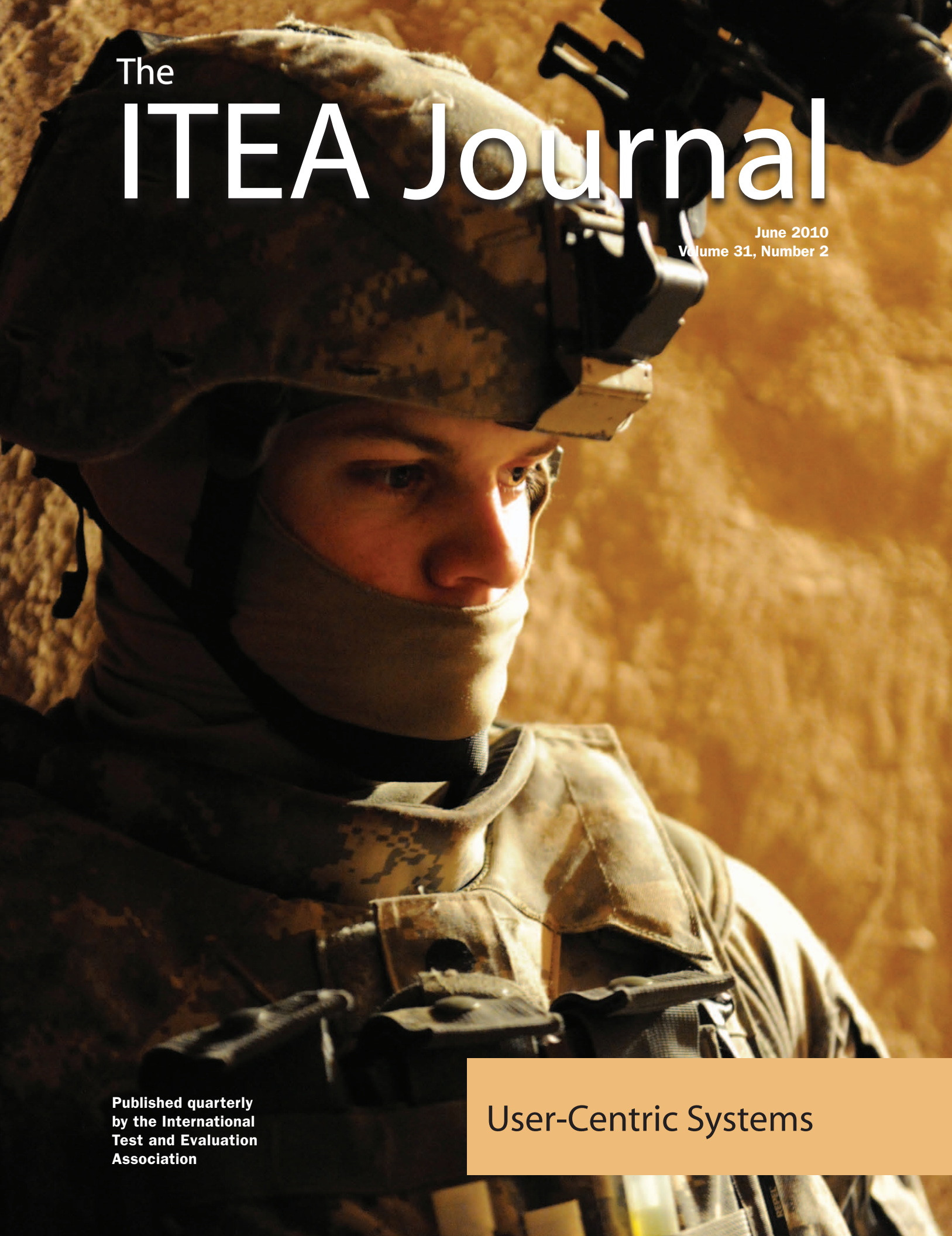


The ITEA Journal

June 2010
Volume 31, Number 2

Published quarterly by the International Test and Evaluation Association

User-Centric Systems

Testing technology in a global environment

Testing in a global environment is a job that our test and evaluation employees take on every day. We bring outstanding technical skills to the toughest problems facing our customers.

Smart people solving hard problems.

www.saic.com

Energy | Environment | National Security | Health | Critical Infrastructure

SAIC®
From Science to Solutions

NYSE:SAI


A new day in
ruggedized network recording
is about to begin...
NetCache™
CALCULEX
Info@calculex.com - 575-525-0131 - www.calculex.com
Organ Mountain Sunrise photo provided by Dan Long

CONTENTS

The ITEA Journal
June 2010
Volume 31, Number 2

TECHNICAL ARTICLES

DEPARTMENTS

ON THE COVER: From the 2008 National Research Council report *Human Behavior in Military Contexts*, (James J. Blascovich and Christine R. Hartel, Editors) "More than 15 years ago, the former commander of the Vietnamese forces against both the French and American armies, General Vo Nguyen Giap, said: 'In war there are the two factors—human beings and weapons. Ultimately, though, human beings are the decisive factor. Human beings! Human beings!' (*New York Times*, 1990, p. 36)."

On the cover a Soldier takes time to survey his environment. Technology is important to mission accomplishment yet the Soldier remains the key element to mission success. Objective test measures must include Soldier-system interface and cognitive performance to address human limitations within complex systems. (Photograph courtesy of Mass Communication Specialist 2nd Class Matthew D. Leistikow, U.S. Navy; provided by the U.S. Army Aberdeen Test Center, Technical Imaging Division).

■ 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.*
■ ITEA is a not-for-profit international association founded in 1980 to further the development and exchange of technical information in the field of test and evaluation.
■ *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 membership dues are $50 for individuals, $25 for full-time students, and $800 for corporations. 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.

MARK YOUR CALENDAR

September 13–16, 2010
Glendale, Arizona

The Future of T&E: Evaluating Operational Effectiveness in a Joint Mission Environment

This symposium will explore the ideas and challenges associated with evaluating the operational effectiveness of systems in the Joint Mission Environment. To do so we will take a look at the policy and requirements to test in a Joint Mission Environment. We will hear from Senior Representatives from Government and Industry to include program managers. The symposium will include tracks and break-out sessions that will delve into the details of lexicon, Joint Mission Environment various methods and process for evaluating the operational effectiveness of the capability.

CURRENT SPONSORS

Platinum: The Boeing Company • Science Applications International Corporation (SAIC) • Scientific Research Corporation (SRC) • Wyle

Gold: BAE Systems • Electronic Warfare Associates, Government Systems Inc. • ManTech International Corporation

Silver: Advanced Systems Development (ASD) • Calculex • CSC • EMRTC / New Mexico Tech

Bronze: Georgia Tech Research Institute (GTRI)

HOTEL ACCOMMODATIONS

Renaissance Glendale Hotel & Spa
9495 W. Coyotes Boulevard, Glendale, Arizona 85305 • www.renaissanceglendale.com

ITEA has a room block at the 2010 Government per diem rate for our attendees. We encourage you to make your reservations early by calling 623-937-3700 or visit the ITEA website for making your reservation online. Deadline August 13.

GOLF

On Monday, September 13, plan to attend the ITEA Golf Tournament at the Arnold Palmer designed course at Arrowhead Country Club, 19888 North 73rd Avenue in Glendale. All the details are available on the ITEA website. Hole sponsorships are welcomed. Visit www.arrowheadccaz.com for more details on the course.

www.itea.org

2010 ITEA ANNUAL SYMPOSIUM

President's Corner

ITEA Journal 2010; 31: 155–156

For this issue of the ITEA Journal, I would like to focus on the *Chapters*. The local ITEA chapters, through their service to their members, are the cornerstone, the heart and the soul of our Association. While I have been dealing with other pressing issues of late (finding the next Director, searching for new revenue streams in a down economy, and understanding the effect of the Government Joint Ethics Regulations interpretations), I have been uplifted by the work being accomplished by our chapters! We have 27 chapters, including three international chapters from Australia, the United Kingdom, and Israel. The chapters contribute to our T&E professionals by providing the means for information exchange through luncheons and meetings. They provide opportunities to publish papers, write articles, and make presentations. Through chapter participation, our members can get to know each other, grow professionally, provide mentorships, network, and be part of a recognized community. Our chapters contribute to the local communities to sustain the art and science of T&E. Several of our chapters have had continuing success through an active scholarship program to local colleges and to area high schools.

Most of our chapters are vibrant and important to their local community. The Live, Virtual, Constructive conference this past January in El Paso, Texas hosted by the White Sands Chapter (Doug Messer, President, with un-tiring assistance from Hank Newton), was an overwhelming success. The White Sands Missile Range leadership, including BG John Regan, actively participated and supported the event. Recently, I accepted an invitation from Chas McKee, the President of the George Washington Chapter, to attend their chapter luncheon at the Army/Navy Club. The luncheon was well attended by Government and industry participants from all over the Washington D.C. metro area who were there to listen to their speaker, Dr. Michael Gilmore, Director, Operational Test and Evaluation (DOT&E). The San Diego Chapter, led by Jack Sears, held its first event last November—again an overwhelming success, which was actively supported by SPAWAR, who provided tours of their C4ISR operations. The San Diego Chapter has now set its sights on partnering with the Fort Huachuca Chapter to host an event next spring. Sandy Webster's success as the President of the Mid-Pacific Chapter has also been contagious. She is experimenting with live streaming and video capture for her luncheon speakers, to maximize ITEA participation from members who cannot physically attend the events in Kauai. My own Atlanta Chapter is now led by four hard-charging young engineers who have already organized three standing-room-only technical luncheons. Our Chapter Presidents' quarterly telecons and breakfast meetings during the Annual Symposium have both been very promising as forums for sharing suggestions and lessons learned. The overarching goal of the telecons is to help all of our Chapters be successful at meeting the educational and networking needs of their membership.

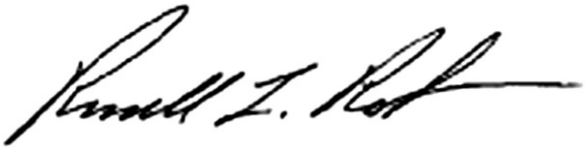
Russell L. (Rusty) Roberts

So what does it take to be a successful chapter? It's not easy but the rewards are worth the effort. It takes a lot of hard work from dedicated volunteers who wish to further the T&E profession. Meetings and events, which capture the interest of your constituents, will provide the foundation for increasing participation. Mark Smith, our Chapter and Individual Membership Development Committee chair, has instituted a program for defining and recognizing a successful chapter—the Chapter Awards for Excellence program. Each Chapter President has a copy of the 14-point checklist aimed at strengthening his/her local organization. Key points include hosting both local and national workshops and conferences, sustaining and increasing membership, and employing an active scholarship program. The ITEA Board of Directors is also looking at better ways to incentivize our chapters as well as getting them involved with national events. The bottom line is that the chapters are where the action is and we have to be mindful as to their importance to our organization.

Three new chapters are being formed this year; an unprecedented event in the recent annals of the Association. Rob Olson has stepped forward to lead the creation of a chapter in Phoenix—the Valley of the Sun

Chapter—which has stepped up to host this year's Annual Symposium, 13–16 September, in Glendale, AZ. Roy Maines is putting the finishing touches on forming the new Charleston Chapter, which will host this year's Technology Review, 19–20 July, in Charleston, SC; their inaugural event. The Southern Nevada Chapter is the third new chapter. President Steve Moraca and the new members will join the Antelope Valley and China Lake Chapters in support of the very popular Test Instrumentation Workshop, 10–13 May, in Las Vegas, NV. All three of these new chapters, ratified by the Board of Directors in March, will bring new enthusiasm and energy to ITEA overall. I am grateful for the dedication and hard work of the new chapters and for all the continued volunteer effort from all of chapters.

As I close, I wish to recognize the contributions of Dr. Michael Gorn. Dr. Gorn, a recognized author of a number of books and articles about the history of spaceflight and aeronautics, has been the author/contributor for the ITEA Journal's Historical Perspective column since 2006. While serving as ITEA's unofficial historian, Dr. Gorn has produced 16 exceptional historical articles for the journal. In bidding farewell to ITEA, he has written his last article for the journal. Thank you for your volunteer effort and outstanding support of our Association.



Issue at a Glance

ITEA Journal 2010; 31: 159

User-Centric Systems views the Warfighter—the system operator—in a new light from the traditional testing perspective. He/she is not just an operator in an operational test; the operator is "interfaced" to the system of systems and must be a central focus in T&E for system performance and mission effectiveness But the traditional focus on the operator is within the domain of operational testing with data collection via subject matter expert observations, questionnaires, and surveys. These are at best subjective and disruptive, and at a very different time, place, and environment than when such data could be collected. Collected in DT, such data could be more accurate, repeatable, objective, timely, and non-intrusive. And of course, system shortfalls associated with operator interface are best discovered early in the T&E process.

The issue opens with the *Guest Editorial* by The Honorable Dr. J. Michael Gilmore, Director of Operational Test and Evaluation. Dr. Gilmore explains rigor and objectivity in test and evaluation and presents his plan for achieving them. John Michelsen describes service virtualization as a unique and compelling approach for reducing cost and accelerating delivery of new functionality across the entire software lifecycle in his *Featured Capability* article. Dr. George Watson of the Office of Air Force History concludes the quarterly features with his colorful history of the A-10 Thunderbolt II, better known as the Warthog.

The contributed articles are led by William Beasley's summary of the Defense Science Board study on Fulfillment of Urgent Materiel Needs, part of which entails formalizing rapid acquisition as part of a dual path process. Dr. Mandy Natter et al. provide a detailed survey of measures for command and control and set the tone for the issue theme. Dr. Robert D. O'Donnell et al. describes a new approach for establishing cognitive performance by addressing cognitive demands of the system under test rather than trying to guess all possible human responses. Dr. Elizabeth Bowman and Dr. Randal Zimmerman discuss integrated analysis of a system of systems by resolution of quantifiable quality of service metrics, system performance characteristics, and human decision making. Diane Eberly presents factors that have the potential to influence how users will utilize a networked system of systems and how well the networked system of systems can support the users' needs. Dr. Cheryl Bolstad and Dr. Haydee Cuevas assess situation awareness as a diagnostic useful in the test and evaluation of new technologies and systems. Dr. Meredith Carroll et al. describe a framework which utilizes physiological sensors to detect indicators of implicit cognitive processing relevant to decision making and accomplish the granularity required to pinpoint and remediate process level issues.

Dr. Jennifer Ockerman et al. illustrate process instrumentation focused on collecting information about human activities for the training and test communities. Dr. Anthony Ries and Dr. Jean Vettel show results from using eye tracking and EEG for continuous measurements of operator performance and the potential for combined use to create a broad-based measure of changes in operator performance. Robert Cassella offers a new approach for developing 21st century leaders that includes methods for developing the flexibility of thought and adaptability needed to meet future operational demands, in a way that will permit behaviorally-based performance assessment tools to be developed and exploited.

In the final article Marcus Shellman, Jr. explains the necessity of addressing and mitigating spectrum supportability and electromagnetic environmental effects issues early during the acquisition process and verifying that these critical issues are achieved through the T&E process to decrease cost and increase mission effectiveness.

Guest Editorial

ITEA Journal 2010; 31: 161–164

Rigor and Objectivity in T&E

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

The Director of Operational Test and Evaluation (OT&E) has begun four T&E initiatives. The initiatives are far reaching and will influence the conduct of T&E across the defense establishment. The initiatives begin with support for our forces: "T&E is to help get the capability needed by our fighting forces to them as quickly as possible." The initiatives extend the boundaries for testers to engage in areas not typically the domain of T&E, such as "Review requirements as they are developed to assess whether they are unambiguous, testable, and relevant to accomplishing missions in combat." They promote new methods to improve T&E efficiency and effectiveness, through integrated testing. Finally, they sustain the important past T&E priorities of working with developmental testers to incorporate a reliability growth curve or software failure profile, reliability tests during development, and evaluation of reliability growth and reliability potential during development in the Test and Evaluation Master Plan (TEMP).

In my confirmation hearing before the U.S. Senate last June, I committed to provide objective evaluations of the effectiveness, suitability, and survivability of weapon systems based on realistic operational testing, with my goal to ensure that the men and women in uniform are provided weapons that they can be confident will work. One senator asked how I defined robust testing. I responded that robust testing is the testing needed to provide operators with high confidence that they understand what the system will do and will not do. And of course, acquisition decision makers must have high confidence to enable them to make proper decisions prior to fielding systems. We can achieve both by ensuring *Rigor and Objectivity in Testing and Evaluation (T&E)*.

J. Michael Gilmore, Ph.D.

The Undersecretary of Defense for Acquisition, Technology, and Logistics recently said, of a particular acquisition program, that the cost estimate on which he relies has a 50 percent chance of being wrong. The Weapon System Acquisition Reform Act now requires 80 percent confidence, or the justification for selecting a confidence level of less than 80 percent. In T&E, let us go about our work, so we consistently provide information on which all users can have high confidence.

I believe the T&E community not only has a legislated responsibility for planning and reporting on tests but also has a broader responsibility: to improve the chances that acquisition systems actually meet the needs of soldiers, sailors, airmen, and marines; in other words, "weapons that work when needed."

After being sworn in on September 23, 2009, and consulting with the senior management of Director, Operational Test & Evaluation (DOT&E), I identified four T&E initiatives and announced them in a memorandum dated November 24, 2009. The memorandum is on the DOT&E Web page available at http://www.dote.osd.mil/.

The four initiatives build upon our proven capability to provide rigorous, objective, and clear information in order to

1. Field New Capability Rapidly,
2. Engage Early to Improve Requirements,
3. Integrate Developmental, Live Fire, and Operational Testing, and
4. Substantially Improve Suitability Before Initial Operational Test and Evaluation (IOT&E).

These initiatives subsume priorities and metrics that have guided the DOT&E organization for the past several years and formed the basis of its annual reports to Congress.

Initiative 1: Field New Capability Rapidly

Secretary of Defense Robert Gates has made clear that his top priority is to get the capability needed by our fighting forces to them as quickly as possible. For T&E to take an initiative supporting the Secretary's top priority should not surprise anyone. The T&E community has been both helpful and unobtrusive in rapidly fielding new capability. We need consider only the Mine Resistant Ambush Protected (MRAP) vehicle as an example of such outstanding T&E support for rapid fielding.

To extend DOT&E's efforts to support rapid fielding as far as possible, the DOT&E staff will review all programs to identify candidates for early fielding or accelerated testing. After many years at war, it appears that there may not be too many systems left to accelerate. But, if testing has already confirmed that a system would be effective and suitable in current theaters of operation, those findings will be identified. If only a small amount of testing remains in order to determine effectiveness and suitability, we will identify opportunities for acceleration of that testing.

As they seek proactively to be involved in early fielding initiatives, the DOT&E staff will assess whether planned testing will be sufficient to identify fully the capabilities and limitations of the program being fielded. The DOT&E staff will also identify opportunities to streamline T&E procedures and processes to support early fielding initiatives. As appropriate, staff will communicate for action their assessments and those opportunities to program offices, the Operational Test Agencies (OTAs), and the DOT&E leadership.

The DOT&E staff should be flexible with respect to T&E procedures to see if they can be expedited. I expect DOT&E staff to be involved in early fielding initiatives, and help—not hinder. If you find we are doing otherwise, please let me know.

The feedback loop from fielding to program development, and later to testing, needs to be strengthened, particularly for rapid fielding initiatives. Thus, the DOT&E staff must work with the OTAs to identify and communicate critical problems with fielded equipment in need of immediate repair to program offices and DOT&E leadership.

The commitment to contribute to the rapid fielding of new capability to our forces will include the T&E of individual systems, efforts to find cases where the delivery of capability can be accelerated, determining that test–fix–test cycles are planned and occur, advocating use of quick reaction testing (QRT) to help develop Tactics, Techniques, and Procedures (TTPs), and the effective and rapid communication of test results to our commanders in the field.

We have seen success in QRT. A recent example is documented in a letter from the Commander, U.S. Central Command, complimenting QRT of procedures for the Joint Entry Control Point/Use of Force Handbook as a model for delivering a timely solution to our warfighters.

T&E must match the commitment of the warfighter. This is our top priority, and it must drive the pace of our daily work.

Initiative 2: Engage Early to Improve Requirements

I probably do not need to tell testers that requirements are seldom if ever perfect. The T&E community has the expertise to help the requirements process. We need to help, and to do so early on in the process—"left" of where testers traditionally engage in the acquisition process. We must do all we can to ensure that systems have realistic, relevant, and testable requirements and to identify programs that fail to meet that standard. To accomplish this initiative, DOT&E will

- Review requirements as they are developed to assess whether they are unambiguous, testable, relevant to accomplishing missions in combat, and operationally and technically realistic;
- Seek opportunities to be involved in reviews of requirements conducted before those requirements are submitted for consideration within the Office of the Secretary of Defense;
- For each project under oversight, review the Test and Evaluation Strategy (TES) and TEMP to ensure they include testing in realistic operational environments initiated during development and continuing through operational testing. This continuum of realistic testing will place increasing stress on subsystem components before final integration into a "full-up" system, thereby identifying problems when they can be fixed cost-effectively;
- Identify operational concerns to program offices at the earliest possible time in order to resolve them in a timely manner;
- Identify test-critical resource shortfalls;
- Ensure that testing in a joint environment is included in TESs and TEMPs where appropriate and feasible;
- Ensure that developers and the operational community share a clear, common understanding

of the planned Concept of Operations (CONOPS) or identify for action by DOT&E leadership inconsistencies in those views. If the CONOPS is not available, work to ensure a representative set of CONOPS is included in TESs and TEMPs; and
- Identify when programs lack a Reliability, Availability, Maintainability-Cost Report providing the rationale for meeting reliability requirements.

We must all strive for requirements that represent mission capability and not system technical specifications—leave those to the system engineers. DOT&E has four action officers who participate in the Joint Chiefs of Staff J8 Functional Capability Boards (part of the joint staff requirements process). I expect them to get the right people engaged to advise the requirements process.

Further, we know the Program Manager does not have control over all the other systems and interfaces needed for his program's mission success. To preclude a limited focus, which short changes the end user, and to step up to the commitment I made to the Senate in my confirmation, I have issued a clarifying policy to make clear that DOT&E will always evaluate a system in the mission context…even when it extends beyond the focus of the Program Manager. Just as important, testers must understand the intended operational context. Only with that information can we properly structure test environments.

Initiative 3: Integrate Developmental, Live Fire, and Operational Testing

Integrated testing is now Department of Defense policy. The legal requirement for a dedicated operational test is also clear. Nonetheless, separation of developmental and operational testing has caused difficulties in the development process that have been documented by a Defense Science Board and the National Academies of Science. Most notably is the lack of operational realism in early testing. Failure modes are hidden and performance limitations become evident only at the end of a program when fixing the problems is expensive, time-consuming, and, often, simply not possible. Therefore, DOT&E action officers will work with their counterparts in the office of Developmental Test and Evaluation (DT&E) and the program offices to incorporate integrated testing into TESs and TEMPs.

The DOT&E staff, as part of its determination of TES and TEMP adequacy, will assess whether the realism included in the early testing is adequate to identify the factors key to understanding whether a new system will actually provide improved military capability, as well as those factors that are not key. Identifying these key factors and screening out unimportant factors is essential to constructing the initial operational test.

In May 2009, the OTA Commanders and DOT&E endorsed Design of Experiments (DOE) as an important means to achieve integrated testing. DOE provides the scientific and rigorous method to plan and execute tests, and evaluate their results. DOE is currently used in some programs to construct individual test events. The DT&E and OT&E offices are working with the OTAs and Developmental Test Centers to apply DOE across the whole developmental and operational test cycle for a program.

Whenever possible, our evaluation of test adequacy must include a rigorous assessment of the confidence level of the test, the power of the test, and the breadth of the test—how well it spans the operational envelope of the system. DOE makes that assessment possible. DOE also will allow DOT&E to make rigorous and objective statements of the confidence we have in the results of the testing. Integrated testing and DOE may save resources, but as the OTA Commanders and DOT&E agreed, it may require more resources to achieve the rigor desired.

I have great expectations for integrated testing and suggest that we all need to work hard to apply it in the planning and execution of our test programs. I have started within DOT&E a task specifically designed to contribute to the whole community's understanding of how to use DOE to do this.

The Office of the Secretary of Defense is helping by collecting best practices for using DOE to enable integrated testing. The OTA Commander's Conference will continue to monitor our progress. In addition, the Defense Acquisition University is developing a continuous learning module to complement their resident courses.

Initiative 4: Substantially Improve Suitability Before IOT&E

Reliability is an essential consideration in our evaluation of system suitability; our forces deserve weapons that work whenever needed. We will encourage programs to have a reliability growth program, and we will track reliability growth during development. DOT&E will

- Assess at appropriate milestones whether programs meet the requirement to have a reliability growth program and identify for action by DOT&E leadership cases where this requirement is not met;

- Work with developmental testers to incorporate a reliability growth curve or software failure profile, reliability tests during development, and evaluation of reliability growth and reliability potential during development in the TEMP; and
- Work with developmental testers to ensure data from the test program are adequate to enable prediction with statistical rigor of reliability growth potential and expected IOT&E results. The rigor should be sufficient to calculate the probabilities of accepting a bad system and rejecting a good system, and those probabilities should be used to plan IOT&E.

For new or restructured programs, DOT&E will not approve TESs and TEMPs lacking a reliability growth curve or software failure profile in order to provide assurance that the system can demonstrate attainment of its reliability requirements.

