shoot the messenger." As noted above, in recent years, there have been far too many programs with developmental delays, many of which could have been avoided through more attention to requirements, technology readiness, and a robust developmental test program in the early stages of development. Nonetheless, the test community, particularly the Office of the Director, Operational Test and Evaluation (DOT&E) in OSD, remains an easy diversionary target for program management problems, where acquisition officials and program managers blame testers for "inventing" requirements, and causing cost growth and schedule delays.

This assertion was made once again in 2010. As a result, an independent team was chartered by the Defense Acquisition Executive (DAE) to "assess concerns that DoD's developmental and operational test communities approach to testing drives undue requirements, excessive cost, and added schedule into programs and results in a state of tension between Program Offices and the Testing Community." The DAE assessment team, led by retired U.S. Navy Admiral Kathleen Paige, looked at 37 programs and "found no significant evidence that the testing community typically drives unplanned requirements, cost or schedule into programs."

Over the last 30 years, DoD and Congress have struggled with the issue of providing objective and effective high-level oversight to acquisition programs during development. In the early 1980s, Congress legislated a new OSD testing oversight organization, the DOT&E, in an attempt to overcome the ongoing problem of fielding weapons systems with serious problems. The DOT&E was given the authority, independence, and related objectivity to oversee, approve, and report on the progress of developmental testing; assess readiness for IOT&E; and report on the adequacy of operational testing, and test results. The DOT&E provided Congress, the Secretary of Defense, and Service acquisition executives with knowledgeable and objective assessments of the acquisition progress throughout development and IOT&E. Initially the Director, Development Test and Evaluation (DDT&E) office in OSD collaborated with DOT&E to oversee and report on the progress of developmental testing in support of program reviews.

Failure to identify and admit to technical issues and solutions, as well as real costs, early in the program cycle is the overwhelming cause for subsequent cost increases and schedule delays.

The DDT&E office went through a series of downgrades and eventual elimination in the late 1990s. After the demise of DDT&E in 1998; DOT&E, whose primary focus was on overseeing the planning for and conduct of operational testing, became the only source of objective information regarding the progress of developmental testing. The problems with a lack of objective oversight of developmental test progress throughout the increasingly long developmental cycle and the resulting major shortcomings are documented for several programs listed previously in this article. The Weapons Systems Acquisition Reform Act (WSARA) of 2009 included an attempt to reinstitute objective oversight of developmental testing by reestablishing the DDT&E function. Prior to passage, the draft legislation was fought by OSD/AT&L, and the DDT&E has not been supported with adequate resourcing or visibility since WSARA passage. To date, no adequately staffed developmental testing function with an independent voice and the requisite staffing to effectively report on weapons systems development exists within AT&L. This fact is supported by ongoing Government Accountability Office (GAO) studies. The primary source in DoD for objective information regarding the measurement of acquisition progress is still DOT&E, the de facto integrated OSD T&E oversight organization.

To optimize the efficiency and effectiveness of DoD system testing, consolidating oversight under DOT&E would insure integrated, unfiltered, more effective oversight through development.

Because of the unprecedented federal deficit created in the last five years, huge cuts to the DoD budget are inevitable. Testing will be a prime target for reductions. This provides an opportunity for the T&E community to take an introspective look and focus on new and more efficient ways of doing business as well as mitigating looming budgetary cuts. There should be two major components of any T&E-related study: 1) test processes and 2) test support facilities. Components of the test process review should include an in-depth look at test requirements and at ways to improve test process efficiencies and reduce cycle time. The test facility component should include an objective assessment of private/public/hybrid ownership and operation of component test facilities; and a framework for consolidation, mothballing, and closure. More flexible methods of providing contractor and government test support personnel should also be addressed, e.g., ways to ramp up to meet short-term surge requirements while maintaining a smaller core cadre of full-time expertise.

An assessment should also be made of test responsibilities at the OSD level. Since OSD does no actual testing, the core goal should be objective, unfiltered oversight and reporting in support of program reviews throughout the acquisition process. This assessment should include a goal of transitioning to a leaner and more effective OSD test organization. To optimize the efficiency and effectiveness of DoD system testing, consolidating oversight of developmental and operational oversight under DOT&E would insure integrated, unfiltered, more effective oversight throughout development. Consolidation would provide improved OSD decision-maker information, beginning with the Request for Proposal (RFP) and continuing through development and production, including significant life cycle upgrades.

In conclusion, the corrective actions required to improve the system development/test process are obvious. Failure to identify and admit to technical issues and solutions, as well as real costs, early in the program cycle is the overwhelming cause for subsequent cost increases and schedule delays. Robust testing early in development and objective assessment of test progress are absolutely essential, to allow for early identification and correction of design deficiencies. It remains to be seen if the current leadership in the Services and the Office of the Secretary of Defense will take the actions necessary to restore the disciplined acquisition process that the warfighters and taxpayers deserve. ❑

JACK KRINGS has been on all sides of the defense acquisition process. He has over 12 years experience as a U.S. Air Force and Air National Guard operational fighter pilot, is a graduate of the U.S. Navy Test Pilot School, worked nearly 30 years with McDonnell as an experimental test pilot, flew the first flight of the F-18 Hornet, and served as Director of Flight Operations and Director of Navy Programs before retiring in 1985. President Reagan appointed Jack as the first Director of OT&E for the DoD, working for both the Secretary of Defense and the Congressional Armed Services Committee. As an Assistant Secretary of Defense for over 4 years, he was a key decision maker in the DoD acquisition process, advisor to the Armed Services Committee, and the policy maker for T&E within DoD. Clients of Jack's consulting firm, Krings Corporation, in Austin, Texas, include key executives in the DoD, most of the prime defense contractors, the Federal Aviation Administration, and NASA. Since 1989, he has provided advice ranging from acquisition strategy, corporate reengineering, and technical systems development to T&E integration and mission assurance. With the transformation of national defense, Jack has been continuously involved as advisor to the Missile Defense Agency Director, the Joint Strike Fighter Program Manager, and the Directors of OT&E and DT&E.

TOM CHRISTIE is a consultant on defense weapons systems and test. From 2001 to 2005, he was the Director for OT&E, a presidential appointee confirmed by the Senate, and the senior policy advisor to the Secretary of Defense on weapon systems test and evaluation. Mr. Christie began his government career in 1955 at the Air Proving Ground Center, Eglin Air Force Base, Florida, and held numerous leadership positions in the Air Force, the OSD, and following retirement from government, the Institute for Defense Analysis. Mr. Christie holds the Presidential Rank, Distinguished Executive Award; the Department of Defense Distinguished Civilian Service Award; and the Air Force Scientific Achievement Award. He graduated from Spring Hill College in 1955 with a bachelor's degree in mathematics, and from the New York University in 1962 with a master's degree in applied mathematics.

CHARLES "PETE" ADOLPH is currently a consultant. He started his career with Convair as a test engineer in 1956. He then spent 3 years as an Air Force officer in flight test at the Air Force Flight Test Center (AFFTC), followed by over 25 years in numerous civilian positions at the AFFTC, advancing to technical director, the senior civilian. From 1987 to 1994, he held several positions in the OSD. For most of that period he was Director of T&E in the Office of the Undersecretary of Defense for Acquisition & Technology. He also served as Acting Director of OT&E and Acting Director of Defense Research and Engineering. He was a senior vice president for Science Applications International Corp. (SAIC) from 1994 to 2000, managing the Test and Evaluation Group. He graduated with a bachelor of science degree in aeronautical engineering from St. Louis University; earned a master of science degree in aeronautical & astronautical engineering from the University of Michigan and a master of science degree in systems management from the University of Southern California; and was a Sloan Fellow at the Stanford Graduate School of Business. E-mail: ccadolph@aol.com

ITEA Journal 2011; 32: 385–387

Military Wargaming: A Brief History of the Operating Environment

COL (R) Wilbur E. Gray, U.S. Army

The need for meaningful and effective testing drives our community to generate test plans and scenarios that represent, as close as possible, real-world operating environments for today's complex systems. Whether evaluating new systems, technology approaches, or developing innovative tactics, techniques, and procedures, the validity and reliability of test outcomes and recommendations build on the foundation of a credible simulated operating environment. An appreciation for these fundamental concerns and desire to "test as we fight" is not unique to our times. In this historical feature article, we explore the history of military wargaming as a consistent approach for achieving this goal.[1]

We now know that military wargames actually evolved from games played principally for fun. The first of these was *Wei-Hai* ("encirclement"), a Chinese game that is usually now called "Go." A later, similar game was the Indian *Chaturanga*, the system from which chess in its various forms came about. Chess itself gave birth to at least one game that more formally depicted armed combat: the 1644 design known as "The King's Game" from one Christopher Weikmann. It included 30 pieces per side of 14 military types, each with a different fixed rate of movement. Like its predecessors, it was played principally for pleasure but differed by its emphasis on the strategic level of war.

The first game to break away from chess, however, was invented by Helwig, Master of Pages to the Duke of Brunswick, in 1780. This game included 1,666 squares, each coded for a different rate of movement depending on the terrain the square represented. Playing pieces now represented groups of men instead of a single soldier, and each unit was rated for different movement (e.g., infantry moved eight spaces, heavy cavalry 12). There were also special rules for such things as pontooneers and the like. In 1795, Georg Vinturinus, a military writer from Schleswig, produced a more complex version of Helwig's game. He modified it in 1798 by using a mapboard that depicted actual terrain on the border between France and Belgium.

Nevertheless, such innovations did not move wargames out of the entertainment world into that of the military until 1811, when a Prussian father–son team began to make their studies known. The father, Baron von Reisswitz, was a civilian war counselor to the Prussian court at Breslau. During the dark days of Prussian domination by Napoleon, Reisswitz introduced a game that used a specific scale (1 : 2,373) and a sand table instead of a map grid. In 1811, the game was observed by two Prussian princes who then showed it to the king. The game immediately became the rage at both the Prussian and Russian courts, but professional soldiers saw little use for it. All that changed in 1824. In that year, Reisswitz's son, Leutnant George Heinrich Rudolf Johann von Reisswitz of the Prussian Guard Artillery, introduced his own version of his father's game. The game was called *Anleitung zur Darstelling Militarische Manuver mit dem Apparat des Kriegsspiels*[2] and included a number of new innovations, the most important of which were the use of actual topographical maps to portray the battlefield and rigid rules that specifically quantified the effects of combat.

The rules were published under the patronage of Prussian Prince Wilhem who became impressed with them after an evening's play. The Prince then recommended the rules to the Chief of the Prussian General Staff, General von Muffling, who finally granted von Reisswitz an audience. One of von Reisswitz's companions, a young officer named Dannhauer, describes the meeting that many believe to be the birth of the military wargame:

> *On our arrival we found the General surrounded by the General Staff officers.*
> *"Gentlemen," the General announced, "Herr von Reisswitz is going to show us something new."*
> *Reisswitz was not abashed by the somewhat lukewarm introduction. He calmly set out his Kriegsspiel map.*

With some surprise the General said, "You mean we are to play for an hour on a map! Very well. Show us a division with the troops."
"May I ask your excellency," replied Reisswitz, "to provide us with general and special ideas for manoeuver, and to choose two officers to be the commanders for both sides. Also it is important that we only give each commander in the special idea the information he would have in reality."
The General seemed rather astonished at the whole thing, but began to write out the necessary idea.
We were allocated as troop leaders to both sides, and the game began. One can honestly say that the old gentleman, so cool towards the idea at the beginning, became more and more interested as the game went on, until he exclaimed, "This is not a game! This is training for war! I must recommend it to the whole army."[2]

The impact of this first military wargame had been significant. Reisswitz's work particularly impressed one Lieutenant Helmuth von Moltke who, in 1828, founded a wargame club called the *Kriegspieler Verein,* which soon began to publish its own periodical. This kept interest in wargames alive, and when von Moltke became Chief of Staff in 1837, he officially pushed wargaming from the top. His influence had the desired effect, and by 1876 another set of German wargame rules was published, this time by Colonel Julius Adrian Friedrich Wilhelm von Verdy du Vernois. Vernois's system was a "free" *Kriegsspiel* as opposed to Reisswitz's rigid variety. This meant that most calculations and die rolling was eliminated in favor of an umpire, who would determine results based on the situation and his own combat experience. Whether "free" or "rigid," however, wargames had become a mainstay of German military training.

Other countries around the world became interested in German wargaming because of the 1870–1871 Franco–Prussian War. In this conflict, the militia and reserve-based armies of Prussia decisively defeated the totally professional army of France, which was then thought to have the finest soldiers in the world. Many believed that wargames, in part, were used to successfully compensate for Prussia's reliance on an army of *Reserven und Landwehren* (i.e., reserve and militia forces).

From that point on, all countries began to build imitations of German systems as well as develop their own. In the United States, Army Major William R. Livermore introduced his "*The American Kriegsspiel*, *A Game for Practicing the Art of War on a Topographical Map*" in 1882. The game was complex and similar to Reisswitz's system but did attempt to cut down on the paperwork involved by the introduction of several training aid–type devices. At the same time, Lieutenant Charles A. L. Totten introduced a game entitled *Strategos: A Series of American Games of War*. Totten's game was as complex as Livermore's, but he appealed to the amateur through the inclusion of a simplified basic set of rules.

Neither was wargaming neglected by the U.S. Navy, thanks to the efforts of William McCarty Little. In 1876, after an accident had forced his retirement from the Navy, Little made his home in Newport, Rhode Island, and assisted in the establishment of the Naval War College. At the same time he made the acquaintance of Major Livermore, who at that time was stationed across the bay at Fort Adams. Under Livermore's influence, and with the help of some very open-minded supervisors like Captain Henry Taylor, president of the Naval War College, Little was able to make wargaming an integral part of the College's curriculum. His efforts practically made the Naval War College into America's unofficial wargaming center. Little produced a ship-on-ship game, a tactical game, and a strategic game; all very accurate (they were able to predict that smaller numbers of big guns on battleships were more effective than large numbers of mixed-caliber weapons) but also very complex. It was, in fact, complexity that encouraged resistance to wargaming within the U.S. Army and elsewhere. Games like Vernois's were introduced to simplify things, but many argued that such umpire-driven systems only replaced arbitrary written rules with arbitrary unwritten rules. Thus, by the turn of the century, there was an increased tendency all over the world to merge the free *Kriegsspiel* with the rigid to produce a semi-rigid system. Even Livermore accepted this as the best solution and often ignored his own tables as much as he consulted them.

The semi-rigid wargame thus became the standard for most military conflict simulations around the world through the First World War. The games proved quite successful, and history abounds with examples of how commanders were defeated as a result of ignoring the result of a wargame. As an example, a Russian wargame in 1914 predicted defeat if General Samsomov's 2d Army did not begin its advance 3 days ahead of General Rennenkampf's 1st Army, "an action not contained in the plans. This change, so clearly indicated in the wargames, was never made in the plans or their execution." The result was the Russian debacle of Tannenburg the same year.

The years between the world wars were notable for the lack of military wargaming activity, particularly in Britain and the United States. In general, most wanted to forget the carnage of the Great War; in addition, not a few noted that the failure of Germany's vaunted Schlieffen Plan in 1914 showed that the wargame was far from perfect. There were exceptions to this general rule of inactivity, of course. Germany still relied on the wargame as a principal training tool, especially since the Treaty of Versailles denied that country the right to field the necessary army appropriate for large-scale training exercises. One must also look to the contribution of F. W. Lanchester who introduced a mathematical formula that predicted attrition rates between two equivalent armies in combat.

In modified form, his two equations are still the basis of many wargames today. Finally, one must note that the Naval War College, in seeming defiance of the other branches of Service, continued and expanded its wargaming efforts. The College's labors were to bear great fruits during the upcoming war against the Axis Powers. Indeed, history records many wargame successes during World War II, but perhaps none was more impressive than America's naval victory over Japan. Our wartime Pacific commander, Fleet Admiral Chester W. Nimitz, explained to a Naval War College class in 1960 that, "the war with Japan had been reenacted in the game rooms here by so many people and in so many different ways that nothing happened during the war that was a surprise—absolutely nothing except the kamikazis towards the end of the war."

From that point on, military wargames followed advances in technology resulting in the complex pilot simulators and computerized strategic systems used around the world today in the armed forces of most advanced countries. Indeed, with the introduction of the U.S. Army's Combat Training Centers, such as Fort Polk, Louisiana, or the National Training Center at Fort Irwin, California, the individual soldier has now become a playing piece. Admittedly, events such as the Vietnam War have shown that wargames are not perfect; they are only as good as the data entered into them by humans. Nevertheless, the history of military wargames is such that most failures seem to occur when the results of a wargame are ignored, not when they are taken seriously. This is a solid record by any measure. ❑

Colonel Bill Gray is a retired Army officer with an exclusive background in Military Intelligence. He holds baccalaureate degrees in both history and political science from Clemson University, as well as graduate degrees in international relations from the University of Southern California and strategic intelligence from the Defense Intelligence Agency's (DIA) Joint Military Intelligence College. His military education includes the DIA Post-graduate Intelligence Program, the U.S. Army Command and General Staff College, and the U.S. Air Force Air War College. He remains active in hobby wargaming and has been a regular contributor to periodicals such as Strategy & Tactics, Napoleon, Wargames Illustrated, *and* Wargamer.com. *He is an officer of both the U.S.-based Historical Miniatures Gaming Society and the United Kingdom–based Pike & Shot Society. He resides with his wife, Paula, in Enola, Pennsylvania, where he works as the Institutional Research Director for Central Penn College. E-mail: hmgs1@hotmail.com*

Endnotes

[1]James Welshans, Ph.D., *The ITEA Journal* Historian. Used with permission by the author, Col. (ret.) Wilbur Gray. Adapted from the author's master's thesis *Playing War: The Applicability of Commercial Conflict Simulations to Military Intelligence Training and Education* (DIA Joint Military Intelligence College, Bolling Air Force Base, Washington, D.C., 1995).

[2]Instructions for the representation of tactical maneuvers under the guise of a wargame.

2012 *ITEA Journal* Themes

The ITEA Publications Committee has established themes for the 2012 issues of *The ITEA Journal* and invites articles in the following areas:

T&E at the Speed of Need (March issue). The speed of need is the time between definition of a user need and initial operation of the capability. The demands of war have shortened the timeline on requirements for military systems. Rapid acquisition and rapid fielding initiatives arising from urgent operational needs have created an entire industry in the defense community. Commensurately hastened is the pace at which transportation security and border protection measures need to be deployed. We see information technology (IT), especially software, change with a frequency of months not years. Test and evaluation (T&E) must be responsive to the acquisition timelines. The Federal Aviation Administration, Border Patrol, law enforcement and many other organizations have adapted to the changing speed of need. This issue takes a candid look at agile software development processes, cyberspace T&E, defense IT acquisition reform, rapid acquisition and fielding, reconfigurable test capability, testing on demand, reuse and other ideas for streamlining the T&E process in support of accelerating deployment of new products, services and capabilities. ***(Manuscript deadline: December 1, 2011)***

Drowning in Data, Thirsty for Information (June issue). Digital technology and Moore's law provide us with the ability to acquire, create, and store data at unprecedented rates and volumes. Literature searches that would have taken days or weeks in the age of library card catalogs can be accomplished in seconds over the internet. Data, or more generally information, has become big business in addition to being the business of T&E. Yet technological solutions also come with their own problems and the morass of data has not led to a commensurate growth of knowledge or ability to exploit the data. This issue examines the plethora of data, which is growing exponentially, and the ever critical demand to extract meaning and value. Articles are invited on such topics as data acquisition, storage, archiving, access, validation, exploitation, and visualization; data as a service; cloud computing; service oriented architecture; metadata syntax and semantics; instrumentation; accelerating the process from acquiring data to implementing a decision; data and sensor fusion; data preservation; distributed and non-relational databases, and related topics. ***(Manuscript deadline March 1, 2012)***

Strategic Partnering: We are Doing More Without More (September issue). Shrinking budgets and accelerating technology development put ever-increasing pressure on product development and T&E organizations: shorter schedules, fewer personnel, less access to facilities. Extreme environments and complex systems add additional demands. To provide the requisite test capabilities, the T&E community must be agile and responsive as well as innovative. The automobile and aircraft industries have long spread liability and gained benefit from seeking standard parts production from independent manufacturers or specific systems development parceled to risk sharing partners. Strategic partnering takes such forms as outsourcing, reuse, and collaboration; common test and training infrastructure; integrating developmental and operational testing; shared facilities among government, industry, and academia. Cooperation comes with its own issues: some loss of control, policy or statutory impediments, protection of proprietary rights, and conflicting goals of leadership. This issue addresses all forms of partnering, allowing testers to do more without having more resources to accomplish their task. ***(Manuscript deadline: June 1, 2012)***

Cultivating the T&E Workforce (December issue). Test and evaluation are professions not academic disciplines and as such we can't merely recruit more as needed. We recruit engineers, physicists, computer scientists, mathematicians, chemists and other degreed professionals and train them in test and evaluation. As technology changes and systems and instrumentation become more complex, T&E professionals need to continue formal education as well as improve T&E expertise. In addition, we need to consistently attract young people to the disciplines of science, technology, engineering, and mathematics. Cultivating the T&E workforce requires asking the question: what should the T&E professional's background consist of today? The internet gives us nearly immediate access to the four W's – who, what when and where – but can't provide why and how, which require human reasoning. Appealing to younger professionals means communicating with them in their preferred mode, such as social networking media, and recognizing that T&E will evolve as IT drives us in new directions and as younger leaders assume their roles in legacy organizations. We need to prepare the future workforce for T&E, and prepare T&E for them. This issue solicits ideas on improving the current workforce and growing the next generation: the use and benefit of certification, such as T&E, modeling and simulation, project management; Science-Technology-Engineering-Mathematics (STEM) initiatives; educating leadership; the role of the service academies in DoD; internships; and many more topics. ***(Manuscript deadline: September 1, 2012)***

Articles and Submission: T&E articles of general interest to ITEA members and *ITEA Journal* readers are always welcome. Each issue includes specialty features, each 2-3 pages long: **"Featured Capability"** describes unique, innovative capabilities and demonstrates how they support T&E; **"Historical Perspectives"** recall how T&E was performed in the past or a significant test or achievement, often based on personal participation in the "old days" of T&E; **"Tech Notes"** discusses innovative technology that has potential payoff in T&E applications or could have an impact on how T&E is conducted in the future. **Interested authors:** Submit contributions to the ITEA Publications Committee Chairman (itea@itea.org, attention: Dr. J. Michael Barton). Detailed manuscript guidelines can be found at **www.itea.org** under the ITEA Publications tab.

ITEA Journal 2011; 32: 389–391

Key Issues Causing Program Delays in Defense Acquisition

J. Michael Gilmore, Ph.D.
Director, Operational Test and Evaluation, Office of the Secretary of Defense, The Pentagon, Washington, D.C.

I recently released a memorandum with the Under Secretary of Defense for Acquisition, Technology and Logistics (USD[AT&L]) summarizing two independent assessments of the key issues causing program delays. The USD(AT&L) had chartered a team to assess concerns from the acquisition community suggesting that testing drives undue requirements, excessive cost, and added schedule into programs. Concurrently, I conducted a systematic review of recent programs to address the questions of whether testing delays programs, what other causes create program delays, what is the duration of the delays, and what is the marginal cost of operational test and evaluation. The results of both studies indicated that testing and test requirements do not cause major program delays or drive undue costs. Our Joint memorandum addressed other problems that were identified in the two studies.

The USD(AT&L) study, based on interviews with senior leaders within the Office of the Secretary of Defense (OSD) and Service leaders with responsibility for program management and oversight, developmental testing, and operational testing, found no significant evidence that the testing community typically drives unplanned requirements, cost, or schedule into programs. The study team found that tensions are often evident between programs and the test community and that for the most part these are normal and healthy; however, the study identified four potential improvements to these relationships and interactions:

1. stronger mechanisms for a more rapid adaptation to emerging facts,
2. a requirements process that produces well-defined and testable requirements,
3. alignment of acquisition and test strategies (i.e., programs lack the budgetary and contract flexibility necessary to accommodate discovery), and
4. open communications between programs and testers, early and often, with constructive involvement of senior leaders.

At the time of writing, we are working with USD(AT&L) to implement changes to the Department of Defense (DoD) acquisition policy (DoD 2008), which we expect will help realize some of the potential improvements listed above.

Causes of program delays

My review examined 67 major programs that experienced significant delays and/or a Nunn McCurdy breach. (The study is available at http://www.dote.osd.mil/pub/presentations.html.) Thirty-six of the 67 programs experienced a Nunn McCurdy breach, and six programs were ultimately canceled. Two of the 36 Nunn McCurdy programs experienced no delays to their schedule. We characterized the programs as exhibiting any of five categories of problems that caused delays:

1. manufacturing and development (to include quality control, software development, and integration issues),
2. programmatic (scheduling or funding problems),
3. performance in Developmental Testing (DT),
4. performance in Operational Testing (OT), and
5. conducting the test (such as range availability, test instrumentation problems, and test execution problems).

Of the 67 programs, we found that 56 programs (84 percent) had performance problems in testing (DT, OT, or both) while only eight programs (12 percent) had issues conducting the tests that led to delays. Only one program had delays solely attributed to the test. (The U.S. Army's Force XXI Battle Command Brigade and Below [FBCB2] operational test was delayed for 1 year because the test unit was deployed.)

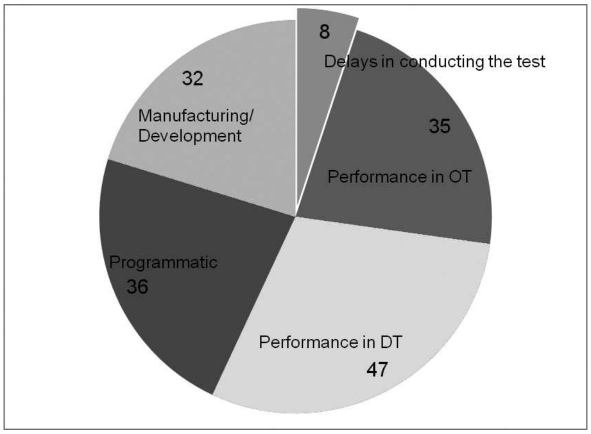


Figure 1. Reasons behind program delays: Programs are more likely to be delayed because of the results of testing vice the testing itself.

Figure 1 shows the distribution of delay cause categories for the 67 programs. There were 158 instances of issues that caused delays for the 67 programs. Many of the programs had multiple problems that fell into more than one of the five categories of reason for delays. There were eight test conduct problems and 82 program performance problems discovered during test—an order of magnitude difference. Clearly, programs are most often delayed because of the *results* of testing, not the testing itself.

Length of delays

The length of delays for the programs examined varied from none (for two of the Nunn McCurdy programs) to 15 years. Thirty-seven programs were delayed by more than 3 years. The delays were measured against the most recent previously published schedule, so in a sense the total delays could be even longer relative to the original planned schedule. Six of the programs were eventually canceled, and one had the Milestone B rescinded.

Cost of OT

Both the USD(AT&L) and the Office of the Director, Operational Test and Evaluation (DOT&E) studies noted that the marginal cost of testing is a small portion of the overall program budget; however, the cost can be a large percentage of the budget in the year(s) in which it occurs. Because the testing occurs at the end of the development process, programs typically have few degrees of freedom (and resources) left to work issues.

We evaluated marginal cost to programs of Operational Test and Evaluation (OT&E) as a percentage of total acquisition cost. A review of 78 recent test programs in the U.S. Army, Air Force, and Navy showed that the average marginal cost of OT&E is 0.65 percent of the total acquisition cost. It also appears that some programs truly have negligible OT&E costs relative to program acquisition costs (OT&E $<$0.1 percent) and that most program OT&E costs are less than one percent. Few programs that we reviewed (seven out of 78) required more than 1.5 percent of program acquisition costs for OT&E. For programs with OT&E costs above average, we found that low program acquisition cost, expense of test articles, and test article expendability were the dominant drivers of high relative OT&E cost. *Figure 2* shows the distribution of the marginal cost of OT&E for the 78 programs we examined.

Summary

The Decker Wagner report commissioned last year by the Secretary of the Army addressed the Army's failure rate of procuring new development programs (Army 2010). The study found that between 1990 and 2010, the Army terminated 22 Major Defense Acquisition Programs (MDAPs) and that 15 of those terminations have occurred since 2001. Further, excluding the Future Combat System (FCS), the Army spent greater than

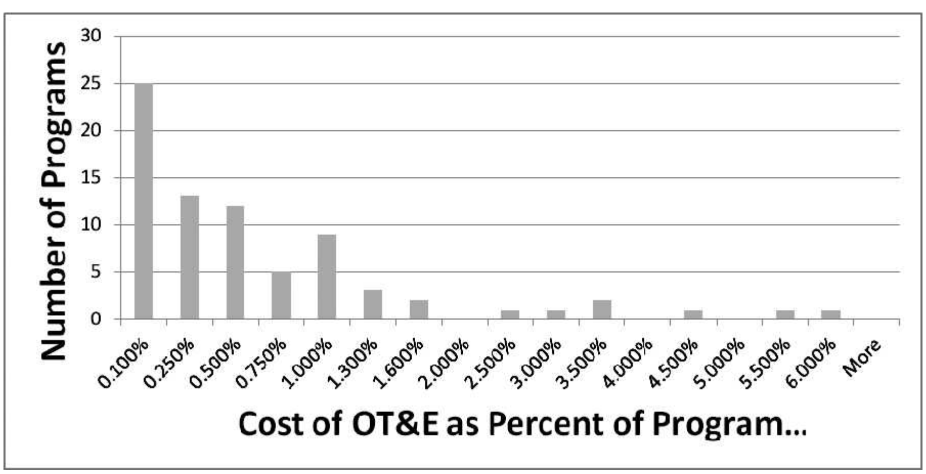


Figure 2. Marginal cost of operational test and evaluation relative to program acquisition cost.

one billion dollars *per year* since 1996 on programs that were eventually canceled before completion. The study cited many reasons for the failed programs, including unconstrained requirements, weak trade studies, and erosion of the requirements and acquisition workforce. However, none of the reasons cited included T&E. In fact, earlier and more robust T&E may have revealed problems and solutions earlier in the program when they would have been less costly to fix, or allowed decision makers to cancel or restructure before wasting billions of dollars.

Finally, in his recent testimony before Congress, Dr. Carter stated his opinion on reducing the average acquisition timeline:

> *"…[acquisition] time is best reduced by ensuring reasonable requirements are set, by being willing to trade away requirements that prove to be excessive, and by controlling requirements creep so that development time can be constrained. I support rigorous developmental and independent operational test and evaluation to provide accurate and objective information on the capabilities and limitations of defense systems to both acquisition executives and warfighters and to ensure contractors deliver products that meet requirements." (Carter 2011)* ❑

DR. J. MICHAEL GILMORE was sworn in as director of Operational Test and Evaluation on September 23, 2009. A presidential appointee confirmed by the United States Senate, he serves as the senior advisor to the Secretary of Defense on operational and live fire test and evaluation of Department of Defense weapon systems. Previously, Dr. Gilmore was the assistant director for National Security at the Congressional Budget Office (CBO). Dr. Gilmore is a former Deputy Director of General Purpose Programs with the Office of the Secretary of Defense, Program Analysis and Evaluation (OSD[PA&E]). Dr. Gilmore served with Program Analysis and Evaluation for 11 years. Earlier, Dr. Gilmore worked at the Lawrence Livermore National Laboratory; Falcon Associates; and McDonnell Douglas Washington Studies and Analysis Group where he became manager, electronic systems company analysis. Dr. Gilmore is a graduate of Massachusetts Institute of Technology, Cambridge, Massachusetts, where he earned a bachelor of science degree in physics. He subsequently earned master of science and doctor of philosophy degrees in nuclear engineering from the University of Wisconsin, Madison, Wisconsin. E-mail: mike.gilmore@osd.mil

References

U.S. Army. 2010. "Army Strong: Equipped, Trained and Ready," Final Report of the 2010 Army Acquisition Review, January 2010. Washington, D.C.: Secretary of the Army.

Carter. 2011. Senate Armed Services Committee. Online at http://armed-services.senate.gov.

ITEA Journal 2011; 32: 392–396

Test As We Fight

Steve Hutchison, Ph.D.
Test and Evaluation Executive
Luanne Overstreet
Deputy Test and Evaluation Executive
Defense Information Systems Agency, The Pentagon, Washington, D.C.

As testers, the phrase "test as we fight" should be a reminder that if the objective of our test is simply to determine whether the product satisfies parameters defined in the requirements document, then we may miss the essence of testing to determine if the product does what the users expect it to do. And, that is an important distinction; while the former is important, the latter is more important!

There are any number of requirements to be met when we design a test of a new product, the obvious ones being operationally realistic conditions, typical users, and production representative equipment, but the list also includes realistic threat, accredited modeling and simulation, a concept of operation or other source of employment doctrine, and a maintenance or system support capability as intended to be available when the product is fielded. When we put all of this together, it sounds like it would satisfy the "test as we fight" mantra, yet it remains possible that regardless of how much more we add to the scope of testing to cover real-world conditions, our soldiers, sailors, airmen, and marines will fight systems differently than the way we tested them. Our challenge is to reduce that differential; to do that, we have to design the test to determine if the product does what the user expects it to do. In the Defense Information Systems Agency (DISA), we have attempted to capture lessons from the information technology (IT) sector to help us craft an improved approach to testing, and these lessons shape the priorities, actions, and tasks we have outlined in the DISA Campaign Plan.

The DISA Campaign Plan

DISA is an operationally focused combat support agency providing Joint and Combined warfighting information technology capabilities and is dedicated to uncompromising support to our national leaders and warfighters. The 2011–2012 DISA Campaign Plan is a roadmap defining the agency mission, vision, and objectives (for a condensed version of the DISA Campaign Plan, see http://www.disa.mil/campaignplan/index.html). The Campaign Plan is a comprehensive document (spanning three lines of operation and nine Joint enablers), which serves as the foundation for planning, prioritizing, allocating resources, and executing the mission. *Figure 1* depicts the lines of operation and Joint enablers. Borrowing from the doctrine described in Joint Pub 5-0, lines of operation "describe and connect a series of decisive actions that lead to control of an objective" (Joint Chiefs of Staff, 2011). The DISA lines of operation identify decisive capabilities that provide the user power to connect and collaborate in an assured, reliable, network-enabled operating environment. The three DISA lines of operation are as follows:

1. Enterprise Infrastructure—an agile, converged enterprise infrastructure that provides a collaborative and trusted environment to enable end-to-end information sharing capable of adapting to rapidly changing conditions;
2. Command and Control and Information Sharing—effective, reliable, secure, agile, national and operational command and control and information sharing capabilities and services that adapt to rapidly changing circumstances; and
3. Operate and Assure—dynamic control and operation of our enterprise infrastructure and the command and control and information sharing capabilities and services.

DISA defines *Joint enablers* as the "processes, disciplines, and governance that support the lines of operation." Testing is one of nine Joint enablers spanning the DISA mission space. The DISA Test and Evaluation (T&E) organization consists of the Office of the T&E Executive (TEO), a staff element located at the DISA headquarters at Fort Meade, Maryland,

Figure 1. Defense Information Systems Agency lines of operation and Joint enablers.

that provides strategic planning and engagement in all Department of Defense (DoD) T&E governance processes, and the Joint Interoperability Test Command (JITC), the test execution arm that provides full-service IT T&E support across all IT test disciplines, including developmental, operational, interoperability, and information assurance testing. The JITC operates in three locations: Fort Huachuca, Arizona; Indian Head, Maryland; and Fort Meade, Maryland; and collectively constitutes the IT testbed in the Major Range and Test Facility Base (MRTFB).

DISA T&E strategic initiatives

The Campaign Plan describes our strategic initiatives in testing in the context of three priorities, with associated actions and tasks, as we strive to reach the ultimate objective: to test like we fight.

The first priority is to "increase effectiveness and efficiencies across T&E with robust proactive virtual methodologies." The actions and tasks supporting this priority are intended to set the conditions for implementing a "testing as a service" construct in a dynamic enterprise infrastructure. To achieve the desired effects, we recognize the need to develop concepts that leverage virtualization and provision testing services to "the cloud" to make them available on demand.

We believe there are tremendous efficiencies to be gained in testing information technologies by exploiting the power of virtualization. As IT capability development shifts to more agile approaches and shorter cycles, testing will shift to a near continuous "on-demand" service. Virtualization is key in an agile environment; it permits us to create "virtual machines" and virtual users, replicate network load and test at boundary conditions, and cleanse sensitive data to reduce the potential for adverse impact to authoritative data sources. When employed in conjunction with a limited set of live users on the operational network, virtualization permits us to reduce the physical T&E burden without compromising realistic operational conditions in test. As we further discover and develop virtualization capabilities, our objective is to promulgate concepts, tools, and methods to other DoD testing agents, with the goal of improving efficiency across the enterprise.

The second priority is to "transform our test organization to be leaders in test-driven development." As IT program developers and requirements owners shift to smaller increments of capability, iteratively building on prioritized requirements, testers cannot afford to bring industrial age test approaches to information age projects. Nowhere is this more apparent than in the field of information technologies. Information technologies evolve so rapidly we can barely keep up, and the test approach we used yesterday is oftentimes not adequate today. In the world of IT, the shortest path to irrelevance is to continue to do what we're doing. Product designers know this; the next generation model is typically well beyond the drawing board the moment the current model hits the marketplace. As testers, we must constantly innovate our processes and methods or we risk becoming irrelevant.

Transformation to a test-driven development approach is as much a challenge for program managers as it is for testers. The DoD acquisition process does not provide much incentive (read resources) for program managers to involve the various government testers early in the product life cycle. Since independent tests compete for scarce resources, program managers opt to perform functions within the program management office and contractor, translating requirements into software design features and executing contractor-planned developmental testing. This often occurs prior to a handoff to external testers to complete initial operational test and other certification testing. As a consequence, independent testing is often a period of defect discovery rather than confirmation that the product works. While that model may work for platforms with long developmental timelines, such an approach will not work well in support of programs following an agile software development approach and will prevent the program from realizing the benefits of test-driven development.

Test-driven development is a technique typically associated with agile software development but could be applied in all programs to significantly reduce the risk of late discovery of critical failure modes. Test-driven development is a means to translate user stories into test cases before development begins; the idea being that when developers understand how the capability will be used, and therefore tested, development improves, and the resulting product has fewer defects. In the systems engineering sense, since testers produce test cases before development begins, test-driven development inserts testing on the left side of the systems engineering "V," a practice brought about after realization that the traditional systems engineering approach was not achieving the desired reduction in product defects.

As leaders in test-driven development, our objective is to remove the barriers between test teams and development teams and create one team with one goal: rapid fielding of enhanced capabilities for the warfighter. To achieve the goal of our second priority, we are working to establish processes that institute the use of user stories in place of typical requirements and parameters. User stories identify what the user expects the capability to do; hence, as stated in the introduction, this approach shifts tester focus from "does it satisfy the requirements in the capability development document" to "does it satisfy user needs?"

In an agile model, testing shifts from being a separate phase late in the acquisition process to a partnership that is closely integrated with every aspect of the product life cycle. Moreover, *agile testing* removes the stovepipes within the test community (developmental testers, operational testers, interoperability testers, and information assurance testers) by integrating the organizations and disciplines to ensure we treat all test activities and test data as a shared resource. Agile testing optimizes the test activity for multiple evaluation and certification functions and provides all stakeholders a high degree of confidence in the results—there is no loss of rigor in agile testing!

The third priority is to "streamline DISA's isolated T&E networks into a single network with the enterprise infrastructure providing core services." As stated in the introduction, to test like we fight, we need to test in operationally realistic conditions, with production systems, users, threat, support, etc. For systems whose operating environment is cyberspace, operationally realistic conditions are typically found only on the operational network; in other words, if a system or service is going to operate on the Global Information Grid (GIG), then it should be tested on the GIG, protected by the computer network defense service provider, and supported by the network operational support services. To achieve this objective, DISA implemented the Defense Information Systems Network (DISN) Test and Evaluation Network (DTEN). This priority—streamlining DISA's networks—serves first to "clean house" by converging multiple DISA networks supporting test activities onto the DTEN, but ultimately to provide the DTEN as a T&E network service to the DoD test community.

DISA's enterprise infrastructure goals are creating a paradigm shift from individual system testing to enterprise net-centric testing. DISA T&E is leading an effort to collapse various DoD networks onto the DTEN. In early FY11, the DTEN obtained initial operational capability after absorbing the DISN–Leading Edge Services (DISN-LES) infrastructure, users, and management processes and is currently transitioning the Combined Federated Battle Lab Network (CFBLNet). Additional network migrations will continue over the course of the next few years, with the goal being substantial cost savings and efficiency gains for the DoD. Ultimately, the DTEN is a critical enabling transport layer that will provide persistent connectivity supporting a federation of DoD test and integration labs, where test agents can find needed test resources, connect them to their test, collect data, and then release the data back to the capability pool when the test is complete.

To achieve the goal of network consolidation, we will work with the DoD T&E community to identify and solicit additional candidates for transition and develop a list of potential partner opportunities to federate capabilities and establish a live, virtual, and constructive test environment on the DTEN. This effort should permit the test community to increase efficiencies, reduce duplication, encourage reciprocity through partnerships, and reduce the time required to establish appropriate configurations needed to support any test. Network management and business processes will ensure on-demand "service" of a converged T&E network.

Lessons from agile software development

Agile software development is about delivery of capability at "speed of need"; it is a highly collaborative framework that is responsive to changing user requirements, and it relies on early and continual involvement of the test team and user. Agile projects typically produce software at a higher rate with fewer defects. The development effort progresses through the product backlog (a prioritized requirements list) in short "time-boxed" iterations or "sprints," during which the product is integrated and thoroughly tested. The objective of each iteration is a potentially shippable product, so when the iteration is complete, the product is demonstrated to stakeholders to determine if it satisfies the user need, and a decision is made whether or not to deploy the product. As development progresses through multiple sprints, the end result is a dynamically defined and improved, fully tested and capable product. *Figure 2* depicts how the agile framework "Scrum" can be applied to the DoD IT requirements and acquisition process.

Agile changes everything; from requirements management, to development, testing, and fielding. Agile

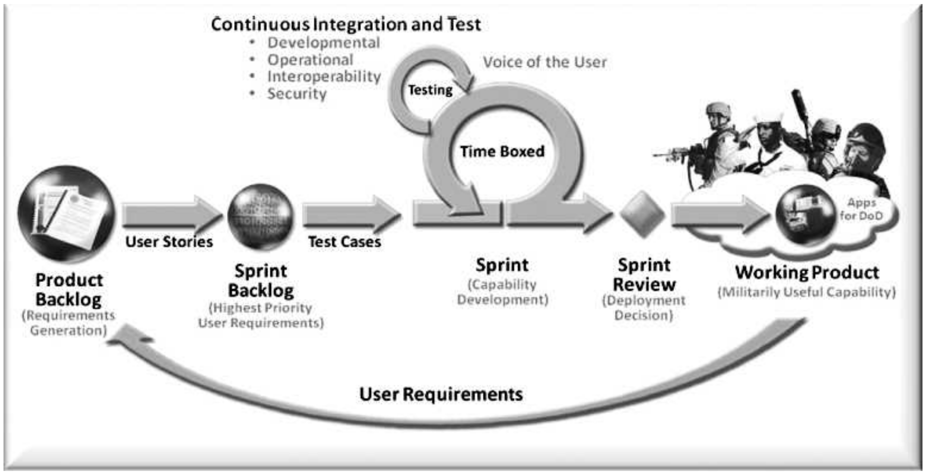


Figure 2. An agile model for the Department of Defense.

software development has several implications for the DoD in general and for testing in particular. Agile projects focus on customer satisfaction. In an agile project, requirements are continuously updated and reprioritized; features are added and removed to reflect that customer needs change, sometimes very rapidly. In agile, a change in the warfighter's needs could be just *one sprint away* from delivery to the field. This marks an essential difference between requirements management in the DoD today and how we would manage requirements for an agile IT project.

For testers, agile brings new challenges. Testers are an integral part of the agile team and are involved in every step of the process from requirements vetting, to product testing, to continuous monitoring once fielded. Testers need to learn and practice test-driven development to translate user stories into executable test cases that focus on how the capability will be used. As features are developed, testers execute the full spectrum of tests to ensure technical, operational, interoperability, and security measures are satisfied. Agile testing maximizes use of automation and virtualization to speed and enhance the test activity. When tests are completed, all data are shared to support independent analysis, leverage reciprocity, and provide comprehensive findings to stakeholders and decision makers. And, all of this is repeated for a new set of features in every sprint to support a product cycle time that is probably three to four times faster than we currently execute. As testers, we need to be prepared for the challenges of the agile future.

Training the future IT T&E workforce

Professionalizing the T&E workforce is a persistent priority that spans all programming cycles. DISA T&E has invested in a major new initiative to improve the training of our IT testers. Over the next year, working with the University of Memphis Systems Testing Excellence Program (STEP), we are developing a certificate program for IT testers. The STEP program is an interdisciplinary research initiative of the University of Memphis FedEx Institute of Technology (http://step.memphis.edu), which views systems testing as a strategic research area encompassing all aspects of testing of business systems, including hardware, software, requirements, and the testing of business rules. Ultimately, we intend to offer the enhanced training program to DoD and other government agency IT testers and seek Defense Acquisition University 200-level equivalence for certification in the T&E career field.

Summary

Our objective in T&E should be mission-focused agility, which is responsive to developers, users, and decision makers. As IT programs become more agile, we need to innovate our processes to provide the ability to rapidly compose test plans, configure the test environment, execute the test, and report findings in a manner that conveys whether the capability is effective, suitable, interoperable, and secure. Our strategic initiatives, with associated actions and tasks, define a roadmap to improve efficiencies, reduce duplication, and provide an operationally realistic environment in which to test and certify network-enabled capabilities the way users will employ them in the field—to test like we fight. ❑

Steven J. Hutchison, Ph.D., a member of the Senior Executive Service, is the Test and Evaluation Executive for the Defense Information Systems Agency. He is

responsible for developing and enforcing T&E policy and procedures, representing DISA to the DoD T&E community, and providing direct support to DISA programs for T&E of IT capabilities. Prior to his arrival at DISA, Dr. Hutchison served in the office of the Director, Operational Test and Evaluation (DOT&E), Office of the Secretary of Defense, as a net-centric warfare systems analyst. While in DOT&E, he had oversight responsibilities for several of the major warfighting information systems, including the Global Command and Control System-Joint, the Service variants of the Distributed Common Ground/Surface Systems, and the Net Enabled Command Capability. Dr. Hutchinson also served for 6 years in the Army Test and Evaluation Command as an evaluation officer and later as the Assistant Director. E-mail: Steven Hutchison@disa.mil.

References

Joint Chiefs of Staff, Joint Operation Planning, *Joint Publication (JP) 5–0*. August 11, 2011. Washington, D.C.: author. Available online at http://www.dtic.mil/doctrine/new_pubs/jp5_0.pdf. Accessed November 2, 2011.

ITEA Journal 2011; 32: 397–401

Myths and Truths of T&E

Amy Markowich
Deputy Department of the Navy Test and Evaluation,
The Pentagon, Washington, D.C.

I am very pleased to be participating in this new feature for *The ITEA Journal*. As the Senior Executive for Department of the Navy Test and Evaluation, I am excited to share our challenges and initiatives and, in turn, gain insight into the perspective of our Service and Office of the Secretary of Defense (OSD) counterparts. In this first installment, I'd like to share what I consider to be some "myths" and "truths" concerning test and evaluation (T&E). I'm hoping they will be great conversation starters and set the stage as we all work to improve the T&E community.

One of the definitions of "myth" from the trusty Google free online dictionary[1] is "a popular belief or story that has become associated with a person, institution, or occurrence, especially one considered to illustrate a cultural ideal." We have several myths that have become part of our T&E identity; some for good reasons and history, some because of misperception and lack of awareness. It can be a valuable exercise to explore the myths of a culture or community and the kernels of truth that lie within them. The saying that the "truth is somewhere in the middle" fits in many cases. One person's myth is another person's truth, so we'll use these terms loosely. Feel free to debate these at the water cooler and come to your own conclusions.

Myth

It has to be an "Us versus Them" mentality.

Truth

We should be one team and we need to act that way from the beginning. There has always been tension between the "T&E Community" and the acquisition community. It's a phenomenon that we all have seen. Some of the culture can be understood if you look back on some of the old ways of doing business. A program manager works diligently to meet his requirements and milestones and when the system is tossed over the fence to T&E, he could get it tossed right back over with a "Hey, your baby is really ugly. Good luck with that!" Or, he meets his requirements but the overall operational capability falls short and we end up with a finger-pointing stand-off. There seem to be two schools of thought. Some say testers get in the way of progress, impose unnecessary requirements, and slow down acquisition. On the flip side, some feel that developers don't fully understand how systems will really be used in the intended operating environment and could deploy items that didn't work if testers weren't there to "catch" them. Both sides have sometimes seen the other as "Them" in our past history and at times still do.

However, this is not the way we are striving to do business anymore. The use of true integrated testing is breaking down barriers, but implementation is inconsistent. It needs to be standard. We are seeing more flexibility in the Operational Test (OT) community; for example, OT being willing to accept modeling and simulation and the early results of developmental testing if planned correctly. In the U.S. Navy, Commander, Operational Test Force (COTF), the Operational Test Agency, is now providing their Integrated Evaluation Framework and Mission Based Scenarios, so programs can build operational realism into the Developmental Tests (DT). For U.S. Marine systems, the Marine Corps Operational T&E Activity (MCOTEA) has a similar approach and is very committed to working with the program manager to plan effective operational tests. The more we work together, the greater chance we have of using results across Contractor Test (CT), DT, and OT.

We are seeing more focus on early development of Concept of Operations (CONOPS) and on better understanding and articulation of testable requirements. Our industry partners would love to have this insight as early as possible. This is like being able to take an open-book test. Who wouldn't like that!? It also gives us the chance to plan seamlessly, building in performance checks, ensuring systemic interoperability, and planning for off-ramps if things aren't going as expected. All of this is a work in progress. The end answer is easy—we're all just "Us." T&E (even Operational Test and Evaluation [OT&E]) is a part and partner of the acquisition process. However, sometimes "Us" and "Them" language and thinking persists, and there is a continued need to work on relationships and communication between all stakeholders. As budgets get tighter, we're going to have to pull together more, be innovative about finding ways to get the data we need as early as possible, and not be afraid to talk about what the data will cost us. This means talking not only about funding risk but also about the risk we run into if we don't do something we need to. This, at times, may need to be a frank conversation with the Department of Defense (DoD) and Service leadership at the highest levels.

Myth

Testing takes way too long.

Truth

The "D" in DT is there for a reason. In DT, the system is being tested and evaluated as part of the development process. The purpose of Developmental Test and Evaluation (DT&E) is to provide insight, not judgment, about the goodness of the technical approach and design. Verification of requirements comes last, and we shouldn't expect to be able to breeze through this phase unless the systems are already mature, which is not usually a reasonable expectation. There are many reasons for this, from extremely complex technology and ill-defined requirements to overly optimistic planning. Many complain about how much time large programs are taking "for testing." There's a feeling that we take years to develop a system and spend unnecessary years in testing. What's actually happening is that the system is still under development, and we're finding big and expensive problems when we enter into testing. Test resources are then devoted to the effort of finding and fixing issues, which takes time and money and keeps large, expensive test teams around longer than originally planned. We need to do a better job of communicating and articulating why T&E is taking as long as it does. This includes early planning, which takes immature technology into account and plans for appropriate resources and test venues to address. This also includes execution metrics that track when testing is finding and fixing design issues and when it is verifying specifications, KPPs, and COIs. This information would give real insight into desired system performance as the program progresses. We need to understand the real issues behind delays—are they really testing issues or are they system maturity issues? The answer to this question changes how we approach finding solutions as a systemic acquisition issue. For quick response requirements, there are policies and methods to streamline traditional processes. Improved understanding of options to streamline testing and improve speed to deployment would benefit the overall acquisition community.

Myth

First flight or sea trial of an aircraft, ship, or sub is a meaningful acquisition milestone and a good contract incentive.

Truth

Data for decisions should be capabilities based. Demonstrating that an aircraft can safely fly or a ship or sub can get safely underway is a necessary developmental milestone, but it is not representative of the weapon system capability as a whole. We should incentivize system maturity and capability, not single developmental events. Leadership and stakeholders would have more insight into system performance if we were demonstrating capability on a "mission performance growth curve" and awards and incentives were based on that. This would force both government and industry to articulate expectations for performance at various phases and would minimize disconnects between meeting contractual requirements and meeting operational requirements. This could change behavior across all stakeholders.

Myth

We just need to let Industry do all the testing and deliver the final product to the government.

Truth

We tried this with Total System Performance Responsibility (TSPR) contracts and found that this approach is extremely expensive, and the end product does not tend to meet mission requirements. We also lose our organic insight into the systems, their design, and their true performance. We cease to be smart buyers when we lose that understanding. No single contractor has the insight or ability to bring the whole capability together, and data cannot be shared easily because of proprietary restrictions. When each system evolves in a closed stovepipe, the individual systems will not integrate or interoperate with the whole, and the end-to-end mission capability does not perform as planned. The government has a role to ensure that all of the various systems, built by various companies, will work together. That role, of ultimate lead systems integrator cannot be delegated. We need to own the various interfaces, to foster competition, and to be able to fully evaluate, so that systems work in a larger system-of-systems construct.

Myth

T&E costs a lot of money and is a likely target in budget reduction drills.