Summary

We must engage in rigorous and objective testing. We must ensure operators have high confidence in their systems—what they will do and will not do—and arm acquisition decision makers with high confidence to make proper decisions prior to fielding systems. I challenge you to join me to field capability rapidly; engage early to improve requirements; integrate developmental, live fire, and operational testing; and, substantially improve suitability before IOT&E. ❑

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 within the Office of the Secretary of Defense, Program Analysis and Evaluation (OSD[PA&E]). Dr. Gilmore's service with Program Analysis and Evaluation covered 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

ITEA Journal 2010; 31: 165–168

You Can't Keep a Good ''Hog'' Down: The Curious Saga of the A-10 Aircraft

George M. Watson, Jr., Ph.D.
Air Force Historical Studies Office, Washington, D.C.

While the A-10 Close Air Support Aircraft (CAS) had been planned for by the U.S. Air Force (USAF) as early as the mid to late 1960s, its true origin can be traced to December 1970 when Secretary of the Air Force Robert C. Seamans, Jr., awarded contracts for two prototypes, both designated the A-X. It would eventually be called the Thunderbolt II, descendent of Republic's famous World War II P-47 Thunderbolt I. But Air Force pilots would also call the Thunderbolt II the "Warthog," "Hog," or "HAWG" because it looked ugly, and because of its tenacious close-in and low-level fighting capabilities, especially during inclement weather.

The "Hog"—built for use against Soviet tanks in Europe during the Cold War—was not welcomed by most Air Force pilots, who felt that it flew much too slowly and clumsily. Indeed, the 1970s marked an era of fighter pilot resurgence after a long period in which strategic airpower and the bomber pilots had overshadowed them. Airmen who had any hope of advancement in the 1970s, however, wanted to fly the F-15 and F-16, both high-speed, highly maneuverable fighter jets. So when the "Hog"—after much testing and flying competition—finally entered the Air Force inventory, it had to fight for acceptance for nearly 2 decades. Not until it won admiration for its spectacular performance in the first Persian Gulf War was it fully accepted. Like the "Ugly Duckling" of nursery fame, the A-10 endured, finally rising to prominence and outlasting most of the aircraft flown by the USAF. At this writing, the A-10s projected lifespan has been extended to 2030, putting it in the category of the USAF's two other great workhorses, the B-52 bomber and the C-130 transport *(Figure 1)*.

The A-10

The A-X/A-10 was the first Air Force development program governed by Design-To-Cost (DTC) principles, in which either a simple budgetary/DTC goal or an average unit flyway cost target would be established by the Secretary of Defense in collaboration with the Air Force. Also initiated was a Competitive Prototype Development phase that consisted of two contractors competing on a prototype development program, culminating 16 months later in full-scale development/production proposals (Watson 1979, 4). Following this paper contest, a flyoff between actual prototypes was planned from which a single contractor would be selected, an approach that placed heavy emphasis on extensive testing of the competing aircraft systems.

Under the rules of the competition, Headquarters Air Force would determine all performance goals. The Air Force's system program office (SPO), in concert with the contractors, was expected to meet these goals. The SPO was also responsible for additional technical projects such as the development of the 30mm gun eventually designated as the GAU-8 (or Gun, Aircraft, Unit-8) being developed by the Air Force's Armament Development and Test Center (ADTC). Studies had indicated that a 30mm gun could best cover the target spectrum of a close air support mission. CAS targets included personnel in foliage and foxholes; moving and fixed armored vehicles including armored trucks and medium and heavy tanks such as the Russian PT-76 light tank, T-54 medium tank, and JS-III heavy tank; the BTR-50P armored personnel carrier; and blast-sensitive, hard point targets, such as small watercraft, and bunkers (Watson 1979, 9).

Six companies submitted proposals for the A-X, and two were selected on December 18, 1970, by Secretary Seamans: the Fairchild Hiller Corporation, Republic Aviation Division from Farmingdale, New York; and the Northrop Corporation, Aircraft division, Hawthorne, California. To distinguish the two competitors, the Air Force designated the Northrop prototype, the A-9, and the Fairchild design, the A-10A. The flyoff was structured to test the two aircraft with difficult flying profiles in an attempt to identify and magnify differences between them. Basically, the trials showed no significant difference in weapons delivery accuracy between the A-9 and A-10A, although the A-10A held a slight edge in strafing on the 15-degree profile.

Figure 1. An A-10 Thunderbolt II close air support aircraft makes its way to the runway during RED FLAG-Alaska (RF-A) 10-1, October 9, 2009, at Eielson Air Force Base, Alaska. RF-A provides participants with 67,000 square miles of airspace, more than 30 threat simulators, one conventional bombing range, and two tactical bombing ranges containing more than 400 different types of targets. The A-10s are assigned to Osan Air Base, South Korea. (U.S. Air Force photo by Staff Sgt. Christopher Boitz. Reprinted with permission.)

On January 17, 1973, the Defense System Acquisition Review Council (DSARC) met to review the A-X program and to select the winning aircraft. Fairchild's A-10A was chosen. Secretary Seamans provided several reasons for the selection. The A-10A had better ground handling capacity, the underside of the wing had easy access, and the aircraft's larger wing size provided more storage for ordnance. Seamans also noted that the A-10A was closer to the production phase than the A-9, which allowed the test program to progress faster with only minor modifications to the prototype. He added that the A-10's simpler design increased the likelihood that its unit recurring flyaway cost would be closer to the desired $1.4 million target than the A-9, although there was no guarantee of achieving that figure. (Later on, the Air Force and the Office of the Secretary of Defense agreed to a flyaway cost of $1.5 million per aircraft.)

A-10/A-7D flyoff

Just as the A-10 appeared ready for production, the program underwent an unexpected diversion. Some members of Congress demanded a second comparative analysis, this time of the A-10 prototype versus the Navy A-7D aircraft. Several factors prompted Congressional interest in a comparison between the two: the A-10's high production cost, the A-10's seeming lack of versatility, and the rivalry between the two manufacturers—Fairchild (A-10) and the LTV Corporation of Dallas, Texas (A-7D). In addition, a September 12, 1973, meeting between Secretary of Defense James R. Schlesinger and Senator Howard W. Cannon (a Democrat representing Nevada) raised the issue of survivability of the A-10 in the European environment and stressed the need for a flyoff between the Fairchild and LTV competitors.

The Office of the Secretary of Defense (OSD), the Air Force, and Fairchild all opposed the flyoff, maintaining that nothing could be learned from it. But the Air Force finally acceded to Congressional pressure and on September 19, 1973, the Air Force Chief of Staff General George S. Brown informed Senator Cannon that the Air Force would comply. The tests were flown between April 15 and May 9, 1974, with aircraft operating from McConnell Air Force Base, Kansas, against ground targets and simulated defenses located at Fort Riley, Kansas. The test and evaluation was managed in two phases. Phase I involved a limited, qualitative evaluation at the respective test sites. Phase II consisted of both qualitative and quantitative evaluations, conducted at Fort Riley. The four fighter pilots who were chosen had no prior experience with either aircraft but had extensive close air support combat experience in either the F-100 or the F-4 (Watson 1979, 23–48).

The evaluation involved 16 missions in all, and each mission included two sorties—one A-10 and one A7D. There were two basic scenarios: the enemy attacks

friendly forces executing delaying actions, and the enemy breaks through an unorganized friendly force. The test results gathered by OSD/WSEG (the Weapons System Evaluation Group) and the Air Force generally demonstrated that the A-10 was the more effective aircraft. The Air Force analysis found that the A-10 achieved greater lethality than the A-7D against CAS targets because of its larger payload and the (projected) antitank capability of the GAU-8 gun. In addition, the Air Force maintained that in a European conflict scenario the A-10 would sustain lower initial losses than the A-7D because of its hardness and superior ability to avoid attrition from air-to-air attacks. Finally, the cost issue was addressed, with the conclusion that the A-10A was less costly than the A-7D, both in terms of acquisition cost and in life cycle cost.

So, on July 31, 1974, Deputy Secretary of Defense William P. Clements authorized the Air Force to proceed with initial production of the A-10 using $39 million for long-lead funding. He approved the procurement of 52 aircraft, subject to the provision that the purchase of an additional 28 aircraft be put on hold until further tests were completed. These tests included the GAU-8 gun and armor-piercing ammunition critical design reviews; completion of the engine qualification test; approach to stall, actual stall, and spin avoidance tests; and in-flight refueling.

Hails study

Despite the go-ahead decision, some residual drama remained for the A-10 program. Once the production phase began, some in the Air Force began to doubt whether Fairchild was capable of producing the promised number of A-10s in a timely fashion. A review team headed by Lt. Gen Robert E. Hails, then Vice Commander of the Tactical Air Command, conducted a study during the period of September 4–22, 1974. They found Fairchild's management and organization inadequate to the task. The firm's last full assembly line effort had been the F-105, which was terminated in November 1964. Since that time the company had mostly done subcontracting. Specifically, the review team found Fairchild's management structure to be too complex and organizationally cumbersome to allow for efficient commitment to the A-10 production program.

The Hails study had definite effects upon both Fairchild and the Air Force. In early 1975, the manufacturer made sweeping changes to management, organization, and equipment; and the Air Force added specific procedures to help institute requirements suggested by the Hails report. As a result, the Air Force stayed with Fairchild and, in helping the company refurbish its machinery, upgrade its facilities, and correct its managerial defects, forged a credible competitor for other major production programs. In the end, the Air Force and its industrial partner succeeded in producing the A-10 on the required schedule (Watson 1979, 23–48).

A-10 performance

The late 1970s is often called the era of the USAF "Hollow Force," when gross underfunding across a range of activities—from operations and maintenance to morale, welfare, and recreation–caused adversities that reduced the morale and effectiveness of the service. Budgetary retrenchment that limited flying hours caused concern among combat pilots who claimed they were not receiving the training and experience necessary to fly and fight. But the A-10 managed to survive this period and later rode the crest of the Reagan defense-spending wave. The Warthog was sent to various units both stateside and in Europe and also was assigned to reserve units without much fanfare.

It wasn't until the Persian Gulf War that the A-10—by then, a 20-year-old weapon system program—got its time to shine. As war appeared increasingly likely, with Iraq and its leader Saddam Hussein contesting the sovereignty of Kuwait, the Western powers began to build up their forces in Saudi Arabia. Among the equipment they assembled was the A-10, from both active and reserve units. Still, there were Air Force commanders such as Lt. Gen. Charles A. Horner from Central Command who didn't want the "Hog" in theater. It took the personal intervention of Army Chief of Staff Gen. Carl E. Vuono and Secretary of Defense Richard Cheney to overrule Horner and allow the A-10s to be used.

During the first stages of the war, the A-10 and the U.S. Navy's AV-8 and the F/A-18 were considered the primary weapon systems preventing an Iraqi invasion of Saudi Arabia. In this period, the A-10 itself flew 175 combat sorties, mainly concentrating on detecting and destroying Iraqi electronic warfare and ground control intercept sites. As the war progressed, however, and the A-10 penetrated farther into Iraq, it destroyed so many targets—trucks, tanks, infantry emplacements, ammunition dumps and storage facilities, as well as artillery and Scud missiles sites—that it caused many of its detractors to change their opinions. Perhaps the most famous convert was General Horner himself who, when assessing the various elements of the Allied air forces then attacking Iraqi targets, said of the A-10, "I take back all the bad things I have ever said about the

Figure 2. A U.S. Air Force A-10 Thunderbolt II aircraft from the 75th Expeditionary Fighter Squadron out of Bagram Air Base, Afghanistan, deploys flares during a combat patrol over Afghanistan December 11, 2008. (U.S. Air Force photo by Staff Sgt. Aaron Allmon. Reprinted with permission.)

A-10. I love them! They are saving our asses!" (Smallwood 1993, 96).

In all, the Warthogs flew over 8,100 missions, an average of 193 missions per day *(Figure 2)*. According to one source, they accounted for over half the confirmed damage inflicted on Iraqi Forces, despite flying only 30 percent of the total sorties. A good assessment of the Warthog's phenomenal success was stated by a "Hog" pilot in Riyadh: "Here we were, a big, slow, strictly low-tech CAS airplane that would have been heading for the boneyard if the war hadn't broken out—and now we're doing BAI [Battlefield Air Interdiction], armed Recee [reconnaissance], and SAM suppression" (Smallwood 1993, 169).

A place at the table

Although used sparingly in Bosnia during *Operation Deliberate Force*, the Warthog continued its superb performance during *Operation Iraqi Freedom* in active duty, reserve, and air guard roles. It has undergone many modifications during its long lifespan, particularly with the addition of global positioning equipment. Many A-10s are also being refurbished with larger wings in order to carry additional ordnance.

Having demonstrated its superiority in direct flyoffs, as well as in several wars, the Thunderbolt II has not only proven its detractors to be mistaken but has managed to endure and to establish itself as an essential ingredient of American airpower. ❑

GEORGE M. WATSON, JR., Ph.D., is a senior historian at the Air Force Historical Studies Office in Washington, D.C. He is the author of The Office of the Secretary of the Air Force 1947–1965 *and co-author of* With Courage: The Army Air Forces in World War II. *He served with the 101st Airborne Division in Vietnam and wrote of his experience in* Voices from the Rear: Vietnam 1969–1970. *In addition, Dr. Watson has contributed to a number of edited volumes and journals and has interviewed numerous key Air Force personnel both military and civilian. E-mail: George.Watson@pentagon.af.mil*

References

Smallwood, William L. 1993. *Warthog: Flying the A-10 in the Gulf War*. Washington, D.C.: Brassey's.

Watson, George M. 1979. *The A-10 Close Air Support Aircraft: from development to production 1970–1976*. Washington, D.C.: Office of History, Headquarters Air Force Systems Command, USAF.

ITEA Journal 2010; 31: 169–178

Review of Report to Congress by Defense Science Board Task Force Assessing the Fulfillment of Urgent Operational Needs

William E. Beasley
Office of the Secretary of Defense,
Joint Rapid Acquisition Cell,
Task Force Executive Secretary

The wars in Iraq and Afghanistan have highlighted the continuing need to improve the Department's ability to rapidly respond to urgent warfighter requirements against a highly adaptive enemy. The Department has created or modified numerous urgent needs processes to assist in countering enemy threats by expediting the fielding of warfighter urgent operational needs. The Duncan Hunter National Defense Authorization Act for Fiscal Year 2009 directed the Secretary of Defense to commission a study and report to Congress to assess the effectiveness of the processes used by the Department of Defense for the generation and fulfillment of urgent operational needs. A Defense Science Board Task Force was established in December 2008 to conduct the study. Its report to Congress in July 2009 included recommendations for the Department of Defense to formalize two paths—one for urgent and the other for normal acquisitions; establish a $3 billion per year fund for rapid acquisition and fielding; and establish a new Defense Agency, the Rapid Acquisition and Fielding Agency, to fulfill the urgent operational needs of the warfighter. Key members of the Department's new leadership have received this report and have incorporated it into their deliberations on how to more effectively support the current war in Afghanistan and Iraq and future operations.

Key words: Capabilities and limitations; Defense Science Board; JIEDDO; Joint Rapid Acquisition Cell; JRAC; JUON; MRAP; rapid acquisition; urgent operational needs.

"The essence of the problem at hand is the need to field militarily useful solutions faster."[1]

The wars in Iraq and Afghanistan have highlighted the continuing need to improve the Department's ability to rapidly respond to urgent warfighter requirements against a highly adaptive enemy. Toward this end, Congress directed the Secretary of Defense to commission a study and report to Congress. The study was to assess the effectiveness of the processes used by the Department of Defense for the generation and fulfillment of urgent operational needs.[2] The Department was instructed to perform the assessment through the use of a federally funded research and development center or the use of an independent commission.

The Defense Science Board, a Department of Defense (DOD) independent commission, was selected to perform this study with representation from the Defense Business Board. To conduct the study, the Defense Science Board (DSB) established a task force in December 2008, which was sponsored by the Under Secretary of Defense for Acquisition, Technology and Logistics, the Vice Chairman of the Joint Chiefs of Staff, the Under Secretary of Defense Comptroller, and the Director, Defense Research and Engineering. Its report was provided to Congress in July 2009.[3]

The Department has created or modified numerous urgent needs processes to assist in countering enemy threats by expediting the fielding warfighter requirements. New organizational structures have been created to fulfill warfighter urgent needs, including the Army's Rapid Equipping Force; Joint Improvised Explosive Device Defeat Organization; Intelligence,

Surveillance, Reconnaissance Task Force; Mine Resistant Ambush Protected Task Force; the Joint Rapid Acquisition Cell; and, more recently, the Counter Improvised Explosive Device Senior Integration Group. These organizations and associated processes have quickly delivered multi-tens of billions of dollars of capability to the warfighter. These processes have garnered their fair share of oversight reviews and assessments by the Department of Defense Inspector General, Service Audit Agencies, the General Accounting Office, and congressionally directed assessments. One recent study reviewed many of the DoD urgent needs processes, assessed the processes with the various process owners, and reviewed the application of the processes by multiple program mangers fulfilling the warfighters' urgent needs. Expedited test and evaluation was an underlying theme throughout this recent assessment.

I served as Executive Secretary to the DSB Task Force, and in the following pages I will provide additional background information, briefly describe the study's findings and recommendations, and provide some insight into the Department of Defense's approach to addressing the recommendations of the study's written report to Congress.

Background information

Members of the DSB Task Force included prior senior DoD and Service acquisition officials, a previous DoD Comptroller (also a member of the DoD Business Board), a recent member of the Office of the Secretary of Defense (OSD) from the prior administration charged with accelerating the transition of technology to acquisition programs, General Officers, and members from industry with expertise on rapidly developing new capabilities, and experts on special acquisition approaches and authorities. In the past, several of these Task Force members participated in studies[4] that examined related issues, many of which made similar recommendations for change focused on expediting the acquisition and fielding of needed equipment.

The Honorable Dr. Jacques Gansler, who chaired the DSB Task Force, wrote in a memorandum to the Chairman of the Defense Science Board:

> *"The accelerated pace of change in the tactics, techniques, and procedures used by adversaries of the United States has heightened the need for a rapid response to new threats. Fielding systems in response to urgent operational needs over the past half decade has revealed that DOD lacks the ability to rapidly field new capabilities for the warfighter in a systematic and effective way."*[5]

In this memorandum he summarized the critical actions needed to address the issue described. He further wrote that the implementation of the recommendations of the DSB Task Force was imperative to supplying the warfighter with the capabilities needed for success.

The DSB Task Force members were well aware that the regular or "deliberate" requirements, budgeting, and acquisition processes were not well suited to meeting urgent needs of the warfighter. Long-standing business practices and regulations are poorly suited to the dynamics of fulfilling urgent needs in a timeframe useful to the warfighter engaged in combat. The DOD is saddled with processes and oversight built up over decades, with managers leading them who are often trained to be risk averse. The "normal acquisition" system is a long chain of demanding, disciplined tasks that can take years and then only respond by exception to rapid changes. The Joint Capabilities Integration and Development System for requirements, the Planning, Programming, Budgeting, and Execution for funding, and the DOD 5000 series for acquisition are examples of the processes that underpin a regular acquisition. Planning is insufficiently anticipatory. Processes are too inward-looking and do not sufficiently leverage the commercial or global market—nor do they sufficiently leverage the public sector—by coordinating with other agencies for solutions to needed capabilities.

The Task Force observed that progress has been made, but DoD's ad hoc "rapid" processes still experience unnecessary and bureaucratic delays in needs generation, vetting, fulfillment, and fielding. These processes continue to lack serious institutional commitment and very little is being built into the Service or other DoD budgets for these programs. The Task Force wrote:

> *"It is hard to criticize the industrious nature of those in the Department who have made something happen when urgent needs have been presented; however, these approaches do not offer a long-term solution to circumstances that will not go away once current contingencies in Iraq and Afghanistan abate. As there is little doubt that the urgent needs from combatant commanders will continue,* the bottom line is that the ability to field critical war fighting needs requires a new approach—a standing acquisition and fielding capability that can fulfill these requirements in a timely way" [emphasis in original].[6]

Findings of the Task Force

The following sections provide summaries of the major findings of the DSB Task Force.

Multiple acquisition goals

All of DOD's needs cannot be met by the same acquisition processes. The Task Force found that the time critical nature or the urgent needs of the warfighter engaged in ongoing military operations, such as Operation Iraqi Freedom and Operation Enduring Freedom, require a requirements, acquisition, and fielding approach markedly different than those associated with the traditional Defense Acquisition System.

The Task Force noted that in the delivery of the 99% solution for traditional acquisitions the "JCIDS processes must be fully satisfied."[7] Development testing, verification and validation, interoperability and supportability assessments, safety evaluations, and operational test and evaluation all are pieces of the processes that must be satisfied before a capability is fielded.

The Task Force found that speed is one of the most important attributes of fulfilling an urgent need. The 75% solution is not only acceptable but welcomed if it provides a capability for operational use sooner. While the Task Force did not discuss at length the need for expeditious testing of capabilities, the urgency of operational evaluation was recognized. In their report the Task Force wrote:

> *"As opposed to traditional acquisition, in which better equipment is often perceived as the only solution, an urgent need may be met with new tactics, new capabilities, new materiel—based on proven technologies—or a combination of these. Also in contrast to traditional acquisition, test and evaluation should not be a pass or fail test, but rather should be used to determine capabilities and limitations—an approach the Army has successfully used to decide whether potential solutions to urgent requirements are good enough to be deployed."*[8]

The Task Force heard presentations by Service operational test agency leaders, multiple program managers delivering urgent needs to the warfighter, and prior and current leadership of the Joint IED Defeat Organization. There was wide agreement among Task Force members that the use of Capabilities and Limitations reports to inform the warfighters' decision to accept capability was needed to expeditiously fulfill the warfighters' most urgent needs.

It was noted by the Task Force that the level of testing is generally tailored to the capability to be provided to the warfighter. The mine resistant ambush protected (MRAP) vehicle received extensive safety, operational, and live fire testing and continued testing after deployment. Counter Radio Controlled Improvised Explosive Device Electronic Warfare systems received extensive operational and interoperability testing prior to deployment and continued surveillance after deployment.

The Task Force found that risks, unless appropriately mitigated, in a traditional acquisition program "[are] perceived as being a show stopper."[9] On the contrary, successful acquisition of an urgently needed capability often involved accepting risk that was "transparent, acknowledged, understood, and weighed against the attendant risk of proceeding along a more deliberate route."[10] Capabilities and Limitations reports provide an appropriate vehicle to characterize risks and enable the warfighter to actively participate in the fielding decisions of the urgent need.

"Rapid" is counter to the traditional acquisition culture

"Rapid" is countercultural and will be undersupported in traditional organizations. The DSB Task Force observed that "the current defense acquisition workforce is rewarded for following complex procedures with accuracy and precision and is punished for bypassing them."[11] The architects of the various successful urgent needs processes developed workarounds and established parallel acquisition paths to the traditional processes. The Task Force found examples in DoD, and in industry, where parallel processes proved effective in achieving the desired development or business outcomes. As examples, the establishment of the Defense Advanced Research Projects Agency to address disruptive technologies was a separate parallel process to Service acquisitions that are focused on more traditional, incremental developments. In industry IBM established its personal computer division separately from its traditional mainframe division.

The Task Force stated that sustaining a rapid acquisition capability in the Department of Defense requires the active support of the requirements, resourcing, and the testing community and the establishment of a parallel acquisition option. They opined that a component of the traditional process will not work and a separate organization is required.

Use of proven technology is essential to rapid response

Any rapid response must be based on proven technology and robust manufacturing processes. The Task Force believes that to achieve rapid deployment of an urgent capability, mature technical solutions are required. Rapid delivery of "Block 1" capabilities with spiral development of additional capabilities is necessary. Where technical maturity

precluded rapid deployment of a capability, the Task Force recommends that the capability be developed, as a high priority, by the defense science and technology community.[12] The Joint IED Defeat Organization exemplifies an organization that both delivers proven technology and, through extensive science and technology efforts, reaches out to industry and academia to develop solutions to quickly evolving enemy threats.[13]

The DSB Task Force agreed that risks must be understood and the use of capabilities and limitations is an important element to the fulfillment of urgent needs. The report states:

> *"While there may be instances in which early fielding of prototypes with contractor logistics support is appropriate, the risks must be well understood and parallel efforts should be in place to mature the technology and to ensure that training and logistics are adequate for the system life cycle. An assessment of capabilities and limitations should be an integral part of the warfighter's acceptance of the system for operational use."*[14]

Ad hoc organizations

Current approaches to implement rapid responses to urgent needs are not sustainable. The DSB Task Force found that many ad hoc processes were established to address urgent needs and that all, with senior level support, used workarounds to "sidestep traditional acquisition and fielding processes."[15] The Task Force found these processes disjointed, with little institutional memory or tracking of lessons learned. Some processes established for specific purposes had no sunset provision, and others appeared to be turning into bureaucratic organizations. The push for fulfilling wartime needs enabled the ad hoc processes to create workarounds to rapidly fielding capability. The Task Force observed that as the wartime push eases, the ability to be rapid will likely be reduced.

Urgent needs will endure beyond today's conflicts, which led the Task Force to recommend the creation of a sustainable organizational capability for rapid acquisition and fielding. Rapid acquisition and fielding capability must "build on the advantages of current ad hoc processes that have found relief from the rigors of the formal, deliberate acquisition bureaucracy."[16]

Lack of integrated triage

An integrated triage process is needed. There is a wide continuum of urgent needs ranging from ill-defined capability gaps to requests for additional supplies of standard equipment. The Task Force recognized that even in a wartime situation, resources are limited, and thus the Task Force found that the triage of urgent needs is an important step. A higher level view of all needs and a wider view of potential solutions are required. The higher level view envisioned by the Task Force enables the allocation of resources to fulfill urgent needs. Game changing capabilities can be fielded through this triage process.[17]

Institutional barriers

Institutional barriers—people, funding, and processes—are powerful inhibitors to successful rapid acquisition and fielding of new capabilities. The primary issue raised by every witness before the Task Force was the availability of dedicated and flexible funds. The competition for funds to address even the most critical urgent needs are affected by institutional barriers established in Service Financial Management and OSD Comptroller processes, rules, and regulations; Office of Management and Budget overall wartime funding priorities; and the Congressional appropriations processes.