Truth

There have been several studies on the cost of T&E over the years. The cost of T&E is actually very small in the big picture. T&E costs tend to be less than 15 percent of RDT&E and less than 5 percent of program cost. OT alone is a fraction of that, coming in at less than 1 percent. Those factual numbers don't change the dynamic, however. Ten percent of a really big number is a big number, and when you also factor in the cost of ranges and facilities, it's admittedly a number we have to be conscious of as we move forward. We, in the T&E community, have to equally be good stewards of taxpayer dollars and continue to look for cost savings and innovative methods.

I have seen willingness from T&E to do its part to increase efficiency and reduce costs. The Department of the Navy has developed a Test Evaluation Improvement Process. Our goal is to improve all aspects of how we execute T&E. This includes consistent implementation of the best practices everyone seems to know but that we execute poorly as a community. Integrated testing is a good example. We all say we're doing it, but in small print you see a "results may vary" disclaimer. We have great examples such as the Air Warfare T&E Enterprise where we do combined combat systems testing for ships, saving cost and duplicate test, but we could probably do this more. Some programs make extensive use of ground test and shore labs and simulations, while others don't. Not all of our program managers know what expertise and T&E facilities exist within DoD, and we sometimes pay to have the wheel rebuilt individually. We are aggressively looking across our ranges and labs for reuse and efficiency opportunities. Our improvement efforts also include workforce development and training, improved policy guidance, and enhancing test strategy development and execution. The T&E community, as with all others, must be sensitive to the DoD budget and our national conversation on the fiscal crisis. We are committed to working with programs, SAEs, Director DT&E (DDT&E), and Director, OT&E (DOT&E) to continually strive for efficiencies without an unacceptable loss of effectiveness of testing.

The current budget struggles and our need to make smart choices in the technologies we pursue and the programs we fund make the role of T&E even more critical. T&E provides insight into performance as we make program and technology decisions and deployment choices. We must fully understand how systems will work in a stressing, realistic operational environment and how reliable, sustainable, and affordable they will be to deploy. T&E must work hand in hand with developers to find problems early enough to fix them and to understand their impact in an operational context. Good data are required to make smart choices effectively. We'll need to find ways to perform our function in new and innovative ways. While existing T&E teams look for new ideas and test strategies, we can't neglect the lessons of the past. Losing rigor in T&E and reducing funding just to meet a budget target is a losing proposition. If we go down that path, we'll likely pay more for those choices later on. Early and operationally realistic CT and DT with a robust integrated test strategy are keys to success. If final OT is a time of discovery, we're definitely doing something wrong.

Myth

The government just needs to hire more people to solve our problems.

Truth

Having a right-sized and trained workforce is key to our future success. After realizing the risk of losing science and engineering expertise, the acquisition community redoubled efforts to ensure our technical skills are not lost. This included hiring in both systems engineering and T&E. Our recent Service T&E self-assessments (2009 and 2010), which were provided to OSD DT&E for their report to Congress, found that our staffing is now adequate. The focus has moved from quantity to quality and toward professionalization of the workforce. Development of robust training for both new and journeymen employees is a priority. We are working with the Defense Acquisition University (DAU) to ensure that acquisition T&E workforce training is effective and useful. We are also taking it a step further by developing more detailed training needed to close skills gaps. It's hard to hire testers; you have to grow them. Testing is not a common subject taught in schools; it requires experience, on-the-job training, hands-on work, and development of good judgment. We are also interested in partnering with organizations such as ITEA to better leverage tutorials in emerging topics such as reliability growth, design of experiments, and use of modeling and simulation. The massive wave of retirements that have been projected has not yet materialized, but we know it's still coming. This drives a need for aggressive mentoring and methods to pass knowledge from the "graybeards" to the "newbies." This is a problem that needs to be worked in both government and industry. We all are struggling to find qualified scientists, engineers, and technicians to grow a new generation of subject-matter experts.

Myth

We have too much duplicative infrastructure, and we need to close things down.

Truth

Over the years, the Department of the Navy has minimized duplication in the T&E infrastructure. We have mothballed and divested some capabilities and are down to a critical core set. We have dedicated funding from the Major Range Test Facility Base (MRTFB) to ensure our core T&E national assets are funded and ready to test DoD programs. Our warfare centers maintain and operate T&E ranges and facilities, and we prioritize investments and modernization in a national manner. Funding is always an issue, and we have been forced to minimize improvement and modernization. Most of our dollars are tied to critical maintenance of radars, telemetry, threats, and ground systems with little left to evolve technology to meet the demands of the future. As dollars get ever tighter, we are working closely with our other Service partners, revitalizing the *reliance* process to look for opportunities to invest jointly and reuse capabilities. We rely heavily on the Test Resource Management Center and the Central Test and Evaluation Improvement Program (CTEIP) to develop advanced Joint test capabilities. To execute effective T&E, we must have advanced labs, facilities, and ranges that let us replicate stressing and operationally realistic environments. We must protect land, sea, and air space from encroachment and spectrum sell-off. We also must continue to evolve our modeling, threat simulations, and targets to allow our systems to be assessed against emerging threats.

While the T&E community has done much to eliminate duplication, looking at cross-Service reuse as well as creating centers of excellence where appropriate, there are still opportunities to explore. In the Department of the Navy, we are broadening the aperture to also look across the analysis, RDT&E, and training continuum where we all invest in models, simulation, scenarios, and environments. They may be used for different purposes, but the things we need have become remarkably similar. How great would it be if we could pay for them once and use them across the acquisition life cycle!? Open, nonproprietary products become very important. We'd also like to team with our Industry partners who don't want to build "white elephant" labs and facilities that are expensive to maintain for just one platform or product line. We've had some great examples of shared investment for mutual benefit. Every contractor who develops complex systems and sensors needs some sort of local integration facilities. However, there comes a time when the "environment" you wrap around a system for testing becomes more complex and expensive. Ideally, this would be reusable to support many programs and have a common framework for evaluation. This is a goal, not a reality, but it's a vision to strive for. We need to move beyond the current model of everyone building their own "hobby shop" environment without taking advantage of the benefit of years of DoD investment. We fully support OSD, DDT&E efforts to encourage use of existing government test facilities vice continuing to rebuild similar capabilities at contractor sites. The Department of the Navy will be offering similar guidance. We are also developing guidance for our test teams on effective use of existing infrastructure and developing a T&E capability roadmap to define current strategy and identify critical gaps for programs in the next 10 to 15 years.

I hope this provided you with a little food for thought. In my career, I've seen these myths in action and how they have subtle influences every day. It's been valuable to see first-hand the issues the acquisition community must deal with and to partner with our program executive offices (PEOs) and program managers to develop robust and efficient T&E strategies. T&E, just like every other community, is being challenged to look ourselves in the mirror and to attack our problems with a new way of thinking. Bureaucracy is not something that any of us intend to be a part of, but it creeps in on insidious soft feet. Testing isn't about the paperwork or about cutting and pasting boilerplate language and getting signatures. It's about the data we need to objectively characterize technical performance and gain insight into the design and systems engineering process. It's also about finding the best ways to get that data to our decision makers. We need to consider the risks if we don't have that data, to explore what the implications are if we only have the 80 percent solution. We need to do this while making absolutely sure that we don't neglect safety or our understanding of the end capability. We need to think about ways to use our test methods and processes to get systems into the theater quicker as a rule not an exception. This is hard stuff. We need those of you reading this journal to be involved, develop good ideas, innovate new test strategies, and surface your thinking.

As we move through the acquisition process, we all have a part to play—DT and OT; Program Management and all who support them; Service and Acquisition Leadership; and OSD oversight agencies. Some roles have a level of independence for assessment and reporting to leadership and to Congress. However, at the end of the day, the biggest myth is that we are working at cross-purposes. The truth is there is no independence of purpose. I know we all have the same goal, to get affordable systems that work in the operational environment out to our warfighters, as efficiently and as quickly as possible. They are depending on us and on you. That's one thing we just can't argue about.

If you have any more myths or would like to suggest future discussions, e-mail me at amy.markowich@navy.mil. ❑

AMY J. MARKOWICH serves as the Deputy, Department of the Navy Test and Evaluation. She is responsible for integration of Test and Evaluation (T&E) across the U.S. Navy, enhancing the T&E workforce and infrastructure, and ensuring programs plan for and complete adequate testing to demonstrate suitable and effective operations in the Joint battlespace. Ms. Markowich was appointed to the Senior Executive Service in June 2009 and has served in an array of engineering and leadership positions over 21 years of experience in T&E and acquisition. Preceding selection as Deputy, Department of the Navy T&E Executive, Ms. Markowich led the Program Executive Office for Unmanned Aviation & Strike Weapons (PEO) (U&W) Common Systems Integration (CSI) program at the Naval Air Systems Command (NAVAIR). She previously served as the Deputy Director, Integrated Battlespace Simulation and Test Department. Her experience across UAS, T&E, M&S, Ground Test, Ranges, and her Joint Service/multinational contacts provide a unique perspective to support future Office of the Secretary of Defense (OSD) and Navy initiatives. Ms. Markowich graduated from the University of South Florida with a bachelor of science degree in electrical engineering and a bachelor of arts degree in mathematics. She is also a graduate of the NAVAIR Senior Executive Management Development Program. E-mail: amy.markowich@navy.mil

Endnote

[1]Google free online dictionary is no longer available. You can use Google Web search to find definitions at http://www.google.com/.

Acknowledgments

Many thanks to colleagues in Assistant Secretary of the Navy for Research Development and Acquisition (ASNRDA), OPNAV N091 and across Department of the Navy T&E who debate myths around test and evaluation often and vocally. Special thanks to RADM Kate Paige (Ret), currently President, Mark India, LLC who completed recent studies on T&E; to VADM Mark Skinner, Principal Deputy (ASNRDA) for program T&E partnership and guidance; and to Mr. Brian Markowich, Naval Aviation Systems Command (NAVAIR) Air Vehicle T&E Deputy Division Head and Associate Fellow for invaluable collaboration on this article.

ITEA Journal 2011; 32: 402–407

Test as We Fight

Dennis O'Donoghue
Vice President
Boeing Test & Evaluation, Seattle, Washington

There are many lessons to be realized from the large-scale Department of Defense (DoD) programs over the past decade, which can improve the success of Test and Evaluation in System-of-Systems (SOS) efforts. These lessons apply beyond DoD programs to other government departments and agencies. SOS verification successes require robust, comprehensive testing of these complex systems. Recent lessons within DoD programs have shown the need for earlier involvement of test engineering within the design/system development process, greater focus on end user and mission utilization of systems beyond the specification-based criteria, and the right sized–right fidelity of modeling and simulation to create the relevant test environment that emulates system use and loading. SOS testing requires a stimulus that drives the interoperability required for verification as well as for producing the emergent behavior necessary to characterize those systems and evaluate their capabilities in relation to end-user mission requirements.

Key words: Acquisition; Boeing T&E; distributed testing; interoperability; live-virtual-constructive, LVC; operational excellence; system of systems; validation; verification.

"A system of systems is a set of interoperable, independent systems that, when integrated, provides capabilities and performance beyond the sum of the individual systems."[1]

During the first decade of the 21st century, many programs have been dubbed "interoperable," "transformational," or, most commonly, "network centric." Much discussion across the test community has outlined the challenges and changes of testing within and across programs in the "test-as-we-fight" paradigm. With budgets tightening on most government programs, we can expect to see decreasing program size and increasing expectations for products to the end users that have been thoroughly tested and meet operational needs, that "do more with less," and that "get it right the first time." The test-as-we-fight maxim is expanding into non–Department of Defense (DoD), large-scale, complex System Of Systems (SOS) efforts, such as the Federal Aviation Administration (FAA) Next Generation Air Traffic Control System Modernization (NextGen) and Department of Energy (DOE) SmartGrid programs.

Boeing Test & Evaluation has extensively studied how we execute SOS testing requirements. We have thoroughly analyzed the data gathered and have discussed what didn't work and how to address shortfalls. In many respects, we have reinforced what the larger test community repeatedly sees: we need Test and Evaluation (T&E) embedded earlier, with a stronger voice, at the start of programs. This need is even more evident in SOS testing. To make the test-as-we-fight paradigm a reality will require developing knowledge bases that transcend programs. Success requires SOS capability that provides a core of subject matter experts and robust processes and tools, including the infrastructure necessary to support programs throughout the product test life cycle.

Shifting the culture: operational excellence

Achieving operational excellence in the context of test as we fight will be tied to how well we develop T&E engineers who can

- think systemically (i.e., who have a system-centric rather than platform-centric perspective);
- work collaboratively within the government and industry T&E community;
- influence how program offices fund resources required for distributed SOS testing in an environment under ever increasing fiscal pressure.

In Boeing Test & Evaluation, we are initiating a true evolution of culture in program structure as it pertains to the testing of complex SOS. Over the past decade, we have developed a type of tester who thinks beyond a respective platform to understanding inter- and intradependencies with other complex systems. Whereas T&E traditionally has had a limited focus within the test life cycle, we are redefining when T&E involvement in this cycle starts. Also, we are placing greater emphasis on earlier collaboration with Systems Engineering, especially on network-enabled programs, with a greater focus on meeting end-user mission requirements.

Boeing Test & Evaluation is executing an internal T&E project of SOS operational excellence development that is integrating SOS engineers who are working on U.S. Army systems with those working on U.S. Air Force systems. In each case, the engineers independently developed SOS skills through the evolution of testing in ever more complex scenarios. This integration required expanded creation of relevant (i.e., live, simulated, emulated) test environments. What we found were similar broad areas of experience, lessons learned, and SOS perspective. Also, in each case, successful T&E of systems is measured increasingly by the ability of the platforms to meet mission needs within these environments. The T&E skill set that will drive test-as-we-fight testing is increasingly found within the SOS programs. This evolving skill set needs to be recognized and cultivated to develop testers who understand the interdependence of programs and the need to integrate test planning jointly across complex SOS.

The test-as-we-fight model brings a further change—how our testers evaluate the system or systems under test. A true SOS has a nearly unbounded set of test criteria. As we strive to scale environments to test multifaceted missions, the tester's focus needs to ascend to the SOS world view and expand beyond traditional system specification testing. BT&E has examples of programs that met product specifications but fell short of meeting mission goals. Conversely, a system may fail specification testing but actually meet mission needs.

As T&E tests SOS in test-as-we-fight environments, we need to be cautious of overemphasizing thread testing. Successful thread testing does confirm that we can pass messaging and data through disparate systems verifying interfaces, but it isn't until we create a complex, relevant environment that we can stress the interdependent individual systems to verify that loading of the test represents how we fight. It isn't always about the scale of the environment, but rather how the system is loaded. Hundreds of simulated assets outside the critical test area of the System Under Test (SUT) create a large-scale environment; but adding, for example, 50 simulated nodes driving the critical chain creates a load on the SUT. The greatest challenge for the test engineer is in evaluating an SOS as the systems align (or fail to align) with each other inside a complex environment, or in how collective systems create unexpected emergent behaviors and interactions. The key to the test-as-we-fight construct continues to be our ability to understand, create, and refine complex testing environments that either align or create emergent behavior within a complex SOS. The next-generation tester must understand SOS interoperability to design tests that will capture the required interoperability data.

The role of the T&E engineer within the test-as-we-fight construct is increasing in importance and urgency. As the number of nodes in an SOS increases, the possible patterns of emergent behavior expand exponentially, creating further complexity and greater challenges for the T&E community in its effort to deliver tested capability to the warfighter. An integrated Developmental Test/Operational Test (DT/OT) government–contractor T&E team will enhance the development of a complex SOS into an effective, adaptable, robust, and continuously evolving system. Increasingly, the new breed of T&E engineers, through their greater understanding of the systems aspects of an SOS, will lead structured testing for integrated functional alignment and detection of unexpected emergent behavior. The responsibility of the T&E engineer to the warfighter has never been greater—the stakes in the performance of our complex systems have never been higher.

Interoperability: interprogram collaboration

Most of us have heard the adage "anything less than war is a test." A true test of the warfighter within a system and across systems (or SOSs) occurs in operation. The T&E community realizes that it must immerse systems under test into environments that will test their respective specification requirements and meet the intended mission.

We hear many success stories within programs of hardware development leveraging simulation, emulation, distributed testing, and robust DT/OT. To a lesser extent, we hear success stories of joint utilization. What we lack is a comprehensive plan within programs to address cross-program needs and support for expanded fidelity, enabling environments that are closer to test as we fight. Without such planning, we often duplicate resources and develop new test facilities (e.g., labs, ranges, system integration laboratories) and simulation tools when such facilities and tools already exist within other isolated programs. The T&E community is faced with increasing requirements for multifaceted mission-based testing. Programs with the most expertise on a system need to share capabilities in support of higher level SOS events, even though they may simply be a participant establishing a higher fidelity test environment for another program.

A key aspect of test as we fight is the asset capabilities available to the T&E community to create relevant complex environments. Often our programs support large-scale events where the fielded assets are brought to the test with limited or no SOS integration beforehand. Typically, our program support labs are funded to meet specific program needs without considering that they could also support distributed interoperability testing across platforms. Programs frequently see their labs as internal only to their own needs. There is an opportunity across government and industry to enhance resources that will connect programs and establish a robust, lab-based SOS testing capability. Such an approach would enable early thread-based testing between programs and give test communities the necessary resources to expand interoperability testing across labs. Bringing these assets into joint testing events would lower risk and enhance benefit. For successful implementation, the T&E community would need to help shape the value proposition of such an investment with program offices early in the product development life cycle. Funding for such SOS testing should be discussed in the context of meeting interoperability requirements within programs during development.

A challenge of test as we fight is found in test control, conduct, and data collection. Just as the SUT may not have been designed to work interoperably, it is likely that the data collection tools were not designed with commonality in mind. Government–industry partnership in data collection tools that are at least trusted, and at most validated, is an area well within the control of the T&E community. An excellent example of such a partnership in tool evaluation is the Central Test and Evaluation Investment Program (CTEIP) Interoperability Test and Evaluation Capability (InterTEC). The focus is on enabling T&E testing of systems that communicate digitally over networks. These common tools bridge phases of testing from component and system level to SOS testing. The partnership is also successfully addressing how diverse distributed government–industry teams can jointly test products within a common tool environment by addressing planning, execution, data collection, and analysis of large-scale distributed testing.

Boeing has been participating in government-led events over the past 4 years. Boeing Test & Evaluation has increased our engineering skill set to integrate, participate, and test in these test events. As a result, Boeing Test & Evaluation is evaluating the unique SOS test processes required to integrate distributed testing into traditional platform-based test plans.

The Joint Interoperability Test Command has certified a number of the InterTEC tools for the greater T&E community to use as a starting point for validation testing. This example of a government-led partnership could expand into the government developing a library of basic tools and models it provides to industry as government-certified and government-supported. This could lower costs across programs and create a consistent T&E evaluation environment. Such an environment is critical to creating a relevant test-as-we-fight environment, commonality in test conduct, and consistent data collection.

Integrating systems T&E with systems engineering

To address lessons learned from past SOS programs and the inherent complexities of designing, integrating, and testing such programs, Boeing is progressing toward implementing an enterprise-wide SOS T&E capability. The Boeing systems approach to test as we fight requires

- a clear understanding of the SOS mission and end-user objectives, requirements, and constraints;
- SOS T&E capability to be integrated and aligned with the systems engineering and integration processes throughout the SOS test life cycle, to develop and provide testable and viable verification methodologies;
- SOS T&E capability to provide the required skills, processes, tools, and test infrastructure to

execute the identified verification methodologies for the SOS being developed.

We must understand the fight before we can determine how to test as we fight. The mission and end-user needs and objectives must be comprehensive and clearly defined for the SOS products: development, integration, verification operations, and training. The needs and objectives can be declared *clear and understood*, when the customer, user, and contractor can clarify:

- what the SOS must accomplish;
- how well the required SOS functions must perform;
- how well the required missions must be performed;
- the environment in which the SOS must perform.

Gaining this level of understanding requires an integrated team of customer, users, program, and functional-engineering experts; not a structure of organizational, programmatic, or functional silos. The SOS engineering, integration, and test processes begin and end with this understanding, and all activities within the SOS test life cycle must satisfy the mission and end-user needs and objectives.

To ensure that we test as we fight and that we meet our customer's mission and end-user needs, Boeing enterprise SOS T&E capability must be integrated and aligned with the systems engineering and integration processes throughout the SOS test life cycle. Whether it is during SOS requirements definition, functional analysis and allocation, design synthesis, or integration of the SOS engineering process, SOS T&E must be able to inform, engage, and collaborate with all key SOS stakeholders to develop, define, and provide testable and viable verification methods. These methods can then be used for executing test for early and persistent discovery and mitigation of program risks throughout the product test life cycle.

To effectively test as we fight, SOS T&E must provide all the requisite verification skills, processes, tools, models, environments, and test infrastructure to execute the identified verification methods for the SOS being developed. In essence, SOS verification becomes the intersection of systems engineering and T&E. The necessary resources for a given SOS program are acquired from, and represent the best of, the Boeing enterprise.

The acquisition process: a new approach for SOS development

The DoD acquisition process identifies operational needs and capability gaps and then evolves requirements to fill those needs and begin the product life cycle. Resulting requirement documents eventually become a key part of the product and work description to which contractors propose solutions. Typically, these documents are sufficiently detailed to describe the system that the customer desires and still allow sufficient flexibility for contractor-unique development processes to bring the system through development and to user evaluation in operational T&E. Thus, the process provides fair competition and, with continued development customer involvement during the selected contractor's development process, works reasonably well in providing the desired systems. However, acquisition history indicates that the process incurs strain when used in the acquisition of very large, complex SOS.

Defining the operational need, analyzing the alternatives, and describing the desired product in contract requirements are well-practiced but time-consuming steps. During the product development process, the contractor performs requirements decomposition to align contractor processes and detail development approaches to the user needs. If the end user's operational need changes, the contract is amended for new or altered requirements to meet the evolving need. This acquisition process works well for many DoD products and especially for those with long product life cycles.

Acquisitions that couple a new system with SOS development are complex. In these acquisitions, the current system under development is only one of the moving parts. Other products that will compose the SOS are often also evolving at varying rates in response to other needs. Changes to these other systems are not part of the new SOS contractor's responsibility, and acquisition policy precludes the contractor from independent actions that could allow fast response to the changes. Thus, at least one of the SOS acquisition process stressors is the time inherent in the traditional acquisition process.

Traditional government assembly and evaluation of SOS based only on existing systems does not incur the overlay stress of concurrent system development. However, some recent acquisitions have coupled complex system development with SOS development in the same acquisition and contracts. This coupling places the contractor and the government customer in a disadvantageous position. When the response time of the acquisition process is overlaid onto an SOS that comprises multiple system constituents responding to multiple change requests, controlling the process is extraordinarily complex. Contractors encounter problems closing the loop to provide the end user with the desired capability when the capability evolves over the product life cycle. The complexity increases as the speed of end-user needs accelerates, yet the development team can change only through the stable, but slow, acquisition process. This delay, necessary to maintain contractual compliance, makes it difficult to meet the warfighters' evolving needs and often results in testing to contractually specified requirements that do not align to the contemporary test-as-we-fight demand.

One strategy to resolve the acquisition complexity, and to allow the response time necessary to improve test as we fight, is to decouple SOS development. The resulting dual acquisition timelines allow each to develop on individual tracks that satisfy their independent requirements. The complex system can be successfully acquired by means of the traditional acquisition and contract methods. Conversely, SOS acquisition, or system integration into an existing SOS, decoupled from large and complex system development can then be performed in a manner that allows more exploration and a quicker response to evaluate the integration of the evolving systems that compose the SOS to meet the user's emerging needs.

Driving the test-as-we-fight paradigm: a proposal for the DoD T&E community

To test as we fight requires bringing all government and industry resources to the table. The challenge is that the acquisition community allocates funds within program stovepipes. The evolution of test within programs results in capabilities being developed during a product life cycle to meet specific program needs. The focus is on a program-level solution, relying on test exercises to bring together significant capabilities from across programs for fielded-system thread testing. But as mentioned previously, although thread testing is good for determining interfaces between platforms of message exchange, it doesn't address the loading of the system. There appears to be a further need not only to connect some level of live assets but also to be able to load these test exercises as more representative of environments relevant to test as we fight.

The missed opportunity is with program development assets (e.g., labs, robust simulations, emulations, data collection tools) being integrated across programs for early interoperability testing, and then being developed for scaling robust interoperability testing in joint test events. This integration will require established programs to fund a level of resource, such as labs, to support across programs. It also requires a level of funding to leverage test engineers from across programs to test jointly during interoperability testing. To achieve that level of testing will require T&E program engineers to consider cross-program interoperability requirements earlier in the test cycle.

Supporting early cross-program interoperability does not require the acquisition community to fund differently in structure, but rather to modify requirement definition. We had an example of a program that had the opportunity to integrate with a government hardware lab through distributed networking. It was an informal demonstration event that was not attempting to evaluate test points. During the integration of a message thread, it was discovered that a fundamental interpretation of a protocol design standard resulted in the two platforms failing to interoperate. Regardless of the interpretation, the fielded system implementation was quickly adopted. Problem reports for this and other results were submitted within the program under development during the 2 weeks the systems were connected. After the demonstration was completed, the two programs didn't interact until further into the development cycle, when such anomalies were more difficult and costly to address. Even though all parties involved with the event realized that interconnecting a fielded systems lab with a system under development had great value in risk reduction and cost savings to the government, the event ended and was not pursued further. Why? Neither program was budgeted for such cooperative testing.

There is great benefit in cost avoidance and risk reduction in connecting new or upgraded programs to existing programs early and periodically during development. Our experience has been with programs having to emulate another platform, rather than being able to fund existing programs to periodically integrate with their higher fidelity assets. The structure for the acquisition community to fund is already in place, but the culture does not consider funding a persistent capability outside the respective program. Test as we fight isn't just about pulling together fielded systems, but about integrating new capabilities with fielded capabilities earlier and throughout the development cycle.

Boeing integrates our test engineers into the program development process much earlier. Boeing Test & Evaluation is coordinating with our systems engineers to right size the test program from the start, taking into consideration the entire product test life cycle. The SOS experience within the test organization has allowed a greater focus on asking the right questions of the government for mission-based testing requirements. In short, Boeing Test & Evaluation has a greater say early in the programs. We acknowledge the greater efforts of the T&E organization within the government to bridge all programs and services. We turn to government T&E for continued early involvement with programs to

- identify how funding choices can be leveraged across programs throughout the delivery cycle;

- influence program offices that fund plans to include capability for interoperability testing;
- continue to lead the greater T&E community in creating relevant joint test environments that greatly increase the ability to test as we fight.

Shifting to the test-as-we-fight paradigm

The role of the contractor appears to be shifting back to a traditional prime role, with the government assuming the responsibility of large-scale integrator. The success of SOS large-scale complementary programs will strongly determine the effectiveness of test as we fight. The government will be faced with the challenge of ensuring accountability across distributed teams with interface control and requirement definition. Lessons from multiple large test programs have revealed a need for greater integration earlier in the program. Within SOS programs, this has been termed the big "I" (for integration), denoting greater complexity when integrating multiple systems. These programs have stated that integration within a complex SOS requires the same level of readiness review as is common in platform-centric T&E. "Integrate as you test" is becoming the precursor of test-as-you-fight; a successful complex SOS starts with integration that considers the mission-level needs to be tested.

Conclusions

Test as we fight is not simply a goal of the T&E community; it is also a cultural shift in test philosophy, increasingly driven by complex SOS testing. It is not enough to deliver a federated program to the warfighter; that program must also be tested for interoperability, incorporating a test-as-we-fight capability that has not previously been required. The key to success always begins with the T&E tester having the needed capabilities, tools, and processes. It also will require a shift in how the acquisition community supports the T&E community with integrated capabilities across programs, as we face changing SOS mission needs throughout the product test life cycle.

In Boeing Test & Evaluation, we are realizing great benefit in addressing how to organize our testers who have been integrating complex SOSs. We continue to expand our best practices in identifying and documenting data collection tools, models, and processes. We are ensuring that T&E is integrated and aligned with the systems engineering and integration processes, for early and persistent discovery and mitigation of program risk to ensure that we meet our customer's mission needs.

We encourage the government to continue defining capability that is approved, certified, validated, supported, and available across the government–industry team. We also recommend that the government consider ways to leverage the acquisition process in support of existing programs to enable SOS distributed testing. This approach could maximize existing government–industry assets with the cost savings of not requiring the development of additional redundant capabilities.

Finally, as we see test-as-we-fight fidelity becoming increasingly relevant earlier in DoD programs, it becomes obvious that the future won't stop there. Outside the DoD, the lessons learned over the past decade have direct application for such complex SOS integrations as, for example, the DOE SmartGrid or FAA NextGen, where the future mantra could be "test as we operate." The success of the greater T&E community in changing the test culture through test as we fight will directly affect the warfighter and will ultimately provide ancillary benefits that contribute to the betterment of the nation at large. ❑

DENNIS O'DONOGHUE is vice president of Boeing Test & Evaluation, part of Boeing Engineering, Operations & Technology. He is responsible for laboratory and flight test operations in support of validation and certification of commercial and defense products. O'Donoghue leads 6,500 engineers, pilots, mechanics, and technicians, who test and evaluate new Boeing aircraft, as well as modifications and upgrades to existing aircraft. O'Donoghue joined Boeing in 1996 as the lead test pilot on the Joint Strike Fighter Concept Demonstrator Aircraft program. He then became deputy project pilot for the Sonic Cruiser and 7E7/787 programs. He was heavily involved in design, development, and test on both programs. In 2004, he became chief pilot, Production Test Operations. O'Donoghue left Boeing in 2005 to serve as the director of Flight Operations and chief test pilot of Eclipse Aviation Corporation's Very Light Jet program. He returned in 2006 as vice president of Flight Operations for Commercial Airplanes, before leading Commercial Airplanes Flight Operations, Test and Validation. He assumed his current responsibilities in 2009. Previously, O'Donoghue was a NASA research test pilot at Lewis Research Center. He also served as a U.S. Marine Corps fighter pilot and test pilot and as a pilot in the U.S. Air Force Reserve. He commanded the 728th Airlift Squadron and the 446 Airlift Wing at McChord Air Force Base and retired at the rank of colonel. O'Donoghue has logged more than 6,000 hours in 81 types of fixed- and rotary-wing vehicles. He holds type ratings in the B-737, B-757, B-767, B-777, B-787, DC-9, G-159, L-300, L-382, NH-T38, T-33, and AV-L39.

Endnotes

[1]Boeing SOS Enterprise Workshop, Definition developed by consensus of people at the workshop. February 2011.

ITEA Best Paper

ITEA Journal 2011; 32: 411–418

Intelligent Test and Evaluation Infrastructure

Michael L. Curry and Nancy M. Sodano

Draper Laboratory, Cambridge, Massachusetts

Noe B. Duarte

Draper Laboratory, Yuma, Arizona

Limited budgets, rapid acquisition strategies, and complex system concepts place a significant burden on the traditional infrastructure for open-air and lab-based test and evaluation (T&E). Test ranges including White Sands Missile Range, Yuma Proving Ground, Natick Soldier Systems Center, and others have identified the need for a common infrastructure in order to perform T&E activities more efficiently. This article takes an information management approach to identify characteristics essential to the design and development of a common infrastructure for T&E. The vision is an "intelligent" infrastructure in which automated tools aid testers, developers, and decision makers in the collection, analysis, and dissemination of test information. This infrastructure provides a foundation for data-driven decision making and development of an integrated, end-to-end T&E strategy. The objective is to enhance the test community's ability to provide high-quality, low-cost capabilities to the warfighter in a time frame consistent with the operational tempo.

Key words: Rapid test and evaluation; Test automation; information management; test technologies; data to decisions.

The Department of Defense (DoD) test ranges have been challenged over recent years with the management of test and evaluation (T&E) information. Traditional methods are too costly with respect to time and resources, and they do not provide the necessary flexibility for rapid acquisition. Current and emerging technologies, however, can provide a cost-effective means to manage T&E information, increase productivity, and leverage prior knowledge to improve the T&E process. This article outlines an architectural approach to develop an effective information management system and identifies some applicable technologies.

In the early 1980s, flight testing of the Global Positioning System user equipment incorporated a very effective, yet manpower intensive, process for data management. The process was strictly controlled and included many checks and balances. A team of test professionals was employed to collect and distribute test data. Teams of analysts including government, military, and contract personnel performed data reduction and analysis to isolate performance anomalies. The success of this process depended on the resources of a large test organization, strict adherence to predefined processes, and significant oversight.

In the late 1980s and early 1990s, the volume of data grew as more subsystems were integrated onto common platforms. Analog data were gradually phased out with the introduction of digital data collection systems. However, digital collection systems only added to the volume of data. It was then possible to capture the entire MIL-STD-1553 data bus traffic, for example, providing all the intersystem communications information for detailed deficiency analysis. The complexities of managing data continued to grow to the extent that data clerks were hired for the sole purpose of coordinating, labeling, and tracking data across several supporting activities. Their role faded, however, with the introduction of robust analysis tools and personal computers, which allowed individuals to process data and analyze system performance as tests were conducted. As a result, information products were created in a distributed fashion, greatly improving the ability of test professionals to perform detailed analysis, but the complexity of the data management problem grew significantly.

After the end of the Gulf War, the DoD began a large drawdown of activities that supported the test

ITEA Best Paper from the Annual ITEA Technology Review, July 19–21, Annapolis, Maryland.

ranges. As a result of personnel reductions, the requirement to manage large volumes of data fell on the immediate test team. Test officers began developing customized databases for tracking and managing data on specific test programs, resulting in the creation of hundreds of highly specialized databases. Recognizing this problem, but still plagued by limited resources, test range management attempted to standardize the databases into a "one-size-fits-all" solution. This approach was received with great resistance by the test community because of the burden it placed on individual T&E professionals. By failing to standardize, programs such as the U.S. Army's Virtual Proving Ground eventually forced the test ranges to find more efficient data management solutions. Aberdeen Test Center (ATC), for example, created a system called VISION that provided tools for management of administrative documents and test data focused on automotive testing. This was a significant breakthrough for ATC as it automated several of the data collection processes for automotive testing. Despite the Army's attempt to make VISION the standard for data management across the Army's developmental test enterprise, the process proved to be too cumbersome for widespread adoption. This was again due to the burden placed on individual T&E professionals.

The attacks of September 11, 2001, created a new era in T&E. Testing that previously required 6 to 12 months to plan and execute needed to be accomplished in weeks or days. Rapid programs, such as Stryker, were successful in leveraging the resources made available by the War on Terror. Existing T&E tools were used in a brute force manner, requiring extensive manpower. The widespread and effective use of the Improvised Explosive Device (IED) in Iraq introduced a new challenge for the test community. Soldiers were being severely injured or killed at an alarming rate. New countermeasures were being developed and tested at a rate that would grow to more than 40 per month. Each system was of a different design, and in many cases, the systems were based on widely different technologies. Financial resources were not a limiting factor, but test ranges were required to act quickly—planning was now down to days. Test managers quickly put in place commonsense practices for managing data, but few leveraged the technologies that were available. The DoD did not expect that the IED threat would persist as a long-term problem; so, little emphasis was placed on data management and archiving. Custom databases were developed to track test logs, photographs, system data, etc. Only in cases where there was widespread testing of a single type of system (e.g., Counter Remote-Controlled IED Electronic Warfare [CREW]) were testers able to standardize data collection and analysis.

The DoD is facing another drawdown of budgets and personnel that will tend to further fragment test activities that are already distributed and somewhat isolated across the test ranges. On the other hand, the DoD is seeking to maintain its position within the geo-political environment through innovation; relying on automation, adaptation, interoperability, and integrated system concepts such as system-of-systems. Technological advances and increased complexity associated with modern defense systems only add to the demands already placed on the test community. The stage is set for the challenges of rapid testing, the realities of limited budgets and personnel, and the complexity of modern defense systems.

In the remainder of this article, the concept of an intelligent T&E infrastructure is introduced and developed with the objective of enabling data-driven decision making within the T&E community without placing the burden on individual test professionals.

Intelligent T&E infrastructure

The intelligent T&E infrastructure described herein is intended to help alleviate the burden placed on the T&E community by limited budgets, rapid acquisition strategies, and complex system concepts. Budgetary constraints are expected to fluctuate as priorities shift, which suggests that a successful T&E infrastructure must be able to adapt to continually changing requirements. Rapid acquisition emerged in response to irregular and asymmetric threats in which our adversaries demonstrated an ability to operate well within traditional acquisition timelines. A successful T&E infrastructure must be able to accommodate the compressed timeline associated with irregular and asymmetric threats. The evolution of modern defense systems has led to systems that are more automated, more adaptive, more interoperable, and more integrated than their predecessors. These systems often conform to tasks and environments through a cooperative association of component capabilities. Although individual components can be tested, their collective capability depends on their context and connectivity, which presents a challenge for traditional T&E processes that seek to test systems in tightly controlled test environments. A successful T&E infrastructure must be able to address the complexity of modern defense systems.

Figure 1 provides an overview of an intelligent T&E infrastructure. The primary activities supported by the infrastructure are testing, analyzing, and reporting. In addition, any new infrastructure for T&E must provide a means to interact with legacy systems such as existing

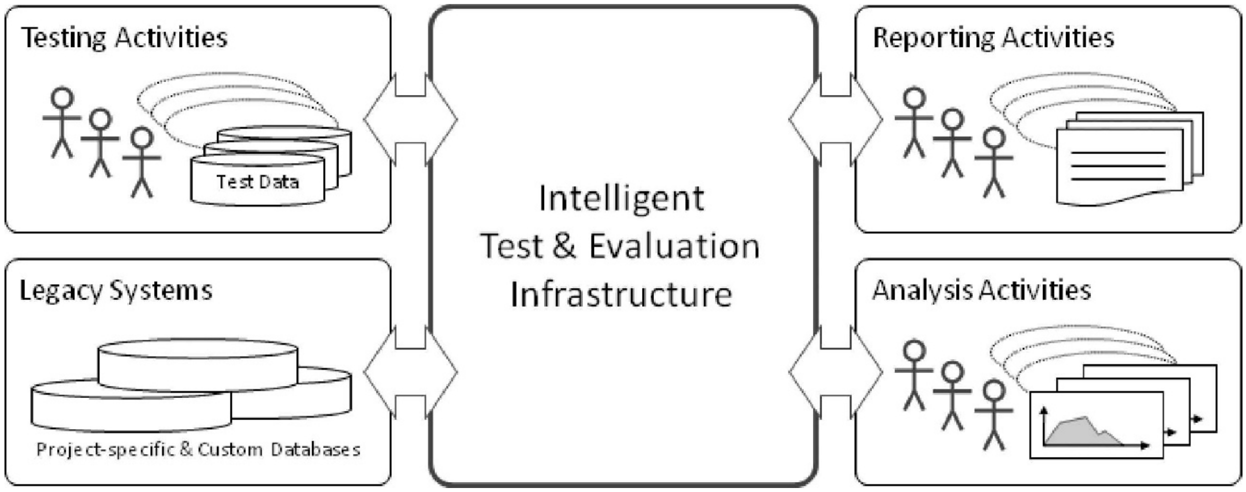


Figure 1. Overview of the intelligent test and evaluation infrastructure.

project-specific and custom databases. Testing activities require a great degree of flexibility. In particular, an intelligent infrastructure must be able to (a) align with best practices within the T&E community, (b) augment the capability of individual test professionals without introducing significant overhead, and (c) adapt to the evolving needs of the test community. Analyzing the data associated with a given test is often complex. For example, the standard for evaluating a system's operational capability is determining the "true" system performance under tactically realistic conditions. Correlation of system performance, ground-truth data such as time-space-position information, meteorological information along with near-earth radio frequency propagation characteristics, and the impact of operations in and around urban features becomes overwhelming without comprehensive data correlation and analysis tools. One of the objectives of the intelligent T&E infrastructure is to facilitate analysis of individual test data, as well as enable analysis across tests. Reporting should be comprehensive and must provide sufficient information to justify the methods as well as the conclusions. Comprehensive reporting often requires additional time and resources, which are typically in short supply. The intelligent T&E infrastructure must ensure that critical information is available in a format suitable for automated and semi-automated reporting. Implementation of an intelligent T&E infrastructure, such as the one depicted in *Figure 1*, is not a "silver bullet" that will revolutionize test and evaluation overnight. It will, however, continue to evolve DoD's T&E enterprise by increasing efficiency, by simplifying the mundane tasks involved in test planning, analysis, and reporting, and by improving collaboration across the full spectrum of T&E organizations.

Current trends within the T&E community suggest that test and evaluation is becoming integrated throughout the development process. Testing will be used increasingly as a tool of design and development, risk identification, and verification and validation. At the heart of these trends is the concept of data-driven decision making. Although the idea of quantifying unknowns through testing is not new, limited resources, accelerated timelines, and increasing complexity suggest the need for an integrated, end-to-end T&E strategy. The intelligent T&E infrastructure is intended to provide a foundation for data-driven decision making and to enable an integrated, end-to-end T&E strategy through the effective management of test information.

Next, each of the primary activities is considered from an information management perspective (i.e., by identifying information requirements, key decision points, and risk potential).

Testing activities

The general structure of testing activities is illustrated in *Figure 2.* Testing activities include test scheduling, developing a test plan, executing the test, and collecting test data. Test activities may also include determination and selection of test objectives, particularly for developmental testing or when critical information is unavailable. The information required to support testing activities includes test objectives (when available), characterization of the system being tested, and characterization of the test environment (including test equipment and test facilities). Additional information including the operational concept, operating environment, and performance requirements may be required, particularly when test objectives are not available. Critical information is also derived

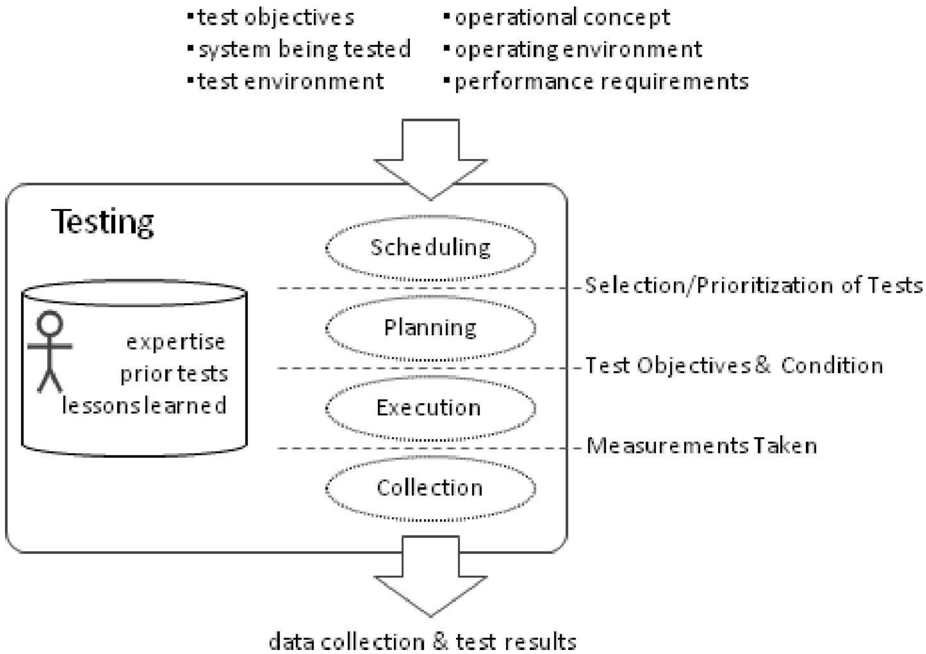


Figure 2. Information management perspective of testing activities.

through the skill and experience of individual test professionals (e.g., technical expertise, prior test results, lessons learned). The key decision points include the selection and prioritization of test activities, the selection of test objectives and test conditions, and the measurements taken during the test. Risks associated with testing activities are significant. If testing is not properly scheduled and planned, the results may not provide the information necessary for an assessment. If test plans are not well designed and executed, critical information may be missed or lost. If proper care is not taken during collection, results may be corrupted with missing or erroneous data.

Successful execution of testing activities frequently depends on the skill, experience, and expertise of individual test professionals. Availability of information regarding the current test, as well as other related tests, provides the best opportunity to augment the capability of test professionals in this area. It is also essential that the information be provided without introducing excessive overhead. In particular, an "intelligent" T&E infrastructure would automatically capture critical information from disparate sources without interfering in ongoing test activities. For example, testers produce products such as daily situation reports or data logs that could be archived automatically and indexed for future retrieval.

Once test information is captured, significant benefit is gained by making it accessible to T&E personnel. Minimal impact on the test community is achieved by using common language queries, which can be initiated without an in-depth knowledge of the underlying databases. For example, a system evaluator developing a plan for a new radar tracking system could search prior tests of similar technologies. Prior tests would aid in the development of a new plan by identifying which analysis methods had been applied in the past along with their associated results. In addition, a record of lessons learned would enable the system evaluator to improve the current plans based on prior testing regardless of where it was conducted and by whom.

Analysis activities

The general structure of the analysis activities is illustrated in *Figure 3*. Analysis activities include formatting the data, developing and applying analysis tools, generating and visualizing results, and interpreting those results to make an assessment and/or recommendation. The information required to support analysis activities include general test results, the test data collected, and observations (if available). Additionally, the context of the analysis is established using information from the test plan (as it was executed), the test objectives, and the particular conditions under which the test was executed. The key decision points include the selection of analysis tools to support the particular test objectives given the data, selection of critical results that elaborate on the test objectives, and prioritization of those results for subsequent reporting. The risks associated with analysis activities can be significant. If formatting and data reduction are performed without a complete understanding of the

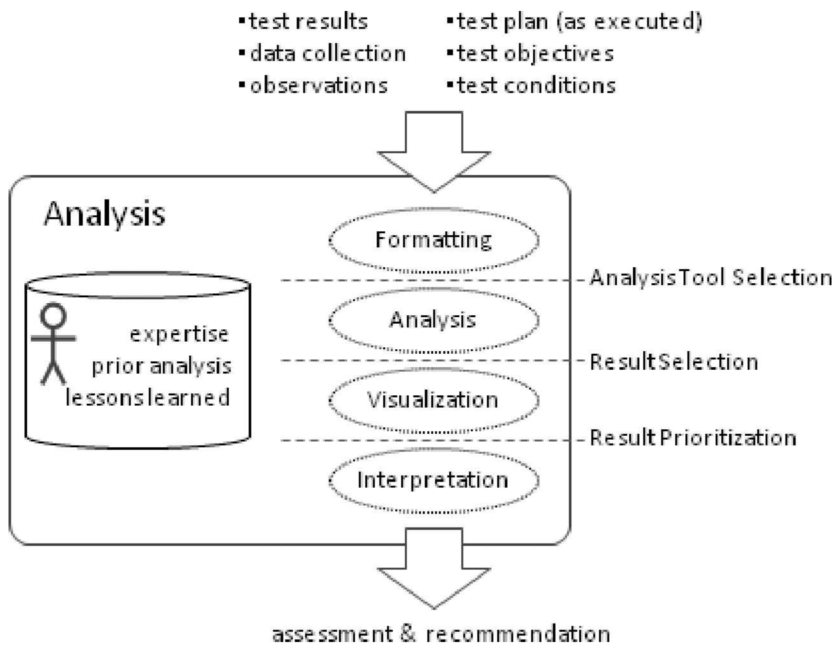


Figure 3. Information management perspective on analysis activities.

test and the test objectives, important data could be lost. If analyses are not carefully planned and executed, critical information and insights could be lost. Finally, if the results are not properly represented or interpreted, the value of the assessment and/or recommendation is diminished.

Analysis activities are critical to the interpretation of test results. Successful analysis depends on the quality of the data and the skill with which the analyst is able to extract information and draw conclusions. Availability of information regarding the current test, as well as related tests and analyses, provides the best opportunity to augment an analyst's ability to transform test data into meaningful information from which valuable assessment and conclusions can be drawn. Inputs to the analysis activities draw from many sources of information throughout the test program in order to enhance knowledge of system performance, isolate vulnerabilities, and identify operational risks. The objective of the intelligent T&E infrastructure to improve availability of test information directly supports data to decision initiatives throughout DoD.

Reporting activities

The general structure of the reporting activities is illustrated in *Figure 4*. Reporting activities include

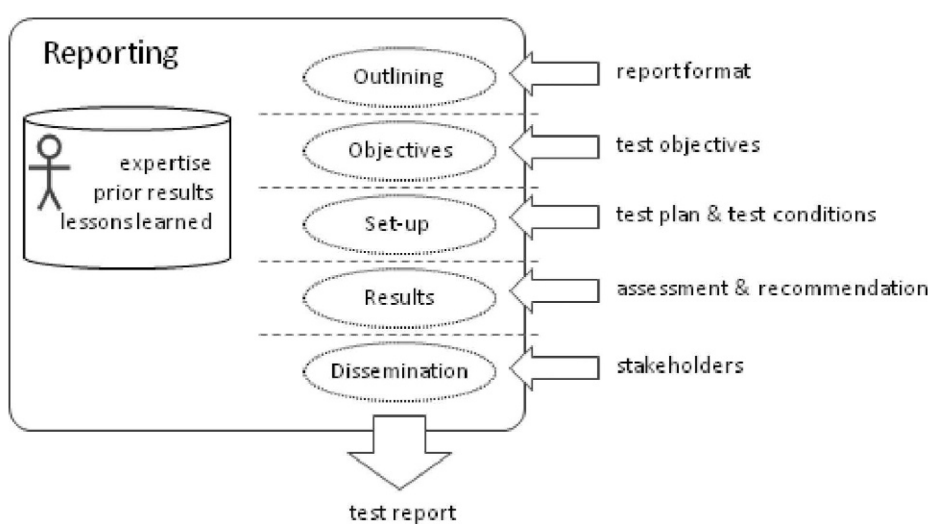


Figure 4. Information management perspective on reporting activities.

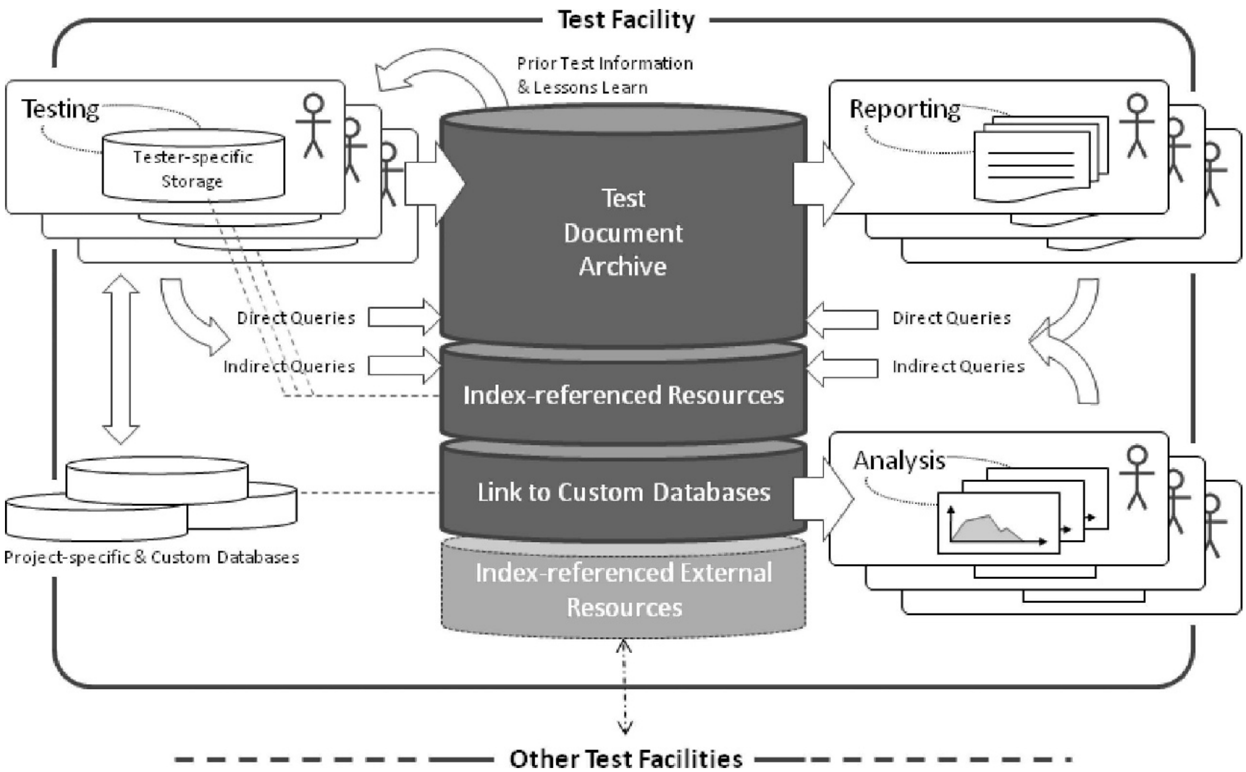


Figure 5. Facility operations with intelligent test and evaluation infrastructure.

development of an outline based on standard reporting practices, identification of the test objectives, description of the test and the conditions under which the test was executed, summarization of the test results, and dissemination to relevant stakeholders. Most of the information required for reporting activities will have been developed as a results of other activities. For example, the reporting activity compiles the following information: (a) the test objectives including any modifications, (b) the test plan (including any variation that occurred during execution), (c) the conditions under which the test was executed, and (d) the associated assessment and recommendation. In addition, the reporting activities must have information on standard reporting practices and the mechanisms for dissemination of the report, including key stakeholders. Key decisions are made throughout the reporting process. It is critical that the test report accurately portray the test, the objectives, and the results. Furthermore, the test report should provide some level of justification for the methods used to execute the test and analyze the results. The risks associated with reporting can have a significant effect if information is not clearly presented to appropriate stakeholders.

Reporting activities are critical, as they produce a document, which frequently becomes the lasting representation of the overall test process. Successful reporting depends on the quality of the information available and the skill with which the report is generated. Availability of information regarding the current test is critical, but supporting information that can be used to provide motivation or justification within the report is also essential. Providing this information, along with tools to automatically, or semi-automatically, generate inputs is the best way to augment reporting activities.

Reporting activities can extend well beyond individual test results. Frequently, custom reports are requested in which specialized information is required. For example, when systems are modified, regression testing requires test professionals to "look-back" at prior tests to determine how the modification affects the system. Another example is when systems that have been tested fail. In this case, inquiry into the particular failure would typically include further assessment, and potentially reevaluation of prior tests. An intelligent T&E infrastructure maintains fundamental relationships among tests, technologies, test environments, and test objectives. Doing so provides the information necessary to review, reconstruct, and reevaluate test information as necessary.

Building an intelligent T&E infrastructure

Figure 5 depicts the intelligent T&E infrastructure within the context of a test facility. This section identifies major functional components of the infrastructure and elaborates on critical interactions with testing, reporting, and analysis activities.

The primary objective of the intelligent T&E infrastructure is to collect, archive, and provide access to test information without increasing the workload on individual test professionals. This objective is accomplished in two parts. First, prior test information and lessons learned are made accessible to test professionals through a combination of direct and indirect queries. Direct queries are used to search a *test document archive* for test-related information. Indirect queries leverage *index-referenced resources* that provide a survey of test-related information not included in the test document archive. The results of indirect queries can be used by test professionals to request information or to add information to the test document archive. The second part of the objective deals with populating the index-referenced resources and the test document archive. Index-referenced resources are populated automatically by searching (as a background process) distributed and disparate sources of test information (e.g., local tester-specific storage). The test document archive is populated automatically using index-referenced resources. Critical documentation identified within the index-referenced resources is automatically uploaded into the test document archive. This approach establishes automated search tools that provide access to extensive test-related information without manually maintaining a centralized database, which can significantly increase the burden on test professionals and is frequently undermined by other priorities.

Access to test information through direct and indirect queries also supports reporting and analysis activities. As a safeguard, to ensure sufficient archiving, reporting activities should draw exclusively from the information within the test document archive. If information critical to the generation of a test report is not accessible within the test document archive, it should be added through indirect queries.

Analysis activities require access to all the same basic test information as testing and reporting activities. However, analysis activities must also access the collected data. Collected data typically require special handling. Even if the data are not classified or proprietary, the format and configuration of the raw data will impose constraints on how the data are stored and used, and what methods are required to process it. The intelligent T&E infrastructure does not attempt (at least initially) to move project-specific data stored in a customized database into a common format. Unit conversion, approximations, reference points, configurations, etc., would tend to impose project-specific process requirements even if the data were centrally located. The intelligent T&E infrastructure addresses this issue by providing a link to the data collected, whether in a file, a custom database, or some other format. Included with the link are comments and caveats associated with the data. A longer-term objective would be to standardize data collection. To the extent that standardized collection practices are adopted, test data could be maintained in a common repository where data from multiple tests can be combined.

Enabling technologies

Many of the capabilities required to construct an intelligent T&E infrastructure can be achieved using currently available or emerging technologies. The challenge is in the selection, implementation, and integration of these technologies to achieve the desired objectives without increasing overhead or introducing additional complexity. *Table 1* provides brief descriptions of technologies that may be used to achieve the desired functionality.

Conclusions

The T&E community tackles some of the toughest challenges associated with fielding of advanced military systems. The need to respond rapidly to warfighter needs, as well as reacting to declining budgets and personnel resources, continues to place a significant burden on test professionals.

This article takes an information management approach to identify characteristics essential to the design and development of a common infrastructure for T&E. The vision of an "intelligent" infrastructure was developed in which automated tools aid testers, developers, and decision makers in the collection, analysis, and dissemination of test information. A critical aspect of this information management approach is flexibility. Test professionals must be able to manage information in accordance with best practices at their individual locations. The broader test community must be able to access this information in a consistent manner regardless of its origin. Creating flexibility without placing a burden on a limited workforce introduces several technical challenges that can be addressed with a combination of existing and emerging capabilities:

- A flexible database schema improves the flexibility with which information can be accessed and annotated.
- Automated tools for populating and indexing data archiving systems provide access to critical information.
- Metadata tags improve traceability by maintaining important relationships.

Table 1. Enabling technologies.

Technologies	Brief description
Human system collaboration	Human system collaboration is an emerging field that incorporates cognitive and psychological knowledge into the design of human interfaces; it provides methods to ensure humans do what they do best, machines do what they do best, and communication is clear and unambiguous.
Workflow management	Tools are available that allow people to define collections of tasks to be accomplished, identify dependencies, and manage progress.
Provenance	"Provenance" refers to information about the origins of an object and the handling it has undergone. In the test domain, provenance may be used to identify non-standard procedures that were performed on a particular test.
Full content search	A full content search will find a query term in the body of a document, as well as in its title.
Latent semantic indexing	Latent semantic indexing is a technique for determining the conceptual content of a document and using that information to provide efficient retrieval of items of interest to the user.
Text analytics	An emerging field, text analysis automates extraction of information and knowledge from text documents, as well as such things as identification of the topic of a document or the author of a document.
Speech & text recognition	Devices and software are available commercially that convert human voice and scanned documents into a machine-recognizable form.
Database technologies	One particular example of database technology is a "flexible schema" in which relations database techniques are used to stored information that doesn't fit into the current data schema (e.g., annotations).
Meta-data tagging	XML is "tag" language that allows development of domain-specific vocabularies. IEEE 1636 is a test domain vocabulary specified in XML.
Visual analytics	Visual analytics provide ways for people to analyze large data sets by interacting with them via visual representations. This is especially useful when the data must be explored in order to draw conclusions.
Security	Many existing and emerging technologies are available to ensure that only people with the correct clearances/approvals are allowed access to particular materials. In addition to physical access control, these include various encryption techniques, authentication, and secure operating system kernels.

- Advanced query capabilities and full-context searches enhance discovery.
- Visual analysis tools ensure usability.

The "intelligent" T&E infrastructure described herein also creates an opportunity to further streamline the T&E process with the addition of decision support systems that aid in risk/opportunity analysis.

Streamlining access to T&E information, reducing the workload of individual test professionals, and utilizing emerging best practices will enhance the ability of the test community to provide high-quality, low-cost capabilities to the Warfighter in a time frame consistent with the operational tempo. ❑

MICHAEL L. CURRY is a principal member technical staff at Draper Laboratory, where he has developed techniques for risk and uncertainty management, architectures for complex systems, and model-based methods for system design and analysis. Prior to joining Draper, Michael developed and analyzed systems for Integrated Air & Missile Defense, Integrated Fire Control, and System of Systems at Raytheon Company. Michael's background includes systems engineering & architecture, system modeling, reasoning under uncertainty, and optimal control systems. E-mail: mike.curry@draper.com

NOE B. DUARTE is a senior member technical staff at Draper Laboratory, where he is a technical liaison to the government working out of Yuma Proving Grounds. His role supports research and development efforts that aid in test and evaluation of military systems including counter improvised explosive device (CIED) systems and intelligence surveillance reconnaissance (ISR) systems. Prior to joining Draper, Noe was a lead radio frequency systems engineer for the Electromagnetic Environment (EME) Generation team at the National Counterterrorism Counterinsurgency Integrated Test & Evaluation Center (NACCITEC). E-mail: nduarte@draper.com

NANCY M. SODANO is a principal member technical staff at Draper Laboratory, where she has extensive experience in development and testing systems for government sponsors, including manned and unmanned vehicles, communication and signal processing systems, and large data storage systems. Her current interests include emerging knowledge representation and processing technologies (e.g., text analytics, ontological concept recognition and inference, information fusion, hypothesis generation/testing, event prediction) and the advantages of integrating them with more traditional methods (e.g., physics-based simulation, predictive models). E-mail: nsodano@draper.com

The Art and Science of Orchestrating Free-Play in Operational Testing

MAJ Cornelius Leonard Allen, Jr.
Assistant Product Manager, Joint Air to Ground Missile (JAGM),
PEO Missiles and Space, Redstone Arsenal, Alabama
MAJ Melvin T. Mitchell
Test and Evaluation Officer,
Army Test and Evaluation Command Forward Operational Assessment (FOA) Team,
Bagram Airfield, Afghanistan
CW4 Kenneth L. Woods
Brigade Aviation Safety Officer/Full Spectrum CAB Operations Officer,
21st Combat Aviation Brigade, Fort Hood, Texas

Over the last decade, the Department of Defense has increasingly emphasized the need for operational testing of major defense equipment in a more relevant environment. However, very little guidance has been given to achieve this goal. This article addresses this deficiency, focusing on the complexity of creating free-play in operational testing across the full spectrum of warfare. Both the art and science of testing must be applied to achieve free-play in an operational-testing combined arms fight. A maneuver weapon system is used as a template (the science) for the employment of test resources: personnel, real estate, and equipment. These resources are addressed in four test elements—test team, test players, test area, and mission vignettes—to achieve operational realism from a bottom-up approach. Most importantly, the art of synchronizing these elements to create a representative combat environment is detailed.

Key words: Free-play; major defense equipment; operational realism; mission vignette; representative combat environment; test planning; test players; test team.

Operational testing (OT) is conducted to validate the system in a representative environment to determine whether a test article meets or does not meet the warfighter's needs. At its essence, OT is both a science and an art. There are four elements of operational testing that must be considered in crafting realism: the test team, test players, test area, and mission vignettes. The test team is responsible for manipulating the remaining three test elements in unison to achieve a desired result. Operating as an "unseen hand," the test team shapes the conditions for both free-play and data collection. Free-play—an exercise to test the capabilities of forces under simulated contingency and/or wartime conditions, limited only by those artificialities or restrictions required by peacetime safety regulations—is the state in which test-player commanders can employ their assets (personnel, equipment, and the test articles) in representative combat scenarios in accordance with their experience, doctrine, and techniques, tactics, and procedures. These conditions can only be established by the test agency's gaining test-player commitment (buy-in), securing the appropriate test area, and developing (and refining) realistic mission vignettes.

Test team—Don't plagiarize

The test team and its test plan are the foundation of a successful OT event. In testing, success is won or lost in the planning phase. Each test must be approached from a fresh perspective, dispensing with the idea that one size fits all. The cardinal sin of operational testing is attempting to duplicate a previous test or copying a test plan blindly. Plagiarism is a pitfall that hinders free-play and initiates complacency. The prevailing reason is that no two conflicts or threats are the same. Threat forces in combat have the cyclic cause-and-effect relationship: As one force adapts to gain a desired effect, the other force modifies its tactical, operational, and/or strategic methods to counter. Therefore, threats, mission vignettes, and environments from past tests are inherently irrelevant and should be referred to for historical reference only. Granted, past tests can provide valuable lessons learned about deficiencies and efficiencies; but direct duplication leads to "check the block" testing and a lost opportunity to stress the system. The representation of the current and emerging threat is the ruler by which the test article is measured. The test community cannot meet its objective by remaining static while the battlefield and threat evolve.

Each test team member must have a clear understanding of his or her roles and responsibilities in the overall test concept. The four primary operational-test positions are test officer, operations officer, observer controller (OC), and test-team liaison officer (LNO). Each member has a unique role that is interlinked with the other team members in crafting realism in an operational test. It is imperative that all test participants understand that the overall objective of testing is to gather usable data about the test article in a relevant environment, not testing players.

The test officer is the overarching authority in the execution of the test, supervising the big picture, charged with overseeing all resourcing, coordination, financial management, and diplomatic activities. Throughout all phases of a test event, the operational test officer is the face of the test agency and on the blame line. As such, the test officer must directly coordinate with the multitude of entities that contribute to a test event. Serving as a diplomat, the test officer must conduct site surveys, deliver readiness briefings to stakeholders, and interact with both civil and military VIPs. Test-officer interaction with VIPs and stakeholders is especially important during the test event to prevent interference and unnecessary changes. On joint programs (multiple services or nations) this is especially important, because the observers can far outnumber the test team and have a political influence.

Though more prestige is placed on the test-officer position, the operations officer produces the preponderance of the effect in achieving realism. The operations officer is the "all-seeing eye" who orchestrates the play-by-play action of the test, controlling every muscle movement from the control room. The operations officer's main objective is to enable and control the tactical convergence of the Red (enemy) and Blue (friendly) forces. From the test control room, he or she shapes the events tactically through military orders, intelligence reports, and battlefield factors. Administratively, he or she controls the initiation and sequence of events, as well as safety through OCs. Reliable lines of communication are the most essential asset of the control room and the execution phase of the test. Redundancy of communication capability is imperative to the operations officer's success.

Operations-officer products

Tactical

- Mission-event matrix (sequence of major events)
- Report formats
  - Intelligence reports
  - Battle-damage assessment
  - Artillery call for indirect fire
  - Medical evacuation request
- Operation orders/fragmentary orders
- Test-event matrix (meetings and movement of test team and support personnel)
- Significant-action tracker (event description, date, time)
- Standing operating procedures (SOP)
  - Unit SOP
  - Tactical SOP or battle book

Administrative

- Emergency contact information
- Equipment and supply status
- Daily range status
- Weather forecast
- Range facility SOP
- Lines of communication
  - UHF/VHF/FM radio
  - Satellite radio (beyond line of sight)
  - Landline phone
  - Cell phone
  - Handheld radio

OCs are the over-the-shoulder administrators and safety officers preventing safety violations from occurring. Serving as the "designers" of the battlefield, they ensure that the conditions are set before a test run. OCs serve as hands-on controllers during the initiation of test runs and contingency events (most dangerous), ensuring that the intent of the operations officer is clearly understood and safely executed. The greatest challenge is that the OC's duties must be conducted while maneuvering tactically, to prevent tipping either force's hand.

The OC's secondary responsibility is to protect the integrity of the test event. Arguably, the American service member's greatest assets are ingenuity and competiveness. These characteristics serve as both a benefit and a potential hindrance to operational testing. The benefit is the discovery of innovative test-article employment that was not envisioned by the developer. The hindrance is the potential skewing of the test data as service members become increasingly familiar with the system and identify ways to manipulate it. Therefore, control measures must be placed on the test player's natural tendency to adapt and overcome. If that ingenuity is not kept within bounds, the data collected will become compromised or invalid.

The last member of the team is an optional position, the LNO. This team member is utilized to reinforce the lines of communication, serving as the link between the organic test team: the test officer, operations officer, and OCs. The position's significance increases exponentially when remote test ranges or multiple test ranges are required. The LNO is the reality check of the test team, verifying that the perception of the operations officer is reality. He or she has the most influence over the battlefield and is positioned to capture data points from a bird's-eye view that would otherwise elude the test team *(Figure 1)*.

The most effective vehicles to achieve test-team cohesion are pre- and post-run coordination meetings. Exclusive test-team coordination meetings are essential in synchronizing the team effort prior to a test run and identifying areas that need improvement after a test run. The pre-run meeting is facilitated by the operations officer and is focused on synchronization of effort. The post-run meeting is led by the test officer; its purpose is to produce refinements to the next test run's mission vignette, schedule, and execution. Mission-vignette refinement enables the test team to maintain unpredictability; incorporate lessons learned, which provide more fidelity; and implement conditions to satisfy unforeseen data requirements and answer the questions of the data-authentication group.

Test area (real estate and airspace)—Don't test on a postage stamp

Geographically, the test area must accommodate the players and their required maneuver space. Beyond maneuver space, the maximum range of each weapon and sensor system must be accommodated. Special attention must be paid to the vertical element, especially in tests that incorporate aviation and artillery assets. Airspace often has dynamic limitations that are not as readily apparent as land limitations. Aviation and artillery must use a three-pronged deconfliction approach: location, altitude, and time. Most importantly, the test range must contain the same environmental characteristics and challenges as the intended environment.

Some characteristics to consider are temperature, humidity, moisture, vegetation, terrain, sand, wind, dust, snow, and illumination (sun- and moonlight). Each of these elements must be scrutinized to a sufficient level. For example, the type of sand (grain) has a largely varying effect on engine components; one desert environment is not necessarily the same as the next. In some cases the test ranges may have identical characteristics to those of the intended theater of operation, which can limit testing opportunities. Depending on the system's sensitivity to these characteristics, an appropriate number of backup range days must be scheduled. For example, since aviation is especially sensitive to adverse weather, one backup day should be scheduled for every five testing days—but this will vary based on location and season.

One scientific method for determining the minimum range size required for operational testing is to sum the maximum sensor acquisition range of each force and multiply by a factor of 2.5. This ensures that, from the onset, the opposing forces cannot observe, range, or engage each other without tactical maneuvering. Additionally, the area must allow for multiple avenues of approach, giving the combatants options and forcing decisions. This practice promotes planning and dividing of resources (economy of force). If the area is improperly selected, the opposing forces will be funneled to the same point for a frontal attack, displaying little maneuver tactics. An appropriate amount of the landscape must be navigable. Heavy armor vehicles cannot traverse mountainous terrain very well, for example.

Site surveys are critical before and after test-site selection. When multiple test ranges or facilities meet test requirements, it is important to apply metrics to test-area selection. Funding and preferences of agencies are involved. An important note to remember is that government test facilities are essentially in competition for each other's business. A test officer can use this to his or her advantage in cost and scheduling considerations. A test officer must be able to logically defend the test-area selection. If a similar test has been conducted in the past at a potential test location, there can be great resistance to change. A detailed scientific selection process with metrics can assist in overcoming this institutional inertia ("That's where we always test"). Training Circular (TC) 25-1 Training Land (2004) and Training Circular (TC) 25-8 Training Ranges (2010) are excellent guidelines for determining the metrics of test-range selection criteria by type of test article.

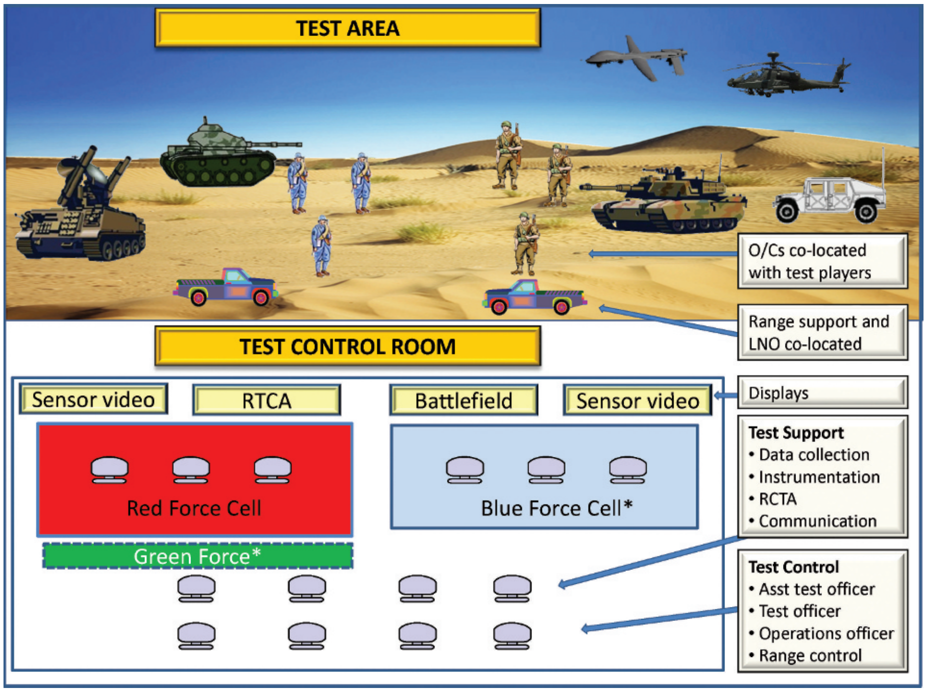


Figure 1. Test-range diagram. Notes: 1. Force-on-force vignettes require the Blue and Red forces' higher commands to be isolated from each other. The counterinsurgent fight, which is driven more by field commanders, allows for the collocation of the Blue and Red forces' higher commands. 2. The Green force can be collocated and overseen by the Red force commander when test resources are limited or the indigenous population is sympathetic to the Red force.

Test players—No golden crews

First and foremost, there should be no "golden crews": Test players and units must be representative of the fielded force (combat crews). These combat crews consist of personnel that habitually train together and are rostered to operate together in combat. Units inherently cross level experience; therefore, testers only need to record the demographic information in order for feedback to be appropriately attributed to the test player. Experience level can be divided into three categories, starting with the lowest: novice (newly trained), senior (journeyman), and expert. Golden crews consist of experts with additional skill sets or training that provides them additional insight and understanding of a system not normally found in a typical unit; examples are Army level standardization pilots, master gunners, and experimental test pilots.

Experience mix is a major consideration in test-player selection of operators of both the test-article (Blue) force and the opposing (Red) force. Characteristics worthy of consideration are age, military education, qualifications, and years of garrison and combat experience. Once the test unit is identified, a read-ahead or desk-side briefing should be conducted to exchange information. In the exchange, the test team will receive the unit SOP and other products that will assist the operations officer in the execution phase of the test (see "Operations-officer products"). Educating the operational unit on the test community and testing is an investment that pays dividends by dispelling misconceptions early. The level of test-player commitment is directly proportional to the information received and understood.

It is important that test players do not deploy with a vague impression of their duties. Each player must have a specific task and purpose to gain buy-in (commitment); otherwise only compliance is gained. This will reduce the occurrence of gamesmanship during test events. Gamesmanship is the manipulation of the system or conditions to gain an advantage that is not available in the normal application of the system. Up front, the test team needs to stress the importance of honest feedback from the users. The implication for the future equipping of the force is the takeaway.

The human factor plays a major role in operational testing. The mindset of the Red force commander is the most essential element of the test-player interaction. As the representation of the threat, the Red force is the medium to gauge the effectiveness of a system or a system of systems. Replicating the degree of competency, commitment, and objective of the Red Force is the key. Levels of commitment may vary within the force. These levels of commitment can be categorized as opportunistic (fight another day), tactical (seek to find, fix, and destroy), and sacrificial (true believer). These commitments can be directly linked to the mission and desired effect on the Blue and Green forces (civilians on the battlefield): harass, disrupt, or destroy. A test-team member—more specifically, an OC—should serve as the highest echelon of the Red force command. The Red force commander must completely assume the role when delivering the mission. A convincing sale of the plan, mission, and cause is required to prevent the Red force from developing a defeatist attitude.

The Blue force commander is responsible for executing the mission while incorporating the test article in a manner consistent with professional training, doctrine, and experience. To achieve realistic employment of the systems, test players must be trained to standard. Training packages should consist of three elements: academic material, hands-on application, and practical exercises. The academic phase consists of classroom instruction on system characteristics, theory, and operating procedures. Hands-on application is a check-on-knowledge phase that consists of test players' interacting with the test article. Practical exercises consist of test players' executing subtasks or mission components to gauge proficiency and understanding. This is an essential part of testing to ensure that test players' unfamiliarity (operator error) is not inhibiting the test article's effectiveness.

Replicate higher command and staff sections. The first-tier command-and-control (C2) element directly interacts with test players and must be organic to the unit. Surrogate test players (stakeholders) or test-team members can perform higher C2 functions that indirectly interact with the test players—which means that the higher's orders are translated into specific missions by the first-tier C2 element. When resource constrained, the first-tier C2 can role-play both functions during the test.

Civilians on the battlefield (COBs), referred to as the Green force in testing, completely reshape the battlefield and therefore the test scenarios. COBs add realism because they enforce the adherence to the rules of engagement. Their presence forces battlefield decision making by restraining test-player actions and providing consequences. They affect the tactical employment and rules of engagement of both Blue and Red forces. An important measure of effectiveness is not just if, but how a unit accomplishes its mission. The winning of hearts and minds is the main objective in the counterinsurgency battle, as opposed to the enemy attrition that it is in the traditional kinetic fight.

COBs are as prevalent as terrain features serving as obstacles and/or aides. Test players must factor into the decision-making process on a tactical, operational, and strategic level the following battlefield realities: historical landmarks and structures, religious customs and ceremonies, and hospitals. The reference to COBs: their density, age, and activity or inactivity serve as indicators of opposing forces' operations. For example, lack of females and/or children could imply an IED emplacement or ambush. Therefore, Green force test players' execution of these roles must be accurate and consistent. In counterinsurgent scenarios, fidelity is gained primarily through the Green force's role-playing.

Mission vignettes—The enemy gets a vote

Often mistaken as an opportunity for the user to play with the test article, OT mission vignettes are the validation of the system and/or system of systems in a representative environment. As such, due diligence must be dedicated to the crafting and execution of the test scenarios, given that mission accomplishment and service members' lives are dependent upon them. Mission vignettes are designed to measure the translation of a test article's functionality into a battlefield capability. Therefore scripted (canned) scenarios must be avoided. Turkey shoots only produce test-player boredom and complacency. This would be akin to poorly crafted gunnery exercises where the combat crews have memorized the location and sequence of the targets, reducing the scenario to a mundane drill. The *U.S. Army Helicopter Gunnery Manual* states, "If crews know exactly where targets are located, their proficiency in target acquisition cannot be evaluated" (Army, 2006). These scenarios have no operational value if the enemy is not a thinking and adaptive entity.

Mission vignettes must be allowed to run their natural course, within the bounds of the test. This freedom must be applied equally to the Blue and Red forces to achieve operational free-play. To maintain mission focus of the test players, the sequence of missions during operational testing must have a natural progression. A logical flow from one test run to the next should correlate to the period and time (or phase) of a combat operation. This provides the test players and all involved with a coherent battlefield picture and storyline. For instance, if a test article requires testing in full-spectrum operations, a slow ramping up in intensity and complexity can be accomplished using the following model: staging, skirmish, force on force, and stability operations.

Because test events for major systems can be long in duration (two months) and intense, the test team must implement appropriate control measures. Administratively, a slight decrease in complexity (workload) at the 85% mark should be planned to compensate for fatigue. A more dramatic or earlier decrease will promote complacency. Alternatively, this can be accomplished by a planned redeployment mission as the last test run. This will maintain mission focus beyond the major event and assist in test breakdown.

Implementation

Steps of operational testing

Pretest run

- Equipment check
- Refine mission vignette
- Issue warning order to Red, Blue, and Green forces
- Issue tentative schedule

Test run

- Test-team coordination meeting
- Issue updated daily schedule
- Issue equipment/uniforms
- Communication check
- Operation order (issue mission)
- Intelligence update
- Positioning
- Initiate vignette (operations officer or OC)
- Initiate contingency initiation (if necessary)
- End vignette

Posttest run

- Personnel and equipment recovery
- Initial data collection
- Mission debriefing/survey
- Data collection
- Test-team coordination meeting
- Issue next mission warning order to Red, Blue, and Green forces
- Issue next day's tentative schedule

Administrative and tactical battle tracking for the operations officer must incorporate the test player (Red, Blue, and Green forces) in addition to test resources (test-team members, range personnel, and equipment). In essence, the test control room is a war room that encompasses two or three independent war rooms. The initiation of a test run begins with the issuance of equipment, uniforms, and mission to the battlefield leader. During equipment issue, the control room conducts checks of systems, equipment, and communication to ensure test readiness. Battlefield leaders conduct a back (confirmation) brief to higher level commander to ensure understanding. OCs and LNOs must ensure that the disseminated information is complete and actionable. OCs then lead the test players to their start positions and report to the operations officer that test conditions are set.

Available battle-tracking assets

- Map board
- Blue Force Tracker
- Position-reporting instrumentation
- Real-time casualty-assessment system
- Air-traffic control (radar)
- Observer controllers
- Liaison officer

Once a test run is initiated, each force is allowed to conduct it mission without outside influence or advice. The only exceptions are for safety consideration and gamesmanship. Since no plan survives first contact, the operations officer must be prepared to interject when and where needed. This is where the art of OT comes into play. If the opposing forces are not converging and/or data requirements are not being met, the operations officer has two options: shaping the conditions and changing the mission.

Shaping the conditions is using battlefield factors to guide the forces down a path of decisions which leads to the desired test conditions. Individually the factors are benign, but when combined with other factors and mission information they drive the actions of the test players. For instance, introducing notional assets such as an unmanned aerial vehicle to conduct reconnaissance of a possible enemy position can result in expedited Blue force movement. A change of mission is the initiation of a preplanned contingency operation. It is a more intrusive approach, in which the test players' priorities or mission are explicitly changed.

Battlefield factors

- Restricted area
- Weather
- Additional assets
- Intelligence report

If there is a shortcoming in data fulfillment, the operations officer can convert a distracter to a shaping condition to initiate another sequence of events or a follow-on mission. Since the distracters have been previously introduced to the test players, this will produce a seamless transition. The goal is a gradual progression of a situation through a series of clues and indicators, not just one intelligence report of an imminent situation or event. Develop the story; build the situational awareness (the mental picture).

Introduction of imperfection and undesirable conditions are a reality of war. In *On War* Clausewitz refers to this as fog and friction (1873). Working against the test team is the range facility itself, because it is designed and maintained to support seamless operations. Therefore, it is important that the operations officer add distracters to make sure that communication, situation reports, and intelligence reports are not perfect.

Distracters

- Follow-on mission
- Imminent bad weather
- Out-of-area significant activity
- Broken radio-transmissions communication
- Incomplete/inaccurate intelligence reports

Conclusion

The importance of operational testing in a more relevant environment is well documented, but there is little guidance on how to achieve this goal. This article addresses the gap by outlining the art and science of crafting realism and enabling free-play in a major defense-equipment test event. The science is the employment of the four operational test elements, which are the test team, test players, test area, and mission vignette. A list of useful guides for the science of operational testing follows. The art is the intangible element of operational testing and can be best described as the deliberate and synchronized employment of operational-test elements to accurately replicate the test article's intended environment and utilization. The test team—more specifically the operations officer—are the artists responsible for orchestrating the remaining three test elements into a battlefield picture and setting the conditions for free-play. Free-play is the state in which test-player commanders can employ their assets (personnel, equipment, and the test articles) in representative combat scenarios in accordance with their experience, doctrine, and techniques, tactics, and procedures.

Useful resources

Test team

- Test Operating Procedures and Methodology (TOPM) 73-15 Test Team Organization and Responsibilities (2006)
- Test Operating Procedures and Methodology (TOPM) 73-110 Test Support Packages (TSPs) (2001)

Test area

- Training Circular (TC) 25-1 Training Land (2004)
- Training Circular (TC) 25-8 Training Ranges (2010)

Test players

- Field Manual (FM) 7-0 Training for Full Spectrum Operations (2008)
- Army Regulation (AR) 385-11 Intelligence Support to Capability Development (2007)
- Training Circular (TC) 90-1 Training for Urban Operations (2008)

Mission vignettes

- Field Manual (FM) 3-0 Operations (2001)
- Field Manual (FM) 3-06.11 Combined Arms Operations in Urban Terrain (2002)
- Field Manual (FM) 3-24.2 Tactics in Counter-insurgency (2009) ❑

Major Cornelius L. Allen Jr. is a U.S. Army acquisition officer serving as the assistant product manager of the Joint Air to Ground Missile program. His concentrations are T&E and aircraft integration. Formerly, he served as the T&E officer of numerous aviation systems, including the AH-64D Apache Block III, Common Infrared Countermeasure, and Mobile Tower System, at Operational Test Command, West Fort Hood, Texas. As an AH-64D aviator, he commanded Charlie Company, 3-3 Attack Reconnaissance Battalion. His awards include Senior Aviator Wings, Combat Action Badge, Meritorious Service Medal, and Air Medal. He has earned a bachelor of science degree in mechanical engineering (Tuskegee University) and a master's degree in aeronautical science (Embry-Riddle Aeronautical University). E-mail: cornelius.allenjr@us.army.mil

Major Melvin T. Mitchell is an acquisition officer assigned to the United States Army, Operational Test Command, Aviation Test Directorate, West Fort Hood, Texas. He is currently serving on an Army Test and Evaluation Command, Forward Operational Assessment Team in Afghanistan. As a UH-60 aviator, he commanded Alpha Company, 1–214th Aviation Regiment, 12th Combat Aviation Brigade. His awards include Senior Aviator Wings, Air Assault Wings, Meritorious Service Medal, Air Medal, Army Commendation Medal, and Army Achievement Medal. He earned a bachelor of architecture from Hampton University.

CW4 Kenneth L. Woods currently serves with 21st Combat Aviation Brigade, Fort Hood, Texas, as the aviation safety officer and full spectrum CAB operations officer. A qualified OH-58D aviator, his previous assignment was with the Operational Test Command as a military test officer; he has also served at positions from Company through Brigade as Standardization, Safety, and Testing. CW4 Woods's awards include the Bronze Star, Air Medal, Meritorious Service Medal, Army Commendation Medal, Army Achievement Medal, Master Aviator Wings, Combat Action Badge, and Air Assault Badge. CW4 Woods

earned a bachelor's degree in aeronautical science from Embry-Riddle Aeronautical University.

References

Army Regulation (AR) 385-11 Intelligence Support to Capability Development. 2007. Headquarters Department of the Army. Washington, D.C. January.

Clausewitz, Carl v. 1873. *On War.* Translated by Col J. J. Graham. London: N. Trubner.

Field Manual (FM) 3-0 Operations. 2001. Headquarters Department of the Army. Washington, D.C. 14 June.

Field Manual (FM) 3-04.140 (FM 1-140). Change 1. 2006. *U.S. Army Helicopter Gunnery Manual*, Headquarters Department of the Army. Washington, D.C. 17 April.

Field Manual (FM) 3-06.11 Combined Arms Operations in Urban Terrain. 2002. Headquarters Department of the Army. Washington, D.C. 28 February.

Field Manual (FM) 3-24.2 Tactics in Counterinsurgency. 2009. Headquarters Department of the Army. Washington, D.C. April.

Field Manual (FM) 7-0 Training for Full Spectrum operations. 2008. Headquarters Department of the Army. Washington, D.C. December.

Test Operating Procedures and Methodology (TOPM) 73-15 Test Team Organization and Responsibilities. 2006. Department of the Army, Headquarters, United States Army Operational Test Command (USAOTC). Fort Hood, TX. November.

Test Operating Procedures and Methodology (TOPM) 73-110 Test Support Packages (TSPs). 2001. Department of the Army, Headquarters, United States Army Operational Test Command (USAOTC). Fort Hood, TX. May.

Training Circular (TC) 25-1 Training Land. 2004. Headquarters Department of the Army. Washington, D.C. March.

Training Circular (TC) 25-8 Training Ranges. 2010. Headquarters Department of the Army. Washington, D.C. May.

Training Circular (TC) 90-1 Training for Urban Operations. 2008. Headquarters Department of the Army. Washington, D.C. May.

Acknowledgments

Ms. Gloria P. Bell, Joint Attack Munitions Systems.

JITC's DoD Interoperability Communications Exercise (DICE):

Joint Interoperability Testing and Certification in a Realistic and Relevant Joint/Combined Tactical and Strategic Environment

Richard Delgado, Jr., Alvin Slarve, and E. N. S. Christopher Brown
Defense Information Systems Agency (DISA) – Joint Interoperability Test Command (JITC), Fort Huachuca, Arizona

The Department of Defense (DoD) Interoperability Communications Exercise (DICE) is performed three times a year and is supported by the Joint Staff and the United States Strategic Command. Joint Interoperability Test Command (JITC) conducts DICE in support of DoD Joint interoperability testing, training, and transformation initiatives. DICE's primary purpose is to mitigate operational risk by testing new communication systems and software releases in a realistic Joint and Combined network. DICE has evolved over the last 20 years to remain the premier testing venue to address interoperability, training, and experimentation of critical communication systems for DoD, military Services, and other government agencies. The most recent exercise, DICE FY11, continues to provide a realistic and relevant Joint and Combined tactical and strategic testing environment, which allows system operators to focus on fighting the enemy, not their communications systems. The BATTLEGROUND should not be the TESTING GROUND!

Key words: Communications systems; interoperability; distributed testing; rapid acquisition; test automation.

In Fiscal Year 2011 (FY11), the Joint Interoperability Test Command (JITC) conducted three major Department of Defense (DoD) Interoperability Communications Exercise (DICE) test events, with the primary purpose of testing and evaluating the Joint interoperability of tactical communication systems crucial to the warfighter. The goal of each event was to ensure that systems met approved interoperability and Joint Staff requirements. Under the watchful eye of the JITC DICE team for FY11-02 *(Figure 1)*, the events provided the military Services, program managers, and other government agencies a venue to test deployable key communication systems designed to enable operators to focus on fighting the enemy, not their communications systems. DICE is dedicated to providing the most effective and valid test resources for exercise participants by replicating the real-world environments in which their systems must operate. DICE exercises have always focused on Joint testing and certification of communication systems in an operationally relevant network that spans the entire range of military operations. For over 20 years, DICE has been the DoD's exercise of choice for interoperability testing, training, and experimentation.

The JITC conducted the first interoperability exercise in 1989 in response to communications interoperability issues found during Operation Urgent Fury. That initial exercise, referred to as the Integrated Tactical Data Network Test (ITDN), focused on ensuring interoperability of Mobile Switch Equipment (MSE), Tri-Service Tactical (TRI-TAC), and Unit Level Circuit Switch (ULCS) switching systems. Even in its infancy, the exercise was focused on providing a realistic environment to test and evaluate mission-critical communication systems *(Figure 2).* Over time, the exercise has evolved to remain focused on relevant technical, operational, and interoperability challenges. From 1990 to 1997, the event was called the Major Switch Interoperability Test (MSIT) and was focused on tactical and strategic switching systems providing Time-Division Multiplexing (TDM) voice capabilities. From 1997 to 2000, the event was named the Defense Internetworking Test (DIT), and its focus was expanded to include networking technologies, cellular systems, and other transport technologies. Finally, in

Figure 1. Joint Interoperability Test Command Action Officers Daniel Granstrom, Kari Fisher, Capt. Patrick Akers (USAF), and Ellen Preiss attend a briefing on updated equipment and current testing efforts during a Department of Defense Interoperability Communications Exercise event.

2001, the exercise was renamed DICE, and since then it has continued to evolve and adapt to meet the DoD's testing needs for Information Technology (IT) and National Security Systems (NSS). Today's DICE test network replicates the operating environments used to support the war on terrorism, peacekeeping operations, humanitarian aid, and natural disaster relief, where complex interoperability issues are addressed among military and other federal, state, county, and local organizations. The DICE network continues to provide a realistic operational environment to address interoperability issues; allow warfighters to refine their Techniques, Tactics, and Procedures (TTPs); and allow military, first-responder, and non-DoD operators to interface with systems that may be called upon to support a real-world emergency. Even though the exercise continues to evolve and adapt, the principal mission and vision have remained constant.

The DICE mission statement is to serve as the premier Joint interoperability testing venue for DoD IT and NSS in support of rapid acquisition and fielding of net-centric warfighting capabilities. By no accident, this fully supports the JITC mission statement and various Defense Information Systems Agency (DISA) campaign plan objectives. JITC conducts DICE in support of DoD Joint interoperability testing, training, and transformation initiatives, with a primary focus on mitigating operational risk by thoroughly testing new NSS, DoD IT systems, and software releases in a realistic, Joint tactical and strategic network. DICE is the only DoD exercise with the primary goal of generating system-level Joint interoperability assessments and certifications to support the expeditious fielding of interoperable systems to the warfighter.

Today, DICE is executed three times per year to provide a constantly revolving window in which systems and organizations can participate based on their interface requirements and system maturity. At any point in the year, the JITC DICE team is completing the test reports from the previous event, executing the current event, or planning the next event. This scheduling process has proven to be very customer friendly and allows for maximum participation during DICE test events.

During DICE FY11, 20 systems participated in one or more of the three events. The efforts of DICE in FY11 are summarized as follows:

- Fourteen systems completed interoperability certifications and/or assessments.
- Five systems supported testing by connecting to systems under test.
- Five systems participated for training.
- Two systems participated for system research and development objectives.

Figure 2. Left, Joint Interoperability Test Command (JITC) Joint Test Facility A-Node set up for a Department of Defense Interoperability Communications Exercise (DICE) event at Fort Huachuca, Arizona. Middle, A U.S. Marine at Camp Pendleton, California, configures equipment used in the Tactical Data Network version 4 Data Distribution System–Modular (TDN v4 DDS-M) interoperability certification test effort. Right, A U.S. Navy sailor programs the AN/USC-60A satellite system in the JITC DICE lab located at Indian Head, Maryland.

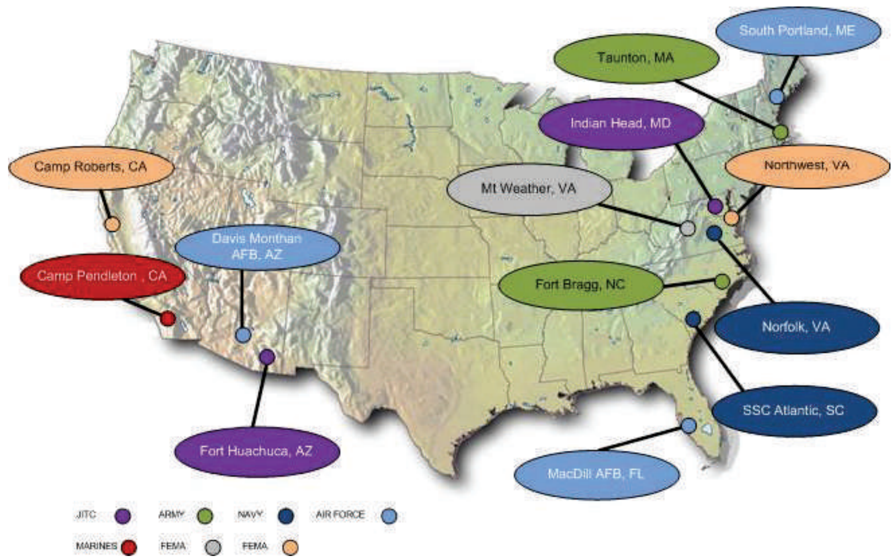


Figure 3. Department of Defense Interoperability Communications Exercise (DICE) distributed network for FY11-03. AFB, Air Force Base; FEMA, Federal Emergency Management Agency; FY, Fiscal Year; JITE, Joint Interoperability Test Command; Mt, Mount; SSC, Space and Naval Warfare Systems Center.

(Note: Some systems participated in multiple DICE events during FY11, and when the outcomes of all three events are combined, a single system may be counted more than once.)

Each DICE event in FY11 offered the significant advantage of distributed testing and encouraged participating systems to connect from their local facilities. In most cases, the JITC testing architecture can emulate most Standardized Tactical Entry Point (STEP) and/or teleport functions to allow point-to-point satellite connectivity. In other cases, the DICE team facilitated remote connectivity by preparing and coordinating Satellite Access Requests (SARs). This distributed testing environment decreased travel costs for all participating systems and contributed to establishing a real-world JTF communications architecture for realistic system testing.

Participants are invited to DICE by a U.S. Joint Forces Command (USJFCOM) tasking message requesting support of the scheduled DICE test events. In accordance with the Secretary of Defense memorandum to disestablish USJFCOM and realign its mission-critical functions, this responsibility was recently transferred to U.S. Strategic Command (USSTRATCOM). During DICE FY11-01 and FY11-02, the United States Ship (USS) *Wasp* and USS *Iwo Jima* were identified by USJFCOM as the U.S. Naval Forces representatives to the JTF communications architecture required for interoperability testing in DICE. The units were available for voice, data, and e-mail information exchanges during the exercise. Both assets provided capabilities and support to DICE for close-to-real-world tests resulting in accurate reports for system interoperability.

DICE FY11-02 included systems from the U.S. Army and Marine Corps, testing from various locations, including Fort Huachuca, Arizona, Fort Bragg, North Carolina, and Camp Pendleton, California. To ensure future interoperability during technology growth, JITC tested Internet Protocol Version 6 (IPv6) capabilities with certifying systems at Marine Corps Tactical Systems Support Activity (MCTSSA) *(Figure 2)*. The results from the voice and data testing over IPv6 native and IPv6/IPv4 mixed networks will be used in system development and additional systems testing for certification. This is a perfect example of sharing available resources to enrich tactical communications testing for participating systems to ensure the best results possible.

As the DICE testing year came to an end, DICE FY11-03 proved to be the most beneficial for U.S. Army systems. JITC was able to provide an integrated and distributed testing architecture, which included JITC facilities at Fort Huachuca, Arizona, and Indian Head, Maryland. This allowed the Warfighter Information Network–Tactical (WIN-T) Increment 1 program manager to access portions of the Coalition Test and Evaluation (T&E) Environment (CTE2) located at Indian Head. The CTE2 is an operationally realistic, distributed testing environment that emulates the Combined Enterprise Regional Information Exchange System (CENTRIXS)–International Security Assistance Forces (ISAF) (CX-I) network used for current operations in Afghanistan. This unique Joint

Table 1. Department of Defense Interoperability Communications Exercise FY12 schedule.

Event	Initial conference	Final conference	Exercise dates
FY12-01	March 16–18, 2011	July 19–21, 2011	October 31–November 18, 2011
FY12-02	July 19–21, 2011	October 18–20, 2011	March 5–23, 2012
FY12-03	October 18–20, 2011	February 7–9, 2012	June 11–29, 2012

and Coalition environment allowed the WIN-T systems to take advantage of an established infrastructure and emulate division to battalion deployment structures in an Afghanistan Mission Network (AMN)-like environment for certification testing. The WIN-T program manager saw DICE as the optimal arena for interoperability testing, evaluation, and system verification while pursuing an interoperability certification of the system. With support from JITC testbed personnel and the DICE testing teams, the WIN-T Increment-2 design team was able to modify the baseline system configuration, ensuring that other developmental changes would not raise potential interoperability issues in the future *(Figure 3)*.

In recent years the utilization of automated equipment to test electronic components and networks has become increasingly more commonplace. Automated test equipment improves the reliability and confidence of test results and provides the ability to fully expose system capabilities and limitations. Throughout FY11, DICE events promoted automated testing to provide a consistent test methodology with repeatable data-gathering procedures. JITC devoted a team of Automated Test Tool Managers (ATTMs) from its Test Instrumentation Branch to provide automated test tool support to systems participating in DICE. The ATTMs supported DICE by identifying the appropriate automated test tools and developing and updating standardized test procedures and test tool scripts. The ATTMs integrated six automated test tools into the DICE testing environment to capture analog and digital signaling as well as protocol-level metrics to provide objective test data. They also ensured that all software, including network control programs, satisfied the established test requirements based on the hardware and software baselines of the Systems Under Test (SUT). These automated procedures reduced costs for participating systems by removing human error and subjectivity from the testing process. Automated tools can generate thousands of call and data transactions in the time it takes to generate hundreds manually. This provides rigor to the testing process and raises the confidence level of the test results, ensuring that the system can and will perform to the standards required to support the users.

Though DICE FY11 events may be over, the planning for DICE FY12 events is well under way. The current schedule for FY12 is provided in *Table 1*.

DICE will continue to evolve during future events to meet the changing needs of our nation in the information domain. For more information, contact the DICE team by e-mail at dicefals@disa.mil. ❑

RICHARD DELGADO JR. is the lead systems engineer for the Battlespace Communications Portfolio at JITC. He oversees and integrates the testing processes, policies, and infrastructures for the certification testing of Unified Capabilities (UC). He has over 24 years of combined military and civilian service and has held various technical and supervisory engineering positions in the Department of the Army and the Department of Defense. Mr. Delgado attended New Mexico State University and graduated in 1991 as a Distinguished Military Graduate (DMG) with a bachelor of science degree in electrical engineering technology. E-mail: richard.delgado@disa.mil

ALVIN SLARVE is the current branch chief of the Integrated Communication Systems Branch at JITC. In this capacity, he is responsible for the planning and execution of DICE and other Joint interoperability testing and certification efforts. He has over 37 years of federal service and has held various technical and supervisory positions with JITC since 1994. He holds various associate of science degrees and graduated Magna Cum Laude from Western International University with a bachelor of science degree in information technology. E-mail: al.slarve@disa.mil

ENSIGN CHRISTOPHER BROWN (USN) is currently assigned as the JITC DICE Operations Officer. He entered Naval Service on August 19, 1998, advanced to chief petty officer on September 16, 2009, and was commissioned on September 1, 2010. Afloat, he has been assigned to USS Abraham Lincoln *(CVN 72) and USS* Chung-Hoon *(DDG 93), where he took part in Operations Southern Watch, Enduring Freedom, Iraqi Freedom, Rim of the Pacific (RIMPAC) exercise and Cooperation Afloat Readiness and Training (CARAT) exercise. Ashore, he has been assigned to Air Test and Evaluation Squadron Nine (VX-9) and Naval Computer and Telecommunications Area Master Station Pacific. E-mail: christopher.brown@disa.mil*

ITEA Journal 2011; 32: 431–438

Distributed Integration and Testing: Empire Challenge

MAJ Lee A. Brinker
Warfighter Support Division,
Joint Interoperability Test Command, Fort Huachuca, Arizona

This article discusses Joint Interoperability Test Command (JITC) involvement in joint-level technology developments and demonstration events, and the nature of distributed integration and testing, using as an example the annual Joint and Combined Intelligence, Surveillance, and Reconnaissance (ISR) demonstration event, Empire Challenge (EC). JITC has served a unique role in EC, which has grown significantly since the event's inception in 2004 at the Naval Air Weapons Station at China Lake, California, and through event relocation to Fort Huachuca, Arizona. JITC not only conducts technical assessments of new and emerging ISR systems and tests of more mature programs of record, such as the Distributed Common Ground/Surface System, but also hosts and supports systems participating in the event and provides test-network resources and expertise, all of which have proven critical to overall event success. This article focuses on two major aspects of JITC mission operations within the scope of the EC ISR demonstration: systems integration across a distributed event and systems testing across a distributed Department of Defense testing virtual private network architecture.

Key words: Distributed-event planning; Empire Challenge; information assurance; interfaces; networks; systems integration.

As technology continues to advance, new avenues for information flow are created, and with these new developments, new opportunities, threats, and vulnerabilities also evolve at a near equally rapid pace. As the volume and speed of information sharing continue to grow, so do the complexity of managing the raw data—both static and dynamic—and the need to ensure that useful information remains intact, securely archived, rapidly retrievable when needed, and able to be manipulated to serve a mission-essential purpose and reintegrated as derivative products.

Regardless of the nature and context of shared information or services, information exchanges, control of information, and access to information and services all continue to grow in complexity. This is especially true when applied to the United States Department of Defense (DoD), as it continues to integrate complex systems into operations—tactical through strategic—in the interest of supporting partnerships between DoD Services, with coalition military forces, and with other governmental and nongovernmental agencies. Underlying this effort is the common need to share information and service resources across an expansive and extremely complex network infrastructure, all the while protecting that information from others who would exploit it to their own advantage and potentially to this nation's detriment.

This article focuses on the macro (enterprise) and micro (individual components and systems) perspectives of distributed integration and testing. The ultimate objective of systems analysis—fully understanding both capabilities and limitations—is to determine the characteristics of a given system, the requirements for the system to function on a given network, and the system's intended products and services, so the system can be rapidly integrated to its full potential on an operational network. This, of course, is not accomplished by simply connecting a system to an active operational network and seeing how well it fares. The potential of developing systems is identified through careful desktop systems analysis, systems screening, preliminary assessment, and integration of systems into an environment that as closely as possible replicates the operational environment in which they will ultimately reside, once fielded.

Empire Challenge (EC) originally began as a demonstration event focused on determining the potential of emerging intelligence, surveillance, and reconnaissance (ISR) systems and methods to meet

DoD needs. The first EC event was conceived and conducted in 2004, under the leadership of the National Geospatial-Intelligence Agency, with intelligence representatives from DoD Services—the United States Army, Air Force, Navy, and Marine Corps—and system-level assessments conducted by the Joint Interoperability Test Command (JITC). Throughout the life of the event, EC has been sponsored by the Under Secretary of Defense for Intelligence to support development of intelligence community-of-interest (COI) tactics, techniques, and procedures as well as technical assessment of intelligence systems—legacy, emerging, and future—in an operationally realistic coalition environment. From the beginning, EC has been characterized as an assessment-driven event.

Since that first year, this annual event has grown considerably; it now includes several coalition partner entities, such as the North Atlantic Treaty Organization, the United Kingdom, Canada, and Australia. The focus has remained on ISR, but also now includes integrating ISR into command, control, communications, and computers systems to more encompass the full spectrum of joint and coalition ISR and such systems. Event leadership recently transitioned from NGA to United States Joint Forces Command, in order to bring more operational relevance to the event and better address operational requirements.

As the Joint Forces Command disbands, the future of EC and similar joint and combined events is still being discussed within the DoD, but the need for such venues is without question. Distributed integration and testing are called for not only at our own national level, but from many of our coalition partners as well, to ensure that when called upon to provide assistance, coalition systems will be capable of rapidly integrating with ours, on common networks, to serve the common purpose. Likewise, when one of our coalition partners calls upon the United States for assistance—whether out of a need for national-defense assistance, humanitarian assistance, or a disaster-relief response—our systems must be capable of rapidly integrating with their systems and networks and operating within their specific requirements. In such pivotal events, rapid response time is critical; and sufficient time for setup, testing, and configuration modification simply isn't available.

Distributed integration—requirements definition

A key consideration when integrating a system into operationally representative network architecture is identifying the requirements for that system and the architecture into which it will be integrated. For DoD acquisitions, these requirements will be defined by several well-established references. Adding coalition considerations can increase requirements considerably.