Task Force members believe that people must work in integrated teams to support the warfighters' urgent needs and that success is achievable only if these integrated teams have "the best and brightest innovative thinkers who are solution-oriented, creative, and uninhibited by bureaucracy."[18]

Needed best practices

The Task Force reviewed solutions to the shortfalls identified in their findings and assessed a number of best practices of the various urgent needs processes. Each of the Best Practices listed in Table 1 reflect attributes of solutions that deserve further evaluation. Note that the Army processes for test and evaluation of urgent needs, the Army Test and Evaluation Command's Capabilities and Limitations process, is viewed as being "good."

Recommendations of the Task Force

The major recommendations of the DSB Task Force are summarized in the following paragraphs.

The Secretary of Defense should formalize a dual acquisition path[19]

The Task Force viewed the deliberate and rapid acquisition processes as incompatible processes with different acquisition goals. They recommend a dual acquisition path, as depicted in *Figure 1*.

Deliberate acquisition process. The goal is a 99% solution, which often translates to delivery in 3 to 11 years or more. It is optimized for delivery of

Table 1. Best practices needed (DSB TF Report, 30).

Best practices needed	Where it's good today
For involving the warfighter from beginning to end of process	Joint Capability Technology Demonstration, Army Asymmetric Warfare Group (AWG), U.S. Special Operations Command (USSOCOM)
For obtaining agile, flexible funding	Joint IED Defeat Organization, Mine Resistant Ambush Protected Vehicle Program (MRAP)
To coordinate status and resolution for each need statement	USSOCOM
For coordinating technology development	Director of Defense Research and Engineering (DDR&E), U.S. Air Force Big Safari
To evaluate effectiveness of the implemented solution	USSOCOM, AWG
For test and evaluation	Army
To determine whether to end or to transition each implementation	–
For a knowledgeable workforce for all rapid acquisitions	U.S. Air Force Big Safari
For business approaches that use existing flexibilities	DDR&E, Defense Advanced Research Project Agency, U.S. Air Force Big Safari, MRAP
For institutionalizing the rapid response process	Navy/Marine Corps
For collaborative innovation	Private sector

complex systems and is scalable to very large military solutions. It uses detailed, extensive, and large-scale oversight and synchronization to ensure success. It includes resources for sustainment and is well adapted to individual Service cultures. Owing to the long time frame, this process often begins by pushing the state of the art of the underlying technologies.

Rapid acquisition process. This process is satisfied with a 75% solution or sometimes less, with the major focus on delivery within 24 months. To be responsive to combatant command timelines, the Task Force recommended that execution be decentralized. Participation by small and nontraditional businesses is sought. Risk is mitigated through the use of proven technology that is rapidly transitioned via competitive prototyping. More advanced or extensive capabilities are provided in subsequent builds through spiral development. Resources for sustainment and training are integrated and delivered in parallel with initial operating capability.

The Task Force also recommended a standard DoD-wide definition be established for an urgent need to enable effective triage of the acquisition path (deliberate or rapid). "The definition should state that an urgent need is one that *if left unfulfilled, will seriously endanger personnel and/or pose a major threat to ongoing or imminent operations*" [emphasis in original].[20]

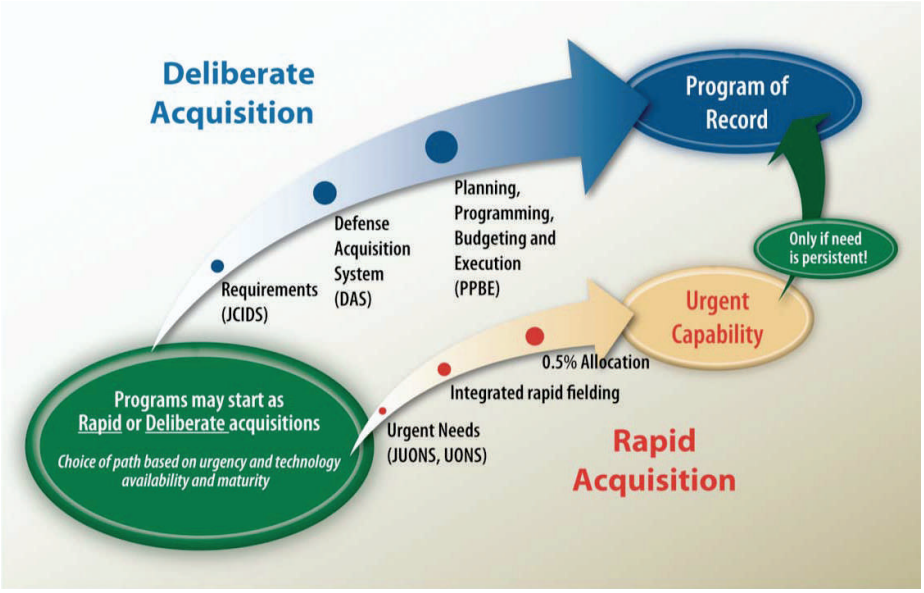


Figure 1. Dual acquisition path proposed (DSB TF Report, 32).

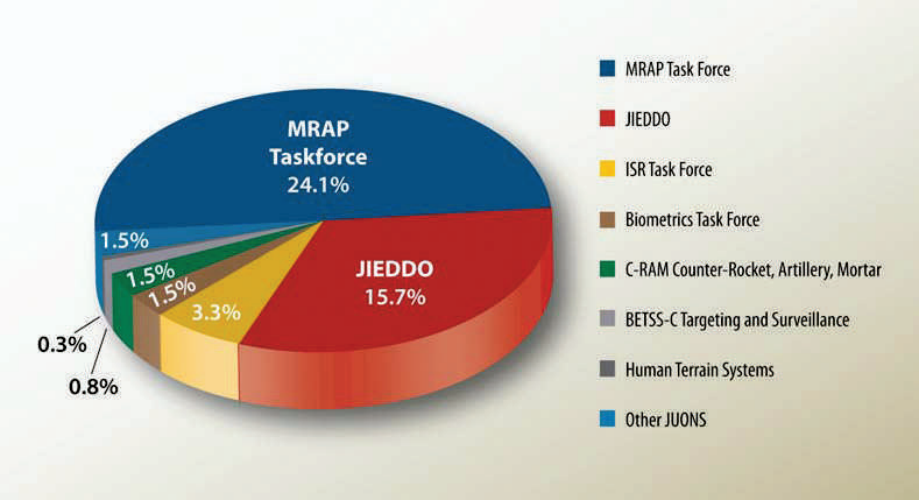


Figure 2. Fifty billion dollars allocated 2005–May 2009 to joint urgent operational needs (DSB TF Report, 11).

Executive and legislative branches must establish a fund for rapid acquisition and fielding[21]

As depicted in *Figure 2*, the Department allocated, in less than 4 years during ongoing wars in Afghanistan and Iraq, approximately $50 billion to fulfill Joint Urgent Operational Needs. The Services, other Defense components (U.S. Special Operations Command, Defense agencies, etc.), and the intelligence community expended significant additional funds to fulfill their specific urgent operational needs. These wartime expenditures for urgent needs informed the DSB Task Force in arriving at a recommended funding level of 0.5% of the DoD budget, roughly $3 billion dollars per year, to support rapid acquisition and fielding.

The Task Force anticipated similar funding needs for the foreseeable future; however, they stressed that the funding should not be contingent upon an ongoing war. In periods without an ongoing war, the funds would support rapid acquisition of capability that is needed more rapidly than the regular requirements, acquisition, and budget processes would allow.

The Secretary of Defense should establish a new agency: the Rapid Acquisition and Fielding Agency[22]

The Task Force recommended that the Secretary of Defense establish a new agency: the Rapid Acquisition and Fielding Agency (RAFA) "focused on speed, utilizing existing technologies and acquisition flexibilities to get a 75 percent solution—initially adequate to address the urgent needs of the warfighter."[23] It was also recommended that each Service establish a rapid acquisition organization that would work in close collaboration with the RAFA. While various organi-

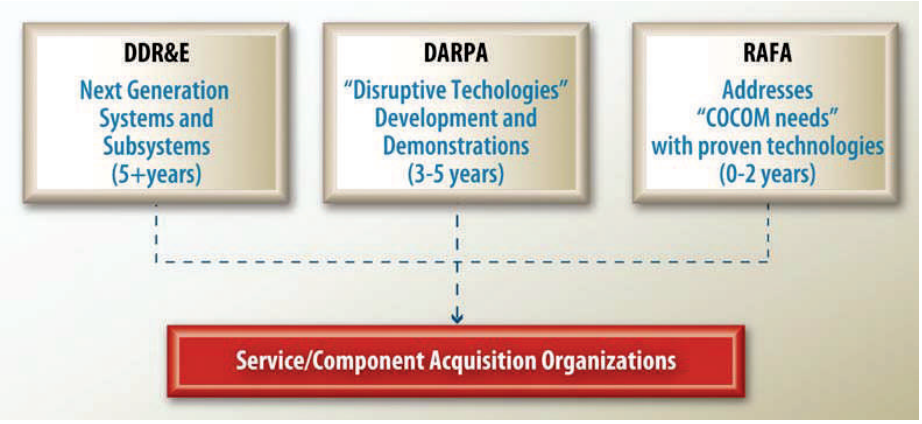


Figure 3. Notional comparisons of organization responsibilities and timelines (DSB TF Report, 34).

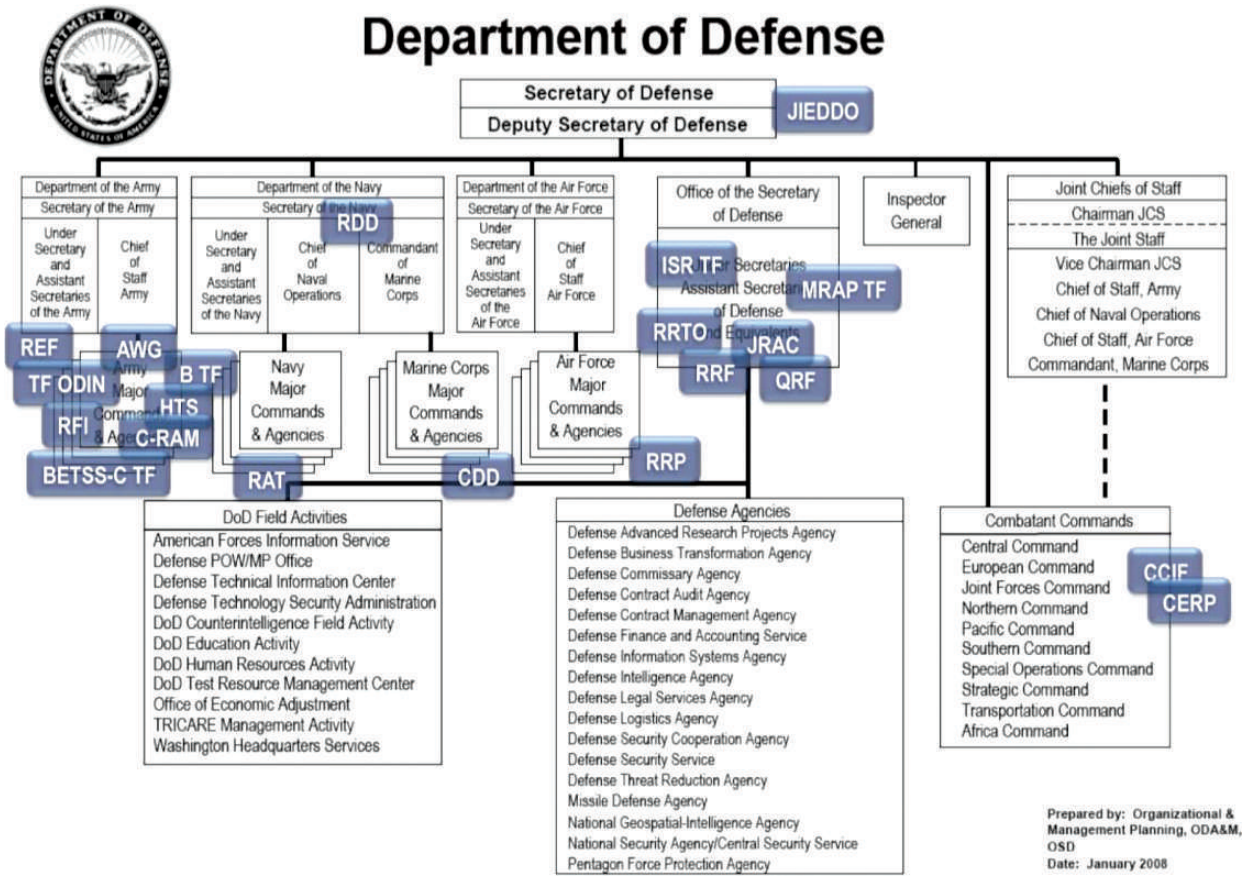


Figure 4. Representative DoD processes, funds, and organizations addressing urgent needs/rapid acquisition (DSB TF Report, 19; see glossary for explanation of acronyms, 57–61).

zational constructs were discussed, no specific internal organization of the RAFA was recommended. It was recommended that the RAFA report to the Under Secretary of Defense for Acquisition, Technology and Logistics and have a dotted line to the Vice Chairman of the Joint Chiefs of Staff. A key recommendation was that the RAFA be headed by a three-star–level active duty officer to maintain a strong and persistent relationship with the warfighter.

This recommendation is not intended to create an organization with responsibilities that overlap with those of other established organizations. *Figure 3* depicts the DSB Task Forces notional view of how the RAFA would fit with two existing organizations and Service and DoD Components acquisition organizations under the purview of the Under Secretary of Defense for Acquisition, Technology and Logistics.

The Task Force recommendation described at some length RAFA's mission "*to rapidly address combatant command needs with proven and emerging technologies in 2 to 24 months*" [emphasis in original][24] and "to *provide integrated triage for incoming needs from combatant commands*" [emphasis in original].[25] The operational assessment of capability provided through the RAFA was not explicitly addressed in the discussion of the RAFA's mission; however, the Task Force, as described elsewhere in their report, recognized the need for expedited operational assessment of an urgent capability upon which the receiving warfighter could make informed decisions to accept the capability.

Initial funding and billets for RAFA will be based on absorbing and integrating existing programs and organizations[26]

The Task Force recognized the potential difficulties with establishing a new Defense Agency. Therefore, they recommended that initial funding and billets for the RAFA should be based on absorbing and integrating existing ad hoc efforts in the OSD. Some of these organizations are depicted in the DoD top-level organization chart in *Figure 4*. Specifically recommended by the Task Force is the use of the Department's Rapid Reaction Fund and Quick Reaction Fund and the billets of the Rapid Reaction Technology Office, the JRAC, and the Joint Concept Technology Demonstration program. It should be

noted that the Task Force recommended the use of existing billets and not necessarily the aggregation of the personnel in the present organizations into the new RAFA. Discussion during Task Force deliberations clearly indicated that the personnel filling the RAFA billets should be specially selected.[27]

DoD should establish a streamlined, integrated approach for rapid acquisition[28]

The Task Force recommended that the RAFA provide continuous oversight of all steps in the urgent needs process and also provide a liaison to the combatant command that authored the urgent need statement. The RAFA director should have acquisition and funding decision responsibility, and RAFA and the combatant command should jointly approve and validate the need, concept of operations, and the proposed initial operating capability. The Task Force recommended tightly coordinated needs, acquisition, and funding steps as a critical feature of the overall process. They further recommended that execution be concurrently tracked while considerations are evaluated and an initial operating capability is approved. Successful completion of these steps leads directly to production and fielding of an initial operating capability and a transition to production or sustainment and operation funding. The Task Force recommended that the RAFA and each Service jointly manage production (as appropriate), and RAFA work with each Service to integrate doctrine, organization, training, materiel, leadership and education, personnel, and facilities and life cycle issues.

Summary and status of implementation of DSB Task Force recommendations

The six major findings supported the five major recommendations of the DSB Task Force on the Fulfillment of Urgent Operational Needs.

1. formalize a dual acquisition path,
2. establish a fund for rapid acquisition and fielding,
3. establish a new agency: the Rapid Acquisition and Fielding Agency,
4. provide initial funding and billets for RAFA from existing programs and organizations,
5. establish a streamlined, integrated approach for rapid acquisition.

Since January 20, 2009, the new administration has appointed new leaders to key DoD positions who have either been briefed or provided copies of the report for their consideration in establishing the way forward to meeting warfighters' urgent operational needs.

Public discussion of the DSB Task Force study, findings, and recommendations occurred in testimony before the House Armed Services Committee Acquisition Reform Panel, October 8, 2009. The Hon. Dov S. Zakheim, a previous DoD Under Secretary of Defense Comptroller, current member of the Defense Business Board, and a member of the DSB Task Force, testified. His testimony strongly supported the findings and the recommendations contained in the DSB Task Force Report, including the need for a separate acquisition path for urgent needs, a separate Defense Agency, and the establishment of a fund to support the fulfillment of urgent needs.[29] The Hon. Dov Zakheim emphasized the following in his testimony:

> *"Put simply, the department needs to field militarily useful solutions more quickly. The current threat environment is one in which the enemy on the battlefield employs easily obtainable, off-the-shelf technology to undermine the effectiveness of U.S. military operations. Yet DoD has made no permanent institutional changes in its acquisition, programmatic and budgetary systems to account for the growing sophistication and flexibility of the threat."*[30]

He went on to testify on the need for support for urgent needs processes by testing and other communities:

> *"The defense acquisition workforce has for many years functioned in an environment that rewards following complex procedures with accuracy and precision, but penalizes those who take shortcuts around those procedures. Yet it is precisely creativity and 'workarounds' that are critical to meeting urgent requirements successfully and in a timely fashion. Sustaining an effective rapid acquisition capability therefore will call for the active support of the testing, resourcing and requirements communities in an unprecedented manner."*[31]

Testimony of a current DoD official argued that the present "deliberate" acquisition system also should become more agile in meeting the Department's requirements. His written statement included the following:

> *"A July 2009 congressionally-directed study by the Defense Science Board Task Force on Fulfillment of Urgent Operational Needs concluded that existing institutions and procedures are incapable of meeting the Department's need*

for rapid and agile acquisition in time of war. Consequently, the study recommended two separate acquisition structures: one for 'deliberate' acquisitions, and one for 'rapid' acquisitions. While the Department continues to review the recommendations of that study, the risk of accepting two distinct structures is a failure to accept that all acquisitions, wartime and peacetime, need to become more agile and responsive in order to keep pace with accelerating development cycles enabled through global access to information and incorporation of commercial technology, especially information and communications systems, in any potential adversary's arsenal. To prepare the Department for the agile threats we must surely anticipate in the future, we need to make our 'deliberate' processes much more relevant to the current fight and capable of responding to urgent needs."[32]

He further stated:

"Under the leadership of the Under Secretary of Defense (Acquisition, Technology and Logistics) Dr. Ash Carter, and his Director of Defense Research and Engineering, Mr. Zach Lemnios, we have restructured the Directorate of Defense Research and Engineering to emphasize the rapid fielding of new technologies, while continuing the invaluable work of discovering and expanding the science for future capabilities. It's not enough to simply respond to new threats. Within the Director of Defense Research and Engineering organization, we consolidated hitherto disparate functions and created a new Rapid Fielding Office charged with discovering the best and most relevant technologies from the commercial and public sector and, when appropriate, facilitating their rapid fielding to theater. This new office is working to better integrate the science and technology with demonstration and prototype efforts throughout the Department and to focus those efforts on supporting the current fight. … The Rapid Fielding Office has also taken over responsibility for the Department's Joint Rapid Acquisition Cell (JRAC), to ensure better synergy between the requirements, acquisition and research communities." [33]

The full set of the DSB Task Force's recommendations is under consideration by the Department of Defense. The study and report to Congress provides a valuable starting point to discuss the future fulfillment of urgent needs. ❑

William Beasley, a career Department of Defense civilian, serves in the Joint Rapid Acquisition Cell (JRAC), Office of the Secretary of Defense. He operated the JRAC from September 2007–March 2009 and also served as a member of the Joint IED Defeat Organization's Senior Resource Steering Group. He graduated from the U.S. Military Academy and holds advanced degrees from Long Island University (MBA), the Massachusetts Institute of Technology (Physics), and the U.S. Army War College (Strategic Studies). He is a member of the Defense Acquisition Corps and is Level III Certified in Systems Planning, Research, Development and Engineering. He has been actively involved, since its inception, with the Joint Staff's Joint Capability Integration and Development System. He has over 35 years' experience in the military and with industry operating, managing, and developing military and cross federal agency information technology capabilities at all levels—tactical through national, in wartime, peacetime, disaster, and continuity of government roles. E-mail: William.Beasley@osd.mil or William.Beasley1@us.army.mil

Endnotes

[1]U.S. Department of Defense, *Report of the Defense Science Board Task Force on the Fulfillment of Urgent Operational Needs*. Washington, DC, Office of the Secretary of Defense for Acquisition, Technology and Logistics, July 2009, 41 http://www.acq.osd.mil/dsb/reports/ADA503382.pdf (Accessed April 23, 2010).

[2]*Duncan Hunter National Defense Authorization Act for Fiscal Year 2009*, Public Law 110–417, 110th Cong., 2nd Sess. (October 14, 2008), Section 801: "Assessment of Urgent Operational Needs Fulfillment."

[3]*Report of the Defense Science Board Task Force on the Fulfillment of Urgent Operational Needs* (DSB TF Report), 61–63.

[4]Key prior studies include:

- U.S. Department of Defense, *Defense Science Board 2006 Summer Study on 21st Century Strategic Technology Vectors*, Vol. I, Washington, DC, Office of the Under Secretary of Defense for Acquisition, Technology and Logistics, February 2007 [http://www.acq.osd.mil/dsb/reports/ADA463361.pdf (accessed April 23, 2010)].
- U.S. Department of Defense, *Report of the Defense Science Board Task Force on Defense Industrial Structure for Transformation – Creating an Effective National Security Industrial Base for the 21st Century: An Action Plan to Address the Coming Crisis*, Washington, DC, Office of the Under Secretary of Defense for Acquisition, Technology and Logistics, July 2008 [http://www.acq.osd.mil/dsb/reports/ADA485198.pdf (accessed April 23, 2010)].
- U.S. Department of Defense, *Report of the Defense Science Board Task Force on Integrating Commercial Systems into the DoD, Effectively and Efficiently – Buying Commercial: Gaining the Cost/Schedule Benefits for Defense Systems*, Washington, DC, Office of the Under Secretary of Defense for Acquisition, Technology and Logistics, February 2009 [http://www.acq.osd.mil/dsb/reports/ADA494760.pdf (accessed April 23, 2010)].
- U.S. Government Accountability Office, *Defense Acquisitions: Perspectives on Potential Changes to Department of Defense Acquisition Management Framework*, Washington, DC, February 27, 2009 [http://www.gao.gov/cgi-bin/getrpt?GAO-09-295R, accessed April 23, 2010].
- U.S. Department of Defense, *Report of the Defense Science Board – Creating a DOD Strategic Acquisition Platform*, Washington, DC, Office of the Under Secretary of Defense for Acquisition,

Technology and Logistics, April 2009, [http://www.acq.osd.mil/dsb/reports/ADA499566.pdf (April 23, 2010)].

- U.S. Department of Defense, *Report of the Defense Science Board 2008 Summer Study on Capability Surprise.*Volume I: Main Report, Washington, DC, Office of the Under Secretary of Defense for Acquisition, Technology and Logistics, September 2009, [http://www.acq.osd.mil/dsb/reports/ADA506396.pdf (accessed April 23, 2010)].
- U.S. Department of Defense, *Report of the Defense Science Board 2008 Summer Study on Capability Surprise,*Volume II: Supporting Papers, Washington, DC, Office of the Under Secretary of Defense for Acquisition, Technology and Logistics, January 2010, [http://www.acq.osd.mil/dsb/reports/ADA513074.pdf (accessed April 23, 2010)].

[5]Hon. Jacques S. Gansler, "Final Report of the Defense Science Board Task Force on Fulfillment of Urgent Operational Needs," memorandum for Chairman of the Defense Science Board, Washington, DC, June 23, 2009. Included in DSB TF Report, iii–iv.
[6]DSB TF Report, 6.
[7]Ibid., 24.
[8]Ibid.
[9]Ibid., 25.
[10]Ibid.
[11]Ibid.
[12]Ibid., 26
[13]Ibid., See Figure B–3, 44.
[14]Ibid., 26.
[15]Ibid., 27.
[16]Ibid.
[17]Ibid., 28.
[18]Ibid., 29.
[19]Ibid., 31–32.
[20]Ibid., 32.
[21]Ibid., 32–33.
[22]Ibid., 33–37.
[23]Ibid., 33.
[24]Ibid., 35.
[25]Ibid., 36.
[26]Ibid., 38.
[27]Ibid.
[28]Ibid., 38–40.
[29]Dov S. Zakheim, Statement before the Defense Acquisition Reform Panel, House Committee on Armed Services, October 8, 2009. http://armedservices.house.gov/pdfs/DAR100809/Zakheim_Testimony100809.pdf. (Accessed April 23, 2010).
[30]Ibid.
[31]Ibid.
[32]Ibid. Thomas P. Dee, Director, Joint Rapid Acquisition Cell, Office of the Under Secretary of Defense (Acquisition, Technology and Logistics), Statement before The House Armed Services Committee Acquisition Reform Panel, October 8, 2009. http://armedservices.house.gov/pdfs/DAR100809/Dee_Testimony100809.pdf. (Accessed April 23, 2010).
[33]Ibid.