Initial requirements derive from the systems' or family of systems' intended mission or use. In DoD acquisitions these requirements are commonly referred to as "sponsor-defined requirements." Sponsor-defined requirements are critical to identifying many of the second- and third-order requirements, such as format and content, as well as interfaces and security enclaves. For example:

- Is the system intended to stand alone as a complete capability, or does it ingest products of another system, modify those products, and generate derivative products to be provided to another system or even shared throughout a COI?
- Does data modification alter the releasability of the product, potentially increasing or decreasing the associated classification level?
- Does the system operate on a single security enclave or multiple domains?

These are just a few of the questions that must be addressed in identifying mission-based requirements.

In addition to sponsor-defined, mission-based requirements, the DoD and the Chairman of the Joint Chiefs of Staff levy overarching requirements detailed in several directives and instructions, along with more specific requirements particular to the type of data or the specific COI. For example, in the area of still imagery, the DoD standard is the National Imagery Transmission Format (NITF). In EC, common imagery formats include NITF, motion imagery, and ground-moving target indicator. Additionally, the Net-Centric Data and Services Strategy (NC DSS) element of the Net-Ready Key Performance Parameter includes a multitude of requirements as specified in Chairman of the Joint Chiefs of Staff Instruction 6212.01E (CJCS, 2008). NC DSS requirements include nonstandard data, usually a form of Extensible Markup Language such as Cursor on Target, that would best be packaged into NITF- or motion-imagery-compliant formats. For less mature systems (generally preacquisition), use of simpler designs and configurations is much more common.

Another consideration for specific format requirements is the associated metadata tagging, which enables information about the data to also be transmitted as part of the packaged product and also enables end-to-end data traceability. This added information includes both mandatory and optional data fields, addressing such variables as the identification and location of the data source, data classification and releasability, and format, to name just a few. Metadata can be quite extensive and must be correctly tied to their associated data and registered with the DoD Metadata Registry to be considered fully functional. In the interest of maximizing network link efficiency, some systems have been known to transport data and their associated metadata as separate packages, but in many cases this can cause significant data loss, due to synchronization or other complications in reuniting the data with their associated metadata later in the information flow.

In addition to DoD general requirements, the network enclave the system is intended to operate on also has requirements that must be considered and addressed. These requirements can be very strict and often include information-assurance (IA) protection measures and specific software versions to ensure adherence to best practices. In many cases, local network security policy, based on DoD policies, dictates requirements for authority to connect, test, and operate.

The process of identifying and complying with network enclave requirements and ensuring full system compliance is the DoD IA certification and accreditation process (DIACAP). This process involves system developers' receiving IA requirements from the designated approval authority for the network enclave and answering an extensive checklist or questionnaire, which represents a substantial portion of the completed DIACAP package. The DIACAP package is reviewed extensively, and for testing purposes yields an interim authority to connect and test. The interim authority to connect authorizes the system to connect to the network enclave, as reportedly configured in the DIACAP package. Before the system can connect to the test network, though, it must also be scanned to identify any security vulnerabilities and potential deviations from the configuration reported in the DIACAP package.

The final consideration in fully defining the requirements for distributed integration is any additional requirements specific to the event itself. These requirements generally derive from specific considerations, such as event location, and simulated or scripted operational considerations within the context of the COI. These requirements also tie into an event operational script referred to as the Master Scenario Event List. All of these aspects should be captured in event-architecture products in order to best enable development of sufficient data-collection and analysis plans, technical assessments, and more specific testing.

Distributed integration—network interfaces

EC is by nature a distributed event, with participant nodes throughout the continental United States, Canada, Europe, and Australia. The network requirements to support such a distributed event are extensive, especially considering the multiple network security enclaves and added interface with operational foreign networks. Once individual system-interface requirements are identified, the task of designing a supporting network architecture within existing network capabilities moves to the forefront of the network-integration process.

To support EC, JITC utilized its Distributed Development and Test Enterprise (DDTE) virtual private network (VPN), which hosted the United States–only enclave. This enclave included 16 points of presence across the continental United States and provided connectivity for most of the programs of record conducting testing throughout the event. *Figure 1* shows the locations of sites the DDTE VPN supported in Empire Challenge 2011 (EC11). The DDTE VPN also interfaced with the Cross-Domain Enterprise All-Source User Repository (CENTAUR), hosted by JITC at Fort Huachuca. The CENTAUR system provided the event's central cross-domain solution—an information-exchange portal enabling data to traverse from one security domain to another, provided they were properly labeled to do so. Additionally, the DDTE provided interface with operational networks as needed to support EC objectives.

The DDTE VPN provided much of the infrastructure for U.S. systems, but additional resources were also extensively used to support the remaining requirements of multinational integration and other security enclaves. The Defense Information Systems Agency's Multi-National Information Sharing Program Management Office provided the Combined Federated Battle-Lab Network, with enclaves to include unclassified and classified for coalition partners. The network added significant infrastructure and complexity to the event, creating an environment that very accurately represented the interfaces of an operational environment. Due to the significant size of imagery files and intelligence products, one inescapable network-infrastructure artificiality was the availability of significantly more bandwidth than would doctrinally be available in an operational environment. This was necessary to best enable the systems to perform without being impeded by other systems also competing for the same network resources.

IA and security

As mentioned previously, IA is a key consideration in any distributed or operational network architecture. The process of IA systems validation can take several months, depending on the review process and the number of agencies involved. Key to success is having as small a review-and-approval chain as possible, to ensure minimal processing time of the many DIACAP

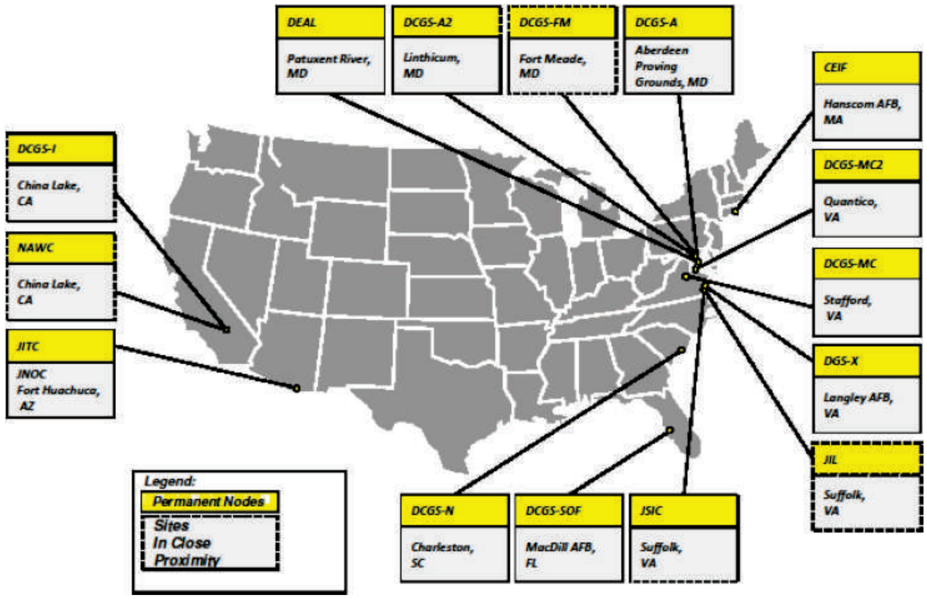


Figure 1. DDTE VPN locations.

packages and systems security-scan results, as well as extensive system details expressed in completed DIACAP submissions. The process of IA begins with initial system-specification reviews and continues through generation and approval of DIACAP packages and into final scans prior to connection of any system to the approved network enclaves.

During EC11 network initial setup and testing, JITC network technicians scanned over 150 systems and components for security vulnerabilities at Fort Huachuca prior to connection to the networks. These scans revealed many vulnerabilities, which were all addressed through software updates, interface-control management, and other mitigation strategies. Although many systems had significant issues identified, no system was denied access to the network enclaves once its vulnerabilities were adequately addressed. JITC also provided access to additional commercial network resources for the purpose of downloading and installing software updates. Once all modifications were made, the security scans were repeated to ensure that the updated, reconfigured systems were safe to connect and operate on EC11 networks.

CENTAUR played a key role in ensuring that the only data that crossed security domains were data that were properly designated to do so. To achieve this, the system implemented a high-speed guard and rule sets as defined by EC event planners. Although the requirements implemented and supported by the CENTAUR system were much more basic than would be used for operational networks, they still provided system developers a realistic idea of the configuration-management considerations they must keep in mind and of the potential impact to ongoing operations if their data were not properly configured to reach the intended destination.

The JITC NC DSS assessment team provided added tools for systems sharing data and services to register and validate their schemas with CENTAUR cross-domain solution rule sets. The NC DSS assessment team also provided a JITC-developed, EC-unique Universal Description, Discovery, and Integration registry, providing much the same functionality as a similar registry on operational networks. These two minor additions provided significant opportunities for capabilities to verify the integrity and viability of their data and services before attempting to test network traffic.

Once the networks were fully established and operational and systems were integrated across the event, standard DoD IA policies and practices were enforced. Monitoring systems and network security was a simplified task due to the isolated nature of the test networks. Still, these networks were required to adhere to all DoD IA standards.

Distributed assessment and testing

Much like the implementation of systems onto networks to facilitate their mission functionality, system assessment or testing is extensively planned based on key fundamentals. In EC, those fundamentals include many of the same considerations mentioned within the earlier discussion of distributed integration—specifically, applicable format and metadata standards, DoD guidance, and sponsor-defined and event-specific requirements.

It is important to mention key differences and similarities between assessment and testing, both of which can occur simultaneously in a distributed event such as EC. Assessment is often a preliminary observation, taking place when systems are still in early development. It is conducted to review a given system's intended functionality, data products, interfaces with other systems, and ability to integrate into a DoD network infrastructure. This is mostly conducted with the intent of identifying a way ahead for system developers to refine the system and compete for integration into DoD operations.

"Failure" is often expected, since most systems are developed commercially without initial consideration for DoD requirements, but the magnitude of that failure does not equate to that of a testing failure. It should also be noted that assessment and test failures often equate to event success. Thus, expectation management is also a critical aspect of planning and providing any feedback to participant systems and their sponsors. Assessment is a critical step for many developing systems to identify their requirements for transition from commercial off-the-shelf technologies to integrated systems that function in concert with DoD networks and other DoD standards-compliant systems.

Testing is generally much more rigid in structure and control and applies to more mature systems, generally programs of record. Tests are more tailored to specific requirements and often can involve isolating systems to identify exactly how they perform and react to given stimuli. In many cases, the system can be tested to the point of system failure purely for the purpose of identifying expected failures once fielded and developing methods for dealing with those anticipated failures before they occur in operational environments.

One could say that the ultimate objective of a tester is to force a system to fail, but the reason for doing this is to ensure that the system will function as designed and meet reliability requirements, as well as the requirements identified by other standards. Testing within a distributed event such as EC has limitations that must be addressed, due to the fact that tested systems are integrated into the same networks as assessed systems; pushing systems to the point of failure can cause a chain reaction, yielding unintended failures in other systems dependent upon interface with the tested systems. Because testing can cause added complications, test objectives must be properly considered within the distributed event to enable test objectives to be met and to avoid any reciprocal impacts on other systems or the networks.

In many cases, the requirements that drive an assessment and those that drive testing can be very similar; it is the results that can be expected to be vastly different. The added benefit of conducting testing within a distributed event such as EC is the ability to conduct end-to-end testing of an entire network of systems, along with enterprise-level testing for which there are few if any other nonoperational venues. With EC, this is illustrated by the flow of raw data from a vast array of sensor assets into managed data archives, access and exploitation of products in data archives, and update of the data archives to include those shared data products and services. Of course, the added complication of multiple security enclaves and multiple coalition partners helps add significant complexity as well, but also provides significant utility.

A last consideration on requirements: Special concern must be given to identifying and addressing conflicting requirements and recommendations to alleviate those conflicts, or at least properly account for them and prevent them from reflecting poorly on participants. One key example from EC was a requirement for basic information to be "burned in" to an image, making it part of the base image and rendering that portion of the image unusable, much like watermarks applied to copyrighted images on the Internet. This practice defies imagery format standards, but was implemented by event organizers to ensure data traceability. Such conflicts can negatively impact assessment observations and should be kept to a minimum if possible.

Table 1 identifies high-level requirements, based on DoD standards that drive assessment and test efforts. Identified are key reference standards for the focus areas of NC DSS, motion imagery, NITF, and ground-moving target indicators. JITC assessment teams used these requirement references to assess or test EC participating systems depending on system maturity according to acquisition category, system sponsorship, and governance objectives. It is critical to point out that testers do not define the requirements to which they test; the requirements are defined by the COI.

These technical format specifications drive assessment, but apply equally—if not more strongly—to testing programs of record. In addition to broad consideration of the requirements identified previously, more mature programs of record have higher-order requirements they must test to. Consideration of these sponsor-driven requirements is paramount to the success of these tests, and enabling that testing is at the heart of distributed testing.

In EC, the technical requirements drove another important aspect of assessment and test planning, the development of evaluation criteria. *Table 2* presents the evaluation criteria JITC standards laboratories used, as applicable, and presents sample results for all systems assessed in EC11. The listed criteria were developed

Table 1. High-level requirements and standards documents.

Assessment area	Requirements source
Net-Centric Data and Services Strategy (NC DSS; Element of Net-Ready Key Performance Parameter)	• Department of Defense (DoD) Information Enterprise Architecture (IEA) Version 1.1 (DoD 2009) • DoD Net-Centric Data Strategy (DoD 2003) • DoD Net-Centric Services Strategy (DoD 2007) • DoD Directive 8320.02, Data Sharing in a Net-Centric Department of Defense (DoD 2004)
Motion Imagery	• Motion Imagery Standards Profile (MISP) V5.5/North Atlantic Treaty Organization (NATO) Standardization Agreement (STANAG) 4609 standards (NATO 2007)
National Imagery Transmission Format (NITF)	• Military Standard 2500, Version C (MIL-STD-2500C), National Imagery Transmission Format Version 2.1 (NITF 2.1) (DoD 2006) • Legacy systems—MIL-STD-2500, Version A (NITF 2.0) (DoD 2006) • NATO Coalition systems—STANAG 4545, NATO Secondary Imagery Format, Version 1.0 (NSIF) (NATO 1998)
Ground-Moving Target Indicator (GMTI) Format	• Systems generating GMTI products—STANAG 4607 and Coalition Engineering Publication Document (AEDP)-7 (NATO 2008) • GMTI products embedded in other formats must be in accordance with the appropriate standards, such as MIL-STD-2500C for NITF or STANAG 4545 for NSIF (NATO 2008)

based on existing requirements, well in advance of the demonstration event, and presented to each participant system through the course of direct coordination of assessment planning.

Operationalizing distributed testing

Once assessment and testing requirements are identified, along with each system's planned operational considerations, development of an adequate stimulation plan is required. Ensuring that each system is sufficiently exercised, receives sufficient source data to perform its mission, and is operationally integrated into event scripting is the next key step in ensuring the achievement of assessment and test objectives for every event participant.

In EC, this function was performed by developing events from the Master Scenario Event List and integrating available simulations into operations whenever possible. The intent of this process was to ensure that each system was sufficiently prompted within the context of its intended mission support and was enabled to provide its products as planned as well as generate the data needed for assessment purposes.

Another key consideration is ensuring an approach that enables repetition as needed for the systems to attain sufficient stimulation. Often due to event complexity and troubleshooting throughout, connections to network enclaves may not function properly until well after the event begins, perhaps not until the final day of the event. A flexible event-operations plan ensures that stimulus can occur even if the system is only operational for a short period. JITC has developed tools for managing and tracking this process. Effective implementation of these tools is paramount. This process includes careful consideration of each system's mission objectives and location within the event and network architecture and the ability to capture sample data products remotely or at the system's physical location.

If requirements are sufficiently considered, the event will properly incorporate network-architecture considerations, data-sharing rule sets, and operational stimuli. Then the job of the assessment or test planner can be fairly simple, identifying required sample data products; location, time, and method of capture for those products; and required network instrumentation and sample data-product collection and analysis methodology. Even if all of these considerations are adequately addressed, a flexible and adaptive assessment or test plan is still required, since any number of nonevent incidents can affect or delay event execution. Test plans must account for all event constraints and limitations and all participant system mission objectives. Sufficiently detailed test plans include data collection, transport, management, reduction, and analysis. It is also very helpful to prioritize assessment and test efforts, with critical assessment objectives that must be met and "stretch goals" that would be good to meet but are not considered as important.

In the case of EC, data collection was conducted both remotely, from JITC laboratories connected to each of the security domains, and at physical locations, when the data needed to be pure for comparative analysis to determine what modifications were made by a particular system along an information pathway and whether those modifications were intended or incidental, constructive or destructive within the scope of the mission-specific requirements.

At the end of all of this planning, event-operations execution, and data collection, one high-interest deliverable is usually produced: the report of how each capability performed planned event participation. This is a highly sensitive product, not to be taken lightly; it

Table 2. Example evaluation criteria and results summary.

System	Net-Centric Data and Services Strategy	Motion Imagery (MI)	National Imagery Transmission Format (NITF)	Ground Moving Target Indicator Format (GMTIF)	Overall Rating	Critical Comments
A	Poor	Excellent	N/A	N/A	Poor	
B	Average	N/A	N/A	N/A	Average	
C	N/A	Excellent	Average	N/A	Average	
…						
X	Poor	Poor	Average	Poor	Poor	• Data not visible to other users • Motion imagery not standards-conformant
Y	Poor	Average	Average	N/A	Poor	
Z	Excellent	Excellent	Excellent	Excellent	Excellent	

LEGEND:

System	Net-Centric Data and Services Strategy	Motion Imagery (MI)	National Imagery Transmission Format (NITF)	Ground Moving Target Indicator Format (GMTIF)	Overall Rating	Critical Comments

Rating	Criteria: NC DSS	MIS	NITF	GMTIF
Excellent	Meets all requirements.	Compliant in all areas (transport stream, video (MPEG 2/H.264), audio, and metadata) as specified by MISP V5.5 NATO Standardization Agreement (**STANAG**) 4609), with added compliant features.	Data products conform to MIL-STD-2500C or STANAG 4545 and contain sufficient metadata to support product dissemination and functionality. No significant compliance issues or metadata deficiencies observed.	Data products conform to STANAG 4607 and contain sufficient metadata to support product dissemination and functionality. No significant compliance issues or metadata deficiencies observed.
Average	Meets some requirements. No critical operational impact.	Compliant in all areas as specified by MISP V5.5/NATO STANAG 4609, with no added features.	NITF data products conform to MIL-STD-2500C or STANAG 4545, but lack metadata to support dissemination or functionality.	GMTIF data products were conformant to STANAG 4607, but lacked metadata to support dissemination or functionality.
Poor	Failed to meet requirements. Critical operational impact.	Non-compliant with MPEG-2 TS, video compression, audio, or metadata, as outlined in MISP V5.5/NATO STANAG 4609.	NITF data products non-conformant to MIL-STD-2500C or STANAG 4545	GMTIF data products were not conformant to STANAG 4607.
N/A	Capability does not perform functions that apply to laboratory's assessment area.			

can directly impact program resourcing, render a potentially advantageous system untenable, or enable a system that should not continue to acquisition to do so. Assessments do not yield certifications of any form; they yield observations and recommendations. Tests can lead to certifications but just as often do not, because they identify additional requirements and considerations calling for further testing—an iterative process tied directly to product development.

In all cases, results must be factual, generating interpretations based on well-defined, known requirements. As an honest broker, JITC has this as one of its charters in supporting such events—ensuring that systems are accurately assessed or tested, based on sufficient integration into a complex network and a sufficient operational framework.

Summary

This article has touched on some key considerations for planning a distributed event focused on systems integration and testing. EC is an example of such a distributed event. This article discussed critical considerations defining requirements to facilitate systems integration and testing across a distributed network environment. Specific requirements applied to EC illustrated just some of the necessary considerations for such an event. Additional discussion presented the aspects of networks and interfaces and of IA. Finally this article discussed the art of using established requirements to generate assessment and test criteria and the need for developing a comprehensive yet flexible assessment plan. Through the course of this discussion, the dynamic of significantly increased complexity remained a common theme and a key consideration in any large-scale distributed event.

Conclusion

Distributed integration and testing are essential elements of ensuring that systems will function as intended as they are developed and make the transition from developmental or commercial off-the-shelf technologies to DoD systems. These systems must implement the proper format and policy configuration management to comply with all requirements. Very few events provide sufficient scope to enable such a diverse method of testing and allow for the flexibility required to effectively test in a distributed environment. EC has evolved into such an event, with all of the right considerations in place and the proper agencies involved in event governance and planning. It improves implementation of these principles with every event and can continue to provide an excellent annual opportunity to conduct distributed integration, assessment, and testing for ISR and command, control, communications, and computers systems. All planning for such an event must include considerations for each system's functionality, interface, and assessment objectives, in concert with the capabilities of the event and the supporting network architecture. ❑

MAJ LEE A. BRINKER is a joint test officer in the Warfighter Support Division, Joint Interoperability Test Command (JITC), at Fort Huachuca, Arizona. He has served in the United States Army for over 14 years. MAJ Brinker has been assigned to JITC since early 2009, and has coordinated JITC's Command-wide support efforts for Empire Challenge for three consecutive years. E-mail: JITCPAO@disa.mil

References

Chairman of the Joint Chiefs of Staff (CJCS), 2008. *Chairman of the Joint Chiefs of Staff Instruction (CJCSI), Interoperability and Supportability of Information Technology and National Security Systems*, CJCSI 6212.01E, December 15, 2008. Washington, D.C.: author. Available online at http://www.dtic.mil/cjcs_directives/cdata/unlimit/6212_01.pdf. Accessed November 14, 2011.

DoD, 2003. *DoD Net-Centric Data Strategy*. Memorandum by John P. Stenbit, May 9, 2003. Available online at http://cio.nii.defense.gov/docs/Net-Centric-Data-Strategy-2003-05-092.pdf. Accessed November 14, 2011.

DoD, 2004. *Data Sharing in a Net-Centric Department of Defense*. DoD Directive 8320.02, December 2, 2004. Available online at http://www.dtic.mil/whs/directives/corres/pdf/832002p.pdf. Accessed November 14, 2011.

DoD., 2006. *Department of Defense Interface Standard: National Imagery Transmission Format (VERSION 2.1) for the National Imagery Transmission Format Standard (NIFTS)* (16 FEB 2011)., MIL-STD-2500C, dated 01 May 2006. Suite of standards available online at http://www.everyspec.com/MIL-STD/MIL-STD+(2000+.+2999)/MIL-STD 2500C_2997/. Accessed November 14, 2011.

DoD, 2007. *Department of Defense Net-Centric Services Strategy*. May 4, 2007. The Pentagon, Washington, D.C.:Author. Available online at http://cio.nii.defense.gov/docs/Services_Strategy.pdf. Accessed November 14, 2011.

DoD, 2009. *Department of Defense Information Enterprise Architecture Version 1.1.* May 2009. The Pentagon, Washington, D.C.:DoD Available online at http://jitc.fhu.disa.mil/jitc_dri/pdfs/dieadoc.pdf. Accessed November 14, 2011.

NATO, 1998. *NATO Secondary Imagery Format (NSIF)*. STANAG No. 4545 (Edition 1) Amendment 1. Brussels, Belgium: NATO Military Agency for Standardization (MAS). Available online in English at http://www.nato.int/structur/ac/224/standard/4545/4545_documents/4545_ed_1amd1.pdf. Accessed November 14, 2011.

NATO, 2007. *NATO Digital Motion Imagery Standard*., June 15, 2007, STNAG 4609, (edition 2). Brussels, Belgium: NATO Standardization Agency. Available online at hppt://www.nato.int.structur/AC/224/standard/4609/4609_documents/STANAG_4069_Ed2.pdf. Accessed November 14, 2011.

NATO, 2008. *NATO Ground Moving Target Indicator Format (GMTIF) STANAG 4607, Implementation Guide*, April 2008 and Allied Publication document AEDP-7. Brussels: Belgium: NATO Standardization Agency . Available online at http://www.nato.int/structur/ac/224/standard/4607/4607e_JAS_ED3.pdf. Accessed November 14, 2011.

ITEA Journal 2011; 32: 439–445

Bringing Operational Interoperability Assessment to the Endeavors

MAJ Robert R. Houston
U.S. Army Joint Interoperability Test Command
Fort Huachuca, Arizona

Joint Interoperability Test Command (JITC) supports three "Endeavor" multinational interoperability testing exercises: Combined Endeavor (CE) for U.S. European Command, Africa Endeavor (AE) for U.S. Africa Command, and Pacific Endeavor for U.S. Pacific Command. The Endeavor exercises prepare multinational forces in each combatant command's (COCOM's) area of responsibility to form and operate as a coalition in support of full-spectrum operations. This year the focus of AE and CE has shifted to include assessing the operational interoperability of the various command, control, communications, and computer (C4) systems utilizing a scenario based on full-spectrum operations; testing of multiple command echelons using the mission thread model; and, for CE, a robust network assessment to include realistic traffic loading and simulated cyber attacks to determine the true capabilities of the coalition network. This major shift to an operational interoperability assessment of C4 systems, as opposed to a point-to-point technical testing focus, will be attempted for the first time at AE and CE in 2011 and will form the foundation of the exercises well into the foreseeable future.

Key words: Command, control, communications, and computer (C4) systems; communications network; information technology; interoperability guide; interoperability testing; multinational test exercises.

Joint Interoperability Test Command (JITC), a subordinate command within the Defense Information Systems Agency, supports three "Endeavor" multinational interoperability testing exercises: Combined Endeavor (CE) for U.S. European Command, Africa Endeavor (AE) for U.S. Africa Command, and Pacific Endeavor for U.S. Pacific Command. The Endeavor exercises help prepare multinational forces in each combatant command (COCOM) to install, operate, and maintain command, control, communications, and computer (C4) systems in support of coalition full-spectrum operations *(Photo 1)*. This is accomplished through technical interoperability testing of C4 equipment; development of interoperable tactics, techniques, and procedures (TTPs); and publication of coalition operation standards for the installation, operation, and maintenance of coalition networks providing services in support of dynamic operational requirements.[1] The Endeavors ensure that nations forming a coalition C4 network can respond rapidly and decisively without requiring excessive time to setup, test, and modify the communications network before becoming fully operational.

JITC has supported the Endeavor programs, beginning with CE, since 1995. The command provides support for the planning and execution of the interoperability assessments and provides third-party verification insuring the reliability, validity, and repeatability of the assessments. This is in accordance with JITC's mission to "professionally test, operationally evaluate, and certify information technology capabilities for joint interoperability, enabling information dominance and increasing warfighter effectiveness for the Nation."[2] The primary focus of the Endeavor programs prior to 2011 has been validating the point-to-point interoperability of the participating nations' C4 systems in several functional areas to include data-transport systems, core services, video, voice, transmission systems, and single-channel radio *(Photo 2)*.

At the end of each Endeavor exercise, JITC provides a summary of the conduct of the exercise as well as the assessment results in the form of a technical or

Photo 1. Combined Endeavor service members set up their work site in Grafenwöhr, Germany, September 16, 2010, during Combined Endeavor, a communications interoperability exercise that prepares international forces' command, control, communications, and computer systems for multinational operations. (U.S. Air Force photo by Staff Sgt. Tim Chacon/Released)

executive summary document. The summary presents the results of all the tests conducted and identifies trends noted by JITC. It also provides the COCOM with analysis on interoperability gaps among nations and provides recommendations for addressing these gaps. This summary is distributed to all exercise participants and forms the basis for the following year's exercise.

The most critical product published at the completion of an Endeavor exercise is the interoperability guide (IOG). The IOG provides communication planners with a tool for designing and engineering

Photo 2. Combined Endeavor service members set up a tactical satellite terminal in Grafenwöhr, Germany, September 16, 2010, during Combined Endeavor, a communications interoperability exercise that prepares international forces' command, control, communications, and computer systems for multinational operations. (U.S. Air Force photo by Staff Sgt. Tim Chacon/Released)

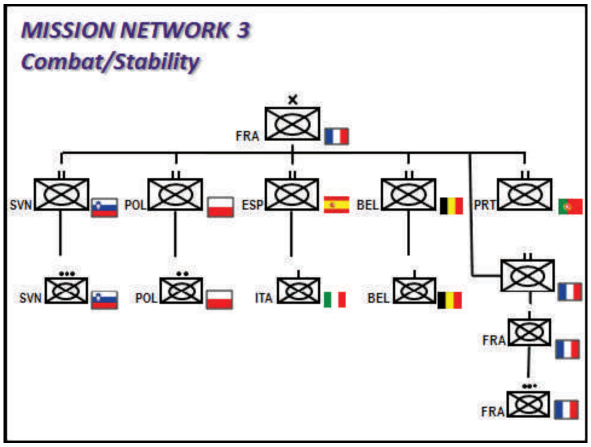


Figure 1. Example order of battle for CE.

coalition C4 networks, and contains all interoperability test data compiled during the respective Endeavor's existence. The IOG is provided to the sponsoring COCOM and participants on high-capacity optical media. A country can use the IOG, with a high degree of certainty that the C4 equipment of each nation will interoperate, to greatly streamline planning multinational operations. JITC as a third-party verification agency ensures the reliability, validity, and repeatability of the assessments contained in the IOG. Over the years, the IOG has been successfully used by many nations to plan and execute coalition networks supporting full-spectrum operations.

This year the focus of AE and CE has shifted from narrowly focused point-to-point testing to a broader assessment of the operational interoperability of coalition C4 systems. Operational interoperability assessments, or OIAs, assess the capability of multinational coalition networks to pass realistic prioritized operational traffic via multiple means and paths. OIAs utilize a scenario that is based on full-spectrum operations; operational interoperability testing of multiple command echelons using a mission thread with realistic message traffic; and, for CE, a robust network assessment to include realistic traffic loading and simulated cyber attacks to determine the true capabilities of the coalition network. This major shift to an operational interoperability assessment of C4 systems, as opposed to a point-to-point technical testing focus, will be attempted for the first time at AE and CE in 2011 and will form the foundation of the exercises for the foreseeable future.

Interoperability assessment of functional areas

Since the inception of the Endeavor exercises, the focus has been on basic interoperability assessments. These assessments began as simple point-to-point testing of predominately tactical radios and evolved over the years to keep pace with the rapid development of C4 capability and the explosive growth of IP-based data networks. Countries would determine which equipment they wanted to test and which countries they wanted to test with. JITC would work closely with the sponsoring COCOM and the participant nations to develop a comprehensive test plan to support the goals and objectives of the Endeavor exercise.

Over time, the Endeavors have evolved to incorporate robust coalition networks providing a full spectrum of C4 services. Individual tests, however, continued to be point-to-point in nature between two participants. As the Endeavors have evolved, so has the concept guiding the testing. Today the nations develop national goals and objectives which drive the determination of what to test and whom to test with. This has led to the incorporation of an overall scenario that tests the ability of the coalition networks to pass operational traffic. Based on the national goals and objectives, the nations are divided into either mission groups or regional groups. These groupings consist of six to 10 nations. Each of these groups may perform a specific type of operation, such as humanitarian support, offensive, defensive, or stability operations. These individual group operations are synthesized into the overall scenario for the particular exercise. See *Figure 1*.

Orders of battle are created based on the scenario and the participating nations within a particular group. Each group has one lead nation with overall responsibility for the network and participants in that group. In the case of AE, the lead-nation role is performed by one of the African nations of the African Union, not by U.S. Africa Command. Assessments are conducted within each of the mission groups or regional groups, and sometimes across groups, but for the most part the focus has still been technical point-to-point testing not synchronized or predicated by the scenario. The testing assessed the ability of the devices to talk to each other, not the ability of the coalition network to handle the transmission of information from one echelon to another during the conduct of full-spectrum operations with a realistic network load. The testing was narrowly focused on determining the interoperability between two pieces of communication equipment. These baseline interoperability tests were evaluated using the criteria found in *Table 1*.

While interoperability testing of certain functional areas determines basic technical interoperability, it does not determine if the tested equipment is interoperable as part of a broader coalition network under operational conditions. Point-to-point interoperability testing does not test the systems under a

Table 1. General test criteria.

Green	**Interoperable:** Achieved full functionality with no or only minor configuration changes. *Minor configuration changes* were changes to a standard configuration for which the tested system's crew was trained, authorized to make, and capable of performing with equipment that was part of the system's kit. The configuration changes must not have required external resources to be completed.
Amber	**Interoperable, but with significant limitations:** Achieved partial functionality or required a major configuration change to achieve at least minimum functionality. *Major configuration changes* were changes for which the tested system's crew was not trained, were not authorized to make, or required external components or support (contractors, vendors, or anyone other than that system's organic crew) to perform.
Red	**Not Interoperable.**
Not Possible	**Deemed not interoperable based on technical specifications prior to the initiation of a scheduled test.**

realistic network load while exchanging information that is time sensitive and realistic in terms of size and priority. For these reasons the Endeavors are evolving to include OIAs as a major part of the testing conducted during the exercises.

Operational interoperability assessment

In 2011, AE and CE will conduct OIAs for the first time. These Endeavors will utilize a tactical scenario to trigger the exchange of operational information, known as information exchange requirements (IERs), between multiple echelons in the regional groups and mission groups. The scenario triggers IERs which form the basis for OIAs that evaluate the technical, procedural, and operational capability of coalition partners to pass information via various C4 capabilities.

The basis for the OIA is the IER. IERs form the fundamental building blocks for larger, more comprehensive mission threads. From the Endeavor perspective, IERs require information to be exchanged between two echelons and include preferred and alternate means of transmittal, timeliness criteria, perishability of the information, and security of the information. IERs are usually exchanged between vertical command and control echelons—for example, from a platoon to a company, a company to a battalion, etc. Fundamentally, IERs are similar in nature to the Endeavor legacy interoperability assessment, as they are point-to-point or point-to-multipoint between adjacent echelons using equipment from a single functional area. Where they differ is in the incorporation of multiple means of communications devices and services, and the additional test-criteria requirements for timeliness and perishability of the information. See *Figure 2*.

In order to test the true operational capability of a large coalition network, it becomes necessary to group interrelated IERs so that they traverse more than one echelon. This is known as a mission thread. A mission thread is a "single end-to-end sequence of activities and events that describes an operationally meaningful process as it actually occurs or is intended to occur. Mission threads can describe tactical processes, logistical processes or support processes."[3] An example mission thread is a situation report from one unit to its higher headquarters of an improvised explosive device explosion. An example of this is depicted in *Figure 3*. Of note is that a series of IERs is combined logically to exchange the information from, in this case, a platoon to the Combined Joint Task Force Headquarters. This figure depicts the priority of preferred means of exchanging the information, the systems used, and the time to send the information at each level. For the Endeavors, the focus is not on the staff processes and actions that are performed, but on the ability of the C4 systems to pass accurate information in a timely manner in order to facilitate command and control during full-spectrum operations. The operational mission thread is the basis for the assessment of coalition C4 interoperability.

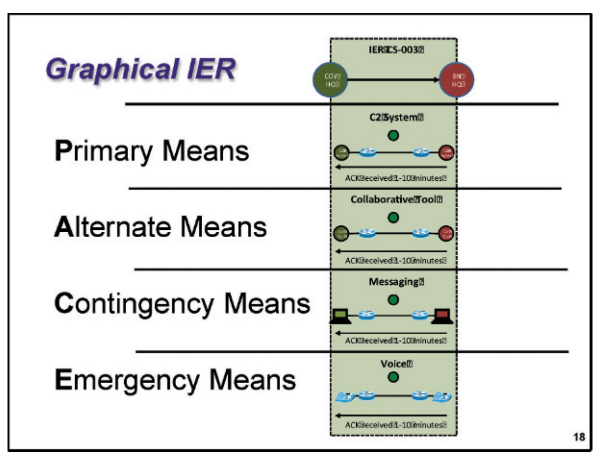


Figure 2. Graphical representation of an Endeavor IER.

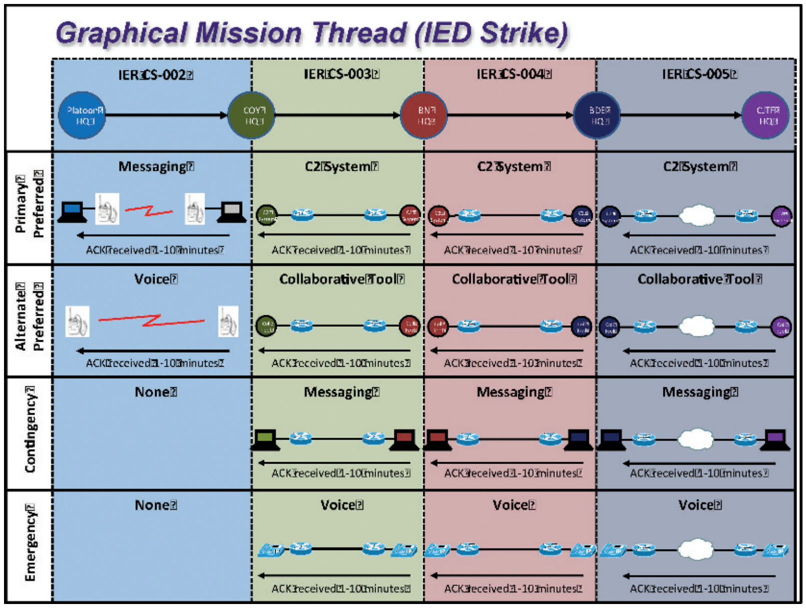


Figure 3. Graphical depiction of a mission thread.

OIAs focus on additional assessment criteria not traditionally used in the Endeavors. Instead of sending arbitrary test data, the focus has now become providing scenario-triggered information exchanges that are timely and accurate.

CE adds one additional element to assessing the operational interoperability of the coalition network: the network assessment. The network assessment determines the ability of the network to handle a true operational load representative of a division-level organization in addition to handling malicious traffic and cyber attacks during the tactical scenario. This realistic stressing of the network is intended to determine if the C4 systems connected to it can truly function in an operational environment as opposed to a sterile and stable laboratory test environment.

The exercise scenario provides the capability to test and validate the TTPs in a coalition network. The testing provides a framework to document these procedures as well as the technical interoperability connections. The network assessment determines if the equipment can operate effectively in an operational environment, as well as determining if the procedures for failover are validated in the event of a network problem caused by network stress that is from standard operational traffic load or due to malicious network attacks. The COCOMs as well as JITC will work to perform these assessments using after-action reviews of the tests, as well as in-depth analysis of the data collected. U.S. Europe Command will attempt this using one of the largest coalition C4 networks ever established. Thirty-eight countries and the NATO Multinational Corps Northeast will come together and install a coalition network providing C4 systems interconnectivity. This network will be established in Grafenwöhr, Germany, with a remote testing site in Lithuania and another in the Slovak Republic. The operational assessment will help to validate not only the technical interoperability of the C4 equipment from these nations, but the procedures used by U.S. Europe Command for the establishment, operation, and maintenance of a coalition network.

The future of the Endeavors

2011 will mark the first year OIAs are conducted as part of AE and CE. The lessons learned during the execution of these events will guide the conduct of exercises for years to come. The Endeavors offer an opportunity seen only in a few other programs to integrate and test interoperability, both human and equipment, in a multinational environment. Not only do these exercises build partnerships between nations, but they provide a unique ability to test TTPs for the installation, operation, and maintenance of large, complex multinational networks.

The complexity of coalition networks continues to increase as technology advances. The warfighter requires more access to information through the C4 networks in order to maintain information dominance on the battlefield, at both the tactical-edge and strategic levels. In order to meet the increasing needs and demands of the participating nations as well as the regional COCOMs, each of the Endeavors strives to grow and evolve. AE is looking to increase its scope from a focus on single-channel radio to building larger data networks incorporating external links to participating nations across the African continent. CE is also looking at expanding to a more distributed execution of the exercise, with multiple remote sites linking together to form a coalition network providing C4 from locations across Europe. This is a more realistic way to assess not only the procedures for the installation of a network, but the ability to exercise command and control over a geographically distributed network in an operational environment. These efforts more accurately replicate real-world situations faced when multinational coalitions support full-spectrum operations in areas of operations such as Iraq or Afghanistan.

In order to allow the warfighter to perform full-spectrum operations, coalition networks need to provide access to vast amounts of information and provide the warfighter with the ability to command and control the units in the areas of responsibility. The networks providing these services need to operate and pass information across an entire coalition. Ensuring interoperability is best accomplished at exercises like the Endeavors, where time and effort can be focused on interoperability, not during operations where time is not available. The Endeavors must continue to evolve to meet these needs, and the incorporation of OIA is just one more step in this evolution ensuring human and technical interoperability.

Conclusion

JITC supports three Endeavor multinational interoperability testing exercises that prepare coalition forces to install, operate, and maintain C4 systems in support of full-spectrum operations. This is accomplished through the technical interoperability assessments of C4 equipment, the development of interoperable TTPs, and the publication of operating standards for the installation, operation, and maintenance of coalition networks providing services in support of dynamic operational requirements. The Endeavors ensure that nations forming a coalition in support of full-spectrum operations can respond rapidly and decisively without concern that the C4 networks they build will require excessive setup, testing, and configuration modifications before becoming operational.

Assessing multinational communications network capabilities in an operational context is important to every communicator. It is clear that unilateral operations are a relic of the past. Almost every military operation conducted over the last twenty years has been a coalition of multinational partners executing the mission, whether that be combat operations, stability operations, humanitarian assistance, or disaster relief. No nation can command and control effectively at the time of execution if communicators have not performed thorough functional interoperability testing of the utilized C4 systems beforehand; furthermore, no coalition can operate successfully if the partner nations have not assessed the operational interoperability of the various C4 systems and most importantly the critical human interoperability that brings it all together.

This year the focus of Africa Endeavor and Combined Endeavor includes assessing the operational interoperability of various C4 systems. OIAs assess the capability of realistic coalition networks formed by multiple nations to pass realistic operational traffic with multiple means of exchanging the information ranked by priority. OIAs utilize a scenario that is based on full-spectrum operations; operational interoperability testing of multiple command echelons using the mission-thread framework with realistic message traffic; and, for CE, monitoring, assessment, and loading of the network to determine its true capabilities. This major paradigm shift to an OIA of C4 systems as opposed to isolated point-to-point testing began at AE and CE in 2011 and will form the foundation and direction of these exercises for years to come as U.S. COCOMs forge ahead with tackling the monumental challenges of coalition interoperability. ❑

MAJOR ROBERT R. HOUSTON is a telecommunications systems engineer (FA24) serving as a combatant command liaison officer at Joint Interoperability Test Command. MAJ Houston graduated from Florida State University in 1997 and was commissioned into the U.S. Army Signal Corps. He served as the battalion signal officer for 2/63rd Armor Battalion from 1998 to 1999, in Vilseck, Germany, and as a platoon leader in B Company/121st Signal Battalion in Kitzingen, Germany, from 1999 to 2000. As a platoon leader he deployed as part of TF Falcon supporting KFOR 1B in Kosovo. He then served as an assistant brigade operations officer in the 3rd Signal Brigade at Fort Hood, Texas, from 2001 to 2003 before taking command of D Company, 57th Signal Battalion, from 2003 to 2005. He deployed his company as part of Operation Iraqi Freedom (OIF) II and supported communications in the international zone, Victory Base, Abu Ghraib, and Al Asad. MAJ Houston took command of Headquarters and Headquarters Company (HHC), 3rd Signal Brigade, from 2005 to 2006. He then deployed to support OIF 06–08 as the 3rd Signal Brigade plans and systems integration officer, and then became the brigade

operations officer upon the 3rd Signal Brigade's return from Iraq until it was inactivated in 2008. He obtained a master's degree in telecommunications from the University of Colorado at Boulder in 2010. He is also a graduate of Signal Captains Career Course, Combined Arms and Services Staff School (CAS3), and Intermediate Level Education (ILE).

Endnotes

[1]Combined Endeavor Collaboration Platform. 2011. Mission page. U.S. European Command. 2 May. Accessed 20 July 2011 at https://www.combinedEndeavor.com/wiki/.

[2]Joint Interoperability Test Command. 2011. *Joint Interoperability Test Command: Command Overview.* Fort Huachuca, Arizona. 24 January.

[3]Combined Endeavor Planning Organization. 2011. *Combined Endeavor 2012 Coalition Operations Concept and Procedures*. Draft, 10 June.

ITEA Journal 2011; 32: 446–451

Modeling to Support Test as We Fight

Jamie L. Pilar, Windsor L. Jones, and Bruce Marihart
U.S. Army Evaluation Center,
Aberdeen Proving Ground, Maryland

Test [and evaluate] as We Fight has become a greater and more pressing reality for today's defense with the prevalence of irregular warfare and leadership's current network-centric focus on sophisticated communication equipment that will enable situational awareness across distances and diverse terrain. In the face of these challenges, the evaluation community has embraced modeling and simulation (M&S) to enable a broader test and evaluation (T&E) of potential systems. M&S is providing advantages to evaluation including the potential to model systems in diverse terrain, changing environments, and populated urban areas. M&S can also provide evaluators the ability to quickly adjust the equipment packages modeled and system performance parameters as programs change. All of this comes with the advantages of repeatability and scalability, allowing the evaluator to perform sensitivity analysis and develop a robust evaluation. This article provides an overview of some recent efforts and future planning to support the T&E of systems in irregular warfare with a network-centric focus.

Key words: Ground Soldier System; Infantry Brigade Combat Team Increment 1; irregular warfare; modeling & simulation (M&S); network.

Staying at the forefront of evaluation and leadership expectations, the U.S. Army Evaluation Center (AEC) has placed emphasis on the use of modeling and simulation (M&S) for the test and evaluation (T&E) of new network and soldier systems that are priorities for the Vice Chief of Staff of the Army and the modernized Army. The guidance of General Peter Chiarelli, the Vice Chief of Staff of the Army, has been to provide the Army with new equipment to increase network functionality in an expeditious manner that does not interfere with soldiers' missions and duties. In April 2011 he spoke before the Senate, stressing the value of the network as a modernization of the Army (Brewin 2011). However, in today's troubled economy the T&E community is being asked to "do more without more" by Under Secretary of Defense for Acquisition, Technology & Logistics Ashton B. Carter and other senior leaders (Carter 2010). As the Army moves forward to obtain the needed future equipment while staying within budget constraints, technologies like M&S are being called on to provide supplemental data for system evaluation and enhanced T&E.

This article covers the work completed for two families of systems that will directly impact the individual soldier's ability to obtain and process more information while on the battlefield. These systems are the Ground Soldier System and the Infantry Brigade Combat Team Increment 1. M&S was implemented during these programs to enable a robust T&E of the systems and overcome test limitations.

Planning is also underway to support future system evaluations and the next step in Army modernization: capability packages. Capability packages will be systems that are identified as needed, top priority, and ready for testing, and will be grouped into a package for integrated T&E and deployment. These future efforts will require evaluation tools that are flexible and network intensive to fully investigate the added value of the systems during evaluation. This article provides an overview of completed AEC M&S activities in support of "test as we fight" evaluations, and provides a glimpse into future planning and development efforts to keep T&E optimal for soldier support within a budget-constraint environment.

Modeling the Ground Soldier System Nett Warrior (NW)

The Ground Soldier System is a collection of systems for ground-combat soldiers to enhance capabilities including command and control, situational awareness, lethality, mobility, survivability, and sustainability, with the benefit of embedded training. Additionally, the soldier will be provided with lighter equipment requiring less space and power, creating a smaller logistical footprint. Nett Warrior (NW), named in honor of the World War II Medal of Honor recipient Col. Robert B. Nett, was the name adopted for the first increment of the Ground Soldier System. During this exercise there were multiple equipment solution sets being examined to determine the best systems to compose the deployed NW system.

In support of needs identified in the NW Test and Evaluation Master Plan, several models were reviewed to identify the best solution to test and analyze identified evaluation measures. The AEC M&S Division selected the Infantry Warrior Simulation (IWARS) as the optimal solution utilizing supporting data and observations from the program's Limited User Test to enhance the simulation. IWARS is codeveloped by the Army Materiel Systems Analysis Activity (AMSAA) and the Natick Soldier Center and functions as a small-unit performance model. It was used to expand live test scenario results to larger-size units (squad and platoon) and experiment with varying battlefield conditions and equipment solutions not examined during the test.

Process

IWARS provides an analytical capability to assess what contribution various soldier systems will have on small-unit operational effectiveness in environments similar to those in which soldiers fight and train. The M&S analysis differed in its objective from the live test by focusing on comparing the capabilities and effectiveness of an NW-equipped force against the capabilities of the current force to investigate the difference in unit performance under selected conditions. IWARS uses predetermined routes, military subject-matter experts' knowledge, and simulation analysts' judgment to develop how soldiers respond to information they receive. Other input data were obtained from supporting models like the AMSAA Brigade and Below Network Model; live tests, such as the Developmental Test held at Aberdeen Test Center; and AMSAA's Joint Data Center, an authoritative source for system characteristics and performance data. Additionally, observations captured during the Limited User Test were used to develop the modeled soldier's responses to situational awareness and command and control.

Three scenarios were selected for comparison purposes: Night Movement to Contact, Night Attack, and Day Cordon and Search. The Day Cordon and Search was located in an urban area, depicted in *Figure 1*. The situation being modeled in this scenario and illustrated in the figure involves a threat force reported to be using one of the buildings in the urban area of operation to store small arms, materials for improvised explosive devices, and possibly mortar rounds. The threat forces will attempt to prevent friendly forces from discovering the weapons cache if they feel it is threatened. The battalion-level mission is to conduct a coordinated offensive operation within the area of operation in order to neutralize threat forces in the area. The company-level mission is to conduct a cordon and search of objective and capture the cache of threat weapons and munitions.

To accurately capture the spectrum of soldier performance in the battlefield, three cases were developed for each scenario, where the soldiers would be provided with (a) no NW capabilities, (b) NW with reduced message-completion-rate capabilities, and (c) NW with full message-completion-rate capabilities. The three scenarios were used to simulate a friendly force approaching a threat-occupied area and analyzing the ability of NW to increase mission success through obtaining and transferring additional situational awareness during the mission.

Results and value added

The completed study using IWARS was able to show that soldiers with fully functional NW systems (or those with a slightly reduced message-completion rate) provided the units with advantages over current equipment. For example, the additional situational awareness could allow an approaching friendly unit in a threat-occupied area the ability to assume a more beneficial approach pattern before encountering resistance. Another example would be the ability of soldiers to communicate a more accurate location of the threat within an urban area to other soldiers within their unit, thus reducing mission time and risk to soldiers. The model was also able to identify some key functions of the NW systems where high performance for the duration of a mission would need to be executed in order to provide value to the user. A comparison of live test observations and data to the model analysis verified that the IWARS results were applicable to a real fight environment and validated the use of the information in system evaluation.

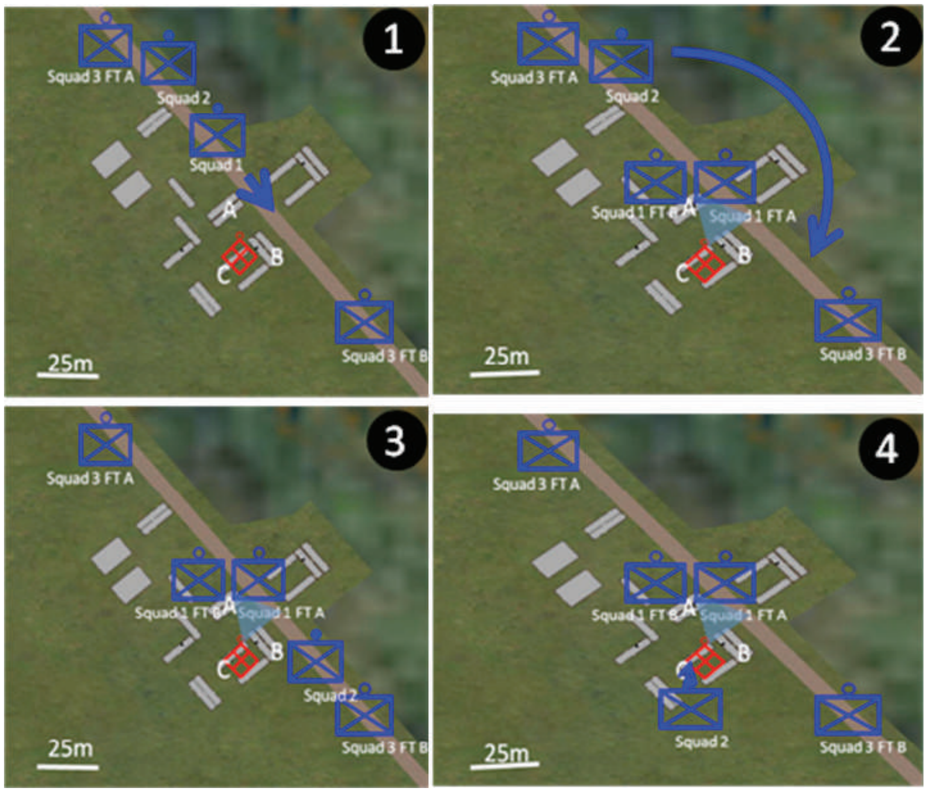


Figure 1. Urban Day Cordon and Search.

Modeling the Infantry Brigade Combat Team Increment 1

The Infantry Brigade Combat Team Increment 1 program contained systems focused on the small unit functioning around urban areas with unmanned systems and network enhancements. The systems capabilities include providing users with greater situational awareness, survivability, command and control, and the ability to use unmanned systems in human-unfriendly environments. While these systems had been looked at in live tests and M&S previously, this subsequent effort was used to refine results based on currently known system performance, expand the system evaluation into new scenarios, and provide evaluators and testers with better information to determine what critical issues need to be addressed during future testing. Again, the AMSAA IWARS model was used to focus on the platoon and squad usage of the systems in irregular warfare environments.

Process

Two scenarios were used in this study: (a) conduct a defense and (b) cordon and search. Focused around an urban area, the scenarios involved both open-area attacks and urban close-quarters battles. *Figure 2* provides model screen shots showing some of the fights occurring inside buildings and the location of a hidden tunnel utilized by the threat forces.

For each scenario, additional variations were explored to stress the systems under test and identify any significant issues or trends. These variations included the methods threat forces used to attack friendly forces, deployment (usage) of equipment, and degradation of network performance (similar to the NW study). During the first scenario, conduct a defense, the friendly forces utilized unmanned ground and air systems to gain earlier situational awareness of the attack. Unlike live testing, the model enabled a more realistic encounter where the threat forces had the capability to shoot down unmanned aircraft systems if they came within range and compromised the threat's mission. In both scenarios, the friendly force had to face urban irregular-warfare situations in which threat forces utilized their knowledge of the urban environment to make use of hiding places and tunnels (urban infrastructure) to gain an advantage. The second scenario focused on urban complexities and included the need to clear a tunnel with multiple 90-degree angles (compared to a straight tunnel in the live test) with minimal threat to soldiers.

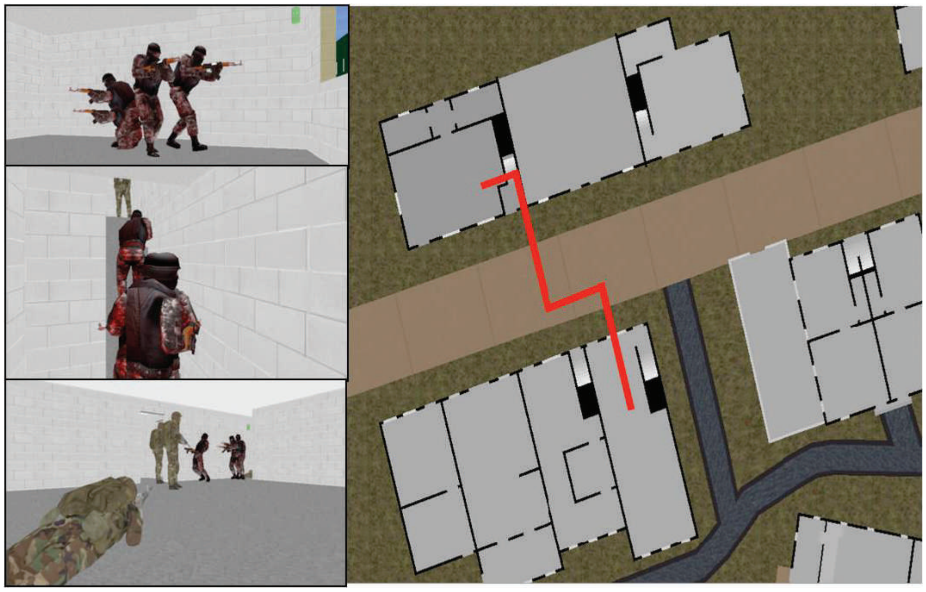


Figure 2. IWARS urban screen captures.

Results and value added

The model results were able to show the conditions under which the systems could provide benefits and value-added capabilities to the soldiers. Analysis showed that sensors enabled the friendly force to have greater situational awareness sooner than with current equipment. This allowed them more time to plan and react to potential threats. Using sensitivity studies in the model, analysis was performed on how altering the altitude of an unmanned aircraft system affected its mission success (ability to acquire) and survivability. These system-of-systems sensitivity studies were also valuable in identifying second- and third-order effects as a result of changing the usage of equipment, such as the survivability of soldiers depending on the additional situational-awareness information and the method of defense (indirect vs. direct fire) used by friendly forces. Similarly, in the building- and tunnel-clearing activities the analysis showed that while it would take longer to accomplish the mission with the additional systems, the benefit was an increase in soldier survivability and interesting observations on what tactics, techniques, and procedures should be utilized in such cases.

Modeling the future capability packages

Planning for the future capability packages is challenging because of the dynamic nature of the program. Systems will be added to or removed from a capability package depending on their readiness and command priorities, allowing changes to occur frequently and up to the last minute. Additionally, the continued focus on the network and how to evaluate performance at a mission level has highlighted some gaps in current testing and modeling capabilities. To meet current gaps and provide adequate T&E for future capability packages, AEC is pursuing an Army Research Laboratory Survivability/Lethality Analysis Division model that promises additional functionality and flexibility.

The System of Systems Survivability Simulation (S4) is a closed-form agent-based model that includes decision-making processes for modeled entities. These decision-making processes enable a more realistic battle environment where a modeled soldier learns, reasons, and makes decisions based on information received via the network, what they see in their surrounding area, and commands provided by superiors. Each simulated entity is also programmed to have a ground truth agent, which will have perfect information and provide analysts with the knowledge of how a soldier would have acted with full situational awareness (vs. how the soldier acted with incomplete knowledge). A simulated soldier may have reduced situational awareness as a result of network issues, a function that is controlled by a supporting model within S4 called the Brigade and Below Propagation and Protocol model. This communications model provides multiresolution modeling for propagation, latency, and communications protocols for both Army legacy communication systems and systems under development.

Process

The S4 model is new for AEC's use in evaluation; therefore in order to verify the use and applicability of the model, a pilot study is being completed in 2011

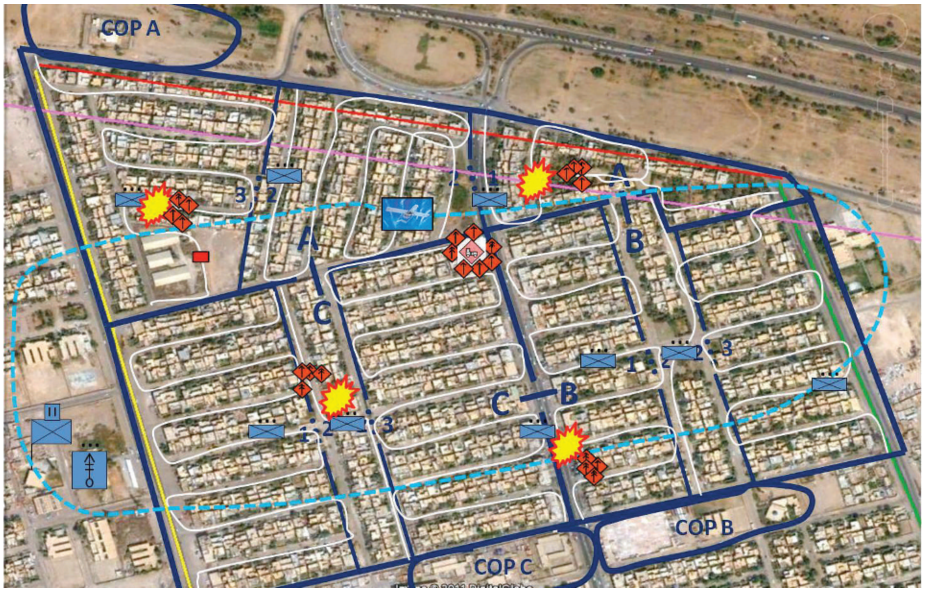


Figure 3. S4 Urban Scenario.

modeling systems that are currently under test. Analysis of the model thus far has been favorable and future planning is underway. The next study will be conducted in 2012 and will focus on potential systems for the next capability package. The 2012 study will include select systems from NW, Infantry Brigade Combat Team, Joint Tactical Radio System, and non-network-centric systems such as the Shadow Unmanned Aircraft System and the Accelerated Precision 120 mm Mortar Initiative.

The proposed scenario will be in an urban area and involve a battalion of three infantry companies. The force will have the mission of searching the urban area for a high-value target and, after finding it, neutralize that threat. *Figure 3* provides a bird's-eye view of the urban area at the beginning of the search. The mission will initially be executed with three variations: (a) a base case with the force equipped with current force capabilities, (b) an upgraded case with the force equipped with the systems under test at full capability, and (c) an upgraded case with slightly reduced network capabilities.

Value added

S4 promises to provide great value to system evaluations and the analysis of how network systems contribute to mission effectiveness. During live testing it is challenging or impossible to collect the required system-test data and allow for free play of the system to fully understand the impact systems will have on the mission as a whole. Using this simulation, it is expected that those broader issues will be evaluated and that a greater statistical confidence will be provided through the use of multiple runs.

The modeling of decision-making processes will also provide greater insight into the human aspect of network evaluation. This will help answer questions such as how will soldiers react when processing information received via the network and how will incomplete information affect the mission.

S4 will also resolve some current M&S gaps identified in IWARS, including the limited force size (platoon or lower in IWARS) and the limited ability to model network effects within urban environments. The 2012 S4 study will provide a complex urban environment not seen in live testing and on a much larger scale than previously used in M&S by AEC evaluators. This development is imperative in enabling T&E of systems in environments like those that current (and future) warfighters are experiencing.

Summary

Testing as we fight is not an easy task, but our current and future forces are in need of quality equipment that is suitable for the changing warfare environment. Through the use of M&S, AEC is striving to provide comprehensive evaluations of network and close-combat systems in complex and urban terrains beyond the limitations of live test. These evaluations will ensure that systems being deployed to the soldiers will keep them safe, facilitate their missions, and keep within the financial constraints of our current economy. AEC is "doing more without more" through the use of modeling in support of evaluation. ❑

Jamie L. Pilar is an operations research analyst for the Army Evaluation Center Modeling and Simulation Division at Aberdeen Proving Ground, Maryland. She holds a master of science in information-technology project management from the University of Maryland University College. E-mail: jamie.L.pilar.civ@mail.mil

Windsor L. Jones is an operations research analyst for the Army Evaluation Center Modeling and Simulation Division at Aberdeen Proving Ground, Maryland. He holds a master of science in operations research from the George Washington University School of Engineering and Applied Science. E-mail: Windsor.l.jones.civ@mail.mil

Bruce Marihart is an operations research analyst for the Army Evaluation Center Modeling and Simulation Division at Aberdeen Proving Ground, Maryland. He holds a master of science in systems engineering/operations and a master of business administration from the Florida Institute of Technology. E-mail: Bruce.g.marihart.civ@mail.mil

References

Brewin, B. 2011. Global network will redefine how the Army fights, vice chief says. *NextGov.com*. http://www.nextgov.com/nextgov/ng_20110406_7920.php (accessed April 19, 2011)

Carter, A. B. 2010. *Better Buying Power: Guidance for Obtaining Greater Efficiency and Productivity in Defense Spending*. Office of the Under Secretary of Defense, 14 September 2010.

Acknowledgments

We would like to thank the model proponents for their support and continued efforts. Special thanks to IWARS lead developer Kevin Guite and analyst Andrew Clark, and S4 project lead Jill Thompson, Army Research Laboratory/Survivability/Lethality Analysis Division and supporting model developers.

ITEA Journal 2011; 32: 452–456

When Testing With Live Agents Is too Risky… Computer Modeling and Simulation May Be a Valuable and Cost-Effective Alternative

Bill Rearick
Alion Science and Technology, Alexandria, Virginia
David Godso
Joint Project Manager, Information Systems, San Diego, California

The rapid pace of advances in weaponry and tactics on today's battlefield greatly compresses the traditional development time available to meet the need for solutions to be placed in the hands of the warrior. As a result, there is a more and more common need to test in conjunction with development, training, and deployment of systems or capabilities. In this article, we explore how recent advances in computer models and simulations are being utilized to conduct more realistic immersive training against chemical agents in a virtual environment. The chemical agent is generated in a virtual overlay of the selected training area. Existing Department of Defense models and tools are used to promulgate and dissipate the plume as well as automatically adjudicate casualties in several levels of severity based upon agent concentrations and the protection status of the participant. An after-action review system allows playback and review following the event and provides hard-data recording for testing purposes. There are obvious areas in which computer modeling and simulation can be used to enhance training and testing while capturing efficiency advantages by leveraging existing systems to meet multiple purposes in a fiscally constrained environment.

Key words: After-action review; chemical agents; computer models and simulations; fiscal constraint; immersive training; realism; virtual environment.

When ground troops want to test new Tactics, Techniques, or Procedures (TTPs), they can use blanks, rubber bullets, or laser emitters to take the place of real bullets. When testing against biological agents or chemical weapons (aka "bugs and gas"), there are very few options for realistic testing, and compromises are made often resulting in skepticism regarding test conclusions. For example, previous efforts have relied heavily on smoke or powder to represent the threat "cloud," despite the fact that the majority of chemical/ biological threats are not visible to the naked eye.

Some progress has been made in the use of stimulants to represent threat agents. These stimulants cause a reaction in test strips and even provide olfactory markers for threat compounds. The drawback with these stimuli is the need for exercise control personnel to interact with and, unfortunately, influence the test participants. Control personnel must remain close to the testers to be able to supply the necessary reactant when testers stop to take a sample. In very short order, testers are negatively trained that "if there is no controller in sight, there must not be anything here to test for" and deviate from the procedure they are supposed to be following because, in effect, they already know the outcome in that area. Personal protective masks are hot and uncomfortable; and in exercises, it has been shown that gear is not always properly employed until the first controller is sighted. The location of sampling is also influenced by the controllers, with participants often approaching the controller, announcing the test they are about to perform, and looking to the controller for the "reading" obtained. While the impact of these constraints and artificialities is difficult to quantify and correct for in

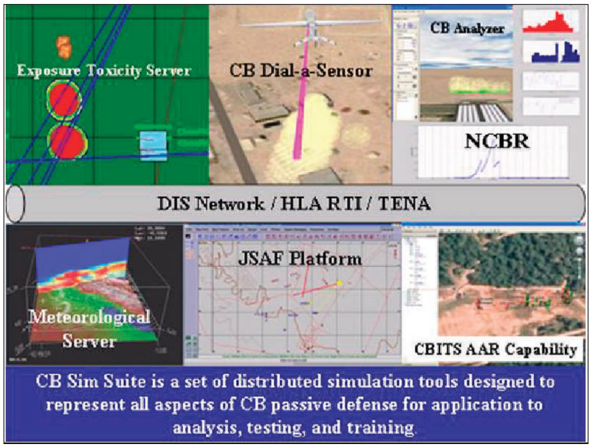


Figure 1. Models and simulations used to create the chemical, biological, radiological, and nuclear hazard.

Figure 2. Modified equipment to interface with the software.

the data collected, there is a universal concurrence that the results would be less "tainted" if it were possible to conduct the tests without the presence of controllers in the test area.

For a number of years, the Services and Combatant Commanders (COCOMs) have used computer Modeling and Simulation (M&S) to train, test, and evaluate TTPs, and to "war game" scenarios set in the future, with modeling allowing the use of weapons systems that were still in development. The U.S. Navy Warfare Development Command utilized a computer simulation to evaluate the effectiveness of tactics and special naval 5-inch gun munitions to combat small boat swarm attacks. A persistent, dedicated network, the Joint Training and Experimentation Network (JTEN), has been established to enable Services and COCOMs to use models and simulations to meet their training and experimentation needs. It also allows for personnel to participate from distributed locations (often their home station), saving the costs normally associated with travel. Utilization of this Joint Live Virtual Constructive (JLVC) federation also greatly reduces the operating expenses and wear and tear on equipment, through extensive use of virtual simulators for flight, vehicle, and ship maneuvers. Being able to integrate Chemical, Biological, Radiological, and Nuclear (CBRN) incidents into Service exercises has often taken a backseat to higher priority requirements on conventional warfare scenarios. The end result is that CBRN training is often left out of Joint Task Force level training events and left to individual units to complete during their individual unit training periods. There has been some pioneering proof of concept work conducted in recent years that is worth highlighting, not so much as a complete solution or capability, but as a means of bringing to light some of the options that are now possible because of the maturation of M&S technologies. One project, developed by the U.S. Joint Forces Command, integrated training chemical sensors with M&S products that could predict the promulgation and propagation of a chemical agent over the battlefield terrain. *Figure 1* provides a sample of the M&S products that have been coupled together to provide a capability that cannot be replicated using conventional training methods.

Coupling this system with a high-frequency testing range that operates on the same frequency as the U.S. Army's Multiple Integrated Laser Engagement System (MILES) training vests would allow CBRN threats to be integrated into the Army's home station training system. This coupling allows the testing/training team to patrol through an area near a potential site in which chemical weapons have been released. *Figure 2* shows how the standard gas mask has been modified to interface with the MILES system, a MILES vest and an Automatic Chemical Agent Detector/Alarm SIMulator (ACADASIM) sensor coupled with a high-frequency radio interface into the MILES system. Personnel in the figure show the minimal impact of the MILES gear on personnel movement and weight loads.

The instrumented training sensors, detecting and displaying information computed by the plume model, alarm when the concentration of the agent in the vicinity of the detector exceeds the detector set point. Upon hearing the detector alarm, personnel in the vicinity don the appropriate protective gear. A differential pressure switch mounted inside the modified filter assembly allows for verification of proper mask sealing (drawing a vacuum across the switch as the tester inhales) and transmits a "mask on" signal to the command and control center. Personnel who fail to

Figure 3. U.S. Army Chemical, Biological, Radiological, and Nuclear School students participating in the military utility assessment.

properly employ their protective mask are adjudicated by the software, accounting for concentration and length of exposure and judged as either incapacitated or killed, with the appropriate data being transmitted to the participant through the MILES interface readout and tones. This process works much the same as if the participant had been "shot" with a laser beam by an enemy participant and "wounded" or "killed."

Integration of this "simulation" with real world command and control (C2) systems was achieved by feeding information developed on the site into the Joint Warning and Reporting Network (JWARN). The Joint Project Manager Information Systems (JPMIS), a component of the Joint Program Executive Office for Chemical and Biological Defense, provided equipment, training, and support for the prototype. Standardized Nuclear-Biological-Chemical (NBC) formatted messages, used by all U.S. and North Atlantic Treaty Organization (NATO) forces, were generated either by input received via voice reports from the survey teams, or directly from sensors tied into JWARN via the JWARN Component Interface Device. JWARN is an official Acquisition Program of Record that is now fielded. Once enough information was received and correlated, JWARN could be used to call up another JPMIS product, the Joint Effects Model (JEM). JEM is able to utilize real-world weather reports or historical weather to predict the promulgation of the hazard plume and create an overlay output that can be displayed on the Common Operational Picture. From this prediction, JWARN is able to identify the units in the path of the hazard and issue a timely warning to allow the units to avoid the plume or adopt a protective posture (MOPP level) before the arrival of the hazard.

The U.S. Army CBRN School at Fort Leonard Wood, Missouri, conducted a Military Utility Assessment (MUA) of the prototype in December 2009. *Figure 3* shows some of the students from the Army CBRN School, who participated in the MUA, passing an ACADASIM sensor placed alongside the road. The findings were generally positive, with the testing audience praising the capability as providing a much more realistic experience than historical methods. The Army used the assessment to provide feedback for future capability development and implementation. The ability to include specialized CBRN reconnaissance vehicles, and biological and radiological hazards, as well as the ability to have officer students operate JWARN as a higher headquarters staff were some of the highest ranked requests.

Figure 4. (a) Prototype control van. (b) Computer workstations inside the control van.

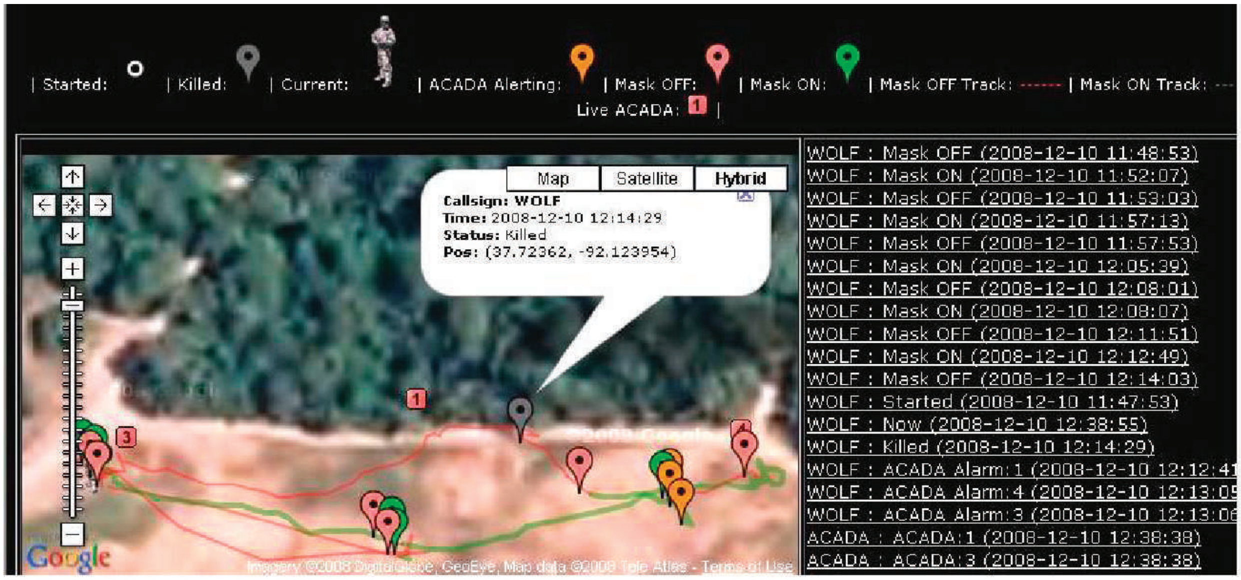


Figure 5. Screen shot of data captured in the after-action review software.

For the prototype, the system was made portable and self-contained, with a recreational vehicle modified to serve as the control van *(Figure 4a)*. Satellite uplink capability as well as onboard power generation allows the prototype to travel to various locations for demonstrations. A practical benefit of the configuration is that the antenna trailer has its own power supply and satellite dish, allowing it to be located remotely from the control van. Instructors at Fort Leonard Wood could execute an event that involves the training/testing audience onsite at Dugway Proving Ground, Utah, or in Korea. A home station training capability could be located in an existing building on base and not require the satellite interfaces, resulting in greatly reduced cost of required hardware. *Figure 4b* shows the computer control stations inside the control van but could easily be set up in a building for fixed sites.

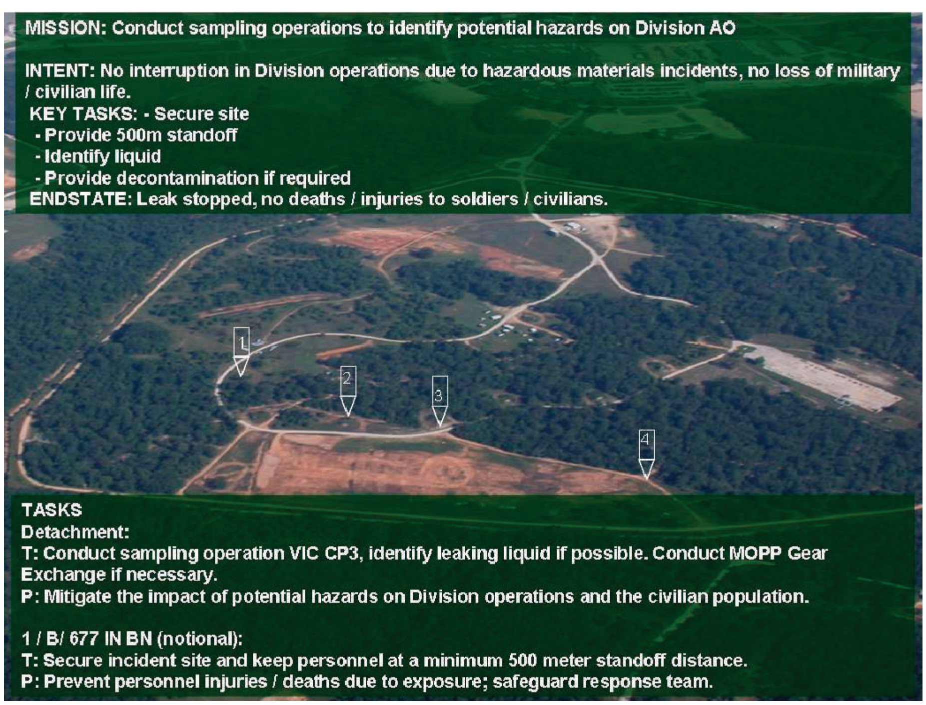


Figure 6. Scenario development and conceptualization.

An additional benefit of the use of computer M&S in the testing and training arena is the ability to capture the results of an event and play them back for review. *Figure 5* shows the after-action review capability, which is nonexistent using current training methods. This level of data capture for postevent analysis is an integral part of testing but has been very hard to capture and utilize in CBRN training events. Being able to display, on a flat screen, the position of personnel and equipment, superimposed over the play-box terrain along with readouts on the levels of contamination, status of protective equipment and the "health" of the player, adds a quality and depth to the events that were nonexistent previously.

Incorporation of computer programs also makes the development of scenarios much easier to visualize and document the initial quantities and location of agent released as well as varying the weather and terrain to test certain conditions or locations. Additionally, having a means of presenting the scenario, as set up on the range, prior to the audience taking to the field allows for more comprehensive pre-event briefings, enhancing participant understanding of the objectives and training mission, as illustrated in *Figure 6*.

In conclusion, this article is intended to inform the reader of capabilities whose existence may not be known, rather than endorse any one application or product. If, during the course of this article, you found yourself saying, "I could use that ability" or, "I wonder if this could be done," then this author has achieved his goal. Computer M&S has the ability to elevate testing and training to a whole new level of realism and fidelity, enabling the end user to more effectively develop the next generation of tools to keep the warfighter safe from CBRN hazards. ❑

BILL REARICK is an Alion Science and Technology, Inc., employee who has worked for the last 8 years in the Chemical, Biological, Radiological, and Nuclear community, initially for the Defense Threat Reduction Agency and more recently for the Joint Project Manager Information Systems. E-mail: william.rearick.ctr@navy.mil

DAVE GODSO has over twenty years of highly diversified and in-depth experience as an innovator, leader, and multidisciplinary engineer in complex information technology and management systems and environments. Mr. Godso's experience includes technical management and program management spanning the product life cycle, including: research and development, architecture, requirements analysis, mission critical distributed and embedded system (hardware, software, and firmware) design and implementation, enterprise information management and automation, information assurance, human systems interfaces, test and evaluation, independent verification and validation, software maturity and quality, training, configuration management, logistics, product fielding, and user support.

Mr. Godso is a recognized leader and subject matter expert in the CBRN community. He currently serves as the chief engineer and chief information officer (CIO) for the Joint Project Manager Information Systems (JPM IS). He is a graduate of Louisiana State University (LSU) with a bachelor of science in mathematics and was also the recipient a Regents Award from LSU. E-mail: David.godso@jpmis.mil

ITEA Journal 2011; 32: 457–462

A Bioinformatics Approach to Finding Candidate Organisms for False Positive Testing of Biological Warfare Agent Detectors

Charlie Holman, Ph.D.
Biology and Mathematics Departments,
Claflin University, Orangeburg, South Carolina
Anika Dodds
Biology Department, Claflin University, Orangeburg, South Carolina

There is an ever present threat and danger from biological warfare and biological terrorism. To mitigate this risk there are several programs to develop enhanced biological warfare agent detection systems. Testing of these systems must be comprehensive. Part of this comprehensive testing is focused on false positives. A false positive is a positive detection to a substance or stimulus other than what the detector was designed to detect. Currently there is no scientifically valid method to identify candidate organisms that are apt to invoke a false detection. This article proposes a method to identify candidate organisms that are apt to invoke a false detection. This method is based on bioinformatics databases and used a bioinformatics program known as Basic Local Alignment Search Tool (BLAST). An illustrative example is provided. BLAST with the Bioinformatics databases provides a useful method to systematically, accurately, and relatively completely identify candidate organisms for false detection testing of biological warfare agent detectors.

Key words: Basic Local Alignment Search Tool (BLAST); bioinformatics; biological warfare agent detection; false detection testing.

As demonstrated by the 2001 anthrax letter attacks in the United States (Canter et al. 2005), the 1984 Rajneeshee *Salmonella* attack in Oregon restaurants (Torok et al. 1997), and the unsuccessful biological terrorism attacks in 1993 and twice in 1995 by the Aum Shinrikyo cult in Japan (Olson 1999), there is an ever present threat and danger from biological warfare and biological terrorism. To mitigate the risk from biological warfare and biological terrorism there are several programs in place to develop enhanced biological warfare agent detection systems.

Detection performance consists of two parts. The first part is how well a detector detects the substance it is intended to detect. These are true detections and are often quantified as the probability of detection of a given concentration. Biological warfare agent detectors are designed to detect biological warfare agents, such as *Bacillus anthracis* and hemorrhagic viruses. The second part is how easy it is to fool a detector into alarming when it is challenged with something that it should not alarm to. These are commonly called false positives, false detections, or false alarms. For biological warfare detectors, a false detection is the detection of a substance other than a biological warfare agent. False positives may be quantified as mean operating hours between false detection in a specific environment or as a list of challenges that produce false detections. Understanding the nature of false detections provides an opportunity to develop tactics, techniques, and producers to mitigate the operational effect of false detections.

We want to know how a biological warfare detector performs in the threat environment when it is operated by actual operators. However, because of the danger to testers and the general population, open air testing with actual live biological warfare agent is illegal and immoral. Actual live biological warfare agent cannot be released into the environment. Testing true detection,

detection performance of biological detectors involves challenging the system under test with both live biological warfare agent and relatively harmless simulated agent in a laboratory chamber, challenging the system under test in the threat environment with relatively harmless simulated agent, and developing an agent–simulant relationship to predict how well the detector would perform in the threat environment when challenged with actual biological agent (Holman and Loerch 2010; Holman and Berkowitz 2009a).

Testing false detection, detection performance involves two types of tests. The first of these tests is operating the system in a specific environment and reporting the mean time between false detections. Typically several different types of threat environments are tested. Information such as the time of year and area activity is reported. The second type of false detection testing is the determination of what types of nonbiological warfare agent challenges produce a false detection. Currently, candidate organisms for false detection testing are chosen more for convenience than for selecting organisms that are apt to produce a false detection. Most often, other biological warfare agents or simulants serve as challenge organisms in false detection testing. Currently, there is no systematic and scientifically valid method to select potential false detection challenge candidate organisms (Chipman et al. 2001; Holman et al. 2008). This article presents a scientifically valid method to identify candidate organisms that are apt to invoke a false detection.

Biological warfare detection technologies

There are three general categories of detectors: nonspecific generic, categorical, and identifier.

1. The nonspecific generic detector alerts when there is a potential biological warfare agent present. The nonspecific generic detector is designed to alert on any potential threat and hence will typically have relatively numerous false detections. Nonspecific generic detectors sense attributes such as particle size and particle florescence, which are not only common to all or at least many biological warfare agents but are also common in much of the biological debris in the air. In the Biological Integrated Detection System-Non-Developmental Item, Biological Integrated Detection System-Pre-Planned Product Improvement, and the Joint Biological Detection System, a generic detector turns on the process of detecting to identify the type of biological warfare agent present. Nonspecific generic detector technology includes aerodynamic particle sizer and florescent aerodynamic particle sizer (Chipman et al. 2001; Holman et al. 2008).
2. The categorical detector identifies biological warfare agent to the major category: bacteria, virus, or toxin. Categorical detector identification technologies include the flow cytometer and the mass spectrometer (Chipman et al. 2001; Holman et al. 2008).
3. The identifier detector identifies the type of biological warfare agent. Bacterial agent identification is to the species such as *B. anthracis*, *Yersinia pestis*, or *Francisella tularensis*. Viral agent identification is to the disease caused by the virus such as smallpox or to the species such as *Variola major*, which causes the more lethal form of smallpox. Toxins are nonliving substances produced by living animals, plants, or microbes. They are identified to toxin type such as botulinum or ricin. A detector that identifies the type of biological warfare agent must be able to distinguish between all types of biological warfare agents and other substances that may be present in the sample. There are two technologies being used to identify the type of biological warfare agent: (a) nucleic acid detection and (b) immunological detection (Iqbal et al. 2000; Ivnitski et al. 2003; Gooding 2006; Pohanka et al. 2007).

Nucleic acid is the inheritable material. It contains the genetic code, which results in protein syntheses and causes organisms to look the way they look. Each species has segments of nucleic acid that are unique to that species. The inheritable material in bacteria is Deoxyribonucleic Acid (DNA). The inheritable material in viruses is either DNA or Ribonucleic Acid (RNA). Purified toxins lack nucleic acid. Hence, nucleic acid methods can be used to identify bacteria and viruses, but not toxins. A molecule of nucleic acid has a unique sequence of bases that is capable of binding to or hybridizing with another nucleic acid molecule that has complementary bases. This property of hybridizing to another molecule of nucleic acid with complementary bases is the mechanism underlying the use of nucleic acids to identify biological warfare agents.

The immune system defends the body from foreign particles such as bacteria, viruses, and toxins. In an immune reaction, the foreign particle is called an antigen. When the immune system encounters an antigen, it produces antibodies that bind to that specific antigen. If the antigen is a specific molecule that is unique to a specific biological warfare agent, then the resulting antibodies will bind only to that type of biological warfare agent. Hence, antibodies can be used to identify biological warfare agents.

*Table 1. The base sequences of two DNA probes. The Ba813 DNA probe is used as an illustrative example of our approach to finding candidate organisms for biological warfare agent detector false-positives test and evaluation.**

Bacillus anthracis gene	Base sequence in DNA probe
Chromosomal gene Ba813	CATTTAGCGAAGATCCAGT
Pag gene in plasmid pXO1	ACGGCTCCAATCTACAAC

A, adenine; T, thymine; C, cytosine; G, guanine.
* DNA probes were developed by Hao et al. (2011).

B. anthracis DNA probe

Two DNA probes designed to provide real-time detection and identification of *B. anthracis* have been advanced by Hao et al. (2011). One of the probes will hybridize to the *B. anthracis* chromosomal gene Ba813. The other probe will hybridize to the *B. anthracis* pag gene in plasmid pXO1. The DNA base sequences of these probes are listed in *Table 1*. The Ba813 DNA probe will be used to provide an illustrative example of the proposed methodology.

Bioinformatics

Bioinformatics is a relatively new field derived from molecular biology, combinational biology, and computer science. In the broadest sense, bioinformatics is the application of computer science to biology; although the focus is on the application of computer science to nucleic acid and protein databases. The National Institutes of Health (NIH) defines bioinformatics as "research, development, or application of computational tools and approaches for expanding the use of biological, medical, behavioral, or health data, including those to acquire, store, organize, analyze, or visualize such data" (Huerta 2000).

Molecular biologists have created large databases on the base sequence of DNA (and RNA for RNA viruses) and the amino acid sequence of proteins for numerous species. These species include human, mouse, fish, worm, insect, bacteria, virus, and many others. Basic Local Alignment Search Tool (BLAST) is the principle tool for comparing a protein or DNA sequence with all the other sequences, including those from other species in the database (Altschul et al. 1990). In a BLAST search, the base sequence or amino acid sequence can be compared with millions of other sequences that are in the database. A BLAST search will find all exact matches and near matches. GenBank, the bioinformatics database at the National Center for Biotechnology Information (NCBI) can be searched by BLAST. GenBank currently has 196,075 different entries for bacteria and 70,955 different entries for viruses. Biologists can use this information to hypothesize the function of an unknown protein or gene based on the functions of matches or near matches. Biologists also use this information to develop phylogenetic or evolutionary relationships.

The bioinformatics false-positive testing opportunity

A species that invokes a false positive on a nucleic acid identification assay must have a sequence of bases in their nucleic acid that is identical or at least very similar to the sequence of bases in the target organism that binds or hybridizes to the nucleic acid probe or assay. The bioinformatics tool, BLAST, provides a method to search databases such as GenBank and find species that have a sequence of nucleic acid that is apt to invoke a false detection.

Method—BLAST the B. anthracis *DNA probe example*

An illustrative example based on the methodology described in Pervsner (2009) is provided to demonstrate how BLAST can be used to find organisms that are apt to evoke a false positive to nucleic acid–based biological warfare detectors. This example uses the DNA probe developed by Hao et al. (2011), which hybridizes to the *B. anthracis* chromosomal gene Ba813. The probe has a base sequence of CATTTAGCGAAGATCCAGT. This example uses NCBI BLAST online software. This example depicts the base case BLAST search. Sensitivity analysis should be employed on any parameter that the results may be sensitive to.

- Step 1. Go to the NCBI BLAST Web address: http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastHome.
- Step 2. Select the BLAST program tab. Use blastn to compare both the complementary strand of DNA in the database and the base sequence of the query.
- Step 3. Specify the sequence of interest. Since DNA is complementary, and BLAST searches both complementary strands, one can simply enter the base sequence of the probe: CATTTAGCGAAGATCCAGT.
- Step 4. Select the database. The default database for blastn is either the human or mouse database

depending upon how a switch is set. Neither of the default databases is appropriate to use when searching for microorganisms. Instead, use the nucleotide collection non-redundant (nr) database to search segments of DNA sequences or nucleotides in GENBANK, European Molecular Biology Laboratory (EMBL), DNA Data Bank of Japan (DDBJ), and Protein Data Bank (PDB) databases. Despite its name, PDB also contains DNA records.

- Step 5. Select organism. This option is used to limit the search to the group of organisms of interest. Limiting the search will improve the speed. For this example, choose bacteria.
- Step 6. Exclude. This option allows certain portions of the database to be excluded from the search. Exclude uncultured environmental samples from the search, since the objective is to identify organisms that can be obtained and used in testing.
- Step 7. Select Entrez query. This option can be used to limit the search to molecular type, molecular weight, the type of organism, and other parameters. Limiting the search will improve the speed. However, leave this field blank.
- Step 8. Select parameters to optimize program. Select "Somewhat similar sequences." This option is slower than the other options but will provide more complete results.
- Step 9. Select the maximum number of aligned sequences to display. It is important that the maximum number of aligned sequences to display is equal to or greater than the number of bases in the search. The three smallest options are 10, 50, and 100. Since there are 19 bases in the probe of interest, choose 50.
- Step 10. Expect. This is the expected number of matches that could occur by chance according to a stochastic model. Lowering the expect threshold will result in fewer "chance" matches being reported. The default is 10. Start the analysis using the default.
- Step 11. Select word size. For nucleotide searches, a word size may range from 7 to 15. A word size of 11 is the default. Lowering the word size increases the accuracy of the BLAST search but will result in slower searches. Since accuracy is more important than speed, set the word size to 7. If the BLAST search does not execute in a reasonable time, increase the word size.
- Step 12. Filter. Choose low-complexity. This will eliminate many cases that have no biological meaning, such as segments of repeating bases.
- Step 13. Mask. Choose mask for lookup table only. This is consistent with the low-complexity filter in step 12 and will eliminate many cases that have no biological meaning, such as segments of repeating bases.
- Step 14. Execute the BLAST search.

Results of the example and discussion

The results produced perfect matches to all six records of *B. anthracis*. A record is equivalent to a strain or at least a genetic variant. This is to be expected. These would undoubtedly result in true detections above the threshold concentration.

The results also produced perfect matches with *Bacillus cereus* and *Bacillus thuringiensis*. Both *Bac. cereus* and *Bac. thuringiensis* if above threshold concentration would most assuredly result in a false detection of *B. anthracis*, and this should be demonstrated by testing.

Bac. thuringiensis can be found in the soil, insects, and tree leaves (Chaufaux et al. 1997; Martin and Travers 1989). It is also used as an insecticide (Roh et al. 2007). Hence, in agricultural environments, blowing soil, blowing leaves, and environments rich in insects, *Bac. thuringiensis* is apt to occur and could result in false *B. anthracis* detections.

Similar to *Bac. thuringiensis, Bac. cereus* can also be found in the soil (Vilain et al. 2006). *Bac. cereus* is common to many environments and is known to cause food poisoning resulting in diarrhea or vomiting (Murray et al. 2002). *Bac. cereus* could contaminate biological samples in the form of blowing soil or infected human contamination.

Under certain conditions, a probe may hybridize to a segment of DNA even if the base pair match is not perfect (Nicholle 2002). Hence, an organism which has a region of DNA that is nearly complementary to the probe may hybridize with the probe and result in a false positive. Consequently, in the search for candidate organisms to use for false-detection testing, it is reasonable to consider the cases in which 18 or 17 bases match the 19-base probe. While it is almost certain that an organism with a perfect base sequence match to the probe will produce a false positive, there is considerable uncertainty regarding whether organisms with very similar base sequences to the probe will evoke a false detection. Hence, testing is desirable.

Haliangium ochraceum has a base sequence that matches 18 out of the 19 bases on the *B. anthracis* probe of interest. *Burkholderia thailandensis* has a base sequence that matches 17 out of the 19 bases on the *B. anthracis* probe of interest. Hence, both *H. ochraceum* and *Bur. thailandensis* are candidate organisms for false-detection testing of this *B. anthracis* probe.

H. ochraceum was isolated from seaweed on sandy beaches in Japan (Fudou et al. 2002). *H. ochraceum* are marine bacteria (Fudou et al. 2002; Zhang et al. 2004). They could inadvertently challenge biological detection systems in marine and littoral environments.

Bur. thailandensis is a soil bacteria from Thailand (Brett et al. 1998). It could be mixed with blowing soil and potentially challenge biological detection systems in Southwest Asia.

A brief description of how each candidate organism could contaminate a biological sample has been provided. This is a critical step in determining which of the candidate organisms should be tested. If it is determined that because of the distribution or other restriction it is virtually impossible for a candidate organism to contaminate a sample, then there is no need to test that candidate organism.

The decision to only look at perfect matches and cases in which 18 or 17 bases match the 19 base probe is based on the highly specific nature of hybridization (Nicholle 2002). However, if both the cases are of 18 and 17 bases matching the 19-base probe, it would be prudent to extend the search for possible false-positive testing candidates to weaker matching criteria.

If there is variation between strains in matching bases to the probe, choose the strain for testing that provides the desired number of base matches.

The focus of this article has been on detection using nucleic acid detection technology. If in fact the amino acid sequence of the antigen is known, then a similar procedure could be applied to immunological detection technology. In the BLAST search, instead of choosing the tab for Blastn, choose the tab for Blastp to execute a protein search. Other parameters will also need to be adjusted.

The complete and accurate identification of organisms that produce false detections in bio detectors has two critical benefits. First, it provides information essential to the evaluation. A Bayesian construct to evaluate biological detector performance based on both true and false detections has been proposed and used (Holman and Berkowitz 2009 b). Second, it allows the development of tactics, techniques, and procedures to mitigate the effects of the false detections.

Conclusion

The Bioinformatics tool, BLAST, with the Bioinformatics databases provides a useful method to systematically, accurately, and relatively completely identify candidate organisms for false-positive testing of biological warfare agent detectors. ❑

Dr. Charlie Holman is an assistant professor of biology and mathematics at Claflin University. He teaches bioinformatics, biostatistics, and data analysis. He recently retired from the U.S. Army Test and Evaluation Command where he served as the Chief of the Chemical and Biological Defense Evaluation Division. He has 21 years of experience in evaluation of chemical and biological defense systems. He is a member of ITEA and the American Society for Microbiologists. He received a bachelor of science degree in biology from the Pennsylvania State University, University Park, Pennsylvania; a master of science degree in biology from Bloomsburg State University, Bloomsburg, Pennsylvania; a master of mathematics degree from the University of South Carolina, Columbia, SC; and a doctor of philosophy degree in biodefense from George Mason University, Fairfax, Virginia. E-mail: chholman@Claflin.edu

Anika Dodds received a bachelor of science degree in bioinformatics from Claflin University, Orangeburg, South Carolina, in May 2011. This fall she will start graduate studies in biostatistics at Georgetown University, Washington, D.C. E-mail: anikadodds@bellsouth.net or adodds@claflin.edu

References

Altschul, S., W. Gish, W. Miller, E. Myers, and D. Lipman. 1990. Basic local alignment search tool. *Journal of Molecular Biology* 215 (3): 403–410.

Brett, P. J., D. DeShazer, and D. E. Woods. 1998. *Burkholderia thailandensis* sp. nov., a *Burkholderia pseudomallei*-like species. *International Journal of Systematic Bacteriology* 48: 317–320.

Canter, D. A., D. Gunning, P. Rodgers, L. O'Connor, C. Traunero, and C. J. Kempter. 2005. Remediation of Bacillus anthracis contamination in the US Department of Justice mail facility. *Biosecurity Bioterrorism* 3: 119–127.

Chaufaux, J., M. Marchal, N. Gilois, I. Jehanno, and C. Buisson. 1997. Investigation of natural strains of Bacillus thuringiensis in different biotopes throughout the world. *Canadian Journal of Microbiology* 43: 337–343.

Chipman, M., C. E. Holman, K. Finanger, J. Fuller, R. Jernigan, L. Wollenberg, B. Rodriguez, et al. 2001. System assessment of the Joint Biological Point Detection System. U.S. Army Test and Evaluation Command (ATEC) Report, Alexandria, VA.

Fudou, R., Y. Jojima, T. Iizuka, and S. Yamanaka. 2002. Haliangium ochraceum gen. nov. and *Haliangium tepidum* sp. nov. :Novel moderately halophilic myxobacteria isolated from coastal saline environments. *Journal of General and Applied Microbiology* 48: 109–115.

Gooding, J. J. 2006. Biosensor technology for detecting biological warfare agent: Recent progress and future trends. *Analytica Chimica Acta* 559: 137–151.

Hao, R., H. Song, G. Zuo, R. Yang, H. Wei, D. Wang, Z. Cui, Z. Zhang, Z. Cheng, and X. Zhang. 2011. DNA probe functionalized QCM biosensor based on gold nanoparticle amplification for *Bacillus anthracis* detection. *Biosensors & Bioelectronics*. Doi: 10.1016/j.bios.2011.01.010.

Holman, C. E., and R. Berkowitz. 2009a. Chemical and biological defense test and evaluation: A hallmark of integrating developmental and operational testing. *ITEA Journal* 30: 381–385.

Holman, C. E., and R. Berkowitz. 2009b. A Bayesian construct to evaluate chemical and biological warfare detector performance. *Phalanx* 41: 6–37.

Holman, C. E., and A. G. Loerch. 2010. Chemical and biological test and evaluation: Detector agent simulant relationship. *ITEA Journal* 31: 525–530.

Holman, C., M. Moe, C. Russell, C. Jennings, R. Bartholomew, and D. Anderson. 2008. Technical note: Whole System Live Agent Test (WSLAT) methodology assessment. U.S. Army Test and Evaluation Command (ATEC) Report, Alexandria, VA.

Huerta, Michael. 2000. BISTIC definition committee. In *Proceedings of NIH Working Definition of Bioinformatics and Computational Biology*. July 17, Bethesda, MD. National Institute of Mental Health. http://www.bisti.nih.gov/docs/CompuBioDef.pdf (accessed March 29, 2011).

Iqbal, S., M. W. Mayo, J. G. Bruno, B. V. Bronk, C. A. Batt, and J. P. Chambers. 2000. A review of molecular recognition technologies for detection of biological threat agents. *Biosensors & Bioelectronics* 15: 549–578.

Ivnitski, D., D. O'Neil, A. Gattuso, R. Schlicht, M. Calidonna, and R. Fisher. 2003. Nucleic acid approaches for detection and identification of biological warfare and infectious disease agents. *BioTechniques* 35: 862–869.

Martin, P. A., and R. S. Travers. 1989. Worldwide abundance and distribution of Bacillus thuringiensis isolates. *Applied and Environmental Microbiology* 55: 2437–2442.

Murray, P., K. Rosenthal, G. Kobayashi, and M. Pfaller. 2002. *Medical Microbiology*. St. Louis, MO: Mosby, 826 pages.

Nicholl, D. S. T. 2002. *An introduction to genetic engineering*. Cambridge: Cambridge University Press.

Olson, K. B. 1999. Aum Shinrikyo: Once and future threat? *Emerging Infectious Disease* 5: 513–516.

Pervsner, J. 2009. *Bioinformatics and functional genomics*. Hoboken, New Jersey: John Wiley & Sons.

Pohanka, M., P. Skladal, and M. Kroea. 2007. Biosensors for biological warfare agent detection. *Defense Science Journal* 57 (3): 185–193.

Roh, J. Y., J. Y. Choi, M. S. Li, B. R. Jin, and Y. H. Je. 2007. *Bacillus thuringiensis* as a specific, safe, and effective tool for insect pest control. *Journal of Microbiology and Biotechnology* 17 (4): 547–559.

Torok, T. J., R. V. Tauxe, R. P. Wise, J. R. Livengood, R. Sokolow, S. Mauvais, K. A. Birkness, M. R. Skeels, J. M. Horan, and L. R. Foster. 1997. A large community outbreak of salmonellosis caused by intentional contamination of restaurant salad bars. *JAMA: the journal of the American Medical Association* 278: 389–395.

Vilain, S., Y. Luo, M. B. Hildreth, and V. S. Brozel. 2006. Analysis of the life cycle of the soil saprophyte *Bacillus cereus* in liquid soil extract and in soil. *Applied and Environmental Microbiology* 72: 4970–4977.

Zhang, Y., Y. Li, B. Wang, Z. Wu, C. Zhang, X. Gong, Z. Qiu, and Y. Zhang. 2004. Characteristics and living patterns of marine myxobacterial isolates. *Applied and Environmental Microbiology* 71 (6): 3331–3336.

Monte Carlo Techniques for Estimating Power in Aircraft T&E Tests

Todd Remund and William Kitto
USAF AFMC 812 TW/EN, Edwards Air Force Base, California

Edwards Air Force Base, as a matter of policy, requires that statistical rigor be a part of test design and analysis. Statistically defensible methods are used to gain as much information as possible from each test. This requires:

- *Identifying statistically defensible methods and applying them to each test,*
- *Setting up tests to maximize scope of inference, and*
- *Determining the power of each test to optimize sample size.*

This article demonstrates how Monte Carlo techniques may be applied to aircraft test and evaluation to determine the power of the test and the associated sample-size requirements. Traditional methods for determining the power of a test are based on distributional assumptions associated with data. Since these assumptions may not be appropriate, a distribution-free Monte Carlo technique is presented for power assessment for tests with (possible) serially correlated data. The technique is illustrated with an example from a target-location error test. Power of the test and appropriate sample sizes are derived using Monte Carlo simulation implemented in R.

Key words: CE90; CEP; circular error; Monte Carlo simulation; power; R; resampling; sample size.

Power is a well-documented and well-researched statistical tool useful for determining the number of samples to gather in planning a scientific experiment. There are many articles and books written on the topic. Many of the more common problems in statistics, such as t-tests, ANOVA, regression, and other linear model methods, seem to have a "closed-form" solution to power calculation and sample-size determination. Despite all these efforts and successes, there is still a wide range of test conditions where statistical-power methodology is not readily available. Conventional methods of determining the power of a test assume that a distributional form for the statistic of interest is known. In some situations this is not the case, and power is difficult to estimate.

In test and evaluation (T&E) there is a need to specify the analysis of radial-error data. Target-location error seems to be one out of many arenas requiring analysis of radial error. There are many methods of calculating upper confidence limits on percentiles of radial error, and of comparing these to specifications of one sort or another. When T&E plans a test, there is an inherent question directly related to the cost: "How many data points do I need?" This leads to the need for power calculations that guide sample-size choice. But the method presents a challenging power calculation if parametric methods based on probability distributions are sought.

A new method of calculating confidence limits for radial error has been proposed (Hurwitz et al. 2011) and is now used as an example of how Monte Carlo methods can be employed to calculate power for sample-size determination. We shall use circular error (CE)—i.e., radial error—as our measure and focus on computing statistics with regard to the 90th percentile of the CE distribution. Obviously CE can be analyzed for any percentile, and the methodology in this article allows for analysis of any percentile and confidence level.

Statistical power

When performing any hypothesis test, or even in day-to-day decision making, there is always a chance that the decision that was made is wrong. If a change is determined to be necessary, it is best to identify a real change—that is, a change of practical significance. A

Table 1. Matrix depicting correct and incorrect decisions.

Truth	Decision	
	H_0: no change	**H_A: change**
H_0: no change	Correct negative	False positive
H_A: change	False negative	Correct positive

new software load may, in fact, degrade point accuracy by a fraction of a degree. This typically has no effect on tactical use of the system, so it can be ignored. There are two types of errors and two types of correct decisions. *Table 1* depicts the problems that are prevalent in any decision.

Each of the four possible outcomes has a corresponding hypothetical probability. It is obviously best to maximize the correct outcomes, subject to cost constraints. In a statistical test, the false positive outcome is fixed at a preselected level, called α. The correct negative outcome has probability in complement to α, $1 - \alpha$, and is called confidence. On the other hand, a false negative is denoted with the Greek letter β, and the corresponding complement is called power. Power is the object of interest. If it is decided that there is reason to declare a change, it is optimal to have high probability of getting it right when it is practical to do so.

Consider the process of sighting in a rifle. You need a certain number of rounds to decide on a correction to the scope or open sights. In order to do this, it is important to realize that there is always going to be some discrepancy in the scope or open sights, and possibly in the rifle as well, that will cause a bias in average hit-point location. The real decision is how big of a discrepancy warrants changing the setting on the aiming apparatus? If 100 yards is the distance of interest, is it really important if the rifle is, on average, 2 inches off? What about 6 inches? This decision determines how many rounds are needed to properly estimate the average hit point and the error surrounding that point, and ultimately amounts to a cost of sighting in the rifle.