ITEA Journal 2010; 31: 179–209

Review of Cognitive Metrics for C2

Mandy Natter, Jennifer Ockerman, Ph.D., and Leigh Baumgart
Johns Hopkins University Applied Physics Laboratory, Laurel, Maryland

Human cognitive knowledge, skills, and abilities are a significant component of complex command and control (C2); hence, measuring cognitive aspects of C2 can provide critical value-added. Cognitive measures provide a consistent gauge to measure C2 cognitive effects. These measures can be used to compare cognitive impacts both between and within systems. Also, these measures help to analyze specific cognitive strengths and weaknesses, so that C2 systems can be improved. Likewise, cognitive measures can be used to analyze training strengths and weaknesses, and to improve training so that it is better suited to user needs. This article summarizes an extensive literature review on macrocognitive metrics that apply to complex C2 assessment. Since a suite of cognitive metrics is required to assess C2 warfighters' actual and perceived effectiveness, guidance is provided on selecting appropriate macrocognitive metrics. Mental constructs, researched in complex C2 domains including workload, situational awareness, decision making, and collaboration, are highlighted. This article defines each construct, provides measurement tools and techniques, and reviews the costs and benefits of each technique. The article concludes with an explanation of how the mental constructs and their metrics are interrelated and suggests using several metrics together to assess and explore C2 in complex endeavors.

Key words: Cognitive measures; collaboration; command and control; decision making; human cognitive knowledge; net-centric operations; situation awareness; warfighter; workload.

Introduction: measures overview

As warfighters in command and control (C2) environments are being required to master a larger set of skills for increasingly complex tasks, performance alone is not all that matters in the evaluation or design of a system. The cognitive demands in net-centric operations become increasingly complex as operators must integrate vast amounts of information with varying content, format, age, and degree of uncertainty. Several measure attributes, such as type of measure, measurement scales, number of participants measured, how the measures are rated, who rates the measures, how timing is measured, and a variety of other factors affect the measure used.

Measurement types

Types of measures include quantitative versus qualitative and objective versus subjective. Simply stated, quantitative measures involve quantities (i.e., numbers), and qualitative measures involve written descriptions (i.e., words). Objective measures are often quantitative and are based on exact external measures of things or concepts that exist. Subjective measures are typically qualitative and are based on personal opinion or judgment. However, the pairing of these terms is not exact; *Figure 1* shows how these types of measures interact.

Measurement scales

Four main types of measurement scales exist: nominal, ordinal, interval, and ratio. Nominal scales are distinct and indicate a difference between entities (e.g., A not B). Ordinal scales are lesser or greater relative scales without reference units. They reflect a difference and the direction of difference. Interval scales include equal measurement intervals and do not have end points. Interval scales depict differences, including the direction and magnitude of the difference. Ratio scales are capable of all mathematical manipulations because they have a "true" or defined zero (e.g., length or speed). Common scales used for cognitive metrics include ordinal Likert-type semantic

	Quantitative	Qualitative
Objective	"The chip speed of my computer is 2GHz"	"Yes, I own a computer."
Subjective	"On a scale of 1-10, my computer scores 7 in terms of its ease of use."	"I think computers are too expensive."

Figure 1. Interaction between quantitative, qualitative, objective, and subjective (Hodgson 2007). Reprinted with permission from Blueprint Usability LLC. Chicago, Illinois. Diagram is available at http://www.blueprintusability.com/topics/articlequantqual.html

scales (e.g., 1 low and 7 high) or ratio scales (e.g., speed and accuracy) (O'Donnell and Eggemeier 1986).

Number of participants measured

The number of participants depends on objectives and available resources. If the goal is to become familiar with the data collection process, and resources are low, a pilot study can be used. A pilot study is small in scope and contains a small number of individuals. If the goal is to provide a high probability of detecting a significant effect size and the magnitude of the effect (if an effect exists), then a power analysis is required. A pilot study can also be used, and a power analysis conducted, to determine the necessary number of participants required to achieve the potential for statistical significance (Bausell and Li 2002).

Another aspect to consider in cognitive metrics is whether the task is individual or team based. Most cognitive metrics address the individual, but team measures are needed for C2.

How measures are analyzed

Typically, descriptive statistics, which describe what the data show, are run on quantitative data if the distribution is normal. If more resources are available, it might be possible to obtain inferential statistics, which are generalizable to a larger population. Qualitative results often reflect trends and patterns in responses.

Raters

Raters can include virtually all involved in a study. There are several cognitive metrics that are self-reports. Other popular raters are subject matter experts (SMEs), and the experimenter can also rate participants.

Timing

Cognitive C2 measures are collected either in real time or post hoc. Real time is usually preferable but can be intrusive. Post hoc is less intrusive but relies on the participants' and raters' memories.

Measurement criteria

Given the diversity of techniques available, it is important that the appropriate technique be chosen based on previous research and practical constraints. O'Donnell and Eggemeier (1986) propose that the criteria in *Table 1* should be met by any technique to assess workload; however, these criteria can be generalized to any cognitive measure and have been used for noncognitive measures as well (O'Donnell and Eggemeier 1986).

In order to adequately choose C2 cognitive assessment techniques, it is important to identify the objective of the measurements. Once the objective is identified, it is important to select measurements fitting the objectives with high validity and repeatability/reliability. Next, the sensitivity, diagnosticity, and selectivity can be considered to down-select appropriate measurements. Practical constraints can be further used to screen assessment techniques, with

Table 1. Measurement criteria.

Criteria	Description
Validity	The extent to which the measure is measuring the mental construct of interest (Zhang and Luximon 2005).
Repeatability/reliability	The ability to obtain the same results of the mental construct when tests are administered more than once (Zhang and Luximon 2005).
Sensitivity	The ability to detect changes in the level of the mental construct imposed by task difficulty or resource demand.
Diagnosticity	The ability to discriminate the amount of the mental construct imposed on different operator resources (e.g., perceptual versus processing versus motor resources).
Selectivity	The ability of a measure to be sensitive to differences only in the cognitive construct of interest (e.g., cognitive demands as opposed to physical workload or emotional stress) (Zhang and Luximon 2005).
Intrusiveness	The tendency of a technique to interfere with performance on the primary task.
Implementation requirements/ convenience	The ease of implementing a specific assessment technique (e.g., instrumentation requirements or operator training).
Operator acceptance	The degree of willingness on the part of operators to follow instructions and actually utilize a particular assessment technique.

intrusiveness and implementation requirements being more heavily weighted than operator acceptance (O'Donnell and Eggemeier 1986). In many cases, the best cognitive assessments are derived using multiple techniques in order to capture the complexity of the operator's task demands.

Cognitive C2 metrics

Currently available cognitive metrics relevant to C2 include workload, situation awareness (SA), decision making, and collaboration. Although each of these can be measured in isolation, they are all related to each other. An appropriate workload level should lead to good SA, which in turn should improve decision making and enhance collaboration. This section will provide examples of each of these groups of cognitive metrics. In many cases, as will be noted, more details can be found in the Appendix.

Workload

The human operator has a limited capacity for collecting, interpreting, processing, and responding to information in their environment. If the demands of a task exceed an operator's limited capacity, decrements in performance and an increased potential for errors often occur. In the C2 environment, such decrements in performance may be substantial, potentially involving errors that lead to the loss of valuable resources and even the loss of life.

Defining workload. Workload is often defined as the portion of the operator's limited capacity that is required to perform a particular task (O'Donnell and Eggemeier 1986). For C2, workload refers to the cognitive demand on the brain and the sensory system (eyes, ears, and skin) due to the task versus physical workload (Zhang and Luximon 2005). The assessment of workload can be used to improve either the tools used by an operator, such as enhancing computer interfaces, increasing automation, and reallocating tasks, or to improve the operator by providing more training.

It is important to understand what optimized workload levels are for particular tasks because nonoptimal levels of workload may induce stress or boredom. Excess stress often results in changes in information processing, possibly increasing the occurrence of errors. Other consequences of nonoptimal workload may include the operator shedding tasks of lower priority, potentially in an unfavorable manner. High levels of workload are not necessarily always "bad." In many environments, low levels of workload coupled with boredom, fatigue, or sleep loss, can have negative impacts on performance as well. The Yerkes-Dodson law demonstrates an empirical relationship between arousal (pressure) and performance (ability to cope) *(Figure 2)*. The law depicts the importance of avoiding extremes of too little or too much workload.

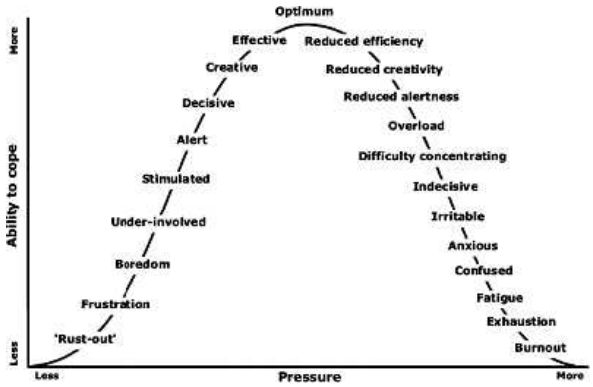


Figure 2. Yerkes-Dodson Law (Haarmann, 2007).

Workload assessment techniques. Over the last 4 decades, a considerable amount of research has been dedicated to understanding and assessing cognitive workload in a variety of domains. Many techniques for workload assessment have been proposed. These techniques generally fall into one of four categories: (a) primary-task measures, (b) secondary-task measures, (c) subjective measures, and (d) physiological measures (O'Donnell and Eggemeier 1986; Wickens and Hollands 2000).

Primary-task measures of workload. Some researchers have used changes in the quality of operator performance on primary tasks as a measure of operator workload. It is often assumed that as workload increases, the resources used by an operator will also increase, resulting in a degradation in operator performance. However, this may not always be the case; in fact, in situations where workload levels increase from very low to moderate, performance may actually improve because the operator becomes less bored and more engaged.

Primary-task workload measures typically include speed and accuracy; however, these measures alone may be insufficient to clearly assess the qualities of the task. For example, good measures of performance on some primary tasks are difficult to define, such as for decision making in the C2 environment. In this environment, the cognitive demand placed on the operator is great, yet the performance outcome result is a function of many other variables aside from the operator's cognitive operations. Further, performance may differ based on data or system limitations, and not on the cognitive demands placed on the operator to perform the task (Wickens and Hollands 2000).

Using primary-task measures of workload alone also makes it difficult to compare operator workload levels across different tasks or different systems. This is especially apparent when workload is extremely low and achieving perfect performance is easy. In these instances, primary-task measures alone would not be able to distinguish workload differences between the two systems.

Secondary-task measures of workload. Secondary tasks are used primarily to simulate realistic workload levels that may not always be experienced in the laboratory or a simulated environment. However, the application of secondary tasks can also be used to assess operator workload. This technique involves imposing a secondary task on an operator in order to measure the residual resources or capacity not utilized in the primary task. Secondary-task performance is assumed to be inversely proportional to primary-task resource demands. Therefore, differences in primary-task resource demand that are not reflected in primary-task performance alone may be revealed.

When using this technique, the investigator's emphasis should be on the primary task, but the variation in secondary-task degradation is measured (Wickens and Hollands 2000). Operators are instructed to avoid degradations in primary-task performance at the expense of the secondary task to ensure that primary-task performance is not affected by the secondary task.

Common examples of secondary tasks to measure workload include the random number generation technique, which involves having the operator generate a series of random numbers (Wetherell 1981). It has been found that the randomness declines as the workload required for primary task increases. Another related technique is to measure an operator's reaction time to certain probes, or questions, asked while they are completing the primary task, as it has been found that reaction time to secondary-task stimulus will increase with increasing primary-task workload (Lansman and Hunt 1982; Wetherell 1981).

A variant to implementing a secondary task is to use a loading task. In this environment, the operators are asked to devote all necessary resources to the loading task, and the degree of intrusion of this task on performance of the primary task is examined. For example, operators could be asked to monitor a gauge and push a button each time it falls below a certain level. In this case, degradations in performance for the primary task provide an indication of the resources demanded.

One difficulty to using secondary tasks is that they often interfere with the performance of the primary task of interest. To overcome this obstacle, many researchers have employed the use of embedded secondary tasks. These are tasks that are a legitimate component of the operator's typical responsibilities, but lower in priority, such as an operator responding to a verbal request from a commander as to the latitude and longitude of a friendly object when their main priority is to closely monitor an unfriendly target for movement. Other researchers have used a chat interface to induce information-seeking secondary tasks during C2 activities (Cummings and Guerlain 2004).

Another method for implementing an embedded secondary task was demonstrated by Raby and Wickens (1994). In their experiment, secondary tasks were divided into "must," "should," and "could" be done. Time spent performing tasks in the three categories were compared between different scenarios of varying workload (defined by time pressure and external communications requirements). This technique provided an accurate means of comparing the differences between workload levels in each of the scenarios (Raby and Wickens 1994).

Using secondary tasks to assess workload provides a high degree of face validity as it helps predict the residual resources an operator will have leftover in the event of a failure or unexpected event. The same secondary task can also be used to compare the workload of two different primary tasks. However, it is important to consider the different kinds of resources (e.g., vision, hearing, touch) required by a primary task before selecting a secondary task. Workload may be underestimated if the resource demands of the secondary task do not match those of the primary task (Wickens and Hollands 2000).

Subjective measures of workload. Subjective workload assessments elicit the subject's perception of cognitive loading during a recently completed task. These assessments take the form of questionnaires or structured/unstructured interviews, and typically subjects submit self-ratings either manually or through automated systems (Cherri, Nodari, and Toffetti 2004). Subjective workload techniques include a predefined rating scale that is either one-dimensional or multidimensional with written or verbal descriptions for each level of the scale. One-dimensional scales, which relate one aspect of workload at a time, are beneficial when diagnosticity is important; while multidimensional scales, which deal with several workload factors at one time, are preferable for a global rating of workload more sensitive to manipulations of task demand (Cherri, Nodari, and Toffetti 2004). Because multidimensional scales touch on several factors, when tasks change, ratings are more

likely to reflect impacts on one or several of the factors than in one-dimensional scales.

Because of the need for cognitive workload assessment and the limitation of other measures, researchers have proposed several subjective workload assessments (Gawron 2000). The three most popular techniques are National Aeronautics and Space Administration–Task Load Index (NASA-TLX), (De Waard 1996), Subjective Workload Assessment Technique (SWAT) (Rubio et al. 2004), and Modified Cooper-Harper (MCH) Scale, (Zhang and Luximon 2005). NASA-TLX and SWAT are multidimensional scales where subjects weigh the dimensions in the scales by perceived priority, complete the task, and then score the dimensions on a 0–100 bipolar scale, where 0 represents virtually no perceived workload and 100 represents high workload. Selected dimensions can be analyzed separately and/or an overall score can be calculated based on the subject's weights. MCH is a one-dimensional scale where subjects make direct estimates of cognitive loading after completing a task (De Waard 1996; O'Donnell and Eggemeier 1986). More details on these methods can be found in the Appendix.

Subjective mental workload methods are popular due to their practical advantages (e.g., low cost, ease of use, general nonintrusiveness), high face validity, and known sensitivity to workload variations (O'Donnell and Eggemeier 1986; Zhang and Luximon 2005). The costs are low, as no particular equipment is required. Subjective workload assessments are easy to employ because they are easily accepted and used by subjects. Low primary-task intrusion is secured as long as the scale is administered after completion of the task. Subjective assessments have high face validity because effort is reported directly from the person experiencing the workload. When biases are low, subjective workload methods can be more sensitive than objective measures (Zhang and Luximon 2005).

The main caveats are confounding factors and short-term memory constraints. One confound is that subjects may confuse different types of task loading (e.g., physical, mental) and personal factors (e.g., mood, fatigue) with cognitive effort required for the task (O'Donnell and Eggemeier 1986). Also, not all of the processing done by an individual is available to conscious introspection, therefore hindering the subjective assessment sensitivity. Further, biases such as dislike or unfamiliarity of the task and hesitations to report difficulties affect workload assessments. In addition, because subjective workload assessments are often administered after the task to prevent intrusiveness, subjects may forget elements of effort expenditures (Cherri, Nodari, and Toffetti 2004; O'Donnell and Eggemeier 1986). Even in instances where subjective assessments are embedded into tasks, they are noncontinuous and do not reflect variations in workload during the task.

Subjective assessments should be considered as global indicators of workload versus highly specified diagnostic techniques. Also, it is beneficial to obtain workload ratings as soon as possible after task performance to minimize degradation due to short-term memory limitations.

Physiological measures of workload. Various physiological measures have been investigated to objectively measure workload. They include eye measures (e.g., pupil diameter, blink rate and duration, saccade number and duration, fixation frequency and duration), cardiac measures (e.g., heart rate variability), neural measures (e.g., background brain wave activity, event-related potentials [ERPs]), and skin measures (e.g., electrodermal response [EDR]). There has been limited success with each of them, but none is currently accurate enough to be used on its own, and a use of several at once is recommended. More details on these measures can be found in the Appendix.

Situation awareness (SA)

It is critical that human operators have an awareness of what is happening in C2 situations, so that they can understand the tasks they are conducting and the context within which they are working. As mentioned earlier, situation awareness often supports decision making, so improving situation awareness can lead to improved decision making.

Defining SA. In the 1950s, the U.S. Air Force coined the winning element in air-to-air combat engagements in Korea and Vietnam as the "ace factor" or what they called having good SA (Spick 1988). Since the term SA originated, it has been expanded to include almost any domain that involves humans performing tasks with complex, dynamic systems. As applications have spread and increased, so have SA definitions and measurement techniques. Some SA definitions are human-centric, others are technology-centric, and some encompass both the human and the technology, but all generally refer to knowing what is going on and what will happen next. SA is important because it frequently guides decision making and action (Gawron 2000). The most widely accepted definition is Endsley's human-centric interpretation that "situation awareness is the *perception* of elements in the environment within a volume of time and space (level 1), the *comprehension* of their meaning (level 2), and the *projection* of their status in the near future (level 3)" (Endsley 1995, p. 85) *(Figure 3)*.

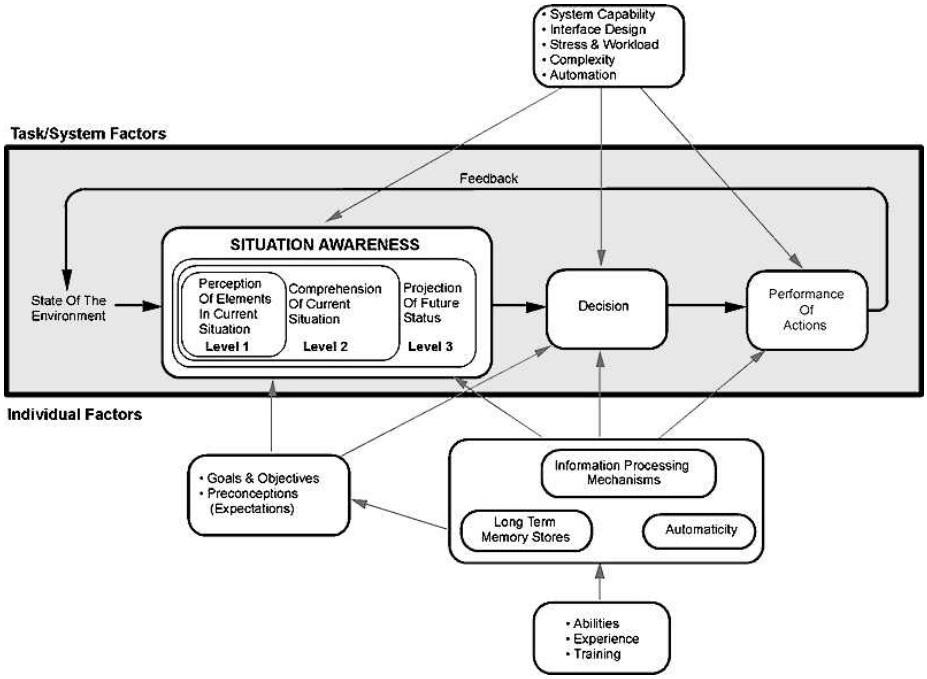


Figure 3. Endsley's model of situation awareness (Endsley 1995).

Military and C2 applications, often refer to SA as knowledge of the physical elements in the environment (equivalent to Endsley's SA level 1), while the other levels (equating to levels 2 and 3) are referred to as situational understanding and assessment (Dostal 2007). The processes involved with arriving at and maintaining situational understanding in C2 are sometimes called sensemaking (Gartska and Alberts 2004). Technology-centric definitions of SA are linked to C2 applications in that they often refer to the quantity and quality of the information provided by the technology and include data visualization. Technology-centric SA addresses technical challenges of SA; for example, information overload, nonintegrated data, rapidly changing information, and high degrees of uncertainty (Bowman and Kirin 2006).

Human-system definitions have recently gained maturity and popularity and relate the information provided by the system to the information needed by the operator. Work by Riese, Kirin, and Peters (2004) and Miller and Shattuck (2004) supply good examples of modeling this relationship. Riese et al.'s model reflects information gleaned from technology (SA_T) being transferred to a human for cognitive situation awareness (SA_C) via an interface *(Figure 4)* (Riese, Kirin, and Peters 2004). Miller and Shattuck's model leverages Endsley's human-centric definition and the lens concept in a multi-step process as shown in *Figure 5* (Miller and Shattuck 2004). The left-hand side illustrates the technology part of situation awareness (SA_T) while the right-hand side represents the human or cognitive situation awareness (SA_C). As can be seen, some amount of the information from the world is detected by sensors and some amount of that information is made available to the human. The human then perceives the information being displayed, comprehends or makes sense of that information, and finally uses that information to predict what will happen in the world.

SA assessment techniques. No matter the definition utilized, SA is challenging to measure. The information that is required at a particular time, in a particular situation, depends on the current goals and objectives of the C2 organization, which are often dynamic. Even when all information is accessible, only a subset of that information is needed to plan and assess the current goals and objectives. Finding the right information at the right time to be aware of what is happening is a challenge, as is leveraging pertinent information to be

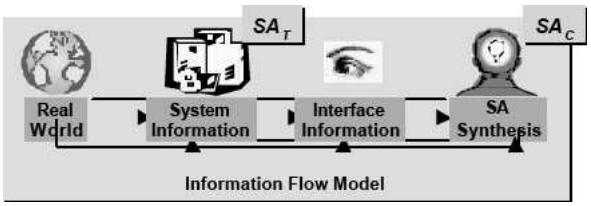


Figure 4. Riese et al. model of situation awareness (SA) (Riese, Kirin, and Peters 2004).

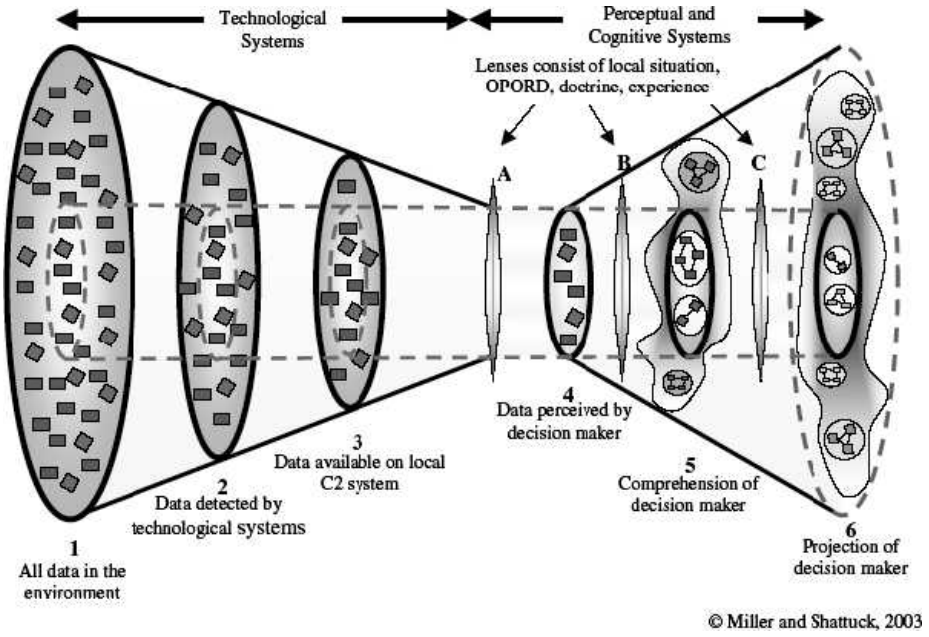


Figure 5. A dynamic model of situated cognition (Miller and Shattuck 2004).

able to make a decision. In complex, real-world scenarios, it is critical that SA measurement questions and methodology are tailored to the domain and context in which they will be used. There are three main categories of measurement strategies: explicit, implicit, and subjective.

Explicit SA measures. Explicit measures assess the users' understanding of what is going on. "Probes," or questions, are administered to prompt subjects to self-report their actual SA. Endsley's three-level information processing definition has a corresponding, validated measurement technique, called SAGAT (Situation Awareness Global Assessment Technique), which is the most commonly used and cited SA metric (Salmon et al. 2006). During SAGAT, the task or simulation is momentarily frozen at various intervals, and subjects are asked a set of predetermined multiple-choice questions that relate to the level of SA they have about the situation at that time. Some critics of SAGAT suggest that it measures recall or information decay versus SA; however, Endsley conducted studies reflecting that the freezes do not impact SA probe responses (Gawron 2000). More recently Endsley's company, SA Technologies, developed a product called The Designer's Situation Awareness Toolkit (DeSAT), which helps to build SA probes and measure SA. Another explicit methodology is to embed opened-ended questions throughout the task or simulation, called "real-time probes." This method is less intrusive and more "naturalistic" than interrupting and stopping the task, but because of the open-ended questions, responses can be inconsistent across participants.