Alternatively, consider the choice of the rifle when making a purchase. Suppose a gun shop allows a would-be buyer to test a rifle previous to buying. Suppose it is desirable that the spread is no larger than a specified radius for 90% of the shots? At the very least, a gun owner would desire to have as precise a rifle configuration as possible, given money constraints. This may influence the decision to buy one rifle configuration over another. In other words, the spread of the holes in the target is not controlled by an adjustment of a knob; it is inherent in the way the rifle was made and the exterior environmental factors that have not been controlled. Suppose the rifle is mounted in a fixed gun stand, and the environmental factors are controlled sufficiently to isolate the cause of the spread to hidden factors in the rifle. The objective at this point is to test certain rifles of interest and determine which rifles satisfy the prescribed radial-error limitation, or at least find which is closest to the limit. This example parallels the target-location error analysis done in many T&E exercises, whether dealing with coordinates from a pod or with actual hit points of a weapon.

In choosing the power and size of the detectable correction, the sample size is automatically determined for the test. Before moving on, it is helpful to give an exact definition of power in mathematical terms:

$$\Pr(\text{revert to Alternate}|\text{Alternate is the truth})$$
$$= \Pr(H_A|H_A) = \text{power}.$$

In words this says, "The probability of reverting to the alternate hypothesis given that the alternate hypothesis is correct." Essentially, this is the probability of saying there is a difference from zero when there really is a difference from zero.

Minimal alternate hypothetical population

The minimal alternate hypothetical population (MAHP) is a concept derived from statistical-hypothesis test philosophy. In statistical hypothesis testing there is a null hypothesis and an alternate hypothesis. Data are gathered for the purpose of testing the null hypothesis to see if it holds. Consider the common one-sample t-test in statistical methodology; we are testing to see if the mean of the data is equal to zero. After a power analysis is performed, a sample size is determined that provides the ability to see a certain size difference between zero and a mean that technically is not zero. This difference, labeled δ, is the *minimal* difference that researchers would like to see with specified power, or probability of correct detection. The researcher might say, "If the mean of the data is at least 4 feet, I would like to see it." This minimal difference is the smallest-size difference the researcher cares about; everything smaller than this is of no practical consequence.

Notice in *Figure 1* that if 80% power is desired and the minimal practical difference of four is determined, all other size differences of greater magnitude have higher power. So the minimal difference is the single value that is detectable across 80% of future samples. All larger differences have a higher chance of detection, and those of smaller size have a lower chance of detection. Hence the mean of the data, which is a sample from the population, may be larger or smaller

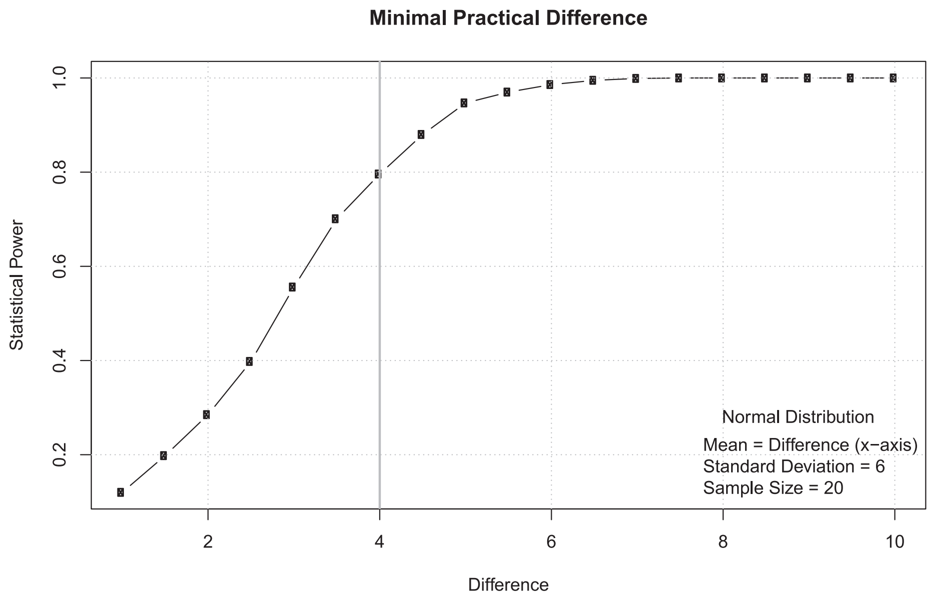


Figure 1. Practical difference and its effect on power.

than this specified difference—if smaller, it is of no consequence and operationally equivalent to zero.

If a difference is detected by the t-test, the null hypothesis is rejected and we revert to the alternate hypothesis. The idea here is to build a hypothetical population that contains the minimal difference inherently; with the case of the one-sample t-test, it becomes the mean of a hypothetical normal distribution. This minimal difference is called the minimal alternate because it is the smallest δ that causes us to revert to the alternate hypothesis in a statistical test with a specific power *(Figure 2)*.

Specifying δ is based on the primary research question; δ represents the smallest practical difference determined by researchers. As in typical power analysis, the uncertainty of the population must be estimated and integrated into the sample-size determination.

The construction of the MAHP is performed by choosing δ, determining a reference distribution—in this case the normal is appropriate—and factoring in uncertainty using an estimated standard deviation from prior sampled data from a real phenomenon. The normal distribution with δ as the mean and measured standard deviation is the MAHP.

Method

It is known at this time that there is truly a difference inherent in the MAHP; this is the minimal alternate or minimal practical difference determined by the researcher. Repeated sampling of the hypothetical population can provide a basis for power estimation. We simply pose candidate sample sizes *n* and repeatedly sample *n* values from the MAHP: We generate 100,000, or 1,000,000 values, then sample using a random sample routine from these simulated values. This is the Monte Carlo simulation.

For each sample from the MAHP, the statistical test of interest is performed and the result is recorded as a zero if the test fails to detect a difference between the sample mean and zero. A one is recorded when the test detects a significant difference from zero. Label this vector $\boldsymbol{v}$. Power is then estimated by summing $\boldsymbol{v}$ and dividing by the number of repeats. If 1,000 repeated samplings of the MAHP were performed, the divisor

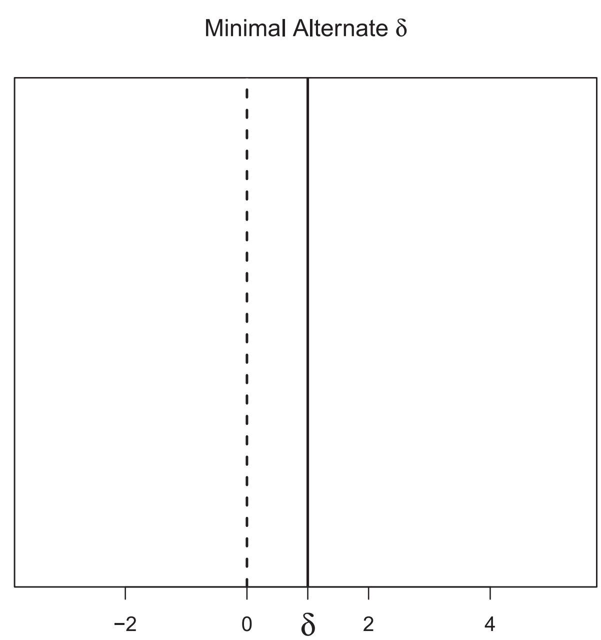


Figure 2. Minimal difference in the hypothetical population.

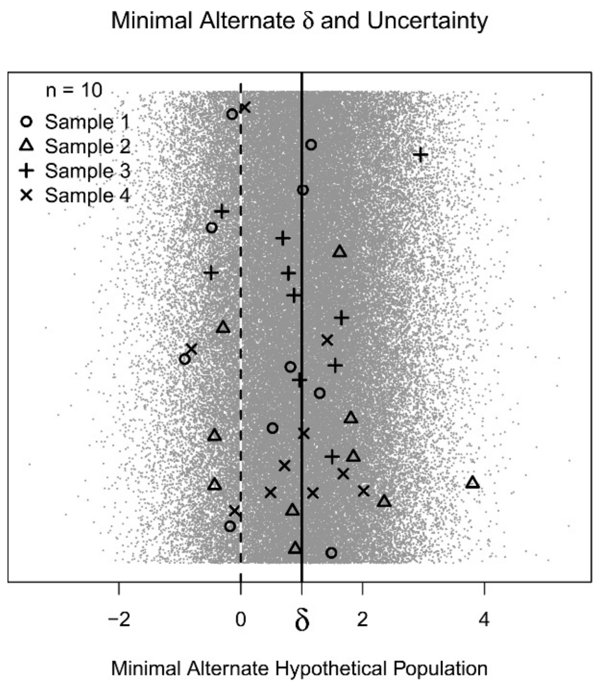


Figure 3. MAHP specification and sampling for power calculation.

would be 1,000:

$$\text{power} = \frac{1}{1{,}000}\sum_{i=1}^{1{,}000} v_i. \tag{1}$$

The MAHP in *Figure 3* was generated with a normal distribution and mean of $\delta = 1$ and standard deviation of $\sigma = 1$. A candidate sample size of 10 achieved the desired power of 80.2%; the target was 80%. In comparison to conventional power calculation, we see a slight deviation: The conventional power value for this situation is 80.31%. The slight discrepancy is due to the nature of the Monte Carlo simulation aspect of the method. The estimated value indicates that 802 of the 1,000 repeated samplings produced significant results in the t-tests.

Having stepped through the process of power estimation via Monte Carlo simulation with respect to a one-sample t-test, we present a more general and concise statement for this method in *Figure 4*.

Serially correlated data

In this section we take one step toward the unknown. If there are two samples for which the means are to be compared, a t-test may be employed. However, if there is serial correlation present in the data, the standard error of the statistic could be grossly

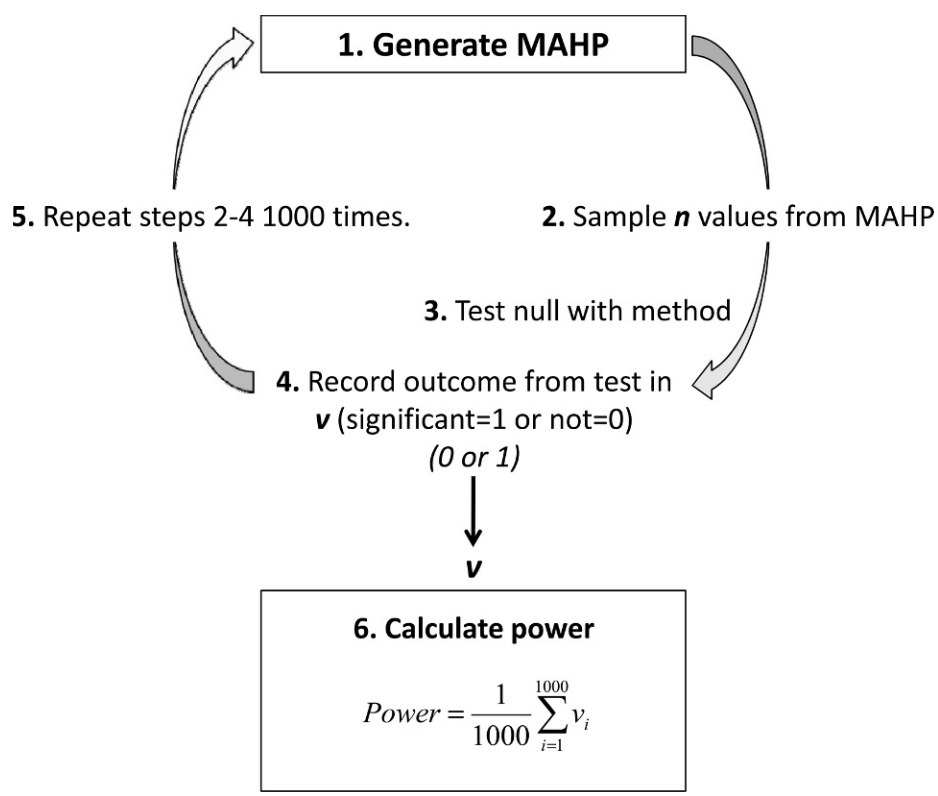


Figure 4. Monte Carlo power-estimation flowchart.

underestimated and the t-test could give misleading results. An adjustment is needed to correct the estimate of standard error.

A simple treatment of serially correlated data is found in Ramsey and Schafer (2002). A description of the general idea is

$$SE(\bar{x}-\bar{y})_{\text{adjusted}} = SE(\bar{x}-\bar{y})\sqrt{\frac{1+r_{pl}}{1-r_{pl}}}. \quad (2)$$

Standard error of the statistic is adjusted to account for serial correlation—this is from a simple case where the time-series process for the error is an order-one autoregressive process, AR(1). The estimate of the autocorrelation coefficient r_{pl} is the pooled autocorrelation between x and y, the two samples. Methods to handle more complex serial correlation patterns are found in many statistics books on time series. The test proceeds by computing a confidence interval for the statistic:

$$CI = \bar{x}-\bar{y} \pm z_{1-(\alpha/2)} SE(\bar{x}-\bar{y})_{\text{adjusted}}. \quad (3)$$

The reference distribution for z is the standard normal, with a mean of zero and standard deviation of one. Significance is detected when the interval does not span zero; if the interval does span zero, then there is no evidence of a significant difference. This is another situation where an MAHP can be created and the Monte Carlo power estimate may be used to compute power and sample size. A δ, which is the desired minimal difference, is chosen. Standard deviations must be figured, and in addition to these two items, the expected autocorrelation for each group must be found. These are the necessary components for generating the MAHP.

This example will be pursued to reveal that the generation of the MAHP needs to be done with great care. If it is done improperly in this case, the serial correlation built into the hypothetical population will be destroyed and the sample-size estimation will be incorrect.

In the previous section, the sampling was done by generating a large number of values from the MAHP, then randomly selecting values from the generated values. For serial correlation this does not work. Instead, the MAHP can be used to generate the 100,000 values, then starting indexes can be randomly drawn. Data are sampled by taking the random start index and then taking the value corresponding to the start index and the $n - 1$ values following. This preserves the serial correlation built into the MAHP. Set δ to a value of 10, meaning the difference in the means is 10 units.

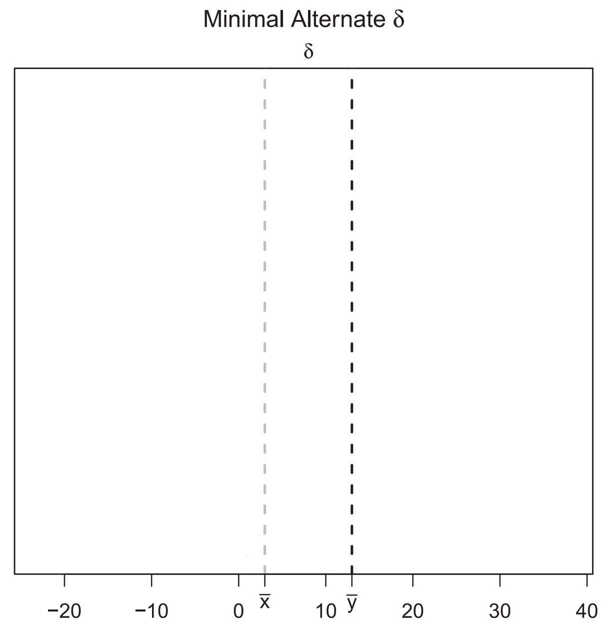


Figure 5. Minimal difference in hypothetical populations, two sample.

In *Figure 5* the difference between the means is set as just defined; the location of the differences across the number scale is not necessarily important—it is determined by what is important for the test under consideration. The next step is to incorporate the uncertainty in each group population. Both have the same standard deviation of one and the same autocorrelation coefficient of .99. Now the distribution can be generated.

It is obvious from a look at *Figure 6* that the random sampling scheme is quite different. In the last example, the samples were from random locations in the MAHP. This scenario samples a consecutive string of samples with random starting location. This is proper and necessary to preserve the serial correlation structure. It also represents real-life sampling. Any single instance of data measurement will start essentially at a random time point in the history of a process; then the samples following will be linked directly to that initial sample. It is apparent that from the last example to this example, the sampling and generation of the MAHP are done differently due to the serial correlation.

Note that the samplings done in this section and in the previous section could be drawn directly from a normal random-number generator, but it is necessary to ensure that it generates normal distributions with serial correlation, or that functionality must be built in. The sampling scheme described will work, given sufficiently large MAHPs; however, it is done in such a way as to portray the philosophy behind the creation

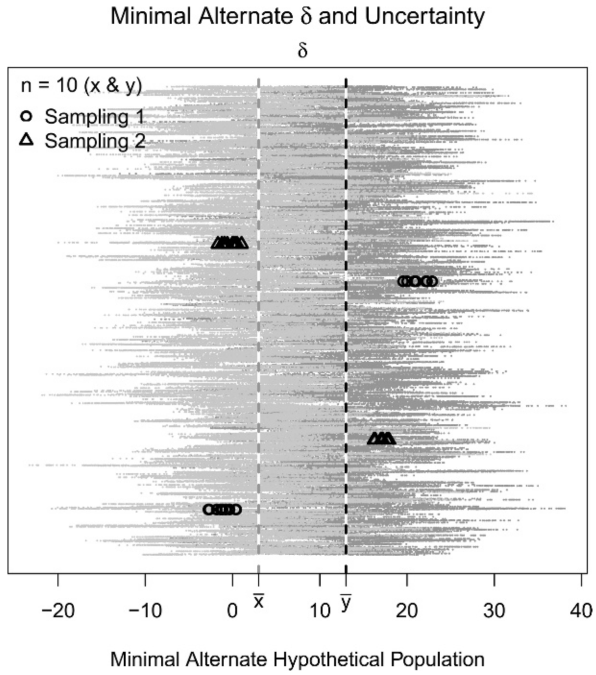


Figure 6. MAHP for the two-sample test of means with serial correlation.

of the MAHP and its use.

$$\frac{(v_1+v_2+\cdots+v_{1,000})}{1,000}=\frac{(1+0+1+\cdots+0+1)}{1,000} = \frac{862}{1,000}=0.86 \quad (4)$$

As Equation 4 shows, for a sample size of 10 the power ends up at 86.2%. That is, 862 of the 1,000 repeated samplings from the MAHP produced significant differences. This is using Equation 1 to calculate power by summing up the $\boldsymbol{v}$ vector containing the indicators of significance and dividing by 1,000. The repeated sampling from the MAHP is done in a statistical-analysis program in a loop. The process is again detailed in *Figure 4*, where the flow of the process is still the same; the MAHP and statistical test are now different from the previous example. As a side note, if the usual two-sample t-test had been employed with $\alpha = .05$ the actual confidence would be in the neighborhood of 24% and the true α for that test procedure would be .76. That is, we would have claimed a confidence of 95% and been quite mistaken. To remedy this situation, the standard error must be adjusted by Equation 2; the confidence would then be near the desired level.

So far the Monte Carlo power estimate has been presented for a one-sample t-test and a two-sample comparison of means in the presence of serial correlation; the same method works for more complex situations.

Circular-error distributions

In T&E, targeting devices are evaluated for accuracy and precision. Circular error, particularly percentiles and their upper confidence bounds, answers questions about whether the targeting device is performing adequately or not. Three methods for calculating the upper confidence bound for the percentile of CE are given by Hurwitz et al. (2011). Power for any one of these methods can be calculated using any or all of the methods presented in this article. The MAHP remains the same for evaluation across any one of these methods. *Figure 4* can be applied, with step 3 set up for the particular application. Everything else then remains the same.

In this article the Monte Carlo approach will be used to demonstrate how to find power for a CE90 (90th percentile) estimation problem. The first step is to create the MAHP. Suppose there is a limit, i.e. a CE which must not be exceeded with a targeting device. The CE percentile must not extend beyond this limit. The estimate of CE90 is sometimes used as a measure against the limit to indicate compliance. This is somewhat unhelpful, because the value itself is simply a point estimate of the actual population percentile, and hence is subject to sample variability. By using the confidence bound or a statistical test for comparison, we attach a probability statement as to whether we met the limit or not. This takes into account the stochastic nature of the data sample. The upper confidence bound with 95% confidence indicates that the true CE90 (an unknown population parameter) is less than this upper bound 95% of the time across many samplings from the actual population.

With this in mind, we can create the MAHP for CE that puts CE90 within a certain vicinity of the limit, say δ feet away. As a hypothetical example, suppose researchers wish to detect a difference between the actual value and the limit if the difference is no less than 2 feet. A specification details the CE limit to be 22 feet. The MAHP must have a CE90 limit that is detectable with power of 80% and must be no greater than 20.

So $\delta = 2$ feet with the limit set at 22 feet. The means of x and y are set to zero, the standard deviation of x is $\sigma_x = 8.87$, and the standard deviation of y is $\sigma_y = 8.87$. This produces a CE90 of 20 feet, which is 2 feet below the limit *(Figure 7)*. The correlation between x and y, $\rho = 0.8$, is used to generate data for the MAHP.

In a research setting, the specification of the standard deviations, means, and correlation come from historical data and/or expert opinion. These parameters should represent real life scenarios so that when data is gathered

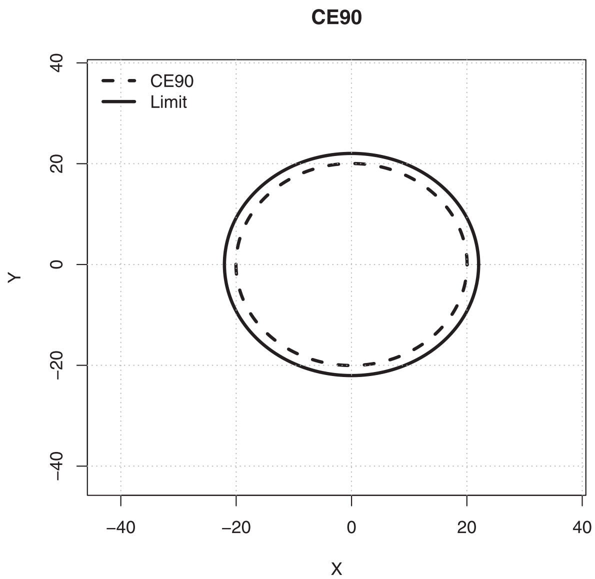


Figure 7. CE90 and the limit.

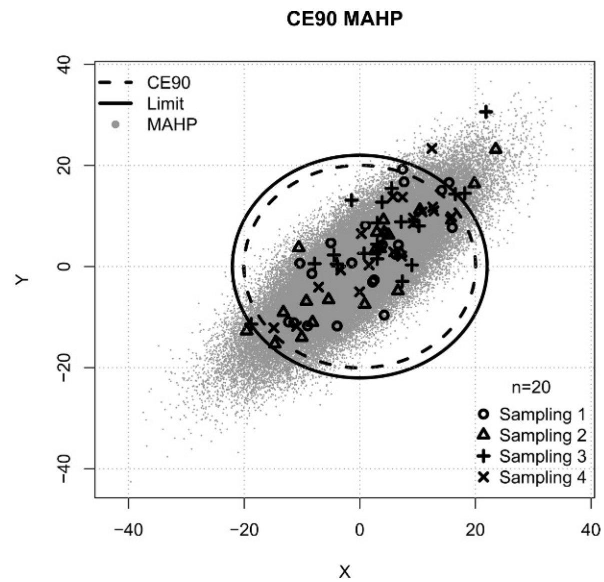


Figure 8. CE90 MAHP and samplings for power calculation.

based on the sample size calculation, actual power in the neighborhood of the planned power will be realized. The parameters described in this example are not related to any historical dataset or expert opinion, but are used for demonstration purposes only.

This is a little different from the t-test and two-sample comparison scenarios. The mean and covariance structure in the data determine δ. It becomes apparent that complete knowledge of the distribution, the statistic, and the testing method must be had before the MAHP can even be created. As the situation becomes more complex, so does the generation of the MAHP.

Generating points from the MAHP is done by drawing 100,000 samples from this bivariate normal distribution *(Figure 8)*, as described below.

$$\begin{pmatrix} x \\ y \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_x^2 & \sigma_x\sigma_y\rho \\ \sigma_x\sigma_y\rho & \sigma_y^2 \end{pmatrix}\right] \quad (5)$$

With a sample size of 300, the power is 76.5%. Again, Equation 1 is utilized in this calculation, after steps 1–5 of *Figure 4* are applied. The vector of ones and zeros used in Equation 1, $\boldsymbol{v}$, is a result of the loop and is applied to the equation. This means that in future samples from the targeting pod we will need 300 samples to ensure that if CE90 is within 2 feet of the limit, we can still detect it across 76.5% of the samples taken from the real population.

Summary

Across all three examples given in this article, it is apparent that the most difficult and time-consuming aspect of the Monte Carlo method for estimating power is the creation of the MAHP. Great care must be taken to ensure that the generation of this hypothetical population is done properly. In situations where this approach may be necessary, in-depth knowledge of the behavior of the data is necessary to ensure that the creation of the MAHP, the sampling from it, and the calculation of power are done correctly.

In all, the method is quite simple, aside from the creation of the MAHP. Create the hypothetical population to have the minimal difference determined, and incorporate the uncertainty. Generate the MAHP and consider many repeated realizations from this distribution—say, 1,000. In each situation, run the same statistical test. The proportion of times the test detects the difference incorporated into the MAHP is the power. ❑

TODD REMUND earned bachelor and master of science degrees in statistics from Brigham Young University. Since 2010 he has worked as a mathematical statistician in the Statistical Methods Group, Edwards Air Force Base. From 2007 to 2010 he worked as a mathematician at Alliant Techsystems in signal processing, data analysis for shuttle flight, and RSRM static test. E-mail: todd.remund@edwards.af.mil

DR. WILLIAM KITTO has worked at the Air Force Flight Test Center since 1976. He earned a doctorate in mathematics from the University of Washington in 1973, and graduated summa cum laude mathematic from the University of Colorado in 1969. His work at Edwards includes support for many different test programs,

beginning with the A10 and including F15, F16, B1, B2, and a variety of classified projects starting in 1980 and continuing to present. He took a year's hiatus in 1981 and taught at California Polytechnic State University, San Luis Obisbo, on an intergovernmental personnel assignment in their mathematics department. His early work was in the creation of software systems (early examples of product-line software systems, e.g., UFTAS) that were built within a flexible framework that was easily reconfigured for changing from one platform (e.g., A10) to another (F15 or F16). He later was an architect of a conceptual policy framework that would put a stamp of statistical validity on all tests conducted at the Air Force Flight Test Center, known as the Statistically Defensible Test Policy. He is now the chairman of the statistics department at the Flight Test Center, a group of consulting statisticians with advanced degrees who teach a variety of statistical courses to the base's engineering population and perform extensive consultant work for all test programs. E-mail: william.kitto@edwards.af.mil

References

Hurwitz, A., W. Kitto, T. Remund, and J. Brownlow. 2011. Estimation of location error for targeting using parametric, Monte Carlo and Bayesian techniques. *The ITEA Journal* 32 (3): in press.

Ramsey, F., and D. Schafer. 2002. *The Statistical Sleuth: A Course in Methods of Data Analysis.* 2nd ed. Pacific Grove, CA: Duxbury.

ITEA Journal 2011; 32: 471–478

A Retrospective Analysis of Lessons Learned in Evaluating Advanced Military Technologies

Craig Schlenoff, Brian A. Weiss, and Michelle Steves
National Institute of Standards and Technology, Gaithersburg, Maryland

For the past 6 years, personnel from the National Institute of Standards and Technology (NIST) have served as the independent evaluation team for two major Defense Advanced Research Projects Agency (DARPA) programs. DARPA ASSIST (Advanced Soldier Sensor Information System and Technology) is an advanced technology research and development program whose objective is to exploit soldier-worn sensors to augment a soldier's situational awareness, mission recall, and reporting capability in order to enhance situational knowledge during and following military operations. TRANSTAC (Spoken Language Communication and Translation System for Tactical Use) is another DARPA program, whose goal is to demonstrate capabilities for rapidly developing and fielding free-form, two-way speech-to-speech translation systems that enable English- and foreign-language speakers to communicate with one another in real-world tactical situations where an interpreter is unavailable. Both of these efforts are concluding, so this article focuses on overall lessons learned in evaluating these types of technologies.

Key words: ASSIST; electronic chronicle; end-user needs; evaluation design; evaluation logistic planning; mission reporting; speech; test environment; test personnel; translation system; TRANSTAC.

Over the past 6 years, the National Institute of Standards and Technology (NIST) has served as the independent evaluation team for two Defense Advanced Research Projects Agency (DARPA) efforts. The first effort, called ASSIST (Advanced Soldier Sensor Information System and Technology), has the objective of exploiting soldier-worn sensors to augment a soldier's situational awareness, mission recall, and reporting capability in order to enhance situational knowledge during and following military operations. The second program, called TRANSTAC (Spoken Language Communication and Translation System for Tactical Use), has the objective of rapidly developing and fielding free-form, two-way speech-to-speech translation systems that enable English- and foreign-language speakers to communicate with one another in real-world tactical situations where an interpreter is unavailable. Between these two efforts, NIST has orchestrated 13 live evaluations involving over 100 military personnel and foreign-language speakers at locations varying from military operations in urban terrain sites to hotel conference rooms.

In this article, we will give a brief description of each of these two DARPA efforts and describe some of the overall lessons learned from our experiences.

DARPA ASSIST and TRANSTAC efforts

This section gives a brief overview of the DARPA ASSIST and TRANSTAC efforts.

ASSIST

Soldiers are often asked to perform missions that can take many hours. Examples of missions include presence patrols (where soldiers are tasked to make their presence known in an environment for a variety of reasons), search and reconnaissance missions, and apprehension of suspected insurgents. After a mission is complete, the soldiers are typically asked to provide a report to their commanding officer describing the most important things that happened during the mission. This report is used to gather intelligence about the environment to allow for more informed planning for future missions. Soldiers usually provide this report based solely on their memory, still pictures, handwritten notes, or grid coordinates that were collected during the mission, provided these tools are available.

Figure 1. Soldiers using the ASSIST technology.

Figure 2. User interface for the ASSIST system.

These missions are often very stressful for the soldiers, and thus there are undoubtedly many instances in which important information is not made available in the report and thus not available for the planning of future missions.

The ASSIST program (Schlenoff 2006) addressed this challenge by instrumenting soldiers with sensors that they can wear directly on their uniform (as shown in *Figure 1*). These sensors include still cameras, video cameras, global positioning systems, inertial navigation systems, microphones, and accelerometers. They continuously record what is going on around the soldiers while on a mission. When soldiers return from their mission, the sensor data are run through a series of software systems that index the data and create an electronic chronicle of the events that happened throughout the time that the ASSIST system was recording (as shown in *Figure 2*). The electronic chronicle includes times that certain sounds or key words were heard, times when certain types of objects were seen, and times that the soldiers were in a specific location or performing certain actions.

With this information, soldiers can give reports without relying solely on their memory. The electronic chronicle will help jog the soldiers' memory on activities that happened that they did not recall during the reporting period, or possibly even make the soldiers aware of important activities that they did not notice when out on the mission. On top of this, the multimedia information that is available in the electronic chronicle is available to the soldiers to include in their reports, which will provide substantially more information to the recipient of the report than the text alone.

Specific technologies being developed include:

- object detection/image classification—the ability to recognize and identify objects in the environment;
- Arabic text translation—the ability to detect, recognize, and translate written Arabic text;
- sound recognition/speech recognition—the ability to identify sound events (e.g., explosions, gunshots, or vehicles) and recognize speech;
- shooter localization/shooter classification—the ability to identify gunshots in the environment; and
- soldier state identification/soldier localization—the ability to identify a soldier's path of movement around an environment and characterize the actions taken by the soldier.

TRANSTAC

The goal of the TRANSTAC program (Schlenoff et al. 2009) is to demonstrate capabilities for rapidly developing and fielding free-form, two-way translation systems that enable speakers of different languages to communicate with one another in real-world tactical situations without an interpreter.

Several prototype systems have been developed under this program, for numerous military applications, including force protection and medical screening. The technology has been demonstrated on smartphone (shown in *Figure 3*) and laptop platforms. NIST was asked to assess the usability of the overall translation system and to individually assess each component of the system (the speech recognition, the machine translation, and the text-to-speech).

All of the TRANSTAC systems work fundamentally the same. Either English speech or an audio file is fed into the system. Automatic speech recognition processes the speech to recognize what was said and generates a text file of the speech. That text file is then translated to another language using machine translation technology. The resulting text file is then spoken to the foreign-language speaker using text-to-speech technology. This same process then happens in reverse

Figure 3. TRANSTAC systems on a smartphone platform.

when the foreign-language speaker speaks. This is shown in *Figure 4*.

Lessons learned

The rest of this article focuses on some of the overall lessons learned while implementing the evaluations of the technologies described previously. Listed are nine lessons, each with brief explanatory text.

Keep your eye on the ball (the ultimate objective of the evaluation) and make sure your decisions along the way reflect that goal

As evaluation planning proceeds and new approaches and constraints are uncovered, it is often easy to get caught up in the minutiae and lose sight of the big picture. Decisions are often made that solve an immediate challenge but take you further away from the goals that are to be accomplished.

As an example in the TRANSTAC effort, one of the metrics that was used to measure the performance of the systems was a high-level concept-transfer metric that gauged how many concepts could be exchanged in a 10-minute period between the speakers using the system. Once the development teams understood this metric, they started making their systems faster at the expense of accuracy. The English- and the foreign-language speakers sometimes spoke over one another, which would have been highly impractical in a fielded environment but helped them to get though more concepts quicker. They determined that they could maximize their score using this approach even though it is not how they envisioned their fielded systems operating.

The evaluation team identified this issue and is now reconsidering using that metric at all. The test subjects in previous evaluations have consistently stated that they would happily sacrifice some translation time for greater accuracy. If this metric were continued, the

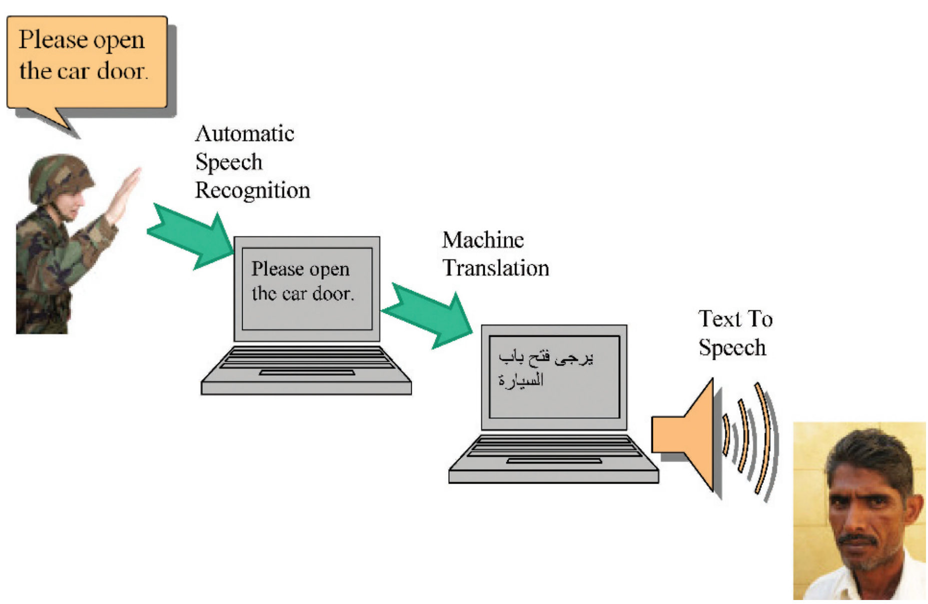


Figure 4. How speech translation works.

TRANSTAC systems would progress in a way not aligned with the goals of the program as a whole.

Deeply understand the needs and wants of the technology end users

It is usually a straightforward process to understand the exact needs and wants of technology end users in the case of testing systems that already have been fielded, where end users can categorically state what they like, what they don't like, and what they would improve. Extracting end-user needs and wants is nontrivial when it comes to testing emerging technologies with end-user groups that have yet to be specifically determined, exact use cases that have yet to be finalized, and precise usage procedures that are unclear. During the evaluation design process, it is critical for evaluation team members to speak with representatives of the intended end-user population to thoroughly understand the related challenges they face without the technology and the constraints they are bound by when presented with a new piece of equipment to carry into the field.

NIST TRANSTAC evaluation team members met with soldiers and marines on many occasions to deeply understand the challenges they faced when communicating with foreign-language-speaking personnel without a machine-translation technology. One of the most significant communication challenges currently faced is unreliable interpreters, including those that don't show up for work on time, are limited in their translation skills, or have ulterior motives when facilitating dialogue between U.S. and foreign forces. Other significant challenges include the general unavailability of interpreters. This leads to soldiers and marines attempting to have conversations with foreign-language speakers using extremely limited vocabularies. All of these challenges can lead to misunderstandings, damaged relationships, and in some instances, injuries or loss of life.

Knowledge of this challenge was also complemented by clear statements from soldiers and marines that they wanted a communication tool that was easy to use, fast and accurate with translations, small, lightweight, and durable enough to stand up to frequent use in harsh environments. This insight provided the evaluation team with a clear idea of the soldiers' and marines' needs and wants.

Realize that utility and technical performance assessments are both very important perspectives

Technology evaluations can take many forms, yielding varying types and amounts of data. Data output can yield two unique types of information: quantitative technical performance and qualitative utility assessments. Each piece of data offers unique insight into a technology's overall behavior, individual functionality, and benefit to the end user. Quantitative evaluations can offer detailed information about a system's overall functionality along with specific performance metrics related to inherent components and capabilities. Determining a technology's means of failure at the system level is an important process. Overall failures can lead to individual component or capabilities testing to identify the point of failure and determine which variables or parameters are responsible for the failure. Quantitative metrics also provide a basis of comparison among multiple evaluations and technologies. Likewise, qualitative metrics enable the evaluation team to assess the perceived worth and value the technology has to the test subjects representative of the target user population. This type of insight complements the quantitative data. For example, a technology could be 100% accurate in its function, yet if it is too heavy to carry, users will seldom use it and will therefore place a low value on it. Individually, both of these data types paint very contrasting pictures. It is important that the data be viewed together to get a complete understanding.

NIST's evaluations of advanced technologies have demonstrated a need to collect both types of data. In both the ASSIST and TRANSTAC programs, evaluations were conducted of technologies that had yet to be finalized and deployed to actual end users. This means that the evaluation team's analysis of the collected quantitative and qualitative data was crucial to informing the technology developers and program sponsors on the current state of the systems, including specific successes and areas for improvement. Across both programs, quantitative data were captured that assessed individual technology components, capabilities, and systems. For example, component-level evaluations of the TRANSTAC systems' automatic speech recognition, machine translation, and text-to-speech demonstrated specifically which of these components produced errors ultimately leading to system errors. Also, both programs captured qualitative data at the capability and system levels. For example, capability-level evaluations of the ASSIST technologies enabled the evaluation team to capture specific feedback from soldiers about which technology capabilities (e.g., real-time data sharing or image annotation) were of the most value, easiest to use, etc. Likewise, this specific information, coupled with the other collected data, enabled the evaluation team to paint a clear picture of the technologies' current state.

The NIST evaluation teams have employed an evaluation approach that captures a range of quantitative and qualitative data. This allows the creation of a definitive picture of the technologies' current successes, shortcomings, and areas that must be improved.

Understand that there are often multiple approaches to evaluating a technology, so it is crucial to identify those that will achieve the overall evaluation goals, given the test constraint

There are many approaches for evaluating systems. For any particular evaluation effort there are also various constraints that much be considered, e.g., logistical, budgetary, and programmatic concerns. Method selection must consider these concerns; otherwise, the assessment effort and results may be compromised in undesirable ways. NIST's evaluation framework advocates identifying evaluation goals and user requirements, and then identifying evaluation methodologies that support those test parameters. Once the set of evaluation methodologies that can support the evaluation have been identified, then method selection can be further refined by other logistical parameters, such as availability of qualified personnel to design and conduct the assessment, type of testing environment needed to execute the test, mechanisms needed to collect the data, and data-analysis considerations, e.g., whether time and resources exist to code many hours of video data. Approaches that do not have contingency avenues for high-risk elements should be avoided if possible. For example, if an approach calls for a specific test environment, e.g., military operations in urban terrain, but there is a high probability that the test will be bumped from the site, a feasible fallback location is needed. If no reasonable fallback location is available, alternate approaches should be considered or a determination should be made that test delays are acceptable.

Understand the interactions of the technology with the test environment and the test personnel to be mindful of the technology's ideal operating conditions and its boundaries

The performance of the system under test is greatly and directly related to the environment in which it is being tested and the personnel that are using the system. Slight changes to either one of these factors can often have a significant effect on how well the system performs. For example, the competency of the end user in operating systems similar to the ones being tested can be the difference between success and failure. In addition, the end user's experience in scenarios where the technology would be useful and understanding of how the technology can be best applied is also a critical factor.

Apart from the individual user, many other variables can play a significant role in how well a system performs. In the case of the TRANSTAC systems, these variables may include background noise, distance between the microphone and the speaker, glare issues, dustiness of the environment, wind conditions, dialects of the speakers, etc. Almost none of these variables are true-false; there are various levels that must be understood.

No matter how familiar one gets with a type of technology, nobody knows a specific system better than its developer. However, the developer also has a vested interest in ensuring that the system works as well as possible. For both the DARPA TRANSTAC and ASSIST efforts, regular interaction occurred between the evaluation team and the developers of the technologies. In every case, the developers provided suggestions for the best ways to test the systems and the most appropriate variables to vary. In parallel with this, the evaluation team always spoke with the end users of the technologies (primarily military personnel) to better understand the environments in which the technologies were expected to be used, including variables such as background noise, temperature, weather conditions, etc. Understanding that the technologies were still under development and not yet ready to be fielded, the evaluation team took both sides into consideration and tried to find the proper balance between realism and the known shortfalls of the systems.

Realize that the background and experience of the test subjects can greatly affect their impression of the systems under test

Test subjects—those individuals using a technology during an evaluation in which qualitative or quantitative data are collected—greatly affect data quality by their actions during the test. Their actions are dictated by both the technology training they receive prior to the evaluation and their specific backgrounds and experiences. The latter may include experiences with similar technologies or experiences within the operating environments within which the technologies under test are envisioned for use.

NIST's involvement in six TRANSTAC technology evaluations from 2007 to 2010 has highlighted the fact that the impressions of the soldiers and marines selected as test subject are greatly influenced by their specific backgrounds and experience. A specific example of this can be seen in assigning evaluation scenarios to marines and soldiers. The evaluation team goes to great lengths to assign test subjects scenarios with which they have intimate knowledge, based upon their own deployment experiences and interactions with foreign personnel. Since the evaluation scenarios are categorized within six domains, the soldiers and marines are queried to see how their experiences correlate. For example, a civil-affairs marine would reasonably be assigned the civil-affairs scenarios and could also be paired with some of the facilities-inspections scenarios, based upon their experiences. Conversely, an infantry officer would most likely be suited for the vehicle-checkpoint/traffic-control-point, combined-training, and combined-operations scenarios. Allowing test subjects to use the TRANSTAC systems to facilitate dialogues they are intimately familiar with supports the capture of targeted feedback. The test subjects will have high confidence in stating what worked well and what needs to be remedied with the technology in order for the system to be successful in an actual situation. Likewise, if test subjects are paired with scenarios with which they have little familiarity, then their dialogue struggles have great potential to negatively influence their perception of the technology.

Be cognizant that the structure and content of the technology training and the feedback requests of the test subjects greatly influence the test subjects' perceptions

Any training provided to subjects on the technology to be tested will have an impact on their interaction with the system and subsequently on their perceptions of the technology. Decisions regarding the amount and type of training required to achieve the test objectives must be made. Complex systems can present additional challenges in attempting to train participants. Some questions to be addressed are: How much training is needed? How long will it take and what is the schedule impact? Where will training take place? If training is conducted in the test environment, will that impact the test results in undesired ways? What training materials are needed, e.g., scenario content or task content? Are the training materials different from or similar to the test materials, and what is the impact of that? Who can provide appropriate, unbiased training? The developers know their systems the best, but they are not unbiased. Testing personnel may not be qualified to conduct training for complex systems.

Removing interactions between system-developer personnel and test subjects can help with controlling those influences on the test subjects; however, there may be advantages of system-developer involvement that lead the evaluation designers to consider having the developers involved during the evaluation period. For example, it may be beneficial to the sponsoring program to have its developers see and learn firsthand how their systems are received and hear subjects' concerns. Also, as mentioned before, the systems may be sufficiently complex that only the system developers can provide adequate training, or be so prototypical in nature that only the developers can set some configuration options (because these controls may not yet have been exposed at the user interface). For off-the-desktop systems, various physical configurations may need to be fitted to each test subject each time the system is deployed. In any of these cases, a simple inquiry of "So, how was it?" and the resulting discussion can have an impact on what the subject ultimately reports in their official assessment feedback. When system developers have access to the test subjects during the testing period, appropriate ground rules need to specified and enforced to control the effect of these influences.

Note that there are often multiple options available to assess specific metrics, so it is critical to identify those options that are optimal for producing the desired assessments

There are typically quite a few measures that can be collected for use in assessing any particular metric. Which measures or assessors are selected may have an impact on what is collected and reported; therefore, careful attention should be paid to these choices. Additionally, some measures are more or less difficult to collect, some are more costly to collect than others in terms of resources needed, some are logistically more difficult to put in place, and so on. Choices here can impact the cost of the assessments as well as the logistical feasibility of completing the data collection and analysis for assessment, so careful attention to these considerations during the measure-selection process is prudent.

For example, when obtaining feedback from subjects, two examples of assessments could be free-form and Likert-type survey responses. Free-form responses typically consist of open-ended responses that need to be coded or categorized for analysis. Likert-type responses to well-formed queries allow quantitative assessment of the data. Assessments for the latter type of data can often be much faster to perform than analysis of free-form responses, and can give quite different perspectives of the same experience interaction.

A case in point is documented in Steves and Morse (2009). In the early stages of the TRANSTAC evaluations, utility data were collected solely via survey instruments. Although a combination of Likert-type response questions and free-form inquiries was used, the free-form responses became repetitive and sparse over the course of the evaluation period. Adding semistructured interviews and the resulting gathered data provided very rich insights into the survey-based data and the user experience overall. However, the cost to collect and analyze the additional data was definitely greater.

Be mindful that your metrics and evaluation approach may need to evolve over time

It is typical for evaluation requirements and concerns to evolve over time, especially if the time span in which the assessments are performed is long or if there are a large number of unknowns at the beginning of the design phase. As more is learned about the system and user requirements, initially envisioned approaches may need to be modified to provide useful assessment of the system. For example, in testing a prototype system, the initial assessment goals may include user testing, but as more is learned, it may be determined that the user interface is not sufficiently developed for users. In this case, another approach could be used, such as expert review, to provide some formative feedback for developers regarding how to move forward to support their eventual users effectively. Understanding of the system, its requirements, its state of development, and user requirements may impact the initial assessment vision, as that vision may not have had the benefit of the understanding gained during the initial design phase.

For example, in both projects the systems were evolving over time. Improvements to existing capabilities were made and new features added between evaluations. This required that changes in what was assessed—and, at times, in how it was assessed—be made. In particular, an early TRANSTAC platform was a laptop; in the field, it was a laptop in a backpack, where the screen could not be viewed and the systems would overheat easily. In the last evaluations, the platform was a smartphone. This meant that field evaluations could be more realistically situated in later evaluations.

Keep the high-level objective of the evaluation in mind and be flexible as modifications need to be made.

Discussion

In this article, we describe the evaluation approach that has been applied to two DARPA-funded efforts over the past 6 years and focus on nine lessons that have been learned during that time. This is not meant to be a comprehensive list of all the factors that should be considered when evaluating these types of systems, but instead represents some of the most critical ones as determined by the authors.

The main lesson described in this article is that additional effort put into the design and logistics planning of the evaluation up front can pay off quite a bit as the evaluation progresses. The design stage of the evaluation is critical, and decisions made during that time have a huge effect on how successful the evaluation will be. Bad decisions in the design can be very difficult to fix later on. This can be compared to the cycle of manufacturing product development: Problems that are identified and resolved in the design stage of a product can cost orders of magnitude less to fix than those same problems if they are not identified until the manufacturing or distribution phases. ❑

CRAIG SCHLENOFF is the acting group leader of the Knowledge Systems Groups in the Intelligent Systems Division at the National Institute of Standards and Technology (NIST). His research interests include performance-evaluation techniques applied to autonomous systems and manufacturing. He has served as the program manager for the Process Engineering Program at NIST as well as the director of ontologies at VerticalNet, Inc. He leads numerous million-dollar projects dealing with performance evaluation of intelligent systems. He received his bachelor's degree in mechanical engineering from the University of Maryland, College Park, and his master's degree from Rensselaer Polytechnic Institute. E-mail: craig.schlenoff@nist.gov

BRIAN A. WEISS is a mechanical engineer at NIST. His focus is the development and implementation of performance metrics to quantify technical performance and assess end-user utility of intelligent systems. His schooling has been through the University of Maryland, including a bachelor of science in mechanical engineering and a professional master of engineering; he is working towards his doctor of philosophy degree in mechanical engineering. His research is targeted at developing the Multi-Relationship Evaluation Design (MRED) framework as a tool that will handle uncertainty in its automatic design of evaluation test plans. E-mail: brian.weiss@nist.gov

MICHELLE STEVES is an information-systems analyst at NIST. Her research interests include utility and usability assessments for desktop and off-the-desktop systems. Recent efforts have included assessments of soldier-worn sensor systems, spoken-language translation systems, a latent-print examination system, and biometrics data-gathering systems. She has a bachelor degree in mathematics and computer science from Western Maryland College, now known as McDaniel College. E-mail: michelle.steves@nist.gov

References

Schlenoff, C. 2006. ASSIST: Overview of the first advanced technology evaluations. In *Proceedings of the 2006 Performance Metrics for Intelligent Systems (Per-*

MIS) Conference, August 21–23, 2010, Gaithersburg, MD, ed. Madhavan, R, and E. Messina, Gaithersburg, MD: National Institute of Standards and Technology.

Schlenoff, C., B. Weiss, M. Steves, G. Sanders, F. Procter, and A. Virts. 2009. Evaluating speech translation systems: Applying SCORE to TRANSTAC technologies. In *Proceedings of the 2009 Performance Metrics for Intelligent Systems (PerMIS) Conference,* September 21–23, 2009, Gaithersburg, Md, ed. Madhavan, R. and E. Messina, Gaithersburg, MD: National Institute of Standards and Technology.

Steves, M., and E. Morse. 2009. Utility assessment in TRANSTAC: Using a set of complementary methods. In *Proceedings of the 2009 Performance Metrics for Intelligent Systems (PerMIS) Conference,* September 21–23, 2009, Gaithersburg, Md, ed. Madhavan, R., and E. Messina, Gaithersburg, MD: National Institute of Standards and Technology.

Acknowledgments

This work was supported by the Defense Advanced Research Projects Agency ASSIST and TRANSTAC programs (program manager: Mari Maeda).

DARPA disclaimer

The views, opinions, and/or findings contained in this article are those of the authors and should not be interpreted as representing the official views or policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the Department of Defense.

ITEA Journal 2011; 32: 479–488

Multi-Relationship Evaluation Design: Formalizing Evaluation-Design Input and Output Blueprint Elements for Testing Developing Intelligent Systems

Brian A. Weiss

National Institute of Standards and Technology, Gaithersburg, Maryland

Linda C. Schmidt, Ph.D.

University of Maryland, College Park, Maryland

Intelligent technologies within the military, law-enforcement, and homeland-security fields are continuously evolving. Testing these technologies is crucial to (a) inform the technology developers of specific aspects for enhancement, (b) request end-user feedback, and (c) verify the degree of the technology's capabilities. Test exercises provide valuable data that both update the state of the technology and present information to the evaluation design team to aid further testing. Evaluation designers have exerted substantial effort in creating methodologies to streamline the test-plan development process. This is particularly evident when producing comprehensive test plans. The Multi-Relationship Evaluation Design (MRED) methodology is being developed to collect input from several source categories and automatically output evaluation blueprints that identify pertinent test characteristics. MRED captures input from three specific categories: personnel stakeholders, the technology state, and the available resources. This information and the relationships among these inputs are merged to feed an algorithm that will output specific test-plan elements. This article will propose a model of developing a technology's state and its influence on the MRED-output. MRED defines the input technology-state category to include the maturity, reliability, and repeatability of a technology under test. The states of these three characteristics evolve as a technology is developed from the conceptual stage to a fully functional system. Likewise, test characteristics evolve to capture the most pertinent data to enhance this development process. In order to ensure that the appropriate test designs are generated, it is critical to understand the relationships between these input and output elements. These relationships are also described in this article. Future efforts will describe and formalize the entire MRED model as relationships are further investigated between all of the inputs and the test-plan output elements.

Key words: Appropriate data; components; input elements; intelligent systems; maturity; metrics; output elements; reliability repeatability; test planning.

Intelligent technologies are continuously being developed for use in military domains, law-enforcement situations, and first-response incidents. These technologies are distinguished by their interactions with human operators and/or robotic elements to achieve specific goals. Assessing these technologies is crucial to update the system creators during the development process and validate the performance of the final systems (Weiss et al. 2010).

Most intelligent technologies are designed by or for the government. It is common for the government to fund these developmental programs on multiyear schedules. These programs are distinct from commercial product-development efforts in that the government organizes its programs in several phases. Each phase usually consists of one or more prescribed test events evaluating technologies created by one or more development teams. It is common for the technology-development and evaluation-design processes to be entwined.