Implicit SA measures. Implicit measures infer SA from objective but indirect evidence. In other words, there is an assumption that an objective measure that is not SA implies SA. The difficulty with this logic is that it holds for only simple responses and behaviors, because as the complexity increases, the relationship weakens. An advantage of implicit measures is that they are often easier to obtain than explicit measures and are frequently less intrusive. A common implicit SA measure is task performance. For instance, a participant who hits many targets is presumed to have awareness of the targets. In a C2 task, Hiniker (2005) defined SA as the proportion of mission critical warfighter platforms (red, blue, or neutral) correctly identified as important by the commander. The study results revealed that teams with a Common Operating Picture (COP) were able to perform the task 10 percent better than control teams, and therefore had better SA than the control group (Hiniker 2005). Another implicit SA measure is physiological data (as described in the previous section). Response patterns from some physiological data may correlate with various states or changes in SA state. Communications have also served as an implicit SA measure. For instance, a study by Artman (1999) reflected that the more successful team sent fewer messages between units than the less successful team (Artman 1999). Another example of an implicit measure is a continuous algorithm developed by Riese et al. (2004) to

measure instantaneous technical SA or SA_T. SA_T is a ratio of the amount of information provided by the technology/system to the amount of information needed by a decision maker. It is a technique that may be used for a system-centric definition of SA. The study concluded that the relative difference between levels of SA_T was related to battle outcome. In other words, information is linked to force effectiveness. The study also concluded that good SA_T does not necessarily lead to good cognitive SA (SA_C) because humans tend to maintain their beliefs and look for evidence to support their predictions (Riese, Kirin, and Peters 2004). More information on calculating SA_T can be found in the Appendix.

Subjective SA measures. Subjective measures of SA are the participant's own rating of his/her SA or someone else's ratings of the participant's SA. Some subjective techniques result in a single overall rating (Likert-type scale with a ratings from 1–7), whereas others use multiple scales and break SA into different elements of interest. Subjective methodologies can be fairly generic and administered to practically any operator or decision-maker or be designed to collect specific SA requirements for a particular role or task. Self-ratings, as the name reflects, are the participant's perception of his/her own SA.

The most commonly used self-rating subjective SA technique is called SART (Situation Awareness Rating Technique) developed by Taylor (Endsley et al. 1998). SART consists of three bipolar scales: (a) demand on operator resources, (b) supply of operator resources, (d) and understanding of the situation. These scales are combined, resulting in an overall score. Critics contend that SART may be confounded by performance and workload because it has correlated with both. Proponents highlight that it does not need to be customized for different domains and can be used in both simulated and real-world contexts.

Another subjective measure is observer ratings, which are specified by someone (usually a subject-matter expert), not involved with the task, observing the subject's performance. Finally, peer ratings are also used to subjectively evaluate SA. These techniques are employed in teaming experiments where one team member provides a subjective rating of another team member's SA and vice versa. The main advantages of subjective ratings, as noted in the subjective workload measures section, are that they are simple to employ and have high face validity. Also, the experimenter can elicit several ratings during a task for comparison across conditions and time. Subjective ratings also gauge perceived quality or confidence in SA, which can be as important as measuring actual SA, since over-confidence or under-confidence in SA may be just as detrimental as errors in actual SA (Sniezek 1992). The disadvantages are that they are usually administered after a task and, therefore, rely on memory. Individual differences and interrater reliability can confound results (e.g., a rating of a "3" to one person may not be a "3" to someone else). In addition, "unknown unknowns" are an issue, for instance when self-raters overestimate their actual SA because they do not know what they do not know.

Correlation between SA measurement techniques. Studies reveal that explicit and implicit SA measures correlate, but subjective and explicit measures do not correlate. Prince et al. (2007) tested 2-person aircrews on SA using both an explicit (modified SAGAT) and implicit (SMEs administered performance probes) method. Results reflected that both methods could be used to reliably measure SA. Both methods correlated slightly with performance and significantly to each other (Prince et al. 2007). Endsley et al. (1998) conducted a study comparing explicit (i.e., SAGAT) and subjective (i.e., SART) measurement techniques within a display evaluation. Both SAGAT and SART were sensitive and diagnostic regarding the effects of the display concept. SART highly correlated with subjective measures of confidence; however, SAGAT and SART did not correlate with each other (Endsley et al. 1998).

Team SA measures. Almost all SA measurement techniques refer to individual versus team SA, in part because of the variety of variables and complexity that impacts team SA. For instance, individuals may have degraded SA because they are ignorant of what they are supposed to do, but team SA may be degraded because of status within the team, lack of control, an expectation that another member of the team will take action, actions by team members that prevent an individual's actions, etc. (Prince et al. 2007). Despite this complexity, C2 environments are composed of teams, which are often distributed, and thus require team SA measurement techniques.

Salmon et al. (2006) evaluated 17 SA measures against a set of human factors criteria for use in command, control, communication, computers, and intelligence (C4i) environments. They concluded that current SA measurements are inadequate for use in assessing team SA, and recommended a multiple-measure approach. They identified three requirements for C4i SA measurement: (a) ability to simultaneously measure SA at different locations, (b) ability to measure SA in real-time, and (c) ability to measure both individual and team/shared SA (Salmon et al. 2006).

Measurement techniques for team SA have been researched and continue to be a subject of study (see Appendix for some examples of team SA definitions). Team/shared SA is in a definition phase; hence, explicit measurement methodologies have not been developed. As the definition matures, it is likely that measurement techniques will advance, but in the meantime, multiple-measure approaches combining collaboration and communication via recording how the team responds to unexpected situations is recommended. A recent human-system SA team measurement approach by Hiniker (2005) called the Tech Team Model of SA, includes the COP as a kind of integral member of the team. In real-world C2 and other stressful environments, implicit measures and subjective measures may be well suited because they are less intrusive than explicit techniques. Regardless of the SA measurement methodology, it is important that it not increase the workload of the participants, because increased workload correlates with degraded SA (Entin and Entin 2000).

Decision making

Complex decision making is the most significant human contribution to C2. Some simple decisions can be automated, but often a human is needed to assess risk, weigh alternatives, and select a course of action (COA). Decision making is a critical component of C2 because decision making supports and directs action. If the decision making process can be better and quicker, there is a higher likelihood that C2 will likewise be improved.

Defining decision making. Decision making is a complex process; not just a result. It involves selecting options from alternatives, where some information pertaining to the option is available, the time allotted is longer than a second, and there is uncertainty associated with the selection (Wickens and Hollands 2000). An integral component to making a choice is the risk involved with decision execution. The risk, assessed via probability or likelihood, and the consequences of a decision need to be taken into account (Wickens and Hollands 2000).

The system (human and technology agents) typically provides the required information, so that decision makers can perceive and interpret that information, generate COAs, evaluate the consequences of alternatives, and select a COA from alternatives (Azuma, Daily, and Furmanski 2006; Endsley et al. 2007; Walker et al. 2006). As with SA, decision making can be decomposed into information, or system-supported components and cognitive, or human-supported, components (Means and Burns 2005).

The goal of the information component is to present a manageable amount of information to warfighters that not only pertains to them but also provides context. Providing the "right" information for decision making depends on having the necessary content with appropriate data attributes. Means and Burns (2005) contend that decisions are information processing tasks and used data attributes to compare information associated with C2 decisions across three Air Force systems (Means and Burns 2005). They used Functional Decomposition Diagrams (FDDs) to depict the major goals, decisions, and information in each system. Then, each decision was rated (high or low) on three axes: (a) dimensionality—size of information that must be processed, (b) temporality—a change in the information that must be processed, and (c) uncertainty—ambiguities and probabilities in the information that the decision maker must consider. This technique can help to compare system supported aspects of decision making and can reveal potential improvements. For instance, when ratings are low across all three dimensions, automation may be implemented; and when ratings are high across all three dimensions, improving visualization may be beneficial to the decision maker (Means and Burns 2005). The other aspect of identifying the "right" information is the required information content. Information requirements for decision making can be derived through cognitive system engineering knowledge elicitation methods (e.g., task analysis [TA], cognitive task analysis [CTA], cognitive work analysis [CWA]). These methods can reveal the critical fragmentary evidence that experienced decision makers use to create a COA (Azuma, Daily, and Furmanski 2006).

Content and data attributes may be used to develop a rational model of decision making that can produce a decision autonomously. The rational approach is highly numeric and consists of Bayesian or other probabilistic models of action assessment. Rational models often take all possibilities into account and are advantageous in situations that are not time critical. Since the analyses are numeric, they are amenable to computer logic but do not map well to how decision makers actually make decisions under time pressure. In addition, rational models frequently do not include a dynamic component, preventing evolution over time. Finally, uncertainty needs to be known a priori to be accounted for in the model. This often leads to compressing uncertainty into probabilities or weightings subjectively assigned, which may obscure the amount, or even presence, of uncertainty. Thus, rational models are difficult in time-critical, uncertain environments like much of C2 execution, but might be beneficial during less

time-critical phases such as planning (Azuma, Daily, and Furmanski 2006).

The decision making strategies employed by humans in C2 more closely match naturalistic decision making (NDM) (e.g., recognition-primed decision making [RPD]) (Klein, Calderwood, and MacGregor 1989). In NDM, the decision maker recognizes the situation based upon features such as expectancies, plausible goals, relevant cues, and typical action. The solutions are not ideal or optimal, but good enough, especially in time-critical situations. Decisions might be based on fragmentary evidence as it is not feasible to fully quantify the situation and find a mathematic solution, and too much information may be detrimental. If an incorrect decision is made, it can often be changed. The person learns from the decision making process of observation and action. In this way, rational processes are outcome oriented, while naturalistic decision making strategies are process oriented. The disadvantage is the assumption with naturalistic models that the past is a good predictor of the future, which is not always true (Azuma, Daily, and Furmanski 2006).

Research applying intuitive (NDM) and analytical methods to C2 suggests that a continuum between the two techniques can be used depending on operational context situational resources such as computational power, information, and time. For instance, in planning situations, where more time is available, analytic strategies should be used, but in execution, intuitive strategies should be employed because less time is available. Bryant et al. (2003) recommend that analytic and intuitive decision making strategies be considered as synergistic styles that when combined can greatly enhance decision making by connecting planning and action (Bryant, Webb, and McCann 2003).

Decision making assessment techniques. Measuring decision making is complicated because defining a "good" decision is difficult, and many factors and dependencies influence decision making. In order for decision making to be a generalized metric, a good decision should be verified and validated. A method of verifying the decision might be to assume that the optimum decision would produce the maximum value if repeated numerous times. However, this requires defining value, which is often personally subjective and contextually dependent. Also, when repeated, there may not always be a single optimal choice because of differing time and other constraints. Finally, a decision maker may be more concerned with minimizing loss versus maximizing gain. A validation technique might be to assume that good decisions are those that produce "good" outcomes. This is a difficult assumption to maintain in uncertain, probabilistic environments like C2. Adversarial decisions and even own force actions can be unpredictable and impact mission success, which confounds the ability to equate decision making with mission effectiveness (Bolia and Nelson 2007).

Because decision making is complex, decision making assessments are usually based on what is known and observable. Observing the decision maker can reveal important aspects of decision making but relies on inferences as to what the decision maker is considering in forming his/her decision. In other words, observations indicate what the decision maker is doing, but not why (i.e., his/her rationale for making the decision, or what major influences led to the decision). Observations or decision products are often used to gauge decisions versus the process because the process is subjective and the products are more objective.

An example of a commonly used observable, objective, result-oriented decision making metric is performance. Decision maker performance is assessed by comparing the decision maker's decision products to ground truth. Performance is analyzed based on speed and accuracy of the decision(s) versus ground truth. The caveat is that ground truth may not always indicate a "good" or "best" decision. In complex situations like C2, there are several dependencies and risks that influence the outcome, and speed and accuracy can be faulty indicators of "good" decision making. Assuming that the "good" or "best" decision is known, if speed to decision is repeatedly low and accuracy of the decision is repeatedly high, when compared with the ground truth "best" decision, then performance can be a useful decision making assessment metric.

Other objective measures previously referred to in the workload section of this article and also relevant in decision making are physiological measures. In the Iowa Card Task, where individuals are given four decks of cards and a loan of $2000, participants are told that some decks will lead to gains and some decks will lead to losses. The participants are instructed to use the money as they see fit with the objective to win as much money as possible. Using galvanic skin response measurements, researchers found that individuals micro-sweated more when choosing cards from the disadvantageous decks before they were able to identify which decks were better. "Advantageous" decision making measured through micro-sweating was detected sooner than when participants could verbally explain their card choices. This reflects that cognitive operations are essential to decision making, and emotions and physiological activity may influence decision

making (Lamar 2006). The physiological measures might also be able to speed up the decision making process by identifying the decision before the decision maker is conscious of his/her decision. Also, these physiological measures could potentially be used as performance speed and accuracy decision making measures.

Another observable decision making metric is an expert decision maker's perception of the decision maker's ability. This is a subjective technique but can be useful because the expert can observe the decision maker's physical processes and deduce whether the decision maker is paying attention to important pieces of information. This technique also reduces the need to have ground truth and a "good" or "best" decision defined because it assumes that the expert will know the "good" or "best" decision. The issue is that experts do not always make "good" or "best" decisions, and they may not always make better decisions than novices. Also, the expert does not know the reason the decision maker made his/her choice.

The issues with these techniques is that they focus more on the decision results than the decision making process. Metrics and measures for decision making should require more than performance results because a correct decision may be the result of chance or luck. "Good decisions may as easily be a fortuitous consequence of ignorance" (Bolia and Nelson 2007, p. B73). Even if the individuals making decisions have similar levels of accuracy, they could have very different perceptions of the situation and justifications for their decisions (Cooke et al. 2000).

Since decision performance alone is not a sensitive indicator of decision making effectiveness, it may be more appropriate to look at ways that different environmental and informational characteristics influence the processing operations and outcomes of the decision process. First, it would be important to consider how the decision maker gathers and assesses evidence, and then it would be beneficial to determine how he/she uses the assessment to make a decision. Information assessment might consist of the sources of information, time to locate the information, and path to locate the information. As mentioned previously, some processing operations that influence decision making are mental constructs such as cognitive workload, SA, and sensemaking, so elements of all should be incorporated in decision making assessments (Bolia and Nelson 2007).

A number of methods exist to obtain information requirements and decision rationale from the decision maker, the most common being CTA. A variety of CTA methodologies have been developed that differ in approach, structure, emphasis, and resource requirements, but all include some sort of knowledge elicitation, analysis, and knowledge representation (Federal Aviation Administration Human Factors Division 1999; Militello and Hutton 1998). (Details on three examples of CTA can be found in the Appendix.) Often the knowledge elicited is not measured quantitatively; however, aspects that can be measured quantitatively are number or type of information/concepts considered and the number or type of consequences considered.

Despite the contrast between CTA techniques, they are useful in revealing C2 decision maker rationale, and all reveal information that could improve C2 processes, or be used to evaluate C2 decision making. The various levels of structure in CTA methodologies parallel the levels of structure in various aspects of C2 (Klein, Calderwood, and Macgregor 1989). Another positive aspect is that many of the CTA techniques are conducted retrospectively, which is less intrusive (Klein, Calderwood, and Macgregor 1989). The caveat to CTAs is that there is "no well-established metric or method for assessing the reliability of cognitive task analysis tools, and yet the issue is an important one." (Militello and Hutton 1998, p. 1634). Also, it is difficult to evaluate differences between CTA methods (Militello and Hutton 1998). This is partially because it is unknown what information is lost versus gained in comparison with other techniques and also because interviewees provide different information each time. Also, CTAs can be very resource intensive. Because individual differences impact how much information interviewees are willing to provide and respond, it is difficult to assess the reliability and validity of CTA methods. Finally, no advanced techniques for team CTA have been developed (Militello and Hutton 1998).

In addition to a lack of verifiable and validated decision making measurement techniques, decision making often involves more than one decision maker or contributor(s) to the decision. This increases the challenge of assessing decision making because, as in SA and workload measures, individual decision making measurement techniques are more mature than team assessments. As described in the next section, the combination of decision making and collaboration assessment techniques are relatively immature.

Communication and collaboration

Communication and collaboration are words that are frequently used interchangeably, but important distinctions exist between them. Communication is expression and may include sharing information but often is from one individual or party to another. Communication is a prerequisite for collaboration, but collaboration involves leveraging the information of others for the purpose of

reaching or meeting a goal or objective. It includes synergy of ideas between two or more parties, and depending on how it is defined, can include aspects of SA, workload, and decision making. Both collaboration and communication involve more than one individual, but collaboration often involves a team.

Collaboration assessment techniques. There are a variety of methods used for collaboration assessment. As in other sections of this article, technological aspects of collaboration will be investigated, and then human aspects of collaboration will be summarized. An example of an assessment of collaboration technology is an analysis of the interconnectivity of team members to information sources, meaning the ability to obtain required information via communication channels. A similar technical support of collaboration that can be assessed is the interconnectivity of team members to each other, in other words whether team members have modes of communication available to collaborate (Freeman and Serfaty 2002). In complex C2, often C2 nodes are distributed and collaborate via the Internet (or soon to be Global Information Grid), so connectivity issues are important to consider. Net-centric metrics such as available bandwidth for communication, communication availability, supported modes of communication, access to communication or collaboration tools, etc., are examples of technological collaboration metrics.

Assessing human aspects of collaboration are usually more complex than technical aspects, depending on the collaboration attribute being analyzed. Relatively simple collaboration metrics to collect and analyze are (a) how much time is spent collaborating, (b) how often various modes of communication are used to collaborate (this may be visualized in a communications usage diagram), and (c) the frequency of collaboration. In C2, these simple metrics can be compared across mission phases. Another common collaboration metric used is "who talks to whom," which is visualized in a form called a social network diagram, consisting of nodes linked by communication lines, often of varying widths to convey frequency or strength (Freeman, Weil, and Hess 2006).

Aptima, Inc., is in the process of testing and refining an automated tool called the Instrument for Measuring and Advancing Group Environmental Situational awareness (IMAGES), which will expand the capability of traditional social network diagrams (Freeman, Weil, and Hess 2006). IMAGES "captures communications, analyzes them, and presents data concerning *the distribution of knowledge* across an organization." *Figure 6* shows a conceptual prototype of some of the functionality (Freeman, Weil, and Hess 2006). The

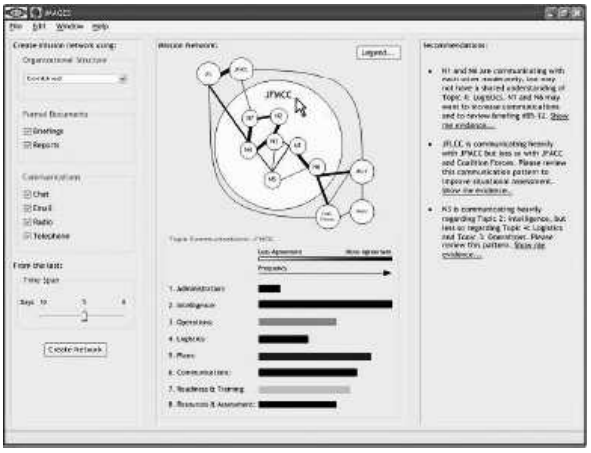


Figure 6. Conceptual prototype of Instrument for Measuring and Advancing Group Environmental Situational (IMAGES) awareness.

mission network in the upper center of *Figure 6* is an example of a social network diagram, and the linkages between nodes reflect communication frequency. An important additional attribute of collaboration included in the prototype is collaboration content (in the right panel).

Analyzing collaboration content can be very challenging for researchers because it usually involves manually categorizing collaboration communications by topic, which can be time consuming. Some information that can be obtained through this type of analysis is learning about types of information that require collaboration and collaboration topics. Some other methods to quantify collaboration content include collaboration communications (e.g., number of instances of paraphrasing others), number of critiques initiated concerning high-priority issues, and number of gaps, conflicts, and untested assumptions identified (Freeman and Serfaty 2002). The IMAGES tool claims to automatically collect collaboration content via Network Text Analysis, which detects the frequency and co-occurrence of terms (Freeman and Serfaty 2002). The prototype demonstrates a capability to visualize conflicts amongst topic areas.

Collaboration techniques that analyze content also have the ability to illuminate patterns of collaboration, knowledge, skills and abilities of nodes, and what information is considered relevant or important to various nodes. Another interesting aspect regarding C2 that may be revealed through content analysis is not *what* is being communicated, but *why* or the purpose of collaboration. For instance, it may be revealed that the purpose of collaboration is to reach team consensus on a decision. In this scenario, the trade-offs and strategies are communicated and compared, which would help system designers learn about C2 team

decision making. Another purpose of collaboration is to reduce uncertainty, and this may assist C2 system designers in automation opportunities. Yet another purpose of collaboration may be to achieve a shared understanding of SA or an understanding of the stress and workload of members of the team. All of these issues are critical in complex C2 and require further study, so that C2 decision making can be improved and action recommendations can be provided to C2 teams.

Applying these concepts to C2 research

Unfortunately, there is no silver bullet or simple answer to how to evaluate cognitive aspects of C2 (Barnes and Beevis 2003). Due to the difficulty of determining what is really occurring within someone's mind, many of the current measurement techniques are still immature, while many others that are more mature are still contested and debated within the cognitive systems engineering community. A second reason is that even the mature and uncontested measurement techniques are mostly geared towards individuals, and there is still a significant amount of research to be done in the area of team evaluation and collaboration. A third reason is the interrelationships between workload, SA, decision making, and collaboration. These relationships make it difficult to evaluate individual aspects of cognition and, in fact, make it undesirable in many cases. This increases the amount of work required to examine cognitive C2 aspects. Finally, many of these measurements cannot be easily collected and analyzed in an automated fashion, which makes them very time consuming, labor intensive, and expensive. As if these difficulties were not sufficient, there is also the obstacle of creating environments that can support the evaluation of C2 when access to actual operations is unavailable or too risky.

However, there are ways to mitigate the effects of these difficulties. First, as has been mentioned numerous times throughout this article, it is best to use a suite of complementary and overlapping measurement techniques to look at the cognitive aspects from various angles as well as validate other measures. Second, taking the time to carefully design not only the evaluation but also the analysis, with the use of pilot studies and the development of automatic data collection (and analysis), can help reduce the amount of labor needed to fully evaluate a C2 environment. There are many tools available for automatic recording of data from audio and video recording to screen recording and usability suites, which allow for correlations between eye, mouse, and screen movements.

This article has briefly summarized an extensive literature review into currently available cognitive C2 metrics (more details can be found in the Appendix). What is abundantly clear is that there is still a significant amount of research required to develop reliable, robust, objective, unobtrusive cognitive measurement techniques. ❑

MANDY L. NATTER is employed as an associate cognitive systems engineer at Johns *Hopkins University Applied Physics Laboratory (JHU/APL), where she integrates the consideration of human tasks, behavior, and cognitive abilities into the design and development of military C2 systems. Mrs. Natter received a bachelor of science degree in engineering psychology and a bachelor of science degree in biomedical engineering from Tufts University. She is currently working on her master's degree in systems engineering at the Johns Hopkins University Whiting School of Engineering. E-mail: Mandy.Natter@jhuapl.edu*

JENNIFER J. OCKERMAN is employed as a senior cognitive systems engineer at JHU/APL, where she applies cognitive engineering techniques to military C2 projects. Dr. Ockerman received a bachelor of science degree in industrial engineering and operations research from Virginia Tech and master of science and doctor of philosophy degrees in industrial and systems engineering, with a specialty in human integrated systems and a minor in cognitive science, from Georgia Tech. She has over 50 publications and is a reviewer for several conferences and journals. She is a member of the Human Factors and Ergonomics Society (HFES), the Association of Computing Machines (ACM), and the International Test and Evaluation Association (ITEA). E-mail: jennifer.ockerman@jhuapl.edu

LEIGH A. BAUMGART is a doctor of philosophy student at the University of Virginia (UVa) in the department of Systems and Information Engineering. Leigh received a bachelor of arts degree in physics at the State University of New York at Geneseo and a masters of science degree in systems engineering at UVa. Prior to entering the Ph.D. program at UVa, Leigh worked as a human-systems integration engineer at JHU/APL. She has conducted research in a variety of domains focused on understanding and modeling human performance and decision making. Leigh is a member of the HFE and the Institute of Electrical and Electronics Engineers (IEEE).

References

Ahlstrom, U., and F. J. Friedman-Berg. 2006. Using eye movement activity as a correlate of cognitive workload. *International Journal of Industrial Ergonomics* 36: 623–636.

Albers, M. J. 1996. Decision making: A missing facet of effective documentation. *ACM Special Interest Group for Design of Communication: Proceedings of the 14th Annual International Conference on Systems Docu-*

mentation: Marshaling New Technological Forces: Building a Corporate, Academic, and User-oriented Triangle, October 19–22, 2006, Research Triangle, NC, 57–65. New York, NY: ACM.

Allanson, J., and S. H. Fairclough. 2004. A research agenda for physiological computing. *Interacting with Computers* 16(5): 857–878.

Artman, H. 1999. Situation awareness and co-operation within and between hierarchical units in dynamic decision making. *Ergonomics* 2(11): 1404–1417.

Azuma, R., M. Daily, and C. Furmanski. 2006. A review of time critical decision making models and human cognitive processes. *Aerospace Conference 2006,* March 4–11, 2006, Big Sky MT. Piscataway, NJ: *Institute of Electrical and Electronics Engineers, Inc.*

Barnes, M., and D. Beevis. 2003. Chapter 8: Human system measurements and trade-offs in system design. In *Handbook of human systems integration*, ed. H. R. Booher, 233–263.

Bass, S. D., and R. O. Baldwin. 2007. A model for managing decision-making information in the GIG-enabled battlespace. *Air and Space Power Journal, 21* (2), 100–108.

Bausell, R. B., and Y. Li. 2002. *Power analysis for experimental research: A practical guide for the biological, medical, and social sciences.* New York, NY: Cambridge University Press.

Berka, C., D. J. Levendowski, C. K. Ramsey, G. Davis, M. N. Lumicao, K. Stanney, et al. 2005. Evaluation of an EEG-workload model in the aegis simulation environment. *Proceedings of SPIE, Volume 5797: Biomonitoring for Physiological and Cognitive Performance during Military Operations*, March 31, Orlando, FL, 90–99. Bellingham, WA: SPIE.

Berka, C., D. J. Levendowski, M. M. Cvetinovic, M. M. Petrovic, G. Davis, M. N. Lumicao, M. V. Popovic, V. I. Zivkovic, R. E. Olmstead. 2004. Real-time analysis of EEG indexes of alertness, cognition, and memory acquired with a wireless EEG headset. *International Journal of Human-Computer Interaction* 17(2): 151–170.

Bolia, R. S., and T. Nelson. 2007. Characterizing team performance in network-centric operations: Philosophical and methodological issues. *Aviation, Space, and Environmental Medicine* 78(5): B71–B76.

Bowman, E. K., and S. Kirin. 2006. Improving platoon leader situation awareness with unmanned sensor technology. *Proceedings of the Command and Control Research and Technology Symposium, June 20–22, San Diego, CA.* Alexandria, VA: CCRP.

Bryant, D. J., R. D. G. Webb, and C. McCann. 2003. Synthesizing two approaches to decision making in command and control. *Canadian Military Journal* 4(1): 29–34.

Cherri, C., E. Nodari, and A. Toffetti. 2004. *Information society technologies (IST) programme: Adaptive integrated driver-vehicle interface review of existing tools and methods* (No. IST-1-507674-IP). European Union: Information Society Technologies (IST) Programme. Available online at http://www.aide.eu.org/pdf/sp2_deliv_new/aide_d2_1_1.pdf, accessed April 20, 2010.

Collet, C., C. Petit, S. Champely, and A. Dittmar. 2003. Assessing workload through physiological measurements in bus drivers using an automated system during docking. *Human Factors* 45(4): 539–548.

Cooke, N. J., E. Salas, J. A. Cannon-Bowers, and R. J. Stout. 2000. Measuring team knowledge. *Human Factors* 42(1): 151–173.

Cummings, M. L., and S. Guerlain. 2004. Using a chat interface as an embedded secondary tasking tool. *Human Performance, Situation Awareness, and Automation: Current Trends and Research*, Vol. 1, ed. D. A. Vincenzi, M. Mouloua, and P. A. Hancock, 240–244.

De Waard, D. 1996. The measurement of drivers' mental workload. Doctoral thesis, University of Groningen, The Netherlands.

Department of Defense. 1999. *Department of defense handbook: Human engineering program process and procedures* (No. MIL-HDBK-46855A). Washington, D.C.: Department of Defense.

Dostal, B. C. 2007. Enhancing situational understanding through the employment of unmanned aerial vehicles. Army transformation taking shape…interim brigade combat team newsletter. No. 01-18. Electronic version available at http://www.globalsecurity.org/military/library/report/call/call_01-18_ch6.html. Accessed April 20, 2010.

Endsley, M. R. 1995. Toward a theory of situation awareness in dynamic systems. *Human Factors* 37(1): 32–64.

Endsley, M. R., R. Hoffman, D. Kaber, and E. Roth. 2007. Cognitive engineering and decision making: An overview and future course. *Journal of Cognitive Engineering and Decision Making* 1(1): 1–21.

Endsley, M. R., S. J. Selcon, T. D. Hardiman, and D. G. Croft. 1998. A comparative analysis of SAGAT and SART for evaluations of situation awareness.

Entin, E. B., and E. E. Entin. 2000. Assessing team situation awareness in simulated military missions. Proceedings of the 44th Annual Meeting of the Human Factors and Ergonomics Society, pp. 73–76.

Freeman, J., S. A. Weil, and K. P. Hess. 2006. Measuring, monitoring, and managing knowledge in command and control organizations. Proceedings of the 2002. Command and Control Research and

Technology Symposium, Monterey, CA. Woburn, MA: Aptima.

Freeman, J. T., and D. Serfaty. 2002. Team collaboration for command and control: A critical thinking model. *Proceedings of Command and Control Research and Technology Symposium, June 11–13, Monterey, CA.* Alexandria, VA: CCRP.

Gartska, J., and D. Alberts. 2004. *Network centric operations conceptual framework version 2.0. Report prepared for the Office of the Secretary of Defense, Office of Force Transformation*. Vienna, VA: Evidence Based Research, 2004.

Gawron, V. J. 2000. *Human performance measures handbook*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Gevins, A., and M. E. Smith. 2003. Neurophysiological measures of cognitive workload during human-computer interaction. *Theoretical Issues in Ergonomics Science* 4(1–2): 113–131.

Gorman, J. C., N. J. Cooke, and J. L. Winner. 2006. Measuring team situation awareness in decentralized command and control environments. *Ergonomics* 49(12–13): 1312–1325.

Haarmann, H. 2007. *Yerkes-Dodson Law*. http://www.extra.research.philips.com/probing_experience/presentations/haarmann.ppt#296,29,Yerkes-Dodson-Law.

Hiniker, P. J. 2005. Estimating situational awareness parameters for net centric warfare from experiments. In *10th International Command and Control Research and Technology Symposium: The Future of C2*, June 13–16, McLean, Virginia, available online at http://www.dodccrp.org/events/10th_ICCRTS/CD/papers/206.pdf. Accessed April 20, 2010.

Hodgson, P. 2007. *Quantitative and qualitative data – getting it straight*. http://www.blueprintusability.com/topics/articlequantqual.html (accessed March 6, 2008).

Kaempf, G. L., G. Klein, M. L. Thordsen, and S. Wolf. 1996. Decision making in complex naval command-and-control environments. *Human Factors* 38(2): 220–231.

Klein, G. A., R. Calderwood, and D. MacGregor. 1989. Critical decision method for eliciting knowledge. *IEEE Transactions on Systems, Man, and Cybernetics* 19(3): 462–472.

Klein, G., B. Moon, and R. R. Hoffman. 2006a. Making sense of sensemaking 1: Alternative perspectives. *IEEE Intelligent Systems* 21(4): 70–73.

Klein, G., B. Moon, and R. R. Hoffman. 2006b. Making sense of sensemaking 2: A macrocognitive model. *IEEE Intelligent Systems* 21(5): 88–92.

Lamar, M. 2006. Neuroscience and decision making. Available online from Triarchy Press at http://www.docstoc.com/docs/28616882/Neuroscience-and-Decisionmaking/download. Accessed April 10, 2010.

Lansman, M., and E. Hunt. 1982. Individual differences in secondary task performance. *Memory and Cognition* 10: 10–24.

Means, C. D., and K. J. Burns. 2005. Analyzing decisions and characterizing information in C2 systems. In *10th International Command and Control Research and Technology Symposium*, June 13–16, 2005, McLean, Virginia, Available online at http://www.dodccrp.org/events/10th_ICCRTS/CD/papers/183.pdf. Accessed April 20, 2010.

Militello, L. G., and R. J. B. Hutton. 1998. Applied cognitive task analysis (ACTA): A practitioner's toolkit for understanding cognitive task demands. *Ergonomics* 41(11): 1618–1641.

Miller, N. L., and L. G. Shattuck. 2004. A process model of situated cognition in military command and control. Proceedings of the 2004 Command and Control Research and Technology Symposium, June 2004, San Diego, CA. Available online at http://faculty.nps.edu/nlmiller/docs. Accessed April 20, 2010.

O'Donnell, R. D., and F. T. Eggemeier. 1986. Workload assessment methodology. In *Handbook of perception and human performance,* Vol. 2, eds. K. Boff, L. Kaufman, and J. Thomas, 42/1–42/9. New York: Wiley.

Poythress, M., C. Russell, S. Siegel, P. D. Tremoulet, P. Craven, C. Berka, et al. 2006. Correlation between expected workload and EEG indices of cognitive workload and task engagement. *Second Annual AugCog International Conference*, San Francisco, California, 32–44.

Prince, C., E. Ellis, M. T. Brannic, and E. Salas. 2007. Measurement of team situation awareness in low experience level aviators. *The International Journal of Aviation Psychology* 17(1): 41–57.

Raby, M., and C. D. Wickens. 1994. Strategic workload management and decision biases in aviation. *The International Journal of Aviation Psychology* 4(3): 211–240.

Rowe, D. W., J. Silbert, and D. Irwin. 1998. Heart rate variability: Indicator of user state as an aid to human-computer interaction. Paper presented at CHI '98 in CHI proceedings, April 18–23, 1998, 480–487. Los Angeles, CA. New York, NY: ACM.

Rubio, S., E. Diaz, J. Martin, and J. M. Puente. 2004. Evaluation of subjective mental workload: A comparison of SWAT, NASA-TLX, and workload profile methods. *Applied Psychology: An International Review* 53(1): 61–86.

Salmon, P., N. Stanton, G. Walker, and D. Green. 2006. Situation awareness measurement: A review of

applicability for C4i environments. *Applied Ergonomics* 37: 225–238.

Sniezek, J. 1992. Groups under uncertainty: An examination of confidence in group decision making. *Organizational Behavior and Human Decision Making Processes*, (52): 124–155.

Spick, M. 1988. *The ace factor: Air combat and the role of situational awareness*. Annapolis, MD: Naval Institute Press.

Van Orden, K. F. 2000. *Real-time workload assessment and management strategies for command and control watchstations: Preliminary findings*. http://www.dtic.mil/matris/sbir/sbir011/Navy89b.doc (accessed March 16, 2006).

Walker, G. H., H. Gibson, N. A. Stanton, C. Baber, P. M. Salmon, and D. Green. 2006. Event analysis of systemic teamwork (EAST): A novel integration of ergonomics methods to analyze C4i activity. *Ergonomics* 49(12–13): 1345–1369.

Wetherell, A. 1981. The efficacy of some auditory-vocal subsidiary tasks as measures of mental load on male and female drivers. *Ergonomics* 24: 197–214.

Wickens, C. D., and J. G. Hollands. 2000. *Engineering psychology and human performance*. Upper Saddle River, NJ: Prentice-Hall.

Zhang, Y., and A. Luximon. 2005. Subjective mental workload measures. *International Journal of Ergonomics and Human Factors* 27(3): 199–206.

APPENDIX

Workload measurement techniques

Subjective

NASA TLX, the most commonly used subjective workload scale, contains six dimensions: (a) Mental demand, which refers to perceptual and cognitive activity, (b) Physical demand, which refers to physical activity, (c) Temporal demand, which refers to time pressure, (d) Performance, which is related to personal goal accomplishment, (e) Effort, which refers to energy expenditure in accomplishing the required level of performance, and (f) Frustration, which is related to feelings of irritation, stress, etc. Subjects complete a pair-wise comparison procedure to weight the dimensions before the task. After completing the task, subjects rate the six dimensions using a 0–100 scale (Hart and Staveland 1988).

The SWAT contains three dimensions: (a) Time (T) load, which reflects the amount of spare time available in planning, executing, and monitoring a task; (b) Mental effort (E) load, which assesses how much conscious mental effort and planning are required to perform a task; and (c) Psychological stress (S) load, which measures the amounts of risk, confusion, frustration, and anxiety associated with task performance. The three SWAT dimensions (T, E, and S) are at three discrete levels (i.e., 1, 2, and 3). Subjects rank the three dimensions and three levels by perceived importance in a 3^3 or 27-card sorting exercise before conducting the task. The card sort results in seven weighting schemes: TES, ETS, SET, TSE, EST, STE, with equal emphasis on T, E, and S. After completing the task, subjects rate the task on the three dimensions (Reid and Nygren 1988).

Wierwille and Casali modified the wording of the validated physically focused Cooper-Harper Rating Scale such that it would be appropriate for assessing cognitive functions such as "perception, monitoring, evaluation, communications, and problem solving" (Wierwille and Casali 1983). The MCH scale maintains a decision tree architecture where participants respond to yes or no questions that lead to options for rating. The rating scale is 1–10, where 1 is very easy and 10 is impossible (O'Donnell and Eggemeier 1986).

Physiological

Eye measures

A variety of studies have shown that various aspects of eye behavior correlate with cognitive workload. One of the most sensitive eye physiological measures is pupil diameter. Pupil diameter increases (dilates) as cognitive workload increases (Ahlstrom and Friedman-Berg 2006). Pupil diameter changes can be dynamic, for instance during comprehension of individual sentences, or sustained during recall of digit span (Van Orden 2000). Although the average pupil diameter changes by as much as 0.6 mm when recalling seven digits, many confounds such as ambient lighting, stimulus characteristics, and even emotional effects can cause pupillary responses that are greater than those from workload alone (O'Donnell and Eggemeier 1986; Van Orden 2000). Also, accurate measurement techniques required may impose constraints on experimentation by requiring the subject to stay in one location or wear a measuring device on his/her head. Pupil diameter measurements are therefore difficult to use in applied settings where the environment and other external factors are not controlled. Finally, research suggests that pupil diameter measurements may be highly responsive to cognitive workload changes, yet not diagnostic because there is little ability to identify the resource (e.g., visual, auditory, etc.) utilized in the task (O'Donnell and Eggemeier 1986).

Generally, blink rate and blink duration decrease as workload increases (Ahlstrom and Friedman-Berg 2006; Van Orden 2000; Veltman and Gaillard 1998).

Boehm-Davis, Gray, and Schoelles (2000) suggest that eye blinks are suppressed when individuals are engaged in cognitive processing; however, eye blinks show great variability (Boehm-Davis, Gray, and Schoelles 2000; O'Donnell and Eggemeier 1986). Blink duration is also somewhat unreliable to gauge cognitive workload because other factors like visual workload confound cognitive workload. A visual tracking task with minimal cognitive load can cause lower blink durations than during a more cognitively challenging flight simulation task (Van Orden 2000). Therefore, eye blinks and blink duration should be considered global indicators of long-term effects versus specified diagnostic techniques (O'Donnell and Eggemeier 1986).

Another potential eye related measurement of workload is number and duration of saccades. The number of saccades, which are a series of small, quick, jerky movements of the eyes when changing focus from one point to another in the visual field, increase as workload increases. Saccade duration, which typically lasts for 20 to 35 milliseconds, decreases as workload increases (De Waard 1996; Poole 2004; Wickens, Mavor, and McGee 1997). While saccades may provide clues about the cognitive strategy employed, studies reflect that prior to voluntary eye movement, attention shifts to the location of interest; therefore saccades may be measures of attention versus cognitive workload (Tsai et al. 2007).

After each saccade, the eyes stay still and encode information in movements called "fixations." Fixation frequency and fixation duration or dwell time both increase as cognitive workload increases, but is task dependent (De Waard 1996; Poole 2004; Van Orden 2000). For example, during a challenging flight simulation, fixation duration correlated with the number of flight rule errors, reflecting a correlation with cognitive workload; however, in a challenging visual search task, search fixation frequency increased and fixation duration did not change (Van Orden 2000). Fixation is particularly sensitive to visual workload, making it more diagnostic than other techniques; however, fixation does not necessarily imply cognition (De Waard 1996).

Scan paths, recurring patterns of saccades and fixations, become less of a pattern between display elements as workload increases (O'Donnell and Eggemeier 1986; Poole 2004). Also, dwell times in each position lengthen and fewer display elements are used. Scanning is typically a global indicator of workload, but scanning may be a diagnostic index of the source of workload within a multi-element display environment if (a) critical information must be gathered from multiple locations, (b) relative importance of data obtained from each location is different, and (c) the subject can adjust or change the imposed load by a change in strategy (O'Donnell and Eggemeier 1986).

In summary, blink rate, blink duration, and saccade duration all decrease, while pupil diameter, the number of saccades, and the frequency of long fixations all increase with increased workload.

The eye is readily accessible to observation and provides rich data that can be assessed, but study results differ in eye physiological measures that correlate highest with cognitive workload. For example, in a mock anti-air warfare task, blink frequency, fixation frequency, and pupil diameter were the most predictive variables correlating eye activity to target density (Van Orden 2000); and in an air traffic controller study, managing traffic during adverse weather conditions, decreased saccade distance, blink duration, and pupil diameter correlated closest to cognitive workload (Ahlstrom and Friedman-Berg 2006). Because of the variability amongst eye physiological measures, it is recommended that multiple techniques be combined (Van Orden et al. 2001). Several confounds, including visual workload, impact results, so eye measures should be used as global indicators of cognitive workload. Eye physiological measures provide more sensitive results in controlled environments.

Cardiac measures

The main cardiac measures studied for sensitivity to cognitive workload include the electrocardiogram, blood pressure, and blood volume. Of these three, measures of electrocardiographic activity show the most promise (Rowe, Silbert, and Irwin 1998). The electrocardiograph produces a graphic called an electrocardiogram, abbreviated ECG or EKG, from the German *elektrokardiogramm*, which records the electrical activity of the heart over time. Surface electrodes are placed on the skin of a subject to identify pulse beats, which are recognizable by a pattern called the QRS complex *(Figure A1)* (O'Donnell and Eggemeier 1986).

Although absolute heart rate has been used as a measure of overall workload, spectral analysis of heart rate variability (HRV) or sinus arrhythmia reflects some correlation to cognitive workload (De Waard 1996; O'Donnell and Eggemeier 1986; Tattersall and Hocky 1995; Veltman and Gaillard 1998). As the name implies, HRV is the variability in heart rate or the variability between R-R intervals *(Figure A1)*. Of the over 30 techniques available for determining HRV (e.g., Fourier transform, autoregressive modeling, time-varying analysis, broadband spectral analysis), most cognitive-loading HRV measures emphasize frequency. Frequency HRV techniques measure the

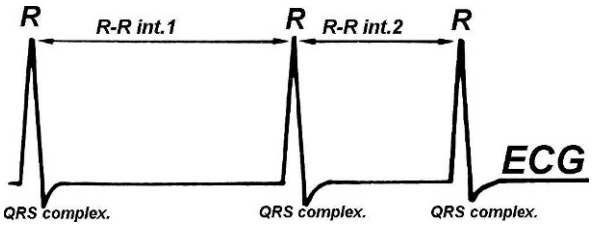


Figure A1. Heart rate and heart rate variability (Dantest Medical System). Reprinted with permission from Dantest Medical System.

amount of variation in different frequency bands. There are three major frequency bands: (a) very low-frequency band (0.0033–0.04 Hz), associated with temperature regulation and physical activity; (b) low-frequency band (0.04–0.15 Hz), associated with short-term regulation of arterial pressure; and (c) high-frequency band (0.15–0.40 Hz), reflecting the influence of respiration. Several studies have suggested that the low-frequency band, and specifically what is called the 0.10 Hz component, indicates cognitive workload (De Waard 1996; Nickel and Nachreiner 2003; O'Donnell and Eggemeier 1986). The 0.10-Hz component reflects short-term changes in blood pressure. A peak of the 0.10-Hz component reflects decreased cognitive workload, and a flattening of the 0.10-Hz component reflects conditions of greater mental workload (Rowe, Silbert, and Irwin 1998).

Nickel and Nachreiner (2003) assessed the diagnosticity (i.e., the ability to differentiate amongst different types of tasks) and sensitivity (i.e., the ability to detect levels of difficulty) of the 0.1-Hz component of HRV for cognitive workload using 14 cognitive tasks (e.g., reaction time, mathematical processing, memory-search, grammatical reasoning task) from an environmental stressors standardized test in a laboratory context. Only one type of task could be discriminated as different from the other types of tasks—that task reflected a cognitive-loading score that matched the cognitive loading expected at rest; however, these results directly conflicted with performance (i.e., performance errors were made) and perceived difficulty (i.e., participants reported mental workload). In terms of sensitivity, the results echoed several other studies that HRV can discern between work and rest, but not to gradations in between (Rowe, Silbert, and Irwin 1998). Because the experimenters noted differences in the 0.10-Hz component when time pressure was involved, they propose that HRV be used as an indicator for emotional strain or time pressure versus cognitive workload (Nickel and Nachreiner 2003).

Research by Hannula et al. (2007) supports the use of HRV as a stress indicator. They applied an artificial neural network analysis to evaluate the relationship between cognitive workload that raises the psychophysiological stress and HRV data in fighter pilots. The Pearson's coefficients between the ECG data and the cognitive workload that increases psychophysiological stress as evaluated by an experienced flight instructor were between 0.66 and 0.69 (Hannula et al. 2007).

There are several caveats to using HRV as a cognitive workload measure. Possibly the most significant disadvantage to HRV is that it has not been validated as a sensitive cognitive workload indicator (O'Donnell and Eggemeier 1986). Also, heart rate and, likewise to a smaller extent, HRV are confounded by psychological processes such as high sense of responsibility or fear, physical effort and speech, and environmental factors such as high G-forces (De Waard 1996; O'Donnell and Eggemeier 1986; Tattersall and Hocky 1995). As indicated in the studies discussed above, HRV is influenced by stress and time constraints. Physical effort will impact HRV results unless it is kept to a minimum and constant across conditions (De Waard 1996). Speech can also confound HRV results if verbalization is longer than 10 seconds and relatively frequent (more than one to five times per minute) (De Waard 1996). Another factor that affects HR measures, and to a lesser degree cognitive workload, is age. If HRV is the primary workload measure, it may be necessary to restrict elderly subjects from participation because HRV may decrease with increasing age. Finally, a last caveat to consider is that operators typically need to act as their own control because of the idiosyncrasies in the measure (De Waard 1996).

Despite the caveats, heart rate measurement is arguably the simplest physiological index to measure, and it has been employed extensively. The ECG signal requires minimal amplifying (approximately 10 to 20 times less than continuous EEG), and if measurements are limited to R-wave detection and registration, then electrode placement is not critical (De Waard 1996). Cardiac techniques are also the most popular physiological techniques used in the last 40 years (Rowe et al. 1998). They are relatively noninvasive and unobtrusive (O'Donnell and Eggemeier 1986). Also, with continuously recorded cardiac measures, research has shown that HRV can indicate within seconds the change from work to rest (Rowe et al. 1998).

Neural measures

Electro-encephalography, the measurement of electrical activity in the brain, is the most common neurophysiological technique used as an indicator of cognitive workload (Berka et al. 2004). It typically involves a noninvasive procedure of placing electrodes on the surface of the head to detect activity through the skull and scalp. In rare instances,

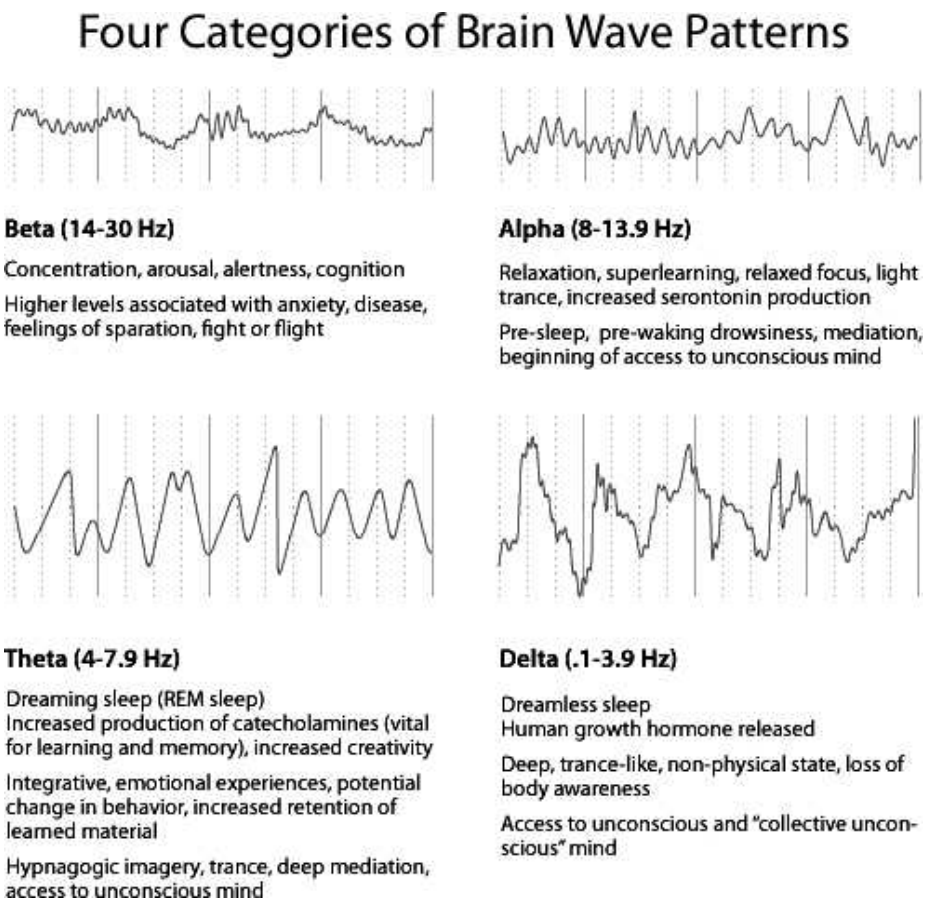


Figure A2. Four categories of brain wave patterns (Harris, 2008). Reprinted with permission from Centerpointe Research Institute, http://www.centerpointe.com.

electrodes are placed subdurally or in the cerebral cortex. The traces of activity that result are called an electroencephalogram (EEG), which represents the electrical signals or postsynaptic potentials from a large number of neurons. In clinical use, the EEG is considered a "gross correlate of brain activity" because instead of measuring electrical currents from individual neurons, it reflects relative voltage differences amongst various brain areas.