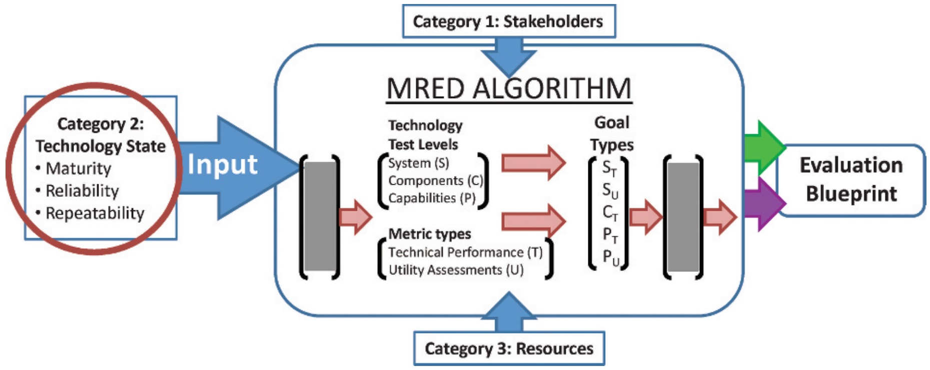


Figure 1. MRED model including inputs and outputs.

Both private and government organizations have expended a considerable amount of effort on the research and development of methods and frameworks to effectively and thoroughly evaluate the capabilities of intelligent technologies. Many of these customized test-design methods have been adequate to evaluate precise technologies and accomplish project-specific objectives. No single method has been recognized as being capable of evaluating quantitative and qualitative performance across a range of prototype and physical technologies, encompassing both human-controlled and autonomous capabilities. Test design can be an arduous and challenging process due to technology complexity. Evaluation designers also face another obstacle in that the test-planning activities are prepared manually, where modifications to the unknown and known information may require them to redesign their test exercises. Many of these test methodologies have been presented in prior work (Weiss et al. 2010; Weiss and Schmidt 2010a; Weiss and Schmidt 2010b). The authors have designed the Multi-Relationship Evaluation Design (MRED) methodology to address these shortcomings. Specifically, the MRED methodology is being created to take multiple inputs from numerous input source categories and automatically output evaluation test plans (also called blueprints).

Technology-state characteristics are challenging to capture in any test-design process, and modeling their influence on test plans is critical to MRED's success. In this article, the MRED model will be presented; detailed definitions and relevant relationships of the technology-state input category will be discussed; the output test-plan elements of technology test levels, metrics, and test environments will be defined and their constraints presented; the technology state's influence on determining the technology test levels and test environments will be discussed; and an example of this cause-and-effect relationship will be highlighted in example test plans for a robot arm.

Multi-Relationship Evaluation Design (MRED)

MRED's objective is to automatically generate evaluation test plans based upon multiple inputs (Weiss and Schmidt 2011). The MRED methodology will take information from three input categories and output one or more evaluation blueprints complete with their own specific test-plan elements. MRED will also characterize the relationships among inputs and the influences inputs have on outputs.

The MRED methodology model describes the important design inputs into the planner and the output evaluation blueprint (Weiss et al. 2010; Weiss and Schmidt 2010a; Weiss and Schmidt 2010b). *Figure 1* presents the overall MRED model. The MRED algorithm will operate on the categories' inputs to generate appropriate evaluation blueprints. The technology-state input category is a main focus of this article. Likewise, the output evaluation blueprint elements of technology test levels, metric types, and goal types are discussed in detail, so they are centrally highlighted, as well.

Input categories

Stakeholders. Stakeholders are classified into six categories of parties interested in a technology's evaluation. Members of these categories have their own motivation for the test-plan results of a technology's performance. Their individual motivations will reflect personal uncertainties manifesting in test-design preferences. The six stakeholder categories are buyers, users and potential users, evaluation designers, evaluators, sponsors and funding sources, and technology developers. These categories are listed in *Table 1* (see Weiss and Schmidt 2011 for more detail).

Technology state. Three factors are selected to describe the technology's anticipated state of development at the time of its test. These factors are presented in *Table 2* and discussed in greater detail later.

Table 1. Stakeholders.

Stakeholder group	Who they are
Buyers	Stakeholders purchasing the technology
Users and Potential Users	Stakeholders that will be or are already using the technology
Evaluation Designers	Stakeholders creating the test plans by determining MRED inputs
Evaluators	Stakeholders implementing the evaluation test plans
Sponsors and Funding Sources	Stakeholders paying for the technology development and/or evaluation
Technology Developers	Stakeholders designing and building the technology

Resources. The final input group is comprised of specific types of material, manpower, and technology to be included in the testing exercise. Resource availability (or lack thereof) and limitations can have a significant impact on the final evaluation design. These categories are shown in *Table 3*.

Output elements

This section presents the output evaluation-blueprint elements that have been specified to date. Technology test levels and test environments are briefly described here and elaborated upon in greater detail in the following sections.

Technology test levels. A system (often called a "technology") is made up of constituent components representing a physical hierarchy or set of levels. Likewise, the system's overall performance is made up of constituent capabilities representing a functional hierarchy or set of levels. There are several terms related to these technology test levels:

- *System:* a group of cooperative or interdependent components forming an integrated whole to accomplish a specific goal;
- *Component:* an essential part or feature of a system that contributes to the system's ability to accomplish a goal; and
- *Capability:* a specific ability of a technology, enabled by either a single component or multiple components working together. A system is made up of one or more capabilities.

Test environments. The setting in which the evaluation occurs, the test environment, can influence the behavior of the personnel and limit the ability to test technology at certain levels of maturity. MRED defines three distinct environments:

- *Lab:* a controlled environment where test variables and parameters can be isolated and manipulated to determine how they impact system performance and/or the users' perception of the technology's utility;
- *Simulated:* an environment outside of the lab that is less controlled and limits the evaluation team's ability to control influencing variables and parameters, since it tests the technology in a more realistic venue; and
- *Actual:* the domain of operations in which the system is designed to be used. The evaluation team is limited in the data they can collect, since they cannot control environmental variables.

Other blueprint elements. Metrics and goal types are two of the remaining outputs from the MRED methodology.

Metrics, or measures, are performance indicators that can be observed, examined, detected, or perceived either manually or automatically. In turn, metrics are the result of the analysis of one or more output measures (Weiss et al. 2010). Specifically, there are two types of metrics:

- *Technical performance:* metrics related to quantitative factors (such as accuracy, precision, time, distance, etc.); and
- *Utility assessments:* metrics related to the qualitative factors that express the condition or status of being useful and usable to the target user population.[1]

Goal types are a dependent variable determined by combinations of technology test levels and desired metrics. There are five goal types that are output from the MRED framework, listed in *Table 4*.

It is important to note that utility assessments cannot be captured in component evaluations. This is

Table 2. Technology state factors.

Factor	Definition
Maturity	Technology's state or quality of being fully developed
Reliability	Technology's ability to perform a required function under stated conditions for a specified period of time
Repeatability	Technology's ability to yield the same or comparable results as in previous test(s)

Table 3. Resources of testing and analysis.

Resource	Description
Personnel	Individuals that will use the technology, those that will indirectly interact with the technology, those that will collect data during the test, and those that will analyze the data following the test(s)
Test Environment	The physical venue, supporting infrastructure, artifacts, and props that will support the test(s)
Data-Collection Tools	The tools, equipment, and technology that will collect quantitative and/or qualitative data during the test(s)
Data-Analysis Tools	The tools, equipment, and technology capable of producing the necessary metrics from the collected evaluation data

because components are defined as parts that technology users are unable to engage or interact with during realistic operations. The remaining output evaluation-blueprint elements are presented in *Table 5*.

Input category: technology-state factors

The technology-state factors are described by three elements: *maturity, reliability, and repeatability*. These three factors must be known (as much as possible) and understood with respect to a given technology to design an effective test plan for that specific technology. A technology's design and construction include that of its components. As components are integrated together, they enable specific capabilities. Some of the technology's capabilities may be operational before the entire system is fully functional. Throughout the technology's development cycle, its maturity, reliability, and repeatability are constantly evolving. For instance, if several components have a nonfunctional maturity, then they cannot be tested. But if the components are functional, yet not fully functional, then it is likely that limited testing can occur.

Component and capability relationships

All intelligent systems are composed of components that are integrated to enable a system to perform one or more capabilities. For example, suppose the system to be tested is an intelligent Cartesian robotic arm (these types of control movements are similar to those of a human using multiple joints in harmony to reach for a cup). This specific example features an arm that is composed of six joints (a combination of revolute and prismatic joints) and an end-effector gripper. The entire assembled arm is considered the system. Further, each of the six joints and the gripper are considered components. The capabilities in this instance would be the x-, y-, and z-translations of the gripper; the roll, pitch, and yaw of the gripper; and the grasping of the gripper.

Table 4. Goal types.

Component-Level Testing—Technical Performance
Capability-Level Testing—Technical Performance
System-Level Testing—Technical Performance
Capability-Level Testing—Utility Assessment
System-Level Testing—Utility Assessment

A distinction critical to this work is that technology end users interact with capabilities, not components. This means that the users are focused on the success of the motions (x, y, z, roll, pitch, yaw, and grasping) of the robotic arm, which are its capabilities—not on any of the components (i.e., prismatic joints, revolute joints, and gripper). To simplify the presentation of this example, the links between the joints and other common elements (drive motors, base, etc.) of the robotic arm are not considered.

Maturity

Maturity must be input into MRED for a technology test level to be considered for testing. The maturity level can be for the system (i.e., the overall technology) and for each individual capability and component that are to be tested. At any time during development, the maturity of the system, its components, and its capabilities will fall into one of the following classes:

- *Nonfunctional:* The technology test level being tested has yet to be developed or is in the process of being developed, so that it is not functional and therefore cannot be tested.
- *Functional:* The technology test level being tested is developed to the point of being functional, yet is not complete (still requires additional development).
- *Fully developed:* The technology test level is developed to the point of being functional and complete.

Maturity data are gathered from the technology developers. These stakeholders are in the best position to provide these data, since they are most familiar with the technology and are likely to have the most up-to-date information.

Reliability

Like maturity, reliability is defined for the system and the individual components and capabilities that are to be tested. Reliability is the probability that a portion of the items will survive under certain conditions for a certain time. Reliability will be represented as either "No Data" (if data have never been collected) or a

Table 5. Other evaluation blueprint elements.

Other element	Definition
Personnel—Evaluation Members	Various individuals and groups are required to perform an effective evaluation. They are classified into two categories: primary (direct interaction) technology users and secondary (indirect interaction or evaluation support). The primary technology users are defined as Tech Users. These individuals directly interact with the technology during the evaluation. Secondary personnel are those that indirectly interact with the technology during the evaluation. This includes Team Members and Participants. Both primary and secondary personnel are discussed in greater detail in the following sections as their selection relates back to the Stakeholders' preferences.
Evaluation Scenarios	The Evaluation Scenarios govern exactly what the technology users will encounter during the test and the challenges within the identified Test Environments. Three types of Evaluation Scenarios are Technology-based, Task/Activity-based, and Environment-based.
Explicit Environmental Factors	The Explicit Environmental Factors are characteristics within the environment that impact the technology and therefore influence the outcome of the evaluation. These factors pertain to the overall physical space, which is composed of Participants, structures, and any integrated props and artifacts. These factors are broken down into two characteristics, Feature Density and Feature Complexity. Together, these two elements determine the Overall Complexity of the environment.
Data Collection Methods	Data Collection Methods are used to capture experimental and ground truth data, depending upon the technology being evaluated and the specified Test Environment. No matter the type of tools used, Data Collection Methods are characterized by factors that influence the techniques being employed.
Personnel—Evaluators	There are three classes of evaluation personnel that are necessary to ensure that the evaluation proceeds according to plan and that the necessary data are captured to evaluate a technology's performance. They fall into the three classes of Evaluators: Data Collectors, Evaluators: Test Executors, and Evaluators: Safety Officers.
Data Analysis Methods	The Data Analysis Methods blueprint element will be a dependent variable that is specified based upon other blueprint elements, including Data Collection Methods and Metrics. These methods are specific to the technology under test and the available resources, and are therefore not specified in greater detail.

numerical value ranging from 0% to 100%. Reliability data are collected from an independent third party, which could be the evaluators or evaluation designers.

Depending upon the prior test data that are known and provided, reliability data will be either directly assessed from quantitative data or extracted from qualitative data. For example, quantitative reliability data can be captured from technical performance evaluations relating to either system- or component-level tests. These are usually represented as a percentage. Qualitative reliability data are captured from utility assessment evaluations completed for either system- or capability-level tests. These data are usually represented on a scale signifying an average perception from test subjects. An example qualitative scale would be 1 = very unreliable, 2 = unreliable, 3 = marginally reliable, 4 = reliable, 5 = very reliable. It would be the evaluation designers' responsibility to correlate the qualitative reliability data to the numerical range of 0% to 100%.

It is important to note that reliability of a system cannot simply be strictly calculated by using component or capability reliability test data. This statement is justified by the following principles:

- "The sum is greater than the parts." Just because components and/or capabilities perform at various reliabilities when individually tested does not mean that they will perform at an aggregated reliability when the entire system is tested.
- "The parts can be greater than the sum." A test subject may have a stronger opinion of a technology in tests that allow a focus on specific capabilities, as compared to tests where the subject is forced to select among or operate multiple capabilities within a system. For example, a test subject could be easily overwhelmed when provided multiple capabilities to employ, as compared to being given a single capability to use.
- "Tests are unique." Component and capability tests, which isolate individual elements, are typically unique in comparison to system tests, where multiple components and capabilities are tested in parallel.

Repeatability

Repeatability is defined as a technology's ability to yield the same or comparable results as those in previous tests. A technology's repeatability can be presented similarly to its reliability: Repeatability can be represented as either "No Data" or a range from 0 to 100%. These data are also gathered by an independent third party.

Repeatability conveys different information from reliability and is measured differently. This is seen in that reliability data can be obtained from a single data set, whereas repeatability must be obtained across multiple data sets. The evaluation designer must consider the scope of the technology and tests when

Table 6. Robotic arm components and capabilities.

	Capabilities						
	Translation			Rotation			
Components	*X* (P_1)	*Y* (P_2)	*Z* (P_3)	Roll (P_4)	Pitch (P_5)	Yaw (P_6)	Grasping (P_7)
Revolute Joint 1 (C_1)	X	X				X	
Revolute Joint 2 (C_2)		X	X		X		
Prismatic Joint 1 (C_3)	X	X	X				
Revolute Joint 3 (C_4)	X		X	X			
Prismatic Joint 2 (C_5)	X	X	X				
Revolute Joint 4 (C_6)				X	X	X	
Gripper (C_7)							X

determining how many data sets are necessary to adequately state the reliability and repeatability of a component, capability, or entire system. Note that repeatability can be measured for almost any type of metric. The following example highlights maturity and reliability. Repeatability will be addressed in future work.

Robotic-arm example

The technology-state factors of maturity and reliability are highlighted in the following example featuring the robotic arm introduced previously. The robotic arm to be tested is comprised of the seven primary components (represented as C_1 through C_7) that produce seven capabilities (represented as P_1 through P_7), whose relationships are shown in *Table 6*. This matrix can be interpreted in several ways. Individual components can be examined to see which capabilities they contribute. In this case, Revolute Joint 2 (C_2) contributes to the capabilities of *y*-translation (P_2), *z*-translation (P_3), and pitch (P_5). Each column of the matrix displays the components necessary to produce a specific capability. For example, yaw (P_6) is controlled by Revolute Joint 1 (C_1) and Revolute Joint 4 (C_6).

Suppose that the seven components of the robotic arm have the various levels of maturity at time *t*, as according to *Table 7*. Note that the maturity levels of these components would be supplied by the technology developers.

This table is split into different regions depending upon the state of the corresponding components and their relationships to the capabilities (capability maturity is dependent upon component maturity).

- Nonfunctional: Grasping, P_7, is a nonfunctional capability because its lone component, C_7, is nonfunctional.
- Nonfunctional to functional: The rotation motions (P_4, P_5, and P_6) may fall anywhere in the range of nonfunctional to functional capabilities. This is because at least one contributing component, C_6, is nonfunctional, while the other contributing components—C_1, C_2, and C_4—are either functional or fully developed. The specific levels of maturity in this instance would be based upon additional queries by MRED of the technology developer.
- Functional: Translations in *x*-, *y*-, and *z*-directions (P_1, P_2, and P_3) are functional capabilities, since their constituent components are either functional (C_3, C_4, and C_5) or fully developed (C_1 and C_2).
- Fully developed: A capability falling into this category would be impacted by components that

Table 7. Influence of component maturity on capability maturity at a given time.

		CAPABILITIES						
		Translation			Rotation			Grasping
COMPONENT MATURITY	COMPONENTS	X (P_1)	Y (P_2)	Z (P_3)	Roll (P_4)	Pitch (P_5)	Yaw (P_6)	(P_7)
Fully-Developed (FD)	Revolute Joint 1 (C_1)	X	X				X	
Fully-Developed (FD)	Revolute Joint 2 (C_2)		X	X		X		
Functional (FN)	Prismatic Joint 1 (C_3)	X	X	X				
Functional (FN)	Revolute Joint 3 (C_4)	X		X	X			
Functional (FN)	Prismatic Joint 2 (C_5)	X	X	X				
Non-Functional (NF)	Revolute Joint 4 (C_6)				X	X	X	
Non-Functional (NF)	Gripper (C_7)							X
	CAPABILITY MATURITY	FN	FN	FN	NF to FN	NF to FN	NF to FN	Non-Functional

Table 8. Influence of component reliability on capability reliability.

		CAPABILITIES						
COMPONENT RELIABILITY	COMPONENTS	Translation			Rotation			Grasping
		X (P_1)	Y (P_2)	Z (P_3)	Roll (P_4)	Pitch (P_5)	Yaw (P_6)	(P_7)
99%	Revolute Joint 1 (C_1)	X	X				X	
98%	Revolute Joint 2 (C_2)		X	X		X		
72%	Prismatic Joint 1 (C_3)	X	X	X				
65%	Revolute Joint 3 (C_4)	X		X	X			
51%	Prismatic Joint 2 (C_5)	X	X	X				
3%	Revolute Joint 4 (C_6)				X	X	X	
No Data	Gripper (C_7)							X
	CAPABILITY RELIABILITY	23.6%	35.6%	23.4%	1.95%	2.94%	2.97%	*No Data*

are all fully developed. This example does not contain any capabilities in this category, although there are several components that are fully developed. This is because no single capability is solely influenced by these fully developed components.

Since the system is the sum of its components and capabilities, it is plausible that the system's maturity could range from nonfunctional to functional. The extent of its functionality would also be ascertained from direct queries to the technology developer.

Table 8 provides an example of component reliability influencing capability reliability. These data assume that capability reliability cannot be measured directly and that it is the product of the reliabilities of those components that influence that specific capability. This means that the reliability of P_1 = Reliability(C_1) $\times$ Reliability(C_3) $\times$ Reliability(C_4) $\times$ Reliability(C_5). The reliabilities of the remaining capabilities would be calculated similarly. If the reliability of a specific capability is available from direct measurement, it is possible this value could differ from that derived by traditional means of calculating system reliability.

Based upon the example information provided in *Table 8*, it is not practical to test any capabilities that are reliant upon component C_6, because this component's reliability is so low (indicated by the stated maturity of nonfunctional), as seen in *Table 7*. Some of the capability reliabilities may appear low in this example, yet these could be reasonable data for those technologies that are undergoing constant development.

Output elements

A majority of the output elements presented in *Figure 1* are influenced by the technology-state factors. A glimpse of this is seen in the earlier section with respect to maturity of the technology test levels. This section will take a deeper look at the relationships among three of the output elements that are impacted by this input category. Specifically, technology test levels and the test environment will be discussed with respect to their influences on one another, and the following section will examine the relationships between them and the technology-state factors. It is important to note that the technology-state factors influence more output elements than these three that are highlighted. Conversely, these two output elements are influenced by more than just the technology-state factors. *Table 9* presents a portion of the overall input-category/output-element relationship matrix.

The input/output relationships presented in this article are highlighted in green in *Table 9*; those highlighted in red are presented extensively elsewhere (Weiss and Schmidt 2011). The remaining relationships will be discussed in future work.

Technology-state factor influence on technology test levels, metrics, and test environments

The technology-state factors impact the available technology test levels and test environments. Evidence of this is seen in the robot-arm example. Given the maturity of the components, capabilities, and system stated in *Table 7*, it is important to identify those technology test levels that can be tested and those that cannot. The maturity data presented in *Table 7* are reorganized in *Figure 2*. The relationships illustrate that the system's maturity is dependent upon the capabilities' maturity, which in turn is dependent upon the components' maturity.

The information provided in *Figure 2* enables the generation of *Figure 3*, which highlights the varying levels of testing that could be performed on the technology test levels. The availability of technology test-level elements for testing is a single example of the numerous evaluation-blueprint characteristics that MRED would output. This example only shows the influence of maturity data. In reality, the reliability and

Table 9. A portion of the overall input-category/output-element relationship matrix.

	INPUT	OUTPUT						
		GOAL TYPES		EVALUATION PERSONNEL			TEST ENVIRONMENT	EVALUATION SCENARIOS
		Technology Levels	Metric Types	Tech Users	Team Members	Participants		
CATEGORY 1: STAKEHOLDERS	Buyers		X	X	X	X	X	X
	User, Potential User			X	X	X	X	X
	Evaluation Designer	X	X	X	X	X	X	X
	Evaluation Executor			X	X	X	X	X
	Sponsor/ Funding Source	X	X	X	X	X	X	X
	Technology Developer	X	X	X	X	X	X	X
CATEGORY 2: TECHNOLOGY STATE	Maturity	X	X	X			X	X
	Reliability	X	X	X			X	X
	Repeatability	X	X				X	X

repeatability data, coupled with stakeholder preferences (e.g., stakeholders only want to test those individual components whose reliability is >70%), have the potential to further delineate which technology test levels should be tested and which should not for a given evaluation.

Relationships involving goal types (combination of technology test levels and metrics) have also been discussed in prior work (Weiss et al. 2010). In summary, the more advanced a technology, the more likely it is capable of operating in an actual environment. Using the robot-arm example, basic tests (at a minimum) should be performed on the individual components to attain a measure of confidence that they will behave as intended when integrated with each other to produce various capabilities and, ultimately, form the entire system. Premature integration can lead to catastrophic failure of multiple components, resulting in unnecessary financial and time loss. It is probable that component testing would take place in a controlled lab environment where a specific input can be produced and component-specific output data are measured. It is not practical (or plausible) to isolate and test an individual joint in a factory setting (i.e., simulated environment) or on a busy assembly line (i.e., actual environment). Advanced testing of the entire system can be performed when the technology is more fully developed. However, it is virtually impossible to isolate a component during system-level testing.

Based upon the information provided in *Figure 3*, it is reasonable to state that MRED would output test plans that call for testing in the lab and/or simulated environment. The actual environment would be a premature test venue, given that the system and several components are nonfunctional at this time. The simulated environment could be a reasonable option, given that several components are either fully developed or fully functional. The lab environment would be a preferred venue to examine individual capabilities and components to isolate specific behaviors and control specific test variables. Of course, stakeholder preferences (discussed in Weiss and Schmidt 2011) and resources (to be presented at a later date) influence the selection of the environment(s).

Conclusions and future work

The simple robot-arm example illustrates MRED's broad potential to be applied to the evaluation design of complex commercial systems. MRED's development

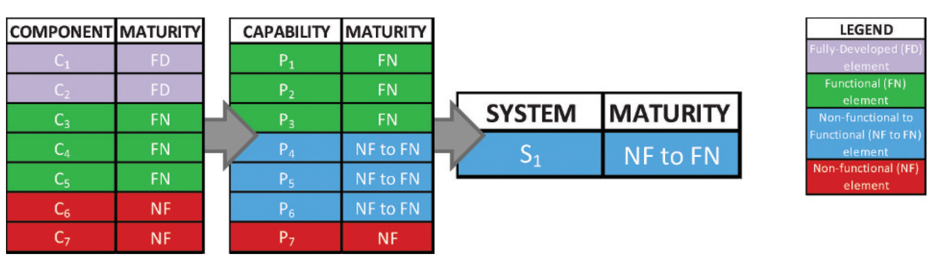


Figure 2. Maturity of the robotic-arm technology test levels.

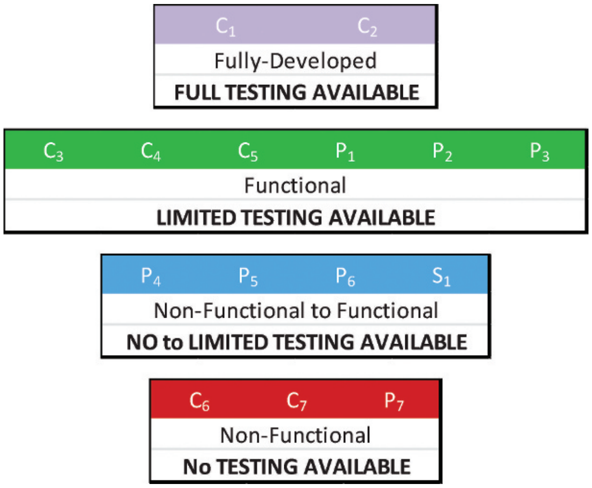


Figure 3. Technology test levels available for testing.

has also been supported by other test efforts, including those sponsored by the government. The National Institute of Standards and Technology and members of the Army Research Laboratory's Collaborative Technology Alliance have collaborated to design and execute evaluations to test multiple pedestrian-tracking algorithms (Bodt et al. 2009). The joint team worked together from 2007 through 2010 to plan and implement numerous test events. This work was used as an example in earlier reporting on the development of MRED (Weiss et al. 2010; Weiss and Schmidt 2010a; Weiss and Schmidt 2010b). The pedestrian-tracking example will continue to be explored using MRED. Upcoming efforts will formalize the relationships between input categories and output evaluation elements. It is anticipated that the expansion of the model shown in *Figure 1* coupled with the input/output relationships shown in *Table 9* will yield a mathematical formalization. This formalization will leverage principles from linear algebra and matrix manipulation to support the development of MRED's driving algorithm.

MRED continues to be defined by detailing the input technology-state factors and their influence on the evaluation-blueprint characteristics of technology test levels and test environment. The robot-arm example will be used to further elaborate upon the metrics and evaluation scenarios along with other MRED output-blueprint elements. Likewise, the input resources category will be explored to see its impact on test blueprints once these data are subsumed into MRED. Further investigation will continue in examining the input categories and output-blueprint elements to build upon the discussed relationships. Ultimately, MRED's model will be solidified and its algorithm defined so that test plans can be generated, given the necessary input data. This will enable evaluation designers, sponsors, etc., to quickly change their evaluation direction and/or test goals in the face of changing requirements. The rapid emergence of advanced and intelligent systems justifies methodologies such as MRED. It is envisioned that this automated test-planning methodology will improve the pace of development and delivery of intelligent systems. ❑

Brian A. Weiss has been a mechanical engineer at the National Institute of Standards and Technology in Maryland since 2002. His focus is the development and implementation of performance metrics to quantify technical performance and assess end-user utility of intelligent systems throughout various stages of development. He has a bachelor of science in mechanical engineering from the University of Maryland and a professional master of engineering from the University of Maryland, and is working towards his doctor of philosophy in mechanical engineering at the University of Maryland. E-mail: brian.weiss@nist.gov

Dr. Linda C. Schmidt is an associate professor at the University of Maryland. She holds a doctorate in mechanical engineering from Carnegie Mellon University and bachelor of science and master of science degrees in industrial engineering from Iowa State University. She is active in teaching design research in theory and practice. She has also coauthored texts on engineering decision making and product development. E-mail: lschmidt@umd.edu

Endnotes

[1]Utility is defined as the status of being useful and usable to the technology user and is not meant in the economic sense.

References

Bodt, B., R. Camden, H. Scott, A. S. Jacoff, T. Hong, T. Chang, R. Norcross, T. Downs, and A. Virts. 2009. Performance measurements for evaluating static and dynamic multiple human detection and tracking systems in unstructured environments. In *Proceedings of the 2009 Performance Metrics for Intelligent Systems (PerMIS) Workshop,* September 21–23, 2009, Gaithersburg, MD, eds. Madhaven, R. & Messina, E., 166–173. New York, NY: ACM.

Weiss, B. A., and L. C. Schmidt. 2010a. The Multi-Relationship Evaluation Design framework: Creating evaluation blueprints to assess advanced and intelligent technologies. In *Proceedings of the 2010 Performance Metrics for Intelligent Systems (PerMIS) Workshop.* September 28–30, 2010, Baltimore, MD, eds. Messina, M. and Tunstel, E., New York, NY: ACM.

Weiss, B. A., and L. C. Schmidt. 2010b. The Multi-Relationship Evaluation Design framework: Producing evaluation blueprints to test emerging, advanced, and intelligent systems. In *Proceedings of the 2010 International Test and Evaluation Association (ITEA) Annual Symposium*. September 13–16, 2010, Glendale, AZ.

Weiss, B. A., and L. C. Schmidt. 2011. Multi-Relationship Evaluation Design: Formalizing test plan input and output elements for evaluating developing intelligent systems. Forthcoming. In *Proceedings of the ASME 2011 International Design Engineering Technical Conferences (IDETC)—23rd International Conference on Design Theory and Methodology (DTM)*, August 28–31, 2011, Washington, D.C. New York, NY: ASME.

Weiss, B. A., L. C. Schmidt, H. Scott, and C. I. Schlenoff. 2010. The Multi-Relationship Evaluation Design framework: Designing testing plans to comprehensively assess advanced and intelligent technologies. In *Proceedings of the ASME 2010 International Design Engineering Technical Conferences (IDETC)—22nd International Conference on Design Theory and Methodology (DTM)*, August 15–18, 2010, 603–616, Montreal, Quebec, Canada. New York, NY: ASME.

Acknowledgments

The authors would like to acknowledge Harry Scott and Craig Schlenoff from the National Institute of Standards and Technology's Intelligent Systems Division for their continued support.

ITEA Journal 2011; 32: 489–496

Reliability Program Scorecard and Reliability Growth Planning and Assessment Models

Nancy Welliver and Marguerite Shepler
U.S. Army Materiel Systems Analysis Activity,
Aberdeen Proving Ground, Maryland

Sustainment costs typically represent 60–70 percent of a program's life-cycle costs. System reliability is a key component of maximizing system availability and minimizing life-cycle costs. Key areas affected by system reliability changes include maintenance labor, repair material, and spares. Overall, an improvement in reliability can represent a significant savings in the support costs for any major weapon system (often in the tens of hundreds or millions of dollars or higher).

The Reliability Program Scorecard and Reliability Growth Planning and Assessment Models are two methods which can be implemented to assist management in meeting system reliability requirements and enable the implementation of both Office of the Secretary of Defense and Army reliability policies (Directive-Type Memorandum 11-003 entitled "Reliability Analysis, Planning, Tracking, and Reporting," 21 March 2011 [DOD 2011], and the Army Reliability Policy, December 2007, respectively [Bolton 2007]).

Key words: Reliability; Reliability Program Scorecard; Reliability growth, planning, tracking, and projection.

In December 2007, the Army Acquisition Executive established a new Reliability, Availability, and Maintainability Policy (Bolton 2007). The policy was developed in response to data that showed a significant number of Army systems failing to demonstrate established reliability requirements. The policy applies to all programs with a Joint Potential Designator of Joint Requirements Oversight Council, "Interest," in accordance with Chairman of the Joint Chiefs of Staff Instruction 3170.01F, dated May 1, 2007 (CJCS 2007). The goal is to cost-effectively increase the reliability of Army systems and encourage use of cost-effective reliability best practices. The policy provides a mechanism to alert key Army leaders when weapon systems are off track with respect to meeting their reliability requirements. One element of this policy is for the U.S. Army Test and Evaluation Command (ATEC) and Army Materiel Systems Analysis Activity (AMSAA) to review the materiel developer's Reliability Program Plan or Reliability Case and other documentation to determine the risk of a system's not achieving its reliability threshold. As a part of the Department of Defense Reliability Improvement Working Group, AMSAA and ATEC developed a Reliability Program Scorecard to assess a system's reliability program.

The Reliability Program Scorecard examines a supplier's use of reliability best practices, as well as the supplier's planned and completed reliability tasks. The purpose of the scorecard is to allow for early evaluation of an acquisition program to identify reliability gaps. One benefit of the scorecard is that it ensures that the program office and contractors think about reliability very early in the acquisition process and throughout a program's life cycle. The scorecard was developed based in part on reliability assessment approaches developed by the Institute of Electrical and Electronics Engineers (IEEE 2008), Raytheon (Raytheon Company, 2007), Alion (Alion 2008), the University of Maryland (Tiku 2005), and others. AMSAA and ATEC expanded and refined the individual assessment areas based on several years of evaluation and reliability program experience. Quantitative risk scores are provided for each assessment area as well as for the overall system. This scorecard is important for tracking the achievement of reliability requirements and rating the adequacy of the overall reliability program. It allows for a standard way of assessing programs.

AMSAA Reliability Scorecard Program Name	To rate an element: Click the cell(s) that you wish to rate (in the *Rating* column), then click the desired rating button (to the right). Macros must be enabled for the Scorecard to function properly.	H	M	L	NE	C

Cat.	#	Element	High Risk Criteria	Medium Risk Criteria	Low Risk Criteria	Rating
Reliability Requirements and Planning	1	**Routinely builds and updates Reliability Case/Reliability Program Plan during product development**	Developer has no Reliability Case/Reliability Program Plan or the plan contains minimal content.	The Reliability Case/Reliability Program Plan does not demonstrate that the developer has an understanding of the reliability requirements, the plan to achieve the requirements is questionable in terms of implementation, and/or progress towards meeting the requirements is documented sporadically. The Reliability Case does not provide the customer assurance that the contractor is pursuing the reliability best design practices or testing activities. The reliability plan is questionable in terms of realistic timelines, testing, and product design activities to produce a product that meets the reliability requirements.	The Reliability Case/Reliability Program Plan documents that the developer has a clear understanding of the reliability requirements, the plan to achieve the requirements is reasonable and achievable, and progress towards meeting the requirements is regularly updated. The Reliability Case provides the customer assurance that the contractor is aggressively pursuing design practices and testing activities consistent with industry high performers. The developer has a reliability program plan that is based upon realistic timelines, testing, and product design activities that will produce a product that meets the reliability requirements.	
	2	**Reliability activities that are integral to design and testing are clearly identified and incorporated into the program Integrated Master Schedule**	Reliability activities that are integral to design and testing are not identified and/or are not incorporated into the program Integrated Master Schedule.	Some reliability activities that are integral to design and testing are identified and some are incorporated into the program Integrated Master Schedule. Delays associated with incorporating corrective actions may not be realistic or are not incorporated into the schedule.	All reliability activities that are integral to design and testing (including realistic delays associated with incorporating corrective actions) are clearly identified and incorporated into the program Integrated Master Schedule.	
	3	**Reliability Program provides delivery dates for specific products and/or analyses to provide progressive assurance (routine evidence) that the program is on track to achieving the reliability requirements**	Reliability Program does not outline a schedule with delivery dates for specified products and/or analyses.	Reliability Program outlines a schedule with delivery dates for some of the specified products and/or analyses. The schedule is not sufficiently detailed to determine whether the program is on track to achieving the reliability requirements.	Reliability Program outlines a detailed schedule with delivery dates for specified products and/or analyses to provide the customer with information to determine if the program is on track to achieving the reliability requirements.	

Figure 1. Selection of Reliability Program Scorecard.

As time progresses and additional programs are evaluated, more data will be gathered which will allow evaluators to compare similar programs. It is intended that engineers and subject-matter experts conduct site visits in order to fully examine program documents and assess the processes in which suppliers intend to do analysis and testing on their products and systems. With respect to requirement areas that are lacking, engineers can recommend analysis or testing solutions and methods that the contractors could use to ensure that they are meeting the reliability standards. Through engagement of subject-matter experts, the scorecard ratings will become more accurate and each program will have a better chance of achieving success. AMSAA and ATEC will be able to re-evaluate and adjust element weighting data on the scorecard to accurately reflect the success of a program according to the scorecard assessment. Scorecard metrics will be collected over time so the Reliability Program Scorecard will continue to be an extremely valuable tool to make an initial reliability projection for a program. One of the ultimate goals is to evolve to a point where the scorecard elements and weights can be adjusted for different system types or phases in the acquisition life cycle.

Scorecard areas

The scorecard evaluates eight critical areas: reliability requirements and planning, training and development, reliability analysis, reliability testing, supply-chain management, failure tracking and reporting, verification and validation, and reliability improvements. Many of the elements in the scorecard were derived from the list of reliability best practices and categorized using the IEEE P1624 *Draft Standard for Organizational Reliability Capability* (IEEE 2008). *Figure 1* shows a snapshot of three of the 10 elements within the Reliability Requirements and Planning section of the scorecard.

An early scorecard may be based solely on a Reliability Program Plan. As a program matures, information is gathered through the Reliability Case, updates to the Reliability Program Plan, site visits, and involvement in technical reviews. A feedback process will occur, allowing all reliability members to understand the perceived risk to that program.

As a part of the feedback process, it is important for suppliers and the program office to have an explanation as to why an element received a particular risk level. The scorecard includes two additional columns to provide such information to the supplier and program office. The first column is for the evaluator to include rationale for assessing the element at that risk level. The second column allows the evaluator to include suggestions to improve the reliability program, based on the risk level assigned to the category or element. Many of the ways to improve reliability come from the TechAmerica Government Electronics and Information Technology Association's Standard 0009, *Reliability Program Standard for Systems Design, Development, and Manufacturing* (TechAmerica 2008), and the IEEE P1624 *Draft Standard for Organizational Reliability Capability* (IEEE 2008).

The first category in the scorecard is Reliability Requirements and Planning. There are 10 elements within this category which are an important part of understanding the customers' reliability requirements, generating reliability requirements, and planning activities that are necessary to ensure that appropriate reliability requirements are met. There are several activities that define the requirements-and-planning section. One is identifying and planning for available resources (personnel, testing, equipment, materials, etc.). Another is creating reliability requirements and allocating them to the subassemblies or components. Identifying the potential suppliers for the product and their history of reliability capabilities is another activity that is evaluated in the requirements-and-planning section. Gathering lessons learned from previous programs and documenting them is essential to not repeating history. Once the lessons learned have been documented, the contractor should detail what the current plan is to avoid the pitfalls of predecessors. This should be updated in the Reliability Program Plan or Reliability Case. An integral part of the reliability program is tying the reliability activities to the program's schedule. The activities to be included within a reliability program need to be clearly identified as to when they will occur, so that when and if the program schedule shifts, the reliability and design team can make adjustments as needed and include these changes within the Reliability Program Plan or Reliability Case. It is important to determine the reliability analysis, reliability testing, and failure-data analysis/tracking needed to ensure the system meets its requirements. These elements should be included in the initial contract with the supplier. The logistics for obtaining feedback on results of reliability activities must also be identified.

Training and Development is the second category in the scorecard. There are four elements that describe the steps necessary to improve people's technical and strategic skills and knowledge so they can properly execute their responsibility in the design, evaluation, and manufacture of a reliable product or system. There are a few activities that help to define the area of training and development. One of these activities is developing and implementing a training plan for both individual contributors and management, including schedule, budget, and identification of training personnel. This could include internal courses, external seminars, college classes, and symposiums. It is important to monitor new developing technologies, modeling and analysis techniques, and trends that impact reliability in order to regularly adjust training.

The Reliability Analysis category includes six separate elements. Among these are conducting thermal and vibration analyses or finite element analysis, characterizing critical loads and stresses, and understanding failure mechanisms and failure sites. One of the activities of the reliability analysis area is identifying the failure implications of components and products, e.g., creating a reliability logic diagram. Some others are identifying potential single points of failure and failure modes, failure mechanisms and their effects, and the criticality of failure modes and mechanisms for a system. Detailed component stress and damage models should be utilized when appropriate. An additional activity for reliability analysis is assessing adherence to design rules that impact reliability derating and to electrical, mechanical, and other guidelines.

The fourth category is Reliability Testing. There are four elements within this category which are important in identifying design weaknesses and exploring design limits and environments. There are several activities that are key practices in the testing area. First, detailed reliability test plans should be created that include sample size for tests and confidence-level specifications. Then, discovery testing must be performed—identifying the design margin and destruct limits for the system. Design verification and reliability testing should be conducted, including ongoing reliability tests, reliability demonstration testing, and accelerated testing. The results from reliability tests should be used to change the design in products before production of the system begins.

Supply Chain Management is the fifth category in the scorecard. There are five elements within this category that will aid in determining that sources of components are identified which will satisfy reliability requirements. Several activities define the supply-chain management section. The first few include creating a list of potential suppliers, selecting a vendor/supplier, and assessing or auditing the supplier. Component qualification is important, including identification and evaluation of key parameters. Another activity is review of component monitoring data from suppliers, including process, quality, reliability testing, accelerated test data, and field-failure data. Checking to ensure that design specifications include the stated reliability requirements is also an important activity in supply-chain management. Another activity that should be present in the Reliability Program Plan is a contingency plan for part obsolescence. In general, the supply-chain management section will need to be evaluated as programs progress toward the production phase, and may include making on-site visits to contractor facilities and gathering performance history of vendors.

The Failure Tracking and Reporting category includes three elements that focus on collecting failure data in order to generate corrective actions and reliability-improvement activities. One of the activities for failure tracking and reporting is conducting statistical analyses of functional-test failure data, manufacturing-test failure data, reliability-test failure data, and field-return failure data. It is critical to track the history of the failed components from production through failure and to prepare failure analysis reports identifying failure modes and mechanisms traced to specific materials or processes. Pareto charts and other statistical reports can be prepared based on failure modes and mechanisms. Lastly, the need to determine appropriate corrective actions and analyze data from prognostic monitoring sensors is addressed in the scorecard. Other activities that should also be included are root-cause analysis and failure confirmation. Time-to-failure data should be collected and analyzed, and engineering-based reliability predictions should be updated based on information gathered.

The seventh area of the scorecard is Verification and Validation. There are four elements within this category that are meant to ensure that planned reliability activities are implemented and to validate that outcomes are consistent with other activities. Activities that apply to verification and validation include conducting internal audits to monitor progress and improve reliability plans and activities; verifying completion of root-cause analysis, taking corrective action, and preventing recurrence of appropriate factory-test and field-return failures; verifying that reliability commitments in supplier agreements have been satisfied; comparing predicted product reliability and failure distributions with actual field reliability and failure distributions; comparing potential failure modes and mechanisms identified in analysis with failure modes and mechanisms from field returns; and comparing actual field reliability with reliability requirements.

Reliability Improvements is the final category in the scorecard. There are four elements included in this category that focus on the identification and implementation of product changes based on lessons learned from testing, failures, technical improvements, and changing operating conditions. Several activities define the improvements area. Two of them are implementing corrective actions identified through failure analysis and monitoring the effectiveness of the corrective actions in improving reliability. Another is identifying and implementing changes that will prevent recurrence of identified failure modes and mechanisms in future products. It is also important to identify changes in the product reliability requirements or their life-cycle application conditions that may require subsequent action. An additional activity in reliability improvements is evaluating developing technologies, modeling and analysis techniques, and trends that could be used to improve reliability of products. Lastly, documenting and implementing corrective actions that will improve reliability in response to product reliability-requirement changes is necessary.

Scoring methodology

Each of the 40 elements in the scorecard is reviewed and rated red, yellow, or green, representing a high, medium, or low risk, respectively. For each element, text is provided in the scorecard to assist the evaluator in assigning the risk levels. An additional rating of gray (representing an element not evaluated) is used for individual elements that are not present and are not a necessary part of the current system's reliability program, given the program's position in the acquisition framework. Once risk levels and not-evaluated (NE) inputs have been assigned, the scorecard will provide a summary of the risk assessment (located in a separate worksheet labeled "Overall Scorecard Results"), which is pictured in *Figure 2*.

The risk-assessment score is calculated based on the individual reliability risk ratings assigned to each element (red equals a risk of 3, yellow equals 2, and green equals 1). All elements that are rated NE are removed from the risk-score calculations. The element-risk scores are then adjusted by weighting factors (1, 2, and 3) that are locked within the scorecard spreadsheet tool. The overall reliability risk is then normalized to a value between 1 and 100.

The "Overall Scorecard Results" worksheet also provides a pie chart, similar to the one pictured in *Figure 3*, for the entire scorecard documenting the number of elements that were rated high, medium,

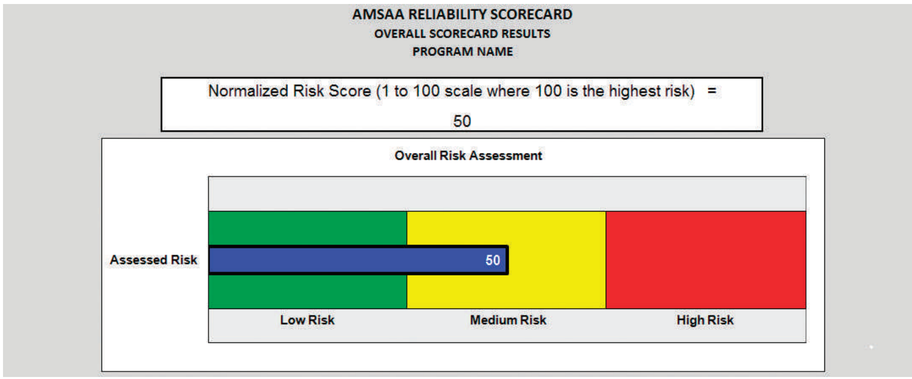


Figure 2. Overall program risk assessment.

low, or NE. Pie-chart results and normalized risk score for each category can also be viewed by clicking on the "Category Results" tab of the Reliability Program Scorecard tool.

As more programs are evaluated and data become available, the weights of the different elements may be refined to produce a final risk-assessment score. Programs that are evaluated early and re-evaluated as the system progresses into the production and deployment phase will help to shape the scorecard. If elements are found to not influence reliability, they can be removed; if they influence reliability more than expected, a higher weight can be assigned to them. If the reliability program scorecard predicts that a program's reliability program is at high risk but the program meets or exceeds the reliability threshold, then the scorecard and the program can be re-evaluated to examine the discrepancy. However, if the scorecard produces acceptable results—in that a program with low risk meets or exceeds the reliability threshold, or a program with initial high risk makes changes based on scorecard results and then produces a product that meets or exceeds the reliability threshold—this will validate the scorecard. Over time, the hope is that the scorecard can be used to evaluate similar systems side by side, and that a variety of scorecards can be produced for different types of systems (e.g., one-shot missile systems) as necessary. This scorecard is important in tracking the achievement of reliability requirements and rating the adequacy of the overall reliability program.

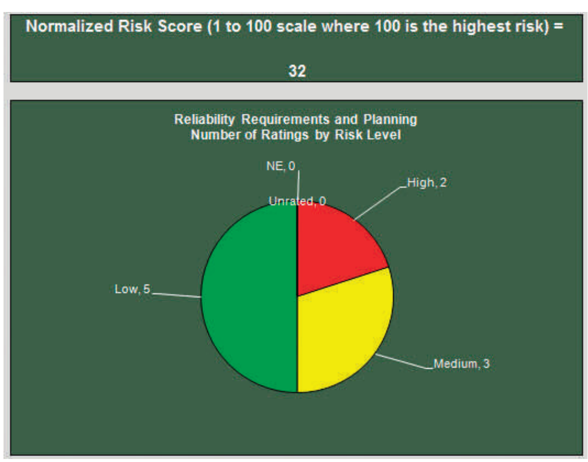


Figure 3. Example of a category with number of ratings by risk level.

Reliability growth planning and assessment models

Reliability growth is the improvement in a reliability parameter over a period of time due to changes in the product design or the manufacturing process. Reliability growth consists of two main areas: planning and assessment. Assessment can be further broken down into tracking and projection. A variety of models can be used to develop reliability growth curves for each of these sections.

Per Directive-Type Memorandum 11-003, entitled "Reliability Analysis, Planning, Tracking, and Reporting," dated March 21, 2011, "Reliability growth curves (RCG) shall reflect the reliability growth strategy and be employed to plan, illustrate and report reliability growth" (DOD 2011).

Reliability growth planning curves (RGPCs) assist management in determining program schedules, allocation of resources, and realism of the test program in achieving the required reliability. RGPCs monitor the reliability of a system in a developmental test program by evaluating the reliability growth that results from the incorporation of design and quality fixes into the system during the test program. Reliability growth projection curves estimate the reliability of a system at a current or future milestone based on planned or implemented fixes, assessed fix effectiveness, and the statistical estimate of problem-mode rates of occurrence.

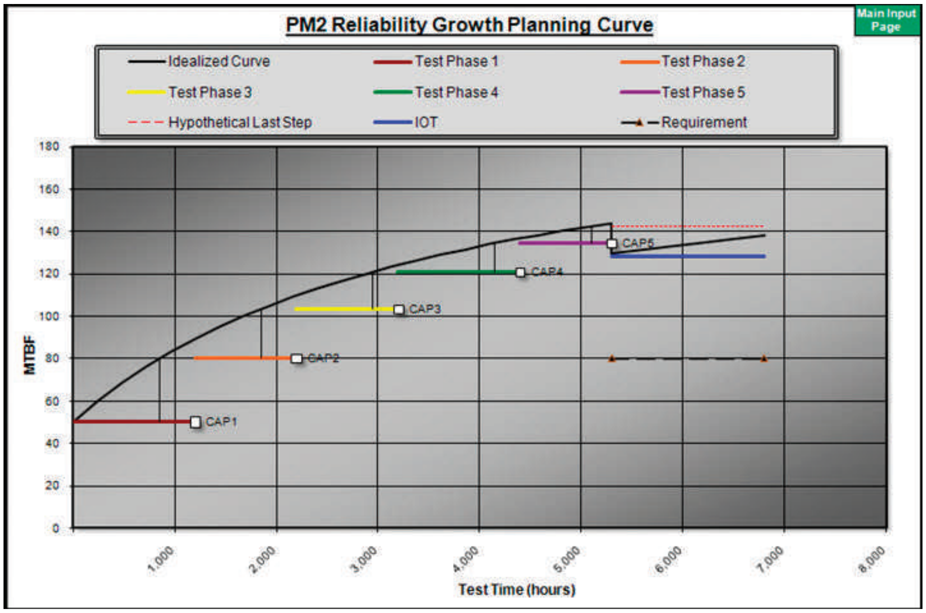


Figure 4. PM2 curve.

The purpose and utility of an RGPC is to portray the planned reliability achievement of a system as a function of test exposure, as well as other important programmatic resources. It also serves as a baseline against which demonstrated reliability values may be compared throughout the test program (for tracking purposes). Finally, an RGPC illustrates and quantifies the feasibility in a potential test program of achieving interim and final reliability goals. The planning curve is constructed early on in a program's life cycle, when little to no reliability data are available. The curve also provides guidelines for the reliability that can be expected at different stages of the program's life cycle. The RGPC gives a basis for evaluating future reliability growth progress.

The planning section of reliability growth begins with a reliability assessment to determine the initial reliability of the system. Other factors that need to be determined are the expected rate of surfacing failure modes, the amount of time needed to analyze and implement corrective actions, the management strategy (fraction of initial failure rate addressed by fixes), and the fix effectiveness factor (fraction by which rate of occurrence of failure modes is reduced). Growth occurs through identifying failure modes and developing corrective actions to improve failure mode, design, or quality. Corrective actions are changes to a system to mitigate the rate of occurrence of failures, not repairs. Growth planning should consider the initial and goal reliability targets, test phases, corrective-action periods, and reliability thresholds (interim goals to be achieved following corrective-action periods). Reliability growth planning should also include realistic management metrics, such as management strategy and fix effectiveness factors. The RGPC should be developed using reliability growth planning models, such as the Planning Model Based on Projection Methodology (PM2). *Figure 4* is an example of a RGPC developed using the PM2 model.

The next part of reliability growth is growth tracking. Growth tracking allows management to assess the system's reliability progress throughout testing. The reliability of the system is now based on actual data collected during the test phase, allowing for a demonstrated reliability metric instead of the initially planned metric. The purpose of reliability growth tracking is to determine if the reliability of the system is growing over time and to what degree growth is occurring, and to estimate the demonstrated reliability at the end of a test phase. This assists management in determining if the program is progressing as planned, better than planned, or not as well as planned. When a program is not progressing as planned, a new reliability strategy may need to be developed. Growth tracking provides a statistical estimate of the system reliability at a given time based on observed test data. Tracking models should be used to update a system's reliability based on actual test data.

The Reliability Growth Tracking Model for Continuous Data (AMSAA, Aberdeen Proving Ground, Maryland) is a tool for assessing the improvement (growth) in the reliability of a system during development for which usage is measured on a continuous scale. This model is useful for monitoring the reliability of a system in a developmental test program by evaluating the reliability growth that results from the incorporation of design and quality fixes into the system during the test program.

Reliability growth projection is the next step in assessing if a program is on track to meet its reliability requirements. Growth projection is an estimate of the reliability at a current or future milestone based on planned or implemented fixes, assessed fix effectiveness, and a statistical estimate of the problem-mode rates of occurrence. Projections are used to help managers decide how to allocate resources prior to entering the next test phase. Projections determine the system reliability at the end of a given (not yet executed or completed) test phase, before or after corrective actions have been implemented, or determine what reliability can be achieved by implementing delayed corrective actions. When a program is experiencing problems, projections can be used to investigate alternative test plans.

AMSAA has produced several reliability tools in Microsoft Excel. These tools assist in system planning and in tracking and projection of reliability growth. A brief description of the purposes of each follows.