Frequency analyses performed on EEG signals are also called epoch analyses or background EEG analyses and usually result in four ranges or wave patterns *(Figure A2)* (De Waard 1996). Although these wave patterns or bands are traditionally used to provide information about the health and function of the brain, some are very responsive to variations in alertness and attention. Specifically, several studies confirm that alpha and especially theta bands are sensitive to aspects of short-term or working memory (Gevins and Smith 2003). When working memory is in use, research reflects decreases in the upper alpha band and increases in the theta band, which become more exaggerated when load increases. This finding suggests that these bands may be indicators of cognitive loading. One study reflected decreased alpha and increased theta activity during dual-task performance when compared with single-task performance. However, individual differences may be significant; for example, a small number of individuals do not generate alpha waves at all (De Waard 1996).

In addition to background EEG, event-related potentials (ERPs) are a neural measurement technique used as a cognitive workload indicator (Berka et al. 2004). Unlike evoked potentials, which are the result of physical stimuli, ERPs may be caused by processes such as memory, attention, expectation, or changes in mental state (Basar-Eroglu and Demiralp 2001). ERPs are any response to an internal or external stimulus but are often considered to be the direct result of cognitive activity. ERPs can be reliably measured from voltage deflections in EEG. For instance approximately 300 milliseconds following unpredictable stimuli, there are positive deflections in voltage called the P300 or more simply, P3. The P3 peaks around 300 milliseconds for very simple decisions, and the amplitude increases with unexpected, task-relevant stimuli. After

Figure A3. Parietal sites for electrodes (Reprinted with the permission of the Department of Kinesiology, University of Waterloo, 2008).

300 milliseconds, the latency of P3 increases with task difficulty and relates to cognitive workload (De Waard 1996). During primary tasks only, the P3 amplitude increases with task complexity; however, when secondary tasks are added, the P3 decreases with primary task complexity (De Waard 1996). An external factor is age because the P3 latency increases between 1 and 2 milliseconds each year of the adult lifespan (Colman 2001). P3s are somewhat difficult to detect, and therefore it is often necessary to present an individual stimulus dozens or hundreds of times and average the results together to cancel out the noise and present the stimulus response clearly. The signal-to-noise ratio can be improved by placing electrodes on the participant's head above the parietal lobe of the brain *(Figure A3)*.

The B-Alert System *(Figure A4)*, a device developed by Advanced Brain Monitoring (ABM), is a commercially available product that claims to have a cognitive workload measurement capability. Originally created to provide early detection of drowsiness, B-Alert advertises real-time workload calculations via a wireless, lightweight EEG headset that can analyze six channels of EEG. The signal processing, including amplification, digitization, and radio frequency transmission of the signals, is built into a portable unit worn with the headset (Berka et al. 2004).

Berka, the president of the B-Alert system, et al. (2005) conducted an experiment to determine if EEG could be used as an indicator of cognitive workload with

Figure A4. B-Alert System (Berka et al. 2005). Reprinted with permission from Advanced Brain Monitoring, Carlsbad, CA.

the B-Alert system. The context of the experiment was a simulated Aegis C2 environment, a combat system with advanced, automatic detect-and-track, multi-function phased array radar. Five participants were trained as identification supervisors (IDSs). IDSs have one of the highest workloads of the 32 operators in a typical Aegis Combat Information Center (CIC). The IDSs were responsible for monitoring multiple data sources, detecting required actions, responding appropriately, and maintaining system status within predefined desirable parameters. Workload measures were calculated in real time for each second of EEG by the B-Alert system. The B-Alert system used signal analysis techniques to decontaminate eye blinks, EMG data, amplifier saturation, and excursions related to movements. Post-hoc analysis identified that the cognitive tasks from most difficult to least difficult were as follows: track selection-identification, alert-responses, hooking-tracks, and queries. High/extreme workload was detected approximately 100 percent of the time during high cognitive-load tasks such as selection-identification and alert-responses, 77 percent of the time for hooking-tracks, and 70 percent of the time for queries. The low false alarm rate of <5 percent reflected that the workload gauge was not overly sensitive (Berka et al. 2005).

Although these results seem encouraging, currently no other known authors have reported such significant success with the B-Alert system and with the signal analysis techniques. A usability evaluation of a prototype of the Tactical Tomahawk Weapons Control System with the B-Alert system reflected that EEG correlated with expected workload when EEG measures were averaged across *all* nine participants. Individual participant EEG measures did not correlate strongly with cognitive workload, reflecting that the EEG was not a very sensitive indicator of cognitive workload. It was also found that EEG did not correlate with expert subjective ratings, but this may be due to insufficient data (Poythress et al. 2006).

The main benefits of EEG measurements are that they are continuous, cognitively unobtrusive, sensitive, moderately diagnostic, and relatively inexpensive. EEG provides a relatively continuous stream of data that can be identified and quantified on a second-by-second basis. Also, the decrease in size of electrodes and the development of wireless systems like B-Alert have led to minimal interference in primary task performance, unlike other functional neuro-imaging techniques that require the subject to be completely immobile and are much more massive (Gevins and Smith 2003). EEG is sensitive to changes in task complexity and task difficulty (Berka et al. 2004). The sensitivity of EEG is high and somewhat diagnostic as it can reflect subtle changes in attention, alertness, and cognitive workload (Berka et al. 2005; Gevins and Smith 2003). Also, the technology required for EEG is fairly low in cost (Gevins and Smith 2003).

However, the cost of the sensitivity is a poor signal-to-noise ratio. EEG can be easily contaminated by electrical signals of the eyes, muscles, heart, and external sources (De Waard 1996). Also the considerable intrasubject and between subject variability makes it difficult to find consistent patterns of physiological change (De Waard 1996). Signal analysis techniques prevent some of the signal-to-noise ratio problems; however, they must be rigorous and are time consuming. Required software and hardware is costly, and wired electrode systems may physically constrain participants. Advances have been made, but there is still a great deal of work to be conducted to accurately apply EEG measurements to cognitive workload assessment (Poythress et al. 2006).

Skin measures

The most common skin measure used as an indicator of cognitive workload is based on the moment-to-moment sweat gland and related activities of the autonomic nervous system. When the body perspires, secreted positive and negative ions change the electrical properties of the skin. This phenomenon, originally called the psychogalvanic reflex (PGR), is currently known as the galvanic skin response (GSR), skin conductance response (SCR), or electrodermal response (EDR). In the remainder of this section, it will be referred to as EDR.

EDR can be performed passively or actively. In passive EDR, weak currents generated by the body itself are measured. Active EDR involves applying a small constant current through two electrodes on the skin such that a voltage develops across the electrodes. The skin acts as a variable resistor, and the effective resistance or conductance of the skin can be calculated by applying Ohm's law (Allanson and Fairclough 2004). The resulting EDR is expressed in terms of conduction or resistance, which are inversely related (De Waard 1996).

EDR is frequently conducted actively when measuring cognitive workload. Electrodes are placed where sweat glands are most abundant such as the fingers, palm, forearm, and soles of the feet (Allanson and Fairclough 2004; De Waard 1996). In order to determine EDR, a baseline measure, which is typically an average of tonic EDR, is required. Tonic EDR is produced from everyday activities. It varies with psychological arousal and rises when the subject awakens and engages in cognitive effort, especially stress. In contrast, phasic EDR is the result of an external stimulus (De Waard 1996). One to 2 seconds following an external stimulus, the skin conductance increases, and it peaks within 5 seconds (De Waard 1996; McGraw-Hill Companies 2007). The amplitude is dependent on the subjective impact of the stimulus, particularly with regard to its novelty and meaning. The wavelike increase in the curve is a fairly reliable indicator of the duration of information-processing and cognitive workload (Collet et al. 2003).

The advantages of EDR are that it is low-cost and relatively simple to implement. The necessary hardware consists of a low-cost device called a galvanometer and inexpensive electrodes. An amplifier can also be added to increase the signal-to-noise ratio. Often only two electrodes are needed, which can be affixed to the skin with adhesive tape.

The main disadvantages to EDR are the latency in response and the insensitivity to cognitive workload when compared with other physiological measures such as HRV and EEG. EDR is confounded by the many factors that impact the sympathetic nervous system and the sweat glands, which include temperature, respiration, humidity, age, gender, time of day, season, arousal, and emotions. Because EDR is related to activities of the sympathetic nervous system, all behavior (emotional and physical) can potentially change EDR (De Waard 1996).

Situation awareness techniques

SA_T calculation

$$SA_T = \frac{\text{Information acquired}}{\text{Information required}}$$

In the study by Riese, Kirin, and Peters (2004), SA_T was based on three elements:

- Location (LOC)—Where is he? Knowing the position of entities or elements to a certain level of accuracy.

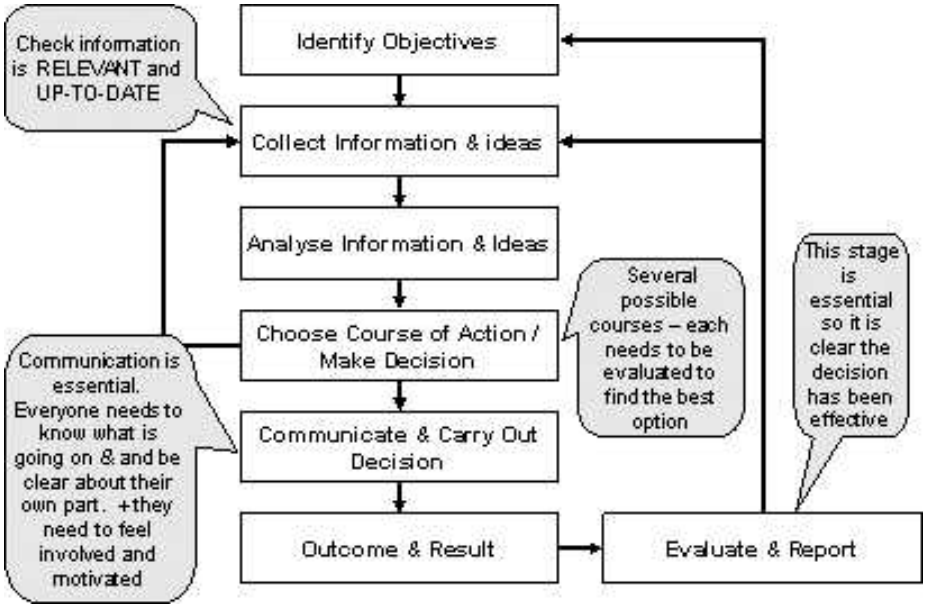


Figure A5. General components of decision making (Finntrack, 2008). Reprinted with permission. Diagram is available online at www.tutor2u.net (2007).

- Acquisition (ACQ)—What is he? Knowing the entity or element type to a particular level of acquisition (detect, classify, identify).
- State (STA)—How healthy is he? Knowing the mission-capable status of entities or elements.

The formula they used for the SA_T is shown below:

$$\text{Instantaneous } SA_T \text{ Score} = \frac{\sum_{i=1}^{n} c_i(W_L D_{L_i} \text{Loc}_i + W_A D_{Ai} \text{Acq}_i + W_S \text{Sta}_i)}{\sum_{i=1}^{n} c_i \times (W_L + W_A + W_S)}$$

The SA_T result is between a 0 and 1, where 0 reflects a complete lack of useful SA information and 1 represents full knowledge of location, type, and state of enemy entities within the defined battlespace.

Team SA

Salmon et al. (2006) evaluated seventeen SA measures against a set of human factors criteria for use in command, control, communication, computers and intelligence (C4i) environments. They concluded that current SA measurements are inadequate for use in assessing team SA and recommended a multiple-measure approach. They identified three requirements for C4i SA measurement: ability to simultaneously measure SA at different locations, ability to measure SA in real time, and ability to measure both individual and team/shared SA (Salmon et al. 2006).

Measurement techniques for team SA have been researched and continue to be a subject of study. Some researchers have aggregated scores from validated SA measurement methods like SAGAT to obtain team SA scores; however, Gorman et al. and others argue that combining individual SA results is not equivalent to team SA because it is missing interaction and critical activities such as coordination, collaboration, and information exchange (Entin and Entin 2000; Gorman, Cooke, and Winner 2006). Gorman et al. developed a measure of SA called the Coordinated Awareness of Situations by Teams (CAST) in which team SA is tested by the group's reaction to unexpected events or "roadblocks" (Gorman, Cooke, and Winner 2006). They suggest that these "roadblock" situations are appropriate for determining shared SA because they require the team to coordinate and collaborate in order to circumvent the obstacle. The five steps to CAST involve (a) identifying roadblocks, (b) documenting primary perceptions, (c) documenting secondary perceptions, (d) documenting coordinated perceptions, and (e) documenting coordinated actions (Gorman, Cooke, and Winner 2006).

Another team SA technique, Situational Awareness Linked Instances Adapted to Novel Tasks (SALIANT), is similar to CAST. It involves how the team responds to problems but requires an a priori script and structure. SALIANT includes five phases: (a) identify team SA behaviors (e.g., demonstrated awareness of surrounding environment, recognized problems, anticipated a need for action, demonstrated knowledge of tasks, demonstrated awareness of information), (b) develop scenarios, (c) define acceptable responses, (d) write a script, and (e) create a structured form with columns for scenarios and responses (Gawron 2000).

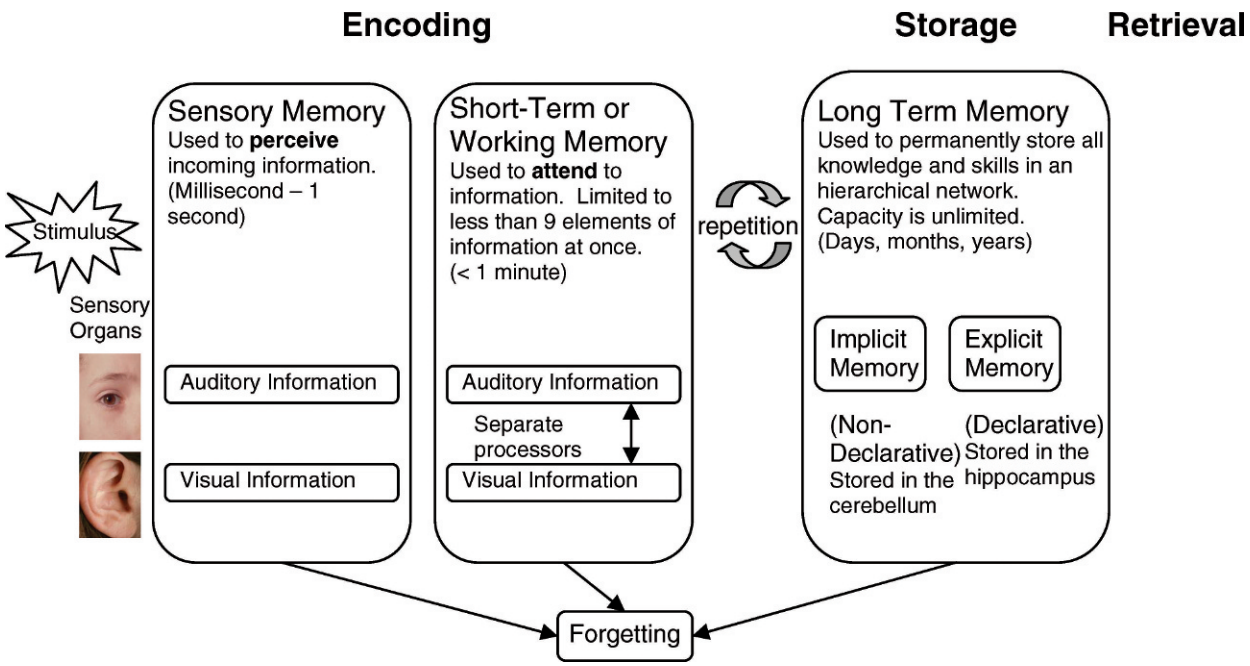


Figure A6. Cognitive aspects of decision making. (Composed with permission from http://www.scientificjournals.org/journals2007/articles/graphics/1038.gif, from http://thebrain.mcgill.ca/flash/i/i_07/i_07_p/i_07_p_tra/i_07_p_tra_2a%20copy.jpg, and from Professor John Sweller, University of New South Wales, Australia. Refer to http://education.arts.unsw.edu.au/staff/sweller/clt/images/CLT_NET_Aug_97_HTML6.gif).

Related to team SA is the concept of distributed situation awareness (DSA) (Stanton et al. 2006). Unlike definitions of team SA that examine shared SA, DSA focuses on the system as a whole and consists of the complementary SA of agents (humans and technology) within the system. Although DSA captures the content of system SA, it does not have a technique for assessing the quality of that SA, other than task performance and SME judgment.

Decision making

In general, the components of decision making, for both system and humans, include identifying objectives, collecting information and ideas, analyzing the information and ideas, choosing a course of action/making a decision, and communicating and carrying out the decision. An outcome results from the action taken and, optimally, there is an evaluation or report reflecting the effectiveness of the decision *(Figure A5)*.

Decision making depends in part on the system providing the "right" information at the "right" time to the "right" person. In order to achieve this goal, an Internet-like system called the Global Information Grid (GIG) is in the conceptual design phase for C2. The Department of Defense defines the GIG as "a globally interconnected, end-to-end set of information capabilities, associated processes and personnel for collecting, processing, storing, disseminating, and managing information on demand to warfighters, policy makers, and support personnel" (Department of Defense, 2002, pg. 8). The GIG will be a repository for military information with the goal of providing information superiority over adversaries. A significant risk of the GIG is inundating warfighters with information and providing noisy data that distracts from mission critical data (Bass and Baldwin 2007). Since net-centric, GIG-type environments are slated for the future, many articles have been written about the GIG and its potential implications. Bass and Baldwin (2007) proposed some rules to direct GIG information flow because information presented at the wrong time, level of detail, and/or lacking proper analysis and interpretation could have devastating effects. Their basic solution is to limit data access to those with authorization and to limit data automatically sent to all users (Bass and Baldwin 2007). The goal is to present a manageable amount of information to warfighters that not only pertains to them but also provides context.

Cognitive aspects of decision making begin when the decision maker attends to information in his/her environment. The brain selects data for cognitive processing and encodes data for storage and later retrieval to support decision making *(Figure A6)*. Encoding, translating stimuli to internal (mental) representations, is the longest component of reasoning and decision making, taking approximately 45 percent of the overall time.

Encoding words takes longer and requires more workload than encoding schematic pictures, so reduc-

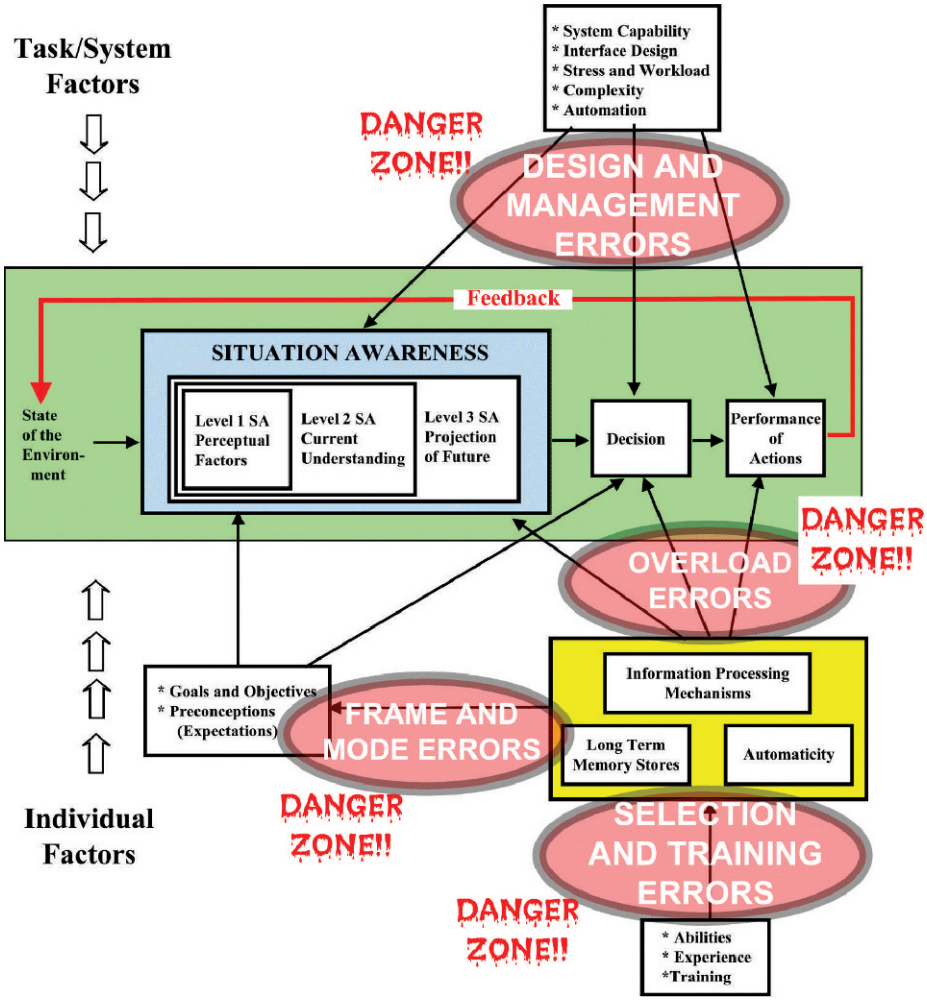


Figure A7. Relationship of situation awareness and decision making aspects (Smith 2007).

ing the amount of text will lead to faster decision making. Time-critical decision making displays should aim to decrease encoding time, for example, by increasing the intuitiveness of symbology and tasking (Azuma, Daily, and Furmanski 2006).

The encoding, storage, and retrieval process reflects the close connections between cognitive workload, SA, and decision making. Part of encoding involves perceiving and comprehending, which are critical components of workload and SA. Cognitive workload is related to the storage process because if encoding takes longer or if the information is encoded incorrectly, it may be because the information is more difficult and requires more cognitive resources to process. SA is also an integral part of the storage process because it directly involves perception and comprehension. Typically, cognitive workload and SA are inversely proportionate: a decrease in workload often results in an increase in SA and vice versa. Hence, if the storage process is improved, there is a higher likelihood that cognitive workload will decrease, SA will increase, and decision making will improve as long as the information is relevant to the decision maker's decision.

The projection aspect of SA provides the foundation for decision making. Humans are adept at detecting and remembering patterns. When they notice trends, they begin to extrapolate from the trends to anticipate what will happen next. Series of trends over time result in the development of mental models. This process is called sensemaking. The Command and Control Research Program's Sensemaking Symposium Final Report defines sensemaking as, "the process of creating situation awareness in situations of uncertainty" (Leedom 2001). Using a data/frame theory as an analogy, data are associated to form a hypothesis called a frame. Preserving and elaborating the frame are like Piaget's assimilation, and reframing is like Piaget's accommo-

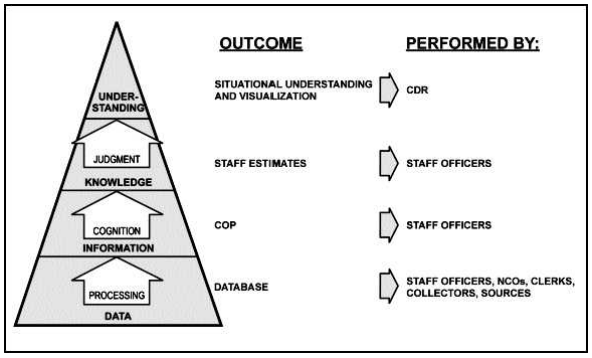


Figure A8. Skill-, rule-, and knowledge-based classification of C2 (Department of the Army, April 2003).

dation. Once anchors are established to generate a useful frame, the frame can be evaluated, reframed, elaborated, or compared with alternative frames (Klein, Moon, and Hoffman 2006b). Sensemaking is similar to Endsley's model of SA except instead of a knowledge state that's achieved, sensemaking is the process of getting to the outcome, the strategies, and the obstacles. "Sensemaking is a motivated, continuous effort to understand connections (which can be among people, places, and events) in order to anticipate their trajectories, and act effectively" (Klein, Moon, and Hoffman 2006a, 71). Decision makers apply their mental models to the tasks they conduct. *Figure A7* provides a model of the relationships between SA and decision making (DM) aspects. In addition, it shows the influence of both the technical/system environment as well as the individual. Finally, it highlights the many areas that can produce errors.

There are several decision making modeling techniques to emulate human decision making. For instance, speeding up decision time provides a tactical advantage in line with John Boyd's high-level sequential model of decision making (consisting of Observation, Orientation, Decision, and Action) called the OODA loop, in which the basic C2 strategy is to execute your OODA loop faster than your opponent (Azuma, Daily, and Furmanski 2006). Several cognitive models in addition to John Boyd's OODA loop have been developed; for example, Adaptive Control of Thought-Rational (ACT-R), Executive Process/Interactive Control (EPIC), Goals, Operators, Methods, and Selection Rules Language (GOMSL), State Operator and Result (SOAR), Improved Performance Research Integration Tool (IMPRINT), and Bayesian networks (Endsley et al. 2007).

Cognitive models tend to generalize decision making; and in uncertain and less familiar situations such as complex C2 environments, individual differences, particularly experience, become more apparent (Lamar 2006). Experience level and prior exposure to similar tasks are discriminating factors between individuals. With novel incoming stimuli, individuals are biased by their past experience, values, and repercussions of the decision (i.e., the cost-benefit analysis) when choosing a course of action (Lamar 2006). Rasmussen described experience using a skill-, rule-, and knowledge-based (SRK) classification (Rasmussen, 1983). Individuals at a skill level have a great deal of experience with the task and the environment and behave virtually automatically. Rule-based behavior involves less experience than the skill level and follows an "if, then" process. Knowledge-based behavior is unskilled and requires more effort because less memory can be leveraged. *Figure A8* shows an application of SRK-based classification to C2.