Planning

1. System Level Planning Model (SPLAN): To construct idealized system RGPCs, identify test time required to improve system reliability, and aid in demonstrating the system reliability requirement with confidence.
2. Subsystem Level Planning Model (SSPLAN): To develop subsystem RGPCs that achieve an objective for system mean time between failure with a specified confidence, and determine the subsystem test times required to meet such an objective with specified confidence. Both growth and nongrowth systems may be included.
3. Planning Model Based on Projection Methodology (PM2) Continuous: To aid in constructing an RGPC over a developmental test program useful to program management, serve as a baseline against which reliability assessments can be compared, and highlight the need to management when reallocation of resources is necessary. PM2 does not have a growth-rate parameter, nor is there a comparable quantity.
4. Planning Model Based on Projection Methodology (PM2) Discrete: To construct detailed reliability growth programs and associated planning curves for discrete systems (systems whose test exposure is measured in terms of discrete trials, shots, or demands, e.g., guns, rockets, missile systems, or torpedoes).
5. Threshold Program (TP): To determine, at selected program milestones, whether the reliability of a system is failing to progress according to the idealized growth curve established prior to the start of the growth test. The program can be used to compare a reliability point estimate (based on actual failure data) against a theoretical threshold value.

Tracking

1. Reliability Growth Tracking Model–Continuous (RGTMC): To assess the improvement in the reliability, within a single test phase, of a system during development for which usage is measured on a continuous scale. The model may utilize both if failure times are known and if failure times are only known to an interval (grouped data).
2. Reliability Growth Tracking Model–Discrete (RGTMD): To track reliability of one-shot systems during development for which usage is measured on a discrete basis, such as trials or rounds.
3. Subsystem Level Tracking Model (SSTRACK): To assess system-level reliability from the use of component, or subsystem, test data.

Projection

1. AMSAA-Crow Projection Model (ACPM): To estimate the system reliability at the beginning of a follow-on test phase by taking into consideration the reliability improvement from delayed fixes.
2. Crow Extended Reliability Projection Model: To estimate the system reliability at the beginning of a follow-on test phase by taking into consideration the reliability improvement from fixes incorporated during testing and those delayed fixes incorporated at the conclusion of the test phase.
3. AMSAA Maturity Projection Model (AMPM): To provide estimates of the following, taking into consideration delayed and nondelayed fixes:
   a. the B-Mode initial failure intensity,
   b. the expected number of B-Modes surfaced,
   c. the percent surfaced of the B-Mode initial failure intensity,
   d. the rate of occurrence of new B-Modes, and
   e. the projected reliability.
4. AMSAA Maturity Projection Model based on Stein estimation (AMPM-Stein): Same as ACPM except that B-Modes are subdivided into BC (fixes incorporated) and BD (fixes delayed).

All of the tools mentioned here are Microsoft Excel spreadsheets that have replaced the older version of the AMSAA Visual Growth Suite CD. The models are available free of charge to U.S. Government personnel and their Defense Department contractors. Those interested should send an e-mail request to amsaa.reltools@us.army.mil.

Summary

The Reliability Scorecard and Reliability Growth Planning and Assessment Models are two methods which can be implemented to assist management in meeting system reliability requirements. System reliability is a key component of maximizing system availability and minimizing life-cycle costs. Overall, an improvement in reliability can represent a significant savings in the support costs for any major weapon system. ❑

Nancy Welliver is a member of the Reliability Analysis Team at the U.S. Army Materiel Systems Analysis Activity. She received her bachelor of science degree in mathematics from New Mexico State University and her master of science degree in business administration from Texas A&M University. E-mail: amsaa.reltools@us.army.mil

Marguerite Shepler is a member of the Electronics Physics of Failure Team at the U.S. Army Materiel Systems Analysis Activity. She earned her bachelor's degree in mechanical engineering at Grove City College in Pennsylvania and her master's degree in operations research at the Florida Institute of Technology. E-mail: amsaa.reltools@us.army.mil

References

Alion. 2008. Reliability maturity assessment procedure. Rome, NY: Alion System Reliability Center (SRC).

Bolton, C. M. Jr. 2007. Reliability of U.S. Army materiel systems. ASA (AT&L) memorandum. Washington, D.C.: OUSD (AT&L). Dec. 6.

CJCS. 2007. Joint capabilities and integration system. Chairman of the Joint Chiefs of Staff Instruction (CJCSI) 3170.01F. Washington, D.C.: CJCS. May 1.

DOD. 2011. Reliability analysis, planning, tracking, and reporting. Directive-Type Memorandum 11-003. Washington, D.C.: Department of Defense, Acquisition, Technology, and Logistics. March 21.

IEEE. 2008. *Draft Standard for Organizational Reliability Capability*. IEEE P1624/Draft 2. 2007. Archived unapproved draft. Available at http://ieeexplore.ieee.org/servlet/opac?punumber=4468763 Piscataway, NJ: IEEE Xplore digital library.

Shepler, M. H. 2010. Raytheon Company. 2007. Excel spreadsheet. McKinney, TX. "Raytheon Reliability Scorecard."

TechAmerica. 2008. *Reliability Program Standard for Systems Design, Development, and Manufacturing*. ANSI GEIA-STD-0009. Preview copy available at http://www.techstreet.com/itaagate.tmpl. Washington, D.C.: TechAmerica.

Tiku, S. 2005. Reliability capability evaluation for electronics manufacturers. PhD diss., University of Maryland.

Chapter News

Antelope Valley Chapter

The Antelope Valley Chapter recently held its leadership election and is pleased to announce the following new slate of officers: **Doyle Janzen**, President; **Nancy Bergren**, Vice President; **Lynnell Parker**, Secretary; and **Wanda Nishitomi**, Treasurer.

■ ■ ■

George Washington Chapter

September Luncheon. At its monthly meeting on September 22nd at the Army Navy Country Club in Arlington, Virginia, the George Washington Chapter heard **Mr. Edward R. Greer**, Deputy Assistant Secretary of Defense for Developmental Test and Evaluation (DASD DT&E), and Director of the Test Resource Management Center (TRMC); discuss "Current Events in Developmental Test and Evaluation." After showing organization charts of both the TRMC and his DASD DT&E office, Greer said staffing options will be considered after an AT&L-directed 90-day organizational study is completed. He described TRMC initiatives concerning Cyber T&E vision, Major Range and Test Facility Base charge policy, inter-agency partnerships, budget certification reports, and biennial infrastructure reviews. He said the TRMC has 112 active projects in 9 technical areas (such as cyber warfare and others). He pointed out that the Central Test and Evaluation Investment Program (CTEIP) budget grew from 1990 until now, and that the Defense T&E facilities budget is approximately $10 billion in fiscal year 2012, but both will likely decline in the years ahead.

Greer addresses GW Chapter luncheon.

In conclusion, Greer observed that readiness for OT is the number one problem due to schedule compression, concurrent development and testing, limited reasources, and other practices. He showed a chart of eleven systems from all four services for which his staff had completed an Assessment of Operational Test Readiness (AOTR) based on developmental T&E results. Seven were assessed as ready for IOT&E while four were not, but in all cases the decision was made to proceed into OT. Thus far, DT&E's AOTRs have a solid track record of predicting success or failure in OT.

In a robust Q&A period, Greer predicted there will be more DT&E oversight of space systems; that T&E professionals will contribute more to the system requirements process; and that the new Milestone A TEMP will address technology readiness for development. Finally, Greer noted that T&E methodology for rapid acquisition of urgent warfighter needs would be a good topic for ITEA to address. After the talk chapter president **Mike Wetzl** presented Mr. Greer a crystal ITEA paperweight as a memento of the occasion.

■ ■ ■

Roadrunner Chapter

In August, the Roadrunner Chapter in Albuquerque, NM held elections. The new leadership is: President, **Erik Thompson**; Vice President, **Mike Chalan**, and Treasurer, **Jeff Olinger**.

Congratulations to the new slate of officers!

■ ■ ■

Rocket City Chapter

On July 28, 2011, at QinetiQ, the Rocket City Chapter held a monthly meeting with guest speaker, Dave Cripps, Deputy Director, Aviation Engineering Directorate. On September 1, 2011, at SAIC, Major Cornelius L. Allen, Jr.-MAJ, AV-PEO MSL & Space Joint Air to Ground Missile, was our guest speaker on the topic of The Role of Simulation in Test and Evaluation. Our most recent event was held on September 21, 2011 at Amtec. It was our annual picnic in combination with the LabVIEW Users Group, tour of facilities, demos of LabVIEW, test equipment, robotics and many other various test & evaluation demos.

We have several activities planned for the remainder of the year including a joint Holiday party with Huntsville Association of Technical Societies at SCIQUEST – a Science Museum.

Tournament Chair Brian Bowden announces the winners at the San Diego Chapter's golf tournament.

Bruce Totty and Carl Caskey compete in putting contest at the San Diego Chapter's golf tournament.

In other news, our Rocket City Chapter was one of five ITEA Chapters awarded the Chapter of Excellence for 2011, and was selected as the Chapter of the Year 2011. **Leigh Christian**, Rocket City Chapter President, was award the ITEA Energizer Award. We are also proud of past Presidents of the Rocket City Chapter, **Mark Brown** and **Mike McFalls**. Mark Brown will serve as ITEA President of the Board and Mike McFalls will serve as Treasurer for ITEA.

■ ■ ■

Valley of the Sun Chapter

The Valley of the Sun Chapter held its Chapter elections and is excited to announce the new leadership, **Mr. Steve Woffinden**, President, and **Mr. Ernie Woodward**, Vice President. **Mr. Warren Smith** will remain as our Secretary and **Ms. Toni Farley** will remain as our Treasurer.

■ ■ ■

Greater San Diego Chapter

The Greater San Diego Chapter hosted its annual golf tournament at Marine Corps Air Station Miramar on Friday, September 16th. Sixty golfers took to the links in a "shotgun start" immediately following morning colors, and spent the next several hours playing a "best ball" format game. At the conclusion, the duffers enjoyed a buffet lunch at the Miramar club, along with prizes.

The event is the Chapter's primary fundraising activity for its scholarship fund, which supports students pursing college degrees in the disciplines leading to a career in test and evaluation. Organized by Chapter Board of Directors member **Bran Bowden** of Epsilons Systems, the tournament was held in conjunction with the San Diego Defense Industry Golf Club.

"This is an important event for us," commented Chapter President **Dan Phelan** of Lockheed Martin. "In addition to giving our chapter visibility within the local defense community, it allows us to support students who will be working in the T&E field in the future."

■ ■ ■

Greater San Diego Chapter President Dan Phelan and Western Regional Vice President Jack Sears get ready to tee off at the Chapter's golf tournament.

Golfers take to the links as the San Diego chapter's annual golf tournament kicks off.

Association News

2011 ITEA Annual Symposium Report

Mark D. J. Brown, Ph.D.
2011 ITEA Symposium Chair

The 2011 ITEA Annual Symposium was held at the Doubletree Hotel in Orlando, Florida 12–15 September 2011. By the feedback we have received, the Symposium was a huge success. This is due to the volunteers and ITEA staff that worked so hard this past year to bring this together. The Technical Program Chairs, **Dr. Dave Brown**, **Dr. Dave Bell**, and **Dr. Suzanne Beers**, did a fantastic job of pulling together presenters, keynotes, and panelists that hit the Symposium's

Chapter Locations

NORTHEAST REGION
Vacant, Vice President

CONNECTICUT & RHODE ISLAND
Narragansett Bay Chapter
Vacant

MASSACHUSETTS
New England Chapter
Michael E. Keller, Sr., President
Boston, MA

NEW JERSEY
South Jersey Chapter
John Frederick, President
Atlantic City, NJ

EAST REGION
Robert A. Vargo, Vice President

OHIO
Miami Valley Chapter
Stephen Tourangeau, President
Dayton, OH

MARYLAND
Francis Scott Key Chapter
https://www.sksi.net/fskitea/index.html
John B. Schab, President
Aberdeen, MD

Southern Maryland Chapter
Bill Darden, President
Patuxent River, MD

DC/NORTHERN VIRGINIA
George Washington Chapter
http://gw-itea.org
Michael Wetzl, President
Washington, DC

VIRGINIA
Tidewater Chapter
Jeanine McDonnell, President
Suffolk, VA

SOUTHEAST REGION
Mike McFalls, Vice President

ALABAMA
Rocket City Chapter
Leigh Christian, President
Huntsville, AL

FLORIDA
Central Florida Chapter
David P. Grow, President
Orlando, FL

Emerald Coast Chapter
http://itea-ecc.org
Robert Guidry, President
Eglin AFB, FL

GEORGIA
Atlanta Chapter
http://iteaAtlanta.org
Christopher Weeks, President
Smyrna, GA

SOUTH CAROLINA
Charleston Chapter
Philip Charles, President
Hanahan, SC

Volunteer Chapter
Nickolas Frederick, President
Arnold AFB, TN

SOUTHWEST REGION
Gregory D. Lamberth, Vice President

COLORADO
Rocky Mountain Chapter
http://www.itea-rmc.org
Christopher Mayette, President
Colorado Springs, CO

ARIZONA
Huachuca Chapter
Jonathan Woodruff, President
Sierra Vista, AZ

Valley of the Sun Chapter
Steve Woffinden, President
Scottsdale, AZ

NEVADA
Southern Nevada Chapter
Court Terry, President
Las Vegas, NV

NEW MEXICO
Roadrunner Chapter
Erik Thompson, President
Albuquerque, NM

White Sands Chapter
Douglas D. Messer, President
White Sands, NM

UTAH
Great Salt Lake Chapter
Jefferey D. Peterson, President
Dugway, UT

WEST REGION
Jack Sears, Vice President

CALIFORNIA
Antelope Valley Chapter
http://www.iteaavchapter.org
Doyle Janzen, President
Edward AFB, CA

Channel Islands Chapter
Christopher J. Weal, President
Point Mugu, CA

China Lake Chapter
Bettye R. Moody, President
China Lake, CA

Greater San Diego Chapter
Daniel Phalen, President
San Diego, CA

HAWAII
Mid-Pacific Chapter
Stu Burley, President
Kekaha, HI

WASHINGTON
Pacific Northwest Chapter
Vacant
Seattle, WA

INTERNATIONAL REGION
Vacant, Vice President

AUSTRALIA
Southern Cross Chapter
Peter G. Nikoloff, President
Edinburgh, South Australia

EUROPE
European Chapter
Steve Lyons, President
United Kingdom

ISRAEL
Israeli Chapter
Aaron Leshem, President
Haifa, Israel

theme "Fostering Partnerships in T&E and Acquisition." We made some significant formatting changes this year as well. In order to differentiate the Annual Symposium from the Annual Technology Review and the other ITEA Workshops, we chose not to break into technical tracks, but keep the Symposium focused on high level by having the symposium conducted in Plenary fashion. To provide the opportunity for individual papers, we held "mini-sessions" in the exhibit hall during the breaks as well as offering the opportunity for poster papers in the exhibit hall. This format seemed to work well, and, as we were told by some long-standing ITEA members, reminded them of how the Symposium was held in its earliest days.

As is Symposium tradition, Monday was dedicated to an outstanding round of golf hosted by the ITEA Central Florida Chapter and the National Center for Simulation. We also held six exceptional tutorials organized by **Dr. Steve "Flash" Gordon**. The tutorials were Reliability Growth Planning, Tracking, and Projection: Methodology and Examples (taught by **Mr. Sean Dobbs** and **Mr. Martin Wayne**); Combinatorial Testing with Design of Experiments (DOE) (**Dr. Mark Kiemele**); Testing 1-2-3: The Fundamentals of T&E (**Mr. Matthew Reynolds**); Mission-Based T&E Strategy: Case Study Tutorial (**Mr. Christopher Wilcox**); T&E of Cyber Systems (**Mr. Josh Davis**); and Using TENA and JMETC to Enable Integrated Testing and Training (**Mr. Gene Hudgins**).

Posting the colors at the 2011 ITEA Annual Symposium.

The Symposium was opened on Tuesday morning by the University of Central Florida's Army ROTC Color Guard with the Presentation of the Colors, followed by **Ms. Jeanine McDonnell** from Command Post Technologies singing the National Anthem. The ITEA President, **Ms. Stephanie Clewer**, then welcomed the Symposium Attendees and officially started the Symposium. After introductions, and some Ph.D. humor by the two Dr. Browns (Mark and David), we began our Keynote Speakers by hearing from the Warfighter.

COL. Michael E. Zarbo, USA, Project Manager, Instrumentation, Targets and Threat Simulators (PMITTS) from the Program Executive Office for Simulation, Training and Instrumentation (PEO STRI) shared with us how PEO STRI is supporting the warfighter as both a material developer and supporting the T&E infrastructure. **Mr. Dennis O'Donoghue**, Vice President, Test & Evaluation with the Boeing Company, shared his perspectives on how the acquisition and T&E communities need to work together in shifting their paradigm and his leadership perspectives from within the commercial sector.

Following Mr. O'Donoghue, the attendees were treated to a firsthand account of how the acquisition and T&E communities have come together to support the warfighter when **SGT Anthony Tipton**, U.S. Army SOFTAC, shared his experiences operating in theater and how his MRAP allowed him and his team to survive an encounter with an Improvised Explosive Device (IED). We also heard from **SGM Patrick Corcoran** on his experiences in theater. SGM Corcoran was joined onstage by COL Zarbo, SGT Tipton, Ms. Clewer and Dr. Mark Brown, as well as **Mr. Bill Criswell** and **Mr. Bill Curdts**, the Chair and Vice Chair (respectively) of the Home at Last program, which builds homes for disabled veterans in the West Orange County Florida area under the West Orange County Habitat for Humanity. The ITEA Symposium provided a $1,000 donation to Home at Last in honor of all our speakers, panelists, tutorial instructors and volunteers.

Following our Warfighter perspectives, we heard from two of our senior leaders about their perspectives on acquisition and T&E: **Mr. Edward Greer**, Deputy Assistant Secretary of Defense, Developmental T&E (DT&E) and Director of the Test Resource Management Center (TRMC), and **Vice Admiral W. Mark Skinner**, Principal Military Deputy, Assistant Secretary of the Navy (Research, Development and Acquisition). Following these distinguished speakers, the Symposium heard from its first panel on the Program Office Perspective on T&E, which was moderated by **the Honorable Sue Payton**, former Assistant Secretary of the Air Force for Acquisition. Joining her panel were **Mr. Stephen Kreider**, Director of the U.S. Army Program Executive Office for Integration; **Mr. C. Scott Anderson**, Senior Principal Economics/Business Analyst with the MITRE Corporation; **Dr. Dai Morris**, Head of Capability, Joint Training, Evaluation and Simulation with the UK Ministry of Defence; and **Maj. Gen. Jeffrey Riemer** (USAF, Retired), Vice President with Indyne.

Top Row, Left: Mark Brown, Ph.D. and SGT Anthony Tipton. Top, Right: Full house at the Doubletree Hotel in Orlando; Second Row Left: Speaker, The Honorable Dr. J. Michael Gilmore; Second Row Right: Dr. Mark Brown and Panel 1: Left to Right: Dr. Dai Morris, Maj. Gen. Jeffrey Riemer (USAF Retired), Moderator, The Honorable Sue Payton, Mr. C. Scott Anderson, and Mr. Stephen Kreider; Right: Speaker, (pointing) Mr. Marc Watson.

Bottom Row: Left: Moderator Mr. Edward Greer and Panel 5, Left to Right Seated: Ms. Leslie Taylor, Dr. Paul A. Alfieri, Mr. Stan Nowakowski, and Mr. Jay Weaver. Bottom Row, Right: Speaker Dr. Mark Brown, Moderator, Mr. David Duma, Speaker, SGT Anthony Tipton, and Speaker, Mr. Dennis O'Donoghue.

Our last plenary panel for Tuesday focused on Real Integrated Testing and was moderated by **Mr. David Duma**, Principal Deputy for the Director, Operational Test and Evaluation (DOT&E). Joining Mr. Duma were **Mr. Stephen Cricchi**, Director of the Integrated Systems Evaluation, Experimentation and Test Department (ISEET) in NAVAIR; **Captain Dr. Keith Joiner**, Director General T&E with the Australian Defence T&E Office; **Maj. Gen. David Eichhorn**, Commander of the Air Force Operational Test and Evaluation Center (AFOTEC); and **Mr. Dave Conrad**, Missile Defense Agency. Tuesday concluded with a grand reception hosted by Mantech, International.

The second day of the Symposium began with **Mr. Keith Sutton**, ITEA Chapter and Individual Membership Committee Chair, recognizing the ITEA Chapters of Excellence: Antelope Valley Chapter - Edwards AFB, California; Emerald Coast Chapter - Eglin AFB, Florida; George Washington Chapter - Washington D.C. Metro; Greater San Diego Chapter - San Diego, California; and Rocket City Chapter-Huntsville, Alabama. The Rocket City Chapter was then recognized as the Chapter of the Year.

Following the awards presentation, the attendees heard from our Featured Speaker, **Mr. Marc Watson**, P.E., President of Imagine Creative Solutions Technology LLC, as he shared his insights from the entertainment and theme park industry. We then heard from two more distinguished keynote speakers on their T&E perspectives: The **Honorable Dr. J. Michael Gilmore**, Director, DOT&E and Maj. Gen David Eichhorn, Commander of AFOTEC.

Outgoing President, Ms. Stephanie Clewer receives an appreciation plaque with a gavel on it.

Following these outstanding presentations, we honored our 2011 ITEA awards recipients at the awards luncheon hosted by **Mr. Al Sciarretta** and Ms. Stephanie Clewer. The awardees were **John B. Foulkes, Ph.D**. (Allen R. Matthews Award); **Mr. Mark E. Smith** (ITEA Board of Director's Award); **Mr. Randon R. Herrin** (ITEA President's Award); Dragon Spear Combined Test Force, Eglin AFB (ITEA Special Achievement Award), represented by **Captain Brad King**, **Lt Eric Lum** and **Ms. Jean Grieve**; **Mr. J. James McCue**, US Naval Test Pilot School (ITEA Richard G. Cross Award); **Dr. Roy Campbell**, **Mr. Tim Owen**, and **Mr. Rob Scott**, DoD High Performance Computing Modernization Program (ITEA Publications Award); Urgent Operational Needs Weapons Analysis Facility Study Team, Naval Undersea Warfare Center/Newport (ITEA Technical Achievement Award) accepted by team lead, **Mr. David Iacono**; **Captain Travis Bryce**, AWACS Joint Test Force (ITEA Junior Achiever Award); and **Ms. Leigh Christian**, Amtec Corporation**,** (Energizer Award).

Following the awards luncheon, we heard from the first Plenary Panel of the day, Emerging Technology Acquisition and T&E. This panel was moderated by **Dr. Charles Watt**, Founder and Chairman of the Board, Scientific Research Corporation. Joining Dr. Watt were **Mr. Derrick Hinton**, Principal Deputy Director at the TRMC; **Mr. Skip Tornquist**, U.S. Army PEO STRI Threat Systems Management Office (TSMO); **Mr. David Last**, Ph.D. Candidate at Auburn University; and **Dr. Jill Gemmill**, Executive Director for Cyberinfrastructure Technology Integration at Clemson University.

The final panel of the day on Wednesday focused on the Role of T&E in Rapid Acquisition. This panel was moderated by **Mr. Brian Simmons**, Technical Director of the U.S. Army Test and Evaluation Command (ATEC). His panel included **Mr. Igor Alonso Portillo**, Director of Strategy Test and Analysis Centre in Gipuzkoa, Spain; **Mr. Marty Meyer**, Director, Requirements & Architectures-Department of Homeland Security-US Customs and Border Protection,; **COL (USA RET). David Lockhart**, Vice President and Business Unit Program Director-SAIC; and **Mr. Douglas K. Wiltsie**, Acting Program Executive Officer for Intelligence, Electronic Warfare & Sensors.

Thursday of the symposium focused on the workforce. In the Plenary session, Mr. Edward Greer, moderated a panel on Improving the Current and Future T&E Workforce. His panel included **Ms. Leslie Taylor**, Director of Flight Test Engineering at Naval Air Systems Command; **Dr. Paul A. Alfieri**, Performance Learning Director for T&E at Defense Acquisition University; **Mr. Stan Nowakowski**, University of South Australia Defence and Systems Institute; and **Mr. Jay Weaver**, Director of Education at ITEA. In parallel with Mr. Greer's panel, ITEA hosted Academia Day, where over 80 area high school students were exposed to opportunities within the T&E career field. They took the time to meet with exhibitors to learn about T&E products and services followed by an excellent program *with the Keynote address* given by **Mr. Bruce Furino**, Director, College of Engineering and Computer Science, University of Central Florida and Chair of the Central Florida STEM Council. A panel discussion followed, moderated by **Mr. Mike Bell**, Academia Day Chair, U.S. Army Test and Evaluation Command with **Mr. George Rumford,** T&E/Science & Technology Program Manager, Test Resource Management Center; **Mr. Al Sweets,** Program Manager, Office of the Associate Dean, Morgan State University – School of Engineering; **Dr. Angela Leimkuhler Moran,** Professor, Mechanical Engineering, United States Naval Academy; and **Mr. Albert A. Sciarretta, PE,** President, CNS Technologies Inc. The students were engaged from the beginning as they did their homework on all of the speakers and submitted questions in advance.

The highly successful Academia Day program was orchestrated by **Mrs. Catherine Reich**, Scientific Research Corporation. As has become the tradition, the 2011 Annual Symposium closed with comments from Mr. David Duma, Principal Deputy, DOT&E. Mr. Duma shared his thoughts on "Solving Problems with Effective Partnerships."

As mentioned earlier, during the breaks a series of "mini technical sessions" were held. These sessions were well attended and the presentations were exceptional. The presentations included Agile T&E: A Workforce Evolution (**Anthony W. Stout**, Chief, Washington Operations, JITC); Managing the Message: Generating Meaningful Test Program Metrics (**Chris Bates** and **John Kovach**, Deloitte Consulting); Practical Ways of Integrating Test & Evaluation and Systems Engineering Standards and Practices to Benefit Acquisitions (**Maureen A. Molz** and **John Frederick**, FAA, William J. Hughes Technical Center); One Way for Collaboration between T&E and Program Management (**Alan Jenkins**, NAVAIR); Integrated Developmental and Operational Testing: Maximizing Efficiency for Systems Design and Development Programs (**Richard Watts**, Northrop Grumman); Value Stream Management of Developmental and Operational Testing (**Bryan Pflug**, Associate Technical Fellow); Real Integrated Testing with Distributed Live-Virtual-Constructive Simulations (**Chip Ferguson**, TRMC, **Dr. Frank Gray** and **Dr. Marion Williams**, IDA); Integrated Rapid Testing at DHS (Joshua Seckel, Blackstone Technology Group); Test Point Ordering Genetic Algorithm with Precedence Constraints (Daniel Carlson and **P. Travis Millet**, Air Force Flight Test Center, Edwards AFB); Joint Mission Environment Test Capability (JMETC) (Chip Ferguson, TRMC); Addressing Radio Frequency T&E Shortfalls in Rapid Acquisition (**Jeffrey Schleher**, SAIC and **J. Mark DelGrande**, TechFlow); Evaluation Frameworks Enable Integrated TE&C of Agile Developed Programs (**Dave Fisher**, JITC); Reducing Space Test Timelines to Support Rapid Acquisition (**Ari Fershtut** and **Clarence Stone** SMC/SDTT); Rapid Assessment to Support Technology Demonstrations (**Jay Zook**, SPAWAR, **Kent Taylor** and **Patrick Cannon**, AEgis Technologies Group, Inc.); Test and Evaluation Competency Model Development Project (**Valerie Outlaw Lee**, FAA); Model-Based Test Planning for Autonomous Systems (**Troy Jones**, Charles Stark Draper Laboratory); and Air Force Acquisition Modeling & Simulation Improvement (**David Young**, AFOTEC).

The Symposium would not happen without the support of the local chapter, volunteers, and the ITEA Staff. We especially want to thank the following for their dedication and support: Gloria Deane for her international contacts and coordination efforts; Carol Ann Dykes for her support of Academia Day; Jeff Galvan; Barry Greeno; Amy Guerin; Amanda Grow; Elizabeth Hood; Henry Jehan; Paul Meckel; Jerry Sirmans, and Ryan Swyers for their behind the scenes support; Roberto Guadalupe for his photography; Mike Morin and Ron Stiles for running a successful golf tournament; Juan Orozco and Lindsey Jones for their untiring efforts as the audio visual coordinators; and Dave Grow, the Central Florida Chapter President for his support throughout the year. We also thank the following for the loan of projectors to support the Symposium: Orbitouch; University of Central Florida Business Incubation Program; ArgonST; Greeno Business Consulting; and David and Amanda Grow. We also want to thank the ITEA Staff for their hard work and dedication: Mr. James Gaidry; Ms. Eileen Redd, Ms. Jean Shivar; Mr. Bill Dallas; Ms. Bonnie Schendell; and Mr. Jay Weaver.

A huge thank you to all our exhibitors and especially to our Sponsors: Platinum: Boeing, Scientific Research Corporation, TASC, and Wyle; Gold: Alion, Calculex, EWA Government Systems Inc., and SAIC; Silver: ASD, Booz|Allen|Hamilton, CSC, Georgia Tech Research Institute, JT3, and Symvionics; Bronze: AEgis Technologies, ArgonST, Rockwell Collins, and TRAX International; Emerald (Small Business): Advanced Test Equipment Rentals and Quadelta; and Ruby (Small Business): Avian Engineering LLC.

We look forward to seeing you at next year's Symposium, September 17 – 20, 2012, in Huntington Beach, California!

■ ■ ■

ITEA 2012 BOARD OF DIRECTORS AND OFFICERS ANNOUNCED

Newly elected President **Dr. Mark D. Brown** of SRC welcomes five new Board members. Five new Board members were introduced at the ITEA Board of Directors annual meeting, which was held during the 2011 Annual Symposium hosted by the Doubletree Hotel and Resort at Universal Studios in Orlando, Florida. New members elected to three year terms were **Mr. Peter Christensen** of MITRE, **Mr. Charlie Garcia** of TRAX International, **Mr. Gene Hudgins** of BAE Systems Technical Services, and **Mr. Tim Chalfant** of the 412th TENG/ENI at Edwards AFB, California. New members appointed to one year terms were **Mr. Peter Nikoloff** of NOVA Systems in South Australia and **Mr. Randy Surch** of The Boeing Company.

President Dr. Mark D. Brown.

During the meeting the Board of Directors also elected as its 2012 officers **Dr. Mark D. Brown** of SRC as President, **Mr. Charles "Chas" McKee** of Alion Science and Technology Corporation as Vice President, Mr. Gene Hudgins of BAE Systems Technical Services as Secretary, and **Mr. Michael T. McFalls** of Avion Solutions as Treasurer.

■ ■ ■

ITEA 2011 CHAPTER AWARDS ANNOUNCED

Rocket City Chapter (Huntsville, Alabama), **Ms. Leigh Christian**, President, is honored as the 2011 Chapter of the Year.

■ ■ ■

ITEA CHAPTER OF THE YEAR

Rocket City Chapter- Huntsville, Alabama

Ms. Leigh Christian, President

The Board of Directors recognizes the Rocket City Chapter of ITEA as the Chapter of the Year for 2011. Chapter officers President Leigh Christian, Vice President of Membership **Janice Isbell**, Vice President of Programs **Timothy Cotter**, Secretary **Dr. Stephen Jones**, and Treasurer **Gloria Power** lead a very active and vibrant Chapter that provides outstanding opportunities for T&E professionals to learn from others and share their own experiences. Through Chapter meetings and events the Rocket City Chapter facilitates the peer networking that test professionals require to expand their knowledge, skills, and abilities, and advance themselves and their profession. Along with the recognition, the Rocket City Chapter will receive $500 for their Scholarship Fund. The ITEA Scholarship Program is designed to foster interest and education in technical fields that may benefit the T&E profession. The program furthers the goals of ITEA by providing financial aid to qualified students. Congratulations to the Rocket City Chapter for their superior achievements!

The Chapter of Excellence Award is presented to Chapters that are vibrant and contributing effectively to ITEA's mission and goals to help ITEA in furthering the professional and technical interests of the T&E community. These Chapters met stringent guidelines intended to measure their level of effectiveness and are recognized at this time for their excellence. Congratulations to the leadership and membership of these Chapters. Recognized as 2011 Chapters of Excellence were:

Antelope Valley Chapter (Edwards AFB, California), **Mr. Michael Berard**, President

Emerald Coast Chapter (Eglin, Florida), **Mr. Robert Guidry**, President

George Washington Chapter (Washington DC area), **Mr. Michael Wetzl**, President

Greater San Diego Chapter (San Diego, California), **Mr. Daniel Phelan**, President

Rocket City Chapter (Huntsville, Alabama), **Ms. Leigh Christian**, President

■ ■ ■

INTERNATIONAL TEST AND EVALUATION ASSOCIATION'S 2011 AWARD RECIPIENTS RECOGNIZED

Special awards luncheon honors individuals, groups, and ITEA Chapters that have made significant contributions to advancing the T&E profession.

The International Test and Evaluation Association (ITEA) annually presents six (6) Professional T&E Awards to individuals and groups that have made significant contributions to advancing the (T&E) profession. In addition, ITEA also presents the Board of Directors' Award, the President's Award, Energizer Award, the Chapter of Excellence Awards, and the Chapter of the Year Award to recognize individuals and Chapters for their significant contributions to the Association.

At this year's awards luncheon, **Ms. Stephanie Clewer**, ITEA Board President, Mr. **Al Sciarretta**, ITEA Awards Committee Chair, and **Mr. Keith Sutton**, ITEA Chapter and Individual Membership Committee Chair, led a special ceremony recognizing the recipients of 2011 ITEA Award during last week's 2011 Annual Symposium, which was held at the Doubletree Hotel at Universal Studios, in Orlando, Florida.

Please join ITEA and others involved in the T&E industry in congratulating the following 2011 award recipients. Also, note that nominations for the 2012 Professional T&E Awards are due by June 1, 2012, for the period of performance from June 1, 2011 to May 31, 2012. Award recipients will be honored during ITEA's 2012 Annual Symposium, which will be held September 17 - 20, 2012, at the Hyatt Regency Hotel and Resort in Huntington Beach, California. Please visit the ITEA Web site at www.itea.org for additional information and nomination forms.

■ ■ ■

Allen R. Matthews Award - Dr. John B. Foulkes

The Allen R. Matthews Award was established in honor of ITEA's founder. The award is presented to an individual for a lasting, significant contribution to the field of T&E, such as the cumulative effect of a distinguished career. It is the highest award bestowed by ITEA. Over a 30 year distinguished professional career, Dr. John B. Foulkes has made outstanding and lasting contributions to the T&E community. He has been directly involved in all facets of T&E ranging from hands-on analytical support thru T&E policy and budget development, and has provided oversight of major defense systems and matters pertaining to the strategic planning and assessment of the nation's critical test range infrastructure. His enormous impact on the T&E community goes beyond government defense, and into academia and the technical association sector. He has, throughout his career, taught the next generation the nuances of T&E and has served as an avid and persistent leader within ITEA. Dr. Foulkes has provided exemplary service to the Department of Defense in support of the T&E mission. He provided executive leadership at the command level within the Office of the Secretary of Defense and previously for the Department of the Army. In support of the Army, his accomplishments include the consolidation of testing organizations that resulted in instituting better business practices together with preserving critical personnel. Dr. Foulkes' contributions while leading the TRMC are the highlight of a career focused on T&E, and his dedication reflects in the enormous body of work accomplished under his direction. His management of the Joint Investment Programs and Policy Division is setting standards for how T&E will be supported in the future. In addition, he was responsible for ensuring the DoD Test Infrastructure is available and ready to support acquisition programs, ensuring Department-wide test infrastructure planning and investment is in place to support future acquisition needs, and proactively federating DoD test infrastructure requirements across Government and Industry boundaries. The ITEA organization and its members salute Dr. Foulkes' career and contributions by presenting him with this award.

Dr. John B. Foulkes receives the Allen R. Matthews Award.

■ ■ ■

ITEA BOARD OF DIRECTOR'S AWARD - Mr. Mark E. Smith

The Board of Directors Award, established in 1997, is presented to an individual who has contributed to the growth, development, goals, and mission of the Association. The ITEA Board of Directors Award is presented to an individual who has contributed to the growth, development, goals, and mission of the Association. For over 16 years, Mark has served ITEA in a selfless manner in various capacities at both the local and international levels. He is a past Vice President of the Roadrunner Chapter, served as panel chair, guest speaker, and tutorial instructor at a variety of workshops, and as Technical Chair for the 2002 and 2005 Annual Symposiums. Mark was elected to the ITEA Board of Directors in 2006, serving two consecutive terms from 2006-2012. He distinguished himself by serving as the Chairman of the Chapter and Individual Membership Development Committee from 2006-2010.

Mr. Mark Smith receives the 2011 ITEA Board of Directors Award.

Mark revitalized the committee by spearheading a number of initiatives to achieve ITEA's strategic goals and to improve vertical and horizontal communications throughout the Association to include: establishing quarterly Chapter President teleconferences and chapter leadership and international participant meetings at annual symposiums; developing and implementing metrics and guidelines to gage chapter vibrancy and revamped the annual chapter recognition program accordingly; and played an instrumental role in developing a Chapter President's Handbook. On top of these numerous accomplishments, Mark also readily volunteered for a remarkable number of "other duties as assigned" while burdened with a heavy workload, both professionally and personally. He recently led a vital Revenue Action Team and played a very active role in the development of the current ITEA Strategic Plan. Since joining the Association, Mark has consistently embraced involvement at every level and, through his many efforts, contributed greatly to the stability and future of ITEA. To say that Mark is a model volunteer and ITEA member would be an understatement. His professionalism and integrity serves as a model that many of us strive for. For that, it is with great pleasure and sincere humbleness that the 2011 ITEA Board of Directors Award is presented to Mr. Mark Smith.

■ ■ ■

ITEA PRESIDENT'S AWARD - Mr. Randon "Randy" R. Herrin

The President's Award was established in 1996 to give the president of ITEA the prerogative of acknowledging individuals whom he or she deems worthy of recognition. The President's Award is reserved for an individual who selflessly steps up to assist the president in the execution of the organization's goals and missions. Improving the quality and relevance of our educational offerings is a key goal in the ITEA Strategic Plan, and ensuring the relevancy of ITEA's educational offerings was President Clewer's primary focus. The Association was extremely lucky to have Randy as the Education Chair to "carry the flag," so to speak, with incredible passion. In spite of significant demands on his time, both professionally and personally, he selflessly achieved more this year than was even hoped. During this time, Randy proactively revamped the Education Committee, facilitated a member survey, and began forging partnerships within the T&E and academic communities that I am certain will enable the Association to meet and exceed the Strategic Plan goals for education. ITEA's success relies heavily on the sheer hard work and determination of our volunteers and Randy definitely went above and beyond. It is with great pleasure that Randy is presented with the 2011 ITEA President's Award for laying the groundwork to ensure ITEA is postured to equip our T&E workforce with the skills and training required to face emerging challenges. Thank you, Randy, for a job well done!

■ ■ ■

SPECIAL ACHIEVEMENT AWARD - Dragon Spear Combined Test Force, Eglin Air Force Base

The 2011 Special Achievement Award The Dragon Spear Combined

Test Force, 780th Test Squadron, 46th Operations Group, 46th Test Wing, Air Armament Center, Eglin Air Force Base, Florida distinguished itself in outstanding service to the United States from 1 August 2010 to 31 May 2011. In order to meet the urgent warfighter need for an enhanced precision strike capability, Special Operations Command made the Dragon Spear Test Program its number one acquisition initiative. A lean core team was formed to expedite parallel development of multiple aircraft systems. The delivery of 12 MC-130W aircraft fitted with multiple precision strike munitions options, enhanced Intelligence, surveillance and reconnaissance capabilities, and a 30 millimeter gun system provided close air support for troops in combat. These capabilities were delivered ahead of schedule while exceeding performance expectations. The spin-off capabilities provided to Department of Defense units have bolstered daily combat capability in direct support of both Operations Enduring Freedom and New Dawn. A final surge of 34 missions in 61 days delivered a highly precise weapon system, allowing the high value target-kill chain to be cut from hours to minutes. The professionalism and devotion to duty of the test force members reflect great credit upon themselves and the United States Air Force. It is with great pleasure that the 2011 ITEA Special Achievement Award is presented to the Dragon Spear Combined Test Force Team.

The Dragon Spear Combined Test Force Team is awarded the 2011 ITEA Special Achievement Award.

■ ■ ■

RICHARD G. CROSS AWARD - Mr. J. James McCue, U.S. Naval Test Pilot School

The Richard G. Cross Award is presented to Mr. J. James McCue in recognition of outstanding achievements in the development and administration of T&E education. This year, Mr. McCue provided over 200 hours of classroom instruction. He has spent the year updating his performance model. The model and associated computer program used for performance data reduction was created by Mr. McCue and has become a widely accepted model throughout the U.S military test community. His constant adaptation to modern challenges and the incorporation of the product of those efforts into the syllabus have a lasting effect on the quality of instruction and the students. In addition to supporting the Test Pilot School's main curricula, Mr. McCue developed lectures and courses specifically targeted at certain issues to aid Navy and Marine Corps test program development. Mr. McCue's work and instruction on high altitude helicopter performance issues and compensation strategies for

The Richard G. Cross Award is presented to Mr. J. James McCue.

operations of heavy helicopters in both extreme hot and cold environments had an immediate impact in preparing Navy and Marine Corps aircrews deploying to Afghanistan for their demanding assignment. Mr. McCue's tireless devotion to development and administration of T&E education reflects great credit upon himself, the US Naval Test Pilot School, and the US Navy. It is with great pleasure that the 2011 ITEA Richard G. Cross Award is presented to Mr. J. James McCue.

■ ■ ■

PUBLICATIONS AWARD - Dr. Roy Campbell, Mr. Tim Owen, and Mr. Rob Scott, DoD High Performance Computing Modernization Program

The Publications Award is presented to Dr. Roy Campbell, Mr. Tim Owen and Mr. Rob Scott for their outstanding contribution to the written body of knowledge and understanding in T&E as the co-authors of, "Joint Sensor: Security Test and Evaluation Embedded in a Production Network Sensor Cloud" published in the December 2010 issue of *The ITEA Journal*. In this article, the authors explain the limitations with traditional

The Publications Award is presented to Dr. Roy Campbell, Mr. Tim Owen, and Mr. Rob Scott.

approaches to network security. The article reveals innovation and insight in merging research and development with production networks to realize more timely technology insertion while ensuring greater cyber security. It is with great pleasure that the 2011 ITEA Publications Award is presented to Dr. Roy Campbell in recognition of his significant contribution to T&E literature.

■ ■ ■

TECHNICAL ACHIEVEMENT AWARD - Urgent Operational Needs Weapons Analysis Facility Study Team, Naval Undersea Warfare Center Division – Newport

It is with great pleasure that the 2011 ITEA Technical Achievement Award is presented to the Urgent Operational Needs (UON) Weapons Analysis Facility (WAF) Study Team of the Naval Undersea Warfare Center Division, Newport for its diligently executed response to an Anti-Submarine Warfare UON statement from Commander, Fifth Fleet.

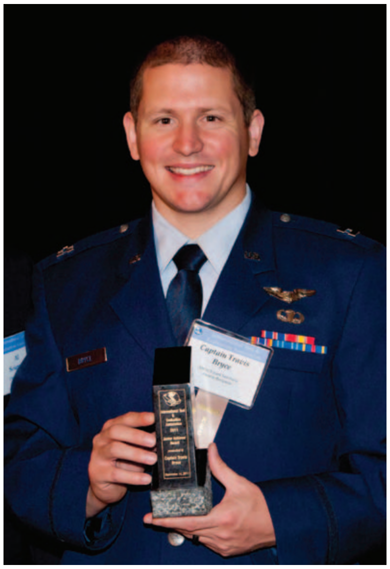
Team members conducted a modeling and simulation (M&S) study to assess heavyweight and lightweight torpedo performance in an emerging threat scenario of high interest to the Submarine Fleet utilizing Division Newport's high-fidelity, hardware-in-the-loop Weapons Analysis Facility (WAF). Evaluating both current Fleet and newer, developmental torpedo variants side-by-side under exactly the same conditions, the team gathered meaningful weapon effectiveness comparison data and identified areas where future torpedo changes could improve performance. This highly dedicated, innovative, and extremely responsive team pursued answers to an urgent Fleet need in a high-tempo, short-deadline situation. It is with great pleasure that the 2011 ITEA Special Achievement Award be presented to the UON WAF Study Team.

The Urgent Operational Needs Weapons Analysis Facility Study Team, *Naval Undersea Warfare Center Division – Newport is awarded the 2011 ITEA Technical Achievement Award.*

■ ■ ■

JUNIOR ACHIEVER AWARD - Captain Travis Bryce, AWACS Joint Test Force

Captain Bryce distinguished himself through outstanding service as the Deputy Chief for flight test at the E-3 Airborne Warning And Control System (AWACS) Joint Test Force, Aerial Air Surveillance Division, Electronic Systems Center, Seattle, Washington from 1 January – 31 December 2010. Captain Bryce skillfully and tirelessly supported seven AWACS test programs; accumulating 15 test sorties and 97.1 flight hours, and becoming a designated instructor Test Director. These programs include Diminishing Manufacturing Sources Replacement of Avionics for Global Operation and Navigation, Transitional Network Capability (TNC), Advanced Modern Electronic Warfare, Next Generation Identification Friend or Foe, AWACS Block 40/45 mission system upgrade, E-3 AWACS, Rivet Joint net centric demo, and Japan Air Self-Defense Force E-767 AWACS. Capt Bryce's superb skill allowed the delivery of TNC two months early and $300K under cost, and seamless transition of the Japanese E-767 AWACS to flight test. The test data that Capt Bryce collected provided top DoD officials to make acquisition decisions on programs totaling over $680M and allowed AWACS



The 2011 ITEA Junior Achiever Award is presented to Captain Travis Bryce.

operators to develop tactics to defeat advanced electronic attack. The distinctive accomplishments of Captain Bryce reflect great credit upon himself and the United States Air Force. It is with great pleasure that the 2011 ITEA Junior Achiever Award is presented to Captain Travis Bryce.

■ ■ ■

ENERGIZER AWARD - Ms. Leigh Christian, Amtec Corporation

The Energizer Award is presented to Ms. Leigh Christian for her tireless efforts both at a local and national level in both ITEA and in the T&E community. She has served as the ITEA President for the Huntsville Rocket City Chapter (RCC) since 2009, and immediately upon joining ITEA in 2005 she was instated as the Vice President of Programs. During her involvement with ITEA RCC, attendance at meetings increased threefold; she nominated two members of the Huntsville chapter to serve as officers on the National ITEA Board; she coordinates ITEA meetings jointly with other local groups' meetings, such as Rocket City LabVIEW User's Group, Women in Defense, and others; exposing many others in the T&E community to ITEA. A member of the ITEA Chapter of Excellence award and the only non-LabVIEW user to be awarded LabVIEW Champion, representing ITEA at the Huntsville Association of Technical Societies as secretary, and assisting with ITEA tutorials at Test Week, the impact Mrs. Christian has is clear on the Huntsville T&E community. Her dedication to each of her responsibilities does not waver, weaving together seamlessly a professional and positive, upbeat attitude. Under her leadership, the ITEA Rocket City Chapter has established a local education committee and is working closely with Redstone Test Center team and Defense Acquisition University to formulate a path for T&E courses in conjunction with ITEA. It is with great pleasure that the 2011 ITEA Energizer Award is presented to Ms. Leigh Christian.

■ ■ ■

ITEA 2011 BOARD OF DIRECTORS ELECTION RESULTS ANNOUNCED

Board welcomes newly elected Directors **Mr. Tim Chalfant**, **Mr. Pete Christensen**, **Mr. Charlie Garcia**, and **Mr. Gene Hudgins**.

The ITEA Elections Committee, chaired by **Mr. Gary Bridgewater**, announced on Monday the results of the recently completed 2011 Board of Directors election. Returning to the Board for a three-year term is Mr. Gene Hudgins, Engineering Manager with BAE Systems of Fort Walton Beach, Florida, and a member of the ITEA Emerald Coast Chapter. Joining the Board for three-year terms are: Mr. Tim Chalfant, Division Chief, Instrumentation Division for the 412th TENG / ENI at Edwards AFB, California, and a member of the ITEA Antelope Valley Chapter; Mr. Pete Christensen, Department Head of the Naval Sea Systems Department for The MITRE Corporation in McLean, Virginia, and a member of the ITEA George Washington Chapter; and, Mr. Charlie Garcia, Vice President of Range and Engineering Services for TRAX International in El Paso, Texas, and a member of the ITEA White Sands Chapter.

The ITEA Board of Directors, the membership, and the ITEA Executive Office staff wish to thank the following outgoing Directors for their dedication, service, and especially the time that they gave to the association while volunteering to serve on the Board: **Mr. Bert Johnston**, Vice President and General Manager of Systems Engineering for Wyle in Lexington Park, Maryland, and a member of the ITEA George Washington Chapter; **Mr. John Smith**, of Commander Operational Test and Evaluation Force (COMOPTEVFOR) in Norfolk, Virginia, and a member of the ITEA Tidewater Chapter; and, **Mr. John Wiley,** Manager of Technical Strategies & Integration for the Federal Aviation Administration (FAA) in Atlantic City, NewJersey, and a member of the ITEA Southern New Jersey Chapter.

■ ■ ■

MS. JEAN SHIVAR, ITEA EXECUTIVE ADMINISTRATOR, RETIRES FROM ITEA

After a professional career lasting over 50 years, including 21 wonderful years of service to the Association, Jean announced her retirement. Jean has been a dedicated and loyal staff member, and tremendous asset to the Association. She will be greatly missed not only in her very visible role running the registration desk at ITEA events, but even more so in her "behind the scenes" hard work in ensuring that ITEA's Executive Office runs efficiently, making sure our members are treated well, and that our Board of Directors, Committees, and Chapters receive outstanding support.

Ms. Jean Shivar, ITEA Executive Administrator.

Jean's official retirement day is Saturday, October 15th, and has graciously agreed to make herself available as reasonably needed to assist the Association as she transitions to retirement, and as we look to fill her very large shoes. Please join me and the rest of the ITEA Staff in congratulating Jean on reaching this milestone in her life and wishing her all the best!

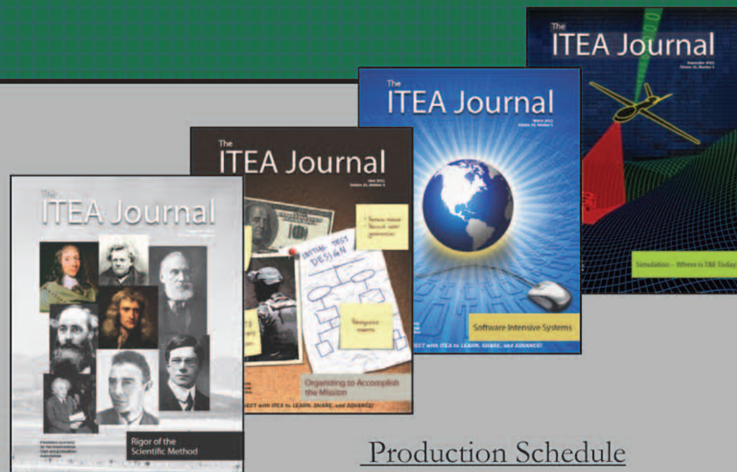
■ ■ ■

THE ITEA JOURNAL AD SALES

Ad Rates

THE ITEA JOURNAL AD SALES CONTACT

Bill Dallas
Ph: 703.631.6226
wdallas@itea.org
4400 Fair Lakes Court,
Suite 104
Fairfax, VA 22033
Fax: 703.631.6221



Production Schedule

Issue	Space Reservation	Ad Material
March	1/12	1/25
June	4/12	4/25
September	7/12	7/25
December	10/12	10/25

Journal Specifications

Trim Size: 8 1/2 x 11 inches. Journal trims 1/8 inch off top, bottom, and outside edge. Live matter should be a minimum of 1/2 inch inside the trimmed edges, and a minimum of 1/2 inch should be allowed for the bind.

1/2-page ad
(Lateral): Width: 6 inches; Height: 4 and 1/8 inches. (Vertical): Width: 3 inches; Height: 9 inches.

1/4-page ad
Width: 3 inches; Height: 4 and 1/8 inches

Graphics should be a minimum of 350 dots per inch. Add 1/4 inch for bleeds.

Formats: We accept TIFF or EPS format for both Macintosh and PC Platforms. We also accept files in the following native application formats: Adobe Acrobat (.pdf), Adobe Photoshop (.psd), Macromedia FreeHand (.fh), Canvas (.cvs), InDesign (.id), QuarkXPress (.qxd), Illustrator, (.ai) CorelPoint (.cdr), PowerPoint (.ppt), and Pagemaker (.pmd).

***

Disclaimer: All claims for errors in advertisements must be made in writing and received within ten days of publication and will be considered only for the first insertion of the advertisement.

AD Type	ITEA Partner Rates (ITEA Corporate Members, Sponsors, and Exhibitors)			Standard Rates		
	1x	2x	4x	1x	2x	4x
Cover 2	$2500	$2200	$2000	$3000	$2700	$2500
Cover 3	$2420	$2120	$1920	$2900	$2600	$2400
Cover 4	$2740	$2440	$2240	$3300	$3000	$2800
	4-Color					
Full page	$2100	$1800	$1600	$2500	$2200	$2000
1/2 page	$1700	$1400	$1200	$2000	$1700	$1500
1/4 page	$1300	$1000	$ 800	$1500	$1200	$1000
2-page spread	$3300	$3000	$2800	$4000	$3700	$3500
	Black/White					
Full page	$1620	$1320	$1120	$1900	$1600	$1400
1/2 page	$1300	$1000	$ 800	$1500	$1200	$1000
1/4 page	$ 900	$ 600	$ 400	$1000	$ 700	$ 500
2-page spread	$2580	$2280	$2080	$3100	$2800	$2600