Another individual difference that affects decision making is physiology. Some individuals are innately better and quicker at decision making than others, but aspects of decision making can be learned. For instance, taxi drivers who memorized a map had a different degree of hippocampal volume, a space involved with learning and memory, when compared with those who had not (Lamar 2006). Expertise leads to differences in function and structure of brain regions required for decision making. Thus, decision making is influenced by both nature and nurture. Another individual difference is personality, but research is ambiguous as to the effect of personality on decision making (Koterba 2004). Another significant individual difference in decision making is confidence (Sniezek 1992). Confidence can have critical consequences as disasters frequently result when decision makers are confidently wrong. Conversely, if decision makers have reached an accurate decision and don't have the confidence in their decision to follow through, then the potential benefits will be lost (Sniezek 1992).

Most individual difference effects are superseded by trained decision making strategies and short-cuts in time-critical environments like C2 (Koterba 2004; Lehner et al. 1997). C2 decision makers often have to follow protocol and procedures they have been trained to execute (Lehner et al. 1997). When C2 decision makers do not have protocol or procedures to follow, they apply rules of thumb or "error-prone heuristics" that result in "good enough" versus exact results obtained from rational decision making outputs (Albers 1996). Decision makers are especially prone to applying heuristic decision processing under stressful conditions (Lehner et al. 1997). This is a natural tendency because humans have limitations on their capacity to process information, and cope by grouping information and applying mental shortcuts.

Kahneman and Tversky pioneered research in this area and began compiling individual cognitive and personal biases in decision making that currently are part of a long list (Tversky and Kahneman 1974). Some examples relevant to C2 follow:

- Availability Bias: Tend to overestimate usual or easy to remember events (Lehner et al. 1997)
- Hindsight Bias: After an answer is known, suggest that they knew all along, when they were unclear at the onset (Lehner et al. 1997)
- Inertia: Be unwilling to change past thought patterns for new situations (Boyer & Robert 2006)
- Automation Bias: Over-rely on automation (Cummings et al. 2007)
- Confirmation bias: Find evidence that supports preconceived conclusion and disregarding evidence that contradicts conclusion (Lehner et al. 1997; Tatarka 2002)

An article from the Military Intelligence Professional Bulletin reports that a large amount of anecdotal evidence suggests the two most dangerous and common biases in military C2-like operations are anchoring and adjustment and confirmation biases (Tatarka 2002). Once an intelligence analyst has anchored (anchoring and adjustment) on an enemy course of action, they seek evidence that confirms their decision and disregard conflicting information (confirmation bias). The authors suggest that military doctrine promotes this bias because of short time constraints. The risk is "cognitive tunnel vision," which is emphasized in high-stress situations like C2 and could lead to devastating effects (Tatarka 2002). Another potentially dangerous and common heuristic in C2 is automation bias. As technology becomes more prevalent and can provide automated solutions, operators may over-rely on the automation, become complacent, and experience a loss of SA (Cummings et al. 2007). These cognitive biases are highly relevant to consider in decision making tasks because they become increasingly resistant to preventative techniques such as training, decision support tools, and devil's advocate approaches, in unfamiliar and stressful environments like complex C2 (Lehner et al. 1997; Tatarka 2002). The heuristics and biases can be better understood by applying techniques such as those in understanding information requirements; for example, cognitive task analysis, cognitive work analysis, etc.

CTA methodologies

A number of methods exist to obtain information requirements and decision rationale from the decision maker, the most common being cognitive task analysis (CTA). A variety of CTA methodologies have been developed that differ in approach, structure, emphasis, and resource requirements, but all include some sort of knowledge elicitation, analysis, and knowledge representation (Federal Aviation Administration Human Factors Division 1999; Militello and Hutton 1998). Often the knowledge elicited is not measured quantitatively; however, aspects that can be measured quantitatively are the number or type of ideas considered and the number or type of consequences considered. The analysis and knowledge representation are the significant parts of the CTA as they can be used to improve processes or systems. If the system and process are in line with the decision maker's mental model derived from conducting a CTA, there is a higher likelihood that the decision maker will be able to use the system more effectively and efficiently to make a decision.

Three examples of CTA are (a) Precursor, Action, Results, and Interpretation (PARI) method, (b) Critical Decision Method (CDM), and (c) Conceptual Graph Analysis (CGA). In the PARI method, subject matter experts identify troubleshooting-use cases to elicit system knowledge (how the system works), procedural knowledge (how to perform problem-solving procedures), and strategic knowledge (knowing what to do and when to do it) from other subject matter experts (Federal Aviation Administration Human Factors Division 1999; Jonassen, Tessmer, and Hannum 1998). "PARI attempts to identify each action (or decision) that the problem solver performs, the *P*recursor (or Prerequisite) to that *A*ction, the *R*esult of that action, and an expert's *I*nterpretation of the Results of that Action" (Jonassen, Tessmer, and Hannum 1998). The PARI technique, developed by Hall, Gott, and Pokorny in 1995, was originally designed for the Air Force to design intelligent tutoring systems. It involves a structured interview during which pairs of subject matter experts probe each other under realistic conditions. The interviews are held during and after troubleshooting, with an emphasis on the reasoning used in making decisions. PARI products include flowcharts, annotated equipment schematics, and tree structures (Federal Aviation Administration Human Factors Division 1999).

An advantage of PARI is that it thoroughly exposes how subject matter experts deal with systems by identifying the technicalities of how the system works (technology focused), demonstrating how to perform problem-solving procedures (human-system interface), and knowing what to do about the problem (cognitive or decision making rationale). PARI is especially strong in revealing troubleshooting and analyzing problem-solving techniques that can be beneficial for

Table A1. CDM Interview Cycles and Cognitive Probes.

Interview Cycles	
Stage	Task
First cycle	Interviewee briefly describes even
Second cycle	Interviewee puts timeline with event
Third cycle	Interviewee uses cognitive probes to fully understand decisions
Fourth cycle	Interviewee compares performance with novice.
	Cognitive Probes
Probe Type	Probe Content
Cues	What were you seeing and hearing
Knowledge	What information did you use in making decision and how was it obtained
Goals	What were your specific goals at that time?
Situation assessment	If you had to describe the situation to someone else at this point, how would you summarize it?
Options	What other course of actions were considered, or were available to you?
Basis of choice	How was this option selected, other options rejected?
Experience	What specific training or experience was necessary or helpful in making this decision?
Aiding	If the decision was not the best, what training, knowledge, or information could have helped?
Hypotheticals	If a key feature of the situation were different, what difference would it have made in your decision?

training. The disadvantage of PARI is that it may focus too much on specific troubleshooting, and it relies heavily on subject matter expertise. Since PARI consists of subject matter experts interviewing each other, there is a risk that some details may be left out because they may be assumed to be included or considered trivial.

CDM, based on Flanagan's critical incident technique developed in 1954, and formalized by Klein in 1993, is a series of semistructured interviews of subject matter experts that focuses on critical, nonroutine incidents requiring skilled judgment (Klein, Calderwood, and Macgregor 1989). The interview is considered semistructured because it falls in between an ongoing verbal protocol, where the decision maker "thinks aloud," and a completely structured interview (Klein, Calderwood, and Macgregor 1989). The theory behind CDM is that probing subject matter experts about difficult activities results in the "richest source of data" to understand decision making of highly skilled personnel as the information gleaned is expertise, not formalized protocol (Klein, Calderwood, and Macgregor 1989). When CDM is conducted, subject matter experts recount a difficult incident, and the interviewer probes to distinguish decision points, critical cues, cognitive strategies, etc. (*Table A1* provides information on interview cycles and example cognitive probes [Federal Aviation Administration Human Factors Division 1999, 126]).

A variety of CDM products can be produced; one of the most common is a narrative account (Federal Aviation Administration Human Factors Division 1999). Another product is a cognitive requirements table that includes cognitive demands of the task and pertinent contextual information. CDM results are usually used to develop system design recommendations or training (Federal Aviation Administration Human Factors Division 1999).

CDM was implemented in a C2 decision making study of anti–air warfare operators on a U.S. Navy AEGIS cruiser to investigate decision maker strategies. Results reflected that the feature-matching strategy, involving recognition of a typical class of situation, was the most used strategy (87 percent of diagnostic strategies). Story building was also used (12 percent of diagnostic strategies), where the situation was novel or where the decision maker built a story from seemingly disparate pieces of information to develop a coherent explanation of the situation. Decision makers did not evaluate 75 percent of the decisions that they implemented, and they considered and compared multiple options in only 4 percent of the cases. In the 4 percent of cases in which multiple options were considered, they were not the most critical decision points. When decision makers did not understand a situation, they prepared for the worst case scenario, probably to avoid risk (Kaempf et al. 1996).

The advantage of CDM is that it reveals expertise and understanding of objectives that would not otherwise be illuminated. The semistructured organization provides flexibility to the decision maker to discuss aspects that might not have been specified a priori. It has also been used in complex C2 environments to determine decision making strategies. The interview cycle approach expands the attributes of information collected and also increases the time and resources required to conduct CDM. A disadvantage of CDM is that it is subjective and reflective on the decision maker's own strategies and basis for decisions (Klein, Calderwood, and Macgregor 1989). Another

disadvantage is that the critical event chosen may be very atypical or rare. Finally, since CDM is less structured, it is more difficult to interpret and analyze the results.

CGA was developed in 1992 by Gordon and Gill and involves generating a visualization of conceptual knowledge structures to conduct CTA (Federal Aviation Administration Human Factors Division 1999). A CGA consists of a formal and detailed collection of nodes (which can be goals, actions or events), relations, and questions (Federal Aviation Administration Human Factors Division 1999; Jonassen et al. 1998). Nodes are connected via arcs, which portray the relationship between nodes. The CGA process begins by exploring any preexisting documentation related to the task to be analyzed. Then, a process called free generation is implemented in which an SME leverages the existing documentation and adds task information requirements. The information is then compiled and visually presented as a draft conceptual graph. Any gaps in the representation are constructed into detailed questions. If there are still gaps after questions are asked, information is filled in from observations (Federal Aviation Administration Human Factors Division 1999). The last step is validating the conceptual graph by having an expert perform the task and check for incorrect or missing information (Jonassen et al. 1998).

An advantage to CGA is that it provides a visual depiction of internal knowledge like a concept map. Clarifying the linkages between concepts causes the interviewer to closely investigate the conceptual relationships that might not be examined through other CTA techniques. Another advantage is that the detailed approach affords a systematic process with more structure than other CTA methods. The structure also yields "specific yet comprehensive" questions. In addition, a variety of automated software tools exist to assist in developing conceptual graphs such as COG-C (De Vries and Gordon, 1994). A disadvantage is the CGA nodes and arcs take time to learn, and a CGA is difficult to develop while an unstructured interview is taking place. Also, while CGA describes concepts well, it is weak at capturing procedural knowledge (Jonassen et al. 1998).

Despite the contrast between CTA techniques, they are useful in revealing C2 decision maker rationale. The PARI technique is considered a traditional cognitive task analysis technique, CDM is considered activity-based analysis, and CGA is considered subject matter/content analysis; however, all reveal information that could improve C2 processes or be used to evaluate C2 decision making. Most of these techniques focus on deviations from standard operating procedures and preplanned responses where C2 decision making and expertise can be exposed (Jonassen et al. 1998). In addition, the various levels of structure in CTA methodologies parallel the levels of structure in various aspects of C2 (Klein, Calderwood, and Macgregor 1989). Another positive aspect is that many of the CTA techniques are conducted retrospectively, which is important in C2 because they are less intrusive (Klein, Calderwood, and Macgregor 1989). The caveat to CTAs is that "no well-established metrics exist" for evaluating CTAs, and it is difficult to evaluate differences between CTA methods (Militello and Hutton 1998). This is partially because it is unknown what information is lost versus gained in comparison to other techniques and also because interviewees and individuals provide different information each. Also, CTAs can be very resource intensive. Because individual differences impact how much information individuals are willing to provide and respond, it is difficult to assess the reliability and validity of CTA methods. Also, another caveat is that no advanced techniques for team CTA have been developed (Militello and Hutton 1998).

References

Ahlstrom, U., and F. J. Friedman-Berg. 2006. Using eye movement activity as a correlate of cognitive workload. *International Journal of Industrial Ergonomics* 36: 623–636.

Albers, M. J. 1996. Decision making: A missing facet of effective documentation. *ACM Special Interest Group for Design of Communication: Proceedings of the 14th Annual International Conference on Systems Documentation: Marshaling New Technological Forces: Building a Corporate, Academic, and User-oriented Triangle*, October 19–22, 2006, Research Triangle, NC, 57–65. New York, NY: ACM.

Allanson, J., and S. H. Fairclough. 2004. A research agenda for physiological computing. *Interacting with Computers* 16(5): 857–878.

Azuma, R., M. Daily, and C. Furmanski. 2006. A review of time critical decision making models and human cognitive processes. *Aerospace Conference 2006,* March 4–11, 2006, Big Sky, MT. Piscataway, NJ: *Institute of Electrical and Electronics Engineers, Inc.*

Basar-Eroglu, C., and T. Demiralp. 2001. Event-related theta oscillations: An integrative and comparative approach in the human and animal brain.

International Journal of Psychophysiology 39(2–3): 167–195.

Bass, S. D., and R. O. Baldwin. 2007. A model for managing decision-making information in the GIG-enabled battlespace. *Air and Space Power Journal, XXI*, (Summer) 100–108.

Berka, C., D. J. Levendowski, C. K. Ramsey, G. Davis, M. N. Lumicao, K. Stanney, L. Reeves, S. Harkness Regli, P. D. Tremoulet, and K. Stibler. 2005. Evaluation of an EEG-workload model in the aegis simulation environment. Proceedings of the Biomonitoring for Physiological and Cognitive Performance during Military Operations. SPIE Defense and Security Symposium 5797, May 2005, Orlando, FL, 90–99. Bellingham, WA: SPIE.

Berka, C., D. J. Levendowski, M. M. Cvetinovic, M. M. Petrovic, G. Davis, M. N. Lumicao, M. V. Popovic, V. I. Zivkovic, R. E. Olmstead. 2004. Real-time analysis of EEG indexes of alertness, cognition, and memory acquired with a wireless EEG headset. *International Journal of Human-Computer Interaction* 17(2): 151–170.

Boehm-Davis, D. A., W. D. Gray, and M. J. Schoelles. 2000. The eye blink as a physiological indicator of cognitive workload. In *Proceedings of the IEA 2000/HFES 2000 Conference*. July 30–August 4, 2000. San Diego, CA, 116–119, Santa Monica, CA: Human Factors Society.

Boyer, M., and J. Robert. 2006. Organizational inertia and dynamic incentives. *Journal of Economic Behavior and Organization,* Vol. 59 p. 324–348.

Collet, C., C. Petit, S. Champely, and A. Dittmar. 2003. Assessing workload through physiological measurements in bus drivers using an automated system during docking. *Human Factors* 45(4): 539–548.

Colman, A. M. 2001. A dictionary of psychology: P300. New York, NY: Oxford University Press.

Cummings, M. L., S. Bruni, S. Mercier, and P. J. Mitchell. 2007. Automation architecture for single operator-multiple UAV command and control. *The International C2 Journal: Special Issue – Decision Support for Network-Centric Command and Control* 1(2): 1–24.

Dantest Medical System. *What is heart rate variability (HRV) analysis?* Retrieved 12/4 2008, from http://www.dantest.com/introduction_what_is_hrv.htm

De Waard, D. 1996. The measurement of drivers' mental workload. Doctoral thesis, University of Groningen, The Netherlands.

Department of Defense. 2002. Department of Defense Directive (DODD) 8100.1, *Global Information Grid (GIG) Overarching Policy*, 19 September 2002. Available online at http://biotech.law.lsu.edu/blaw/dodd/corres/pdf/d81001_091902/d81001p.pdf. Retrieved April 19, 2010.

Department of Kinesiology University of Waterloo. 2008. *Parietal lobe*, http://ahsmail.uwaterloo.ca/kin356/dorsal/parietal.jpg (accessed January 13, 2008).

Department of the Army. (April 2003). *Field manual 3-21.21: The Stryker brigade combat team infantry battalion.*

De Vries, M. J., and S. Gordon. 1994. COG-C: A tool for estimating cognitive complexity and the need for cognitive task analysis. In Proceedings of the Human Factors and Ergonomics Society 38th Annual Meeting, October 1994, Nashville, TN, 993–944. Santa Monica, CA: Human Factors Society.

Endsley, M. R., R. Hoffman, D. Kaber, and E. Roth. 2007. Cognitive engineering and decision making: An overview and future course. *Journal of Cognitive Engineering and Decision Making* 1(1): 1–21.

Entin, E. B., and E. E. Entin. 2000. Assessing team situation awareness in simulated military missions. Proceedings of the 44th Annual Meeting of the Human Factors and Ergonomics Society, pp. 73–76.

Federal Aviation Administration Human Factors Division. 1999. *Department of defense handbook: Human engineering program process and procedures* (No. MIL-HDBK-46855A). Washington, D.C.: Department of Defense.

Finntrack. 2008. *Decision making process*, http://www.finntrack.com/hnc_hnd/bus-decision.htm (accessed February 12, 2008).

Gawron, V. J. 2000. *Human performance measures handbook*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Gevins, A., and M. E. Smith. 2003. Neurophysiological measures of cognitive workload during human-computer interaction. *Theoretical Issues in Ergonomics Science* 4(1–2): 113–131.

Gorman, J. C., N. J. Cooke, and J. L. Winner. 2006. Measuring team situation awareness in decentralized command and control environments. *Ergonomics* 49(12–13): 1312–1325.

Hall, E. M., S. P. Gott, and R. A. Pokorny. 1995. *A Procedural Guide to Cognitive Task Analysis: The PARI Methodology*. Air Force Technical Report, AL/HRTR-1995-0108. Brooks Air Force Base, TX: Air Force Armstrong Laboratory.

Hannula, M., J. Koskelo, K. Huttenen, M. Sorri, and T. Leino. 2007. Artificial neural network analysis of heart rate under cognitive load in a flight simulator. Paper presented at the IASTEO International, Conference on Artificial Intelligence and Applications, 25th Multiconference on Applied Informatics, Innsbruck Austria, February 12–14, 2007. Calgary, AB, Canada: ACTA Press.

Harris, B. 2008. *A revolution in neuroscience: Tuning the brain*, http://www.centerpointe.com/about/articles_research.php (accessed January 13, 2008).

Jonassen, D. H., M. Tessmer, and W. H. Hannum. 1998. *Task analysis methods for instructional design*. Mahwah, NJ: Lawrence Erlbaum Associates.

Kaempf, G. L., G. Klein, M. L. Thordsen, and S. Wolf. 1996. Decision making in complex naval command-and-control environments. *Human Factors* 38(2): 220–231.

Klein, G. A., R. Calderwood, and D. Macgregor. 1989. Critical decision method for eliciting knowledge. *IEEE Transactions on Systems, Man, and Cybernetics* 19(3): 462–472.

Klein, G., B. Moon, and R. R. Hoffman. 2006a. Making sense of sensemaking 1: Alternative perspectives. *IEEE Intelligent Systems* 21(4): 70–73.

Klein, G., B. Moon, and R. R. Hoffman. 2006b. Making sense of sensemaking 2: A macrocognitive model. *IEEE Intelligent Systems* 21(5): 88–92.

Koterba, N. T. 2004. *APL internal report: The effect of personality on decision making*. unpublished internal report, Johns Hopkins University Applied Physics Laboratory.

Lamar, M. 2006. Neuroscience and decision making. Available online from Triarchy Press at http://www.docstoc.com/docs/28616882/Neuroscience-and-decisionmaking/download. Accessed April 10, 2010.

Leedom, D. K. 2001. *Sensemaking symposium final report*. Washington, D.C.: Command and Control Research Program.

Lehner, P., M. Seyed-Solorforough, M. F. O'Connor, S. Sak, and T. Mullin. 1997. Cognitive biases and time stress in decision making *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans* 27(5): 698–703.

McGraw-Hill Companies. 2007. *Science and technology encyclopedia, 5th edition: Electrodermal response*. http://www.answers.com/topic/galvanic-skin-response (accessed May 18, 2007).

Militello, L. G., and R. J. B. Hutton. 1998. Applied cognitive task analysis (ACTA): A practitioner's toolkit for understanding cognitive task demands. *Ergonomics* 41(11): 1618–1641.

Nickel, P., and F. Nachreiner. 2003. Sensitivity and diagnosticity of the 0.1-Hz component of heart rate variability as an indicator of mental workload. *Human Factors* 45(4): 575–590.

O'Donnell, R. D., and F. T. Eggemeier. 1986. Cognitive processes and performance. In *Handbook of perception and human performance,* Vol. 2, eds. K. Boff, L. Kaufman, and J. Thomas, 42/1–42/9. Oxford, England: John Wiley & Sons.

Poole, A., and L. J. Ball. 2004. Eye Tracking in Human-Computer Interaction and Useability Research: Current Status and Future Prospects. In Ghaui, Claude (ed). *Encyclopedia of Human Computer Interaction.* Available online at http://www.alexpoole.info/academic/bookchapter.html. Accessed April 21, 2010.

Poythress, M., C. Russell, S. Siegel, P. D. Tremoulet, P. Craven, C. Berka, et al. 2006. Correlation between expected workload and EEG indices of cognitive workload and task engagement. *Second Annual AugCog International Conference*, San Francisco, California, 32–44.

Rasmussen, J. 1983. Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models. *IEEE Transactions on Systems, Man, & Cybernetics* 13(2): 257–266.

Rowe, D. W., J. Silbert, and D. Irwin. 1998. Heart rate variability: Indicator of user state as an aid to human-computer interaction. Paper presented at CHI '98 in CHI proceedings, April 18–23 1998, Los Angeles, CA, 480–487. New York, NY: ACM Press.

Salmon, P., N. Stanton, G. Walker, and D. Green. 2006. Situation awareness measurement: A review of applicability for C4i environments. *Applied Ergonomics* 37: 225–238.

Smith, D. J. 2007. *Situation(al) awareness in effective command and control*. http://www.smithsrisca.demon.co.uk/situational-awareness.html (accessed January 3, 2008).

Sniezek, J. 1992. Groups under uncertainty: An examination of confidence in group decision making. *Organizational Behavior and Human Decision Making Processes* 52: 124–155.

Tatarka, C. J. 2002. Overcoming biases in military problem analysis and decision-making. [Electronic version]. *Military Intelligence Professional Bulletin*. Jan–March 2002. Accessed December 12, 2008.

Tattersall, A. J., and G. R. J. Hocky. 1995. Level of operator control and changes in heart rate variability during simulated flight maintenance. *Human Factors* 37(4): 682–698.

Tsai, Y., E. Viirre, C. Strychacz, B. Chase, and T. P. Jung. 2007. Task performance and eye activity: Predicting behavior relating to cognitive workload. *Aviation, Space, and Environmental Medicine: Operational Applications of Cognitive Performance Enhancement Technologies* 78(5s): B176–B185.

Tversky, A., and D. Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *Science* 185: 1124–1131.

Van Orden, K. F. 2000. *Real-time workload assessment and management strategies for command and control watchstations: Preliminary findings*. http://www.dtic.mil/matris/sbir/sbir011/Navy89b.doc (accessed March 16, 2006).

Van Orden, K. F., W. Limbert, S. Makeig, and T. P. Jung. 2001. Eye activity correlates of workload during a visuospatial memory task. *Human Factors* 1: 111–121.

Veltman, J. A., and A. W. K. Gaillard. 1998. Physiological workload reactions to increasing levels of task difficulty. *Ergonomics* 41(5): 656–669.

Wickens, C. D., A. S. Mavor, and J. P. McGee. 1997. ed. *Flight to the future: Human factors in air traffic control*. Washington, D.C.: National Academy Press.

ITEA Journal 2010; 31: 210–216

Integrating Cognitive Assessment Into the Test and Evaluation Process

Robert D. O'Donnell, Ph.D., Samuel Moise, Ph.D., Douglas Eddy, Ph.D., and Regina Schmidt, Ph.D.
NTI, Inc., Fairborn, Ohio

In the total assessment of any system, there is a need to consider that the human is capable of a wide range of cognitive adaptations in managing system anomalies. Yet, it is virtually impossible to anticipate all of these potential adaptations. One solution is to focus on the cognitive demands of the system and to determine the minimum capacity in each cognitive skill the person must have to meet those demands. This approach demands an integrated cognitive assessment involving definition of the type and level of cognitive skills required by the system, and evaluation of the impact of various levels of human cognitive capacity on system performance. This approach has been instantiated in several efforts supported by the Department of Defense and NASA. It involves a taxonomy of cognitive skills, an armory of cognitive tests, and mathematical treatments to relate a person's capacity to system demands.

Key words: Adaptation; cognitive testing; degrading systems; human cognitive capacity; readiness for duty; situational awareness.

There is a paradigm shift occurring in the cognitive performance assessment areas of operational test and evaluation. From the vantage point of the twenty-first century, there is an increasing realization that human cognition in complex, technological environments is an extremely plastic entity. The human is capable of interacting with the system by employing a variety of cognitive skills in a variety of combinations. While we may design a system to be operated on by the human in a particular way, and may test it based on that design, a person frequently finds ways to operate within the system that utilize a significantly different mix of cognitive skills than we anticipated. It is becoming clear that we need to move beyond assessment under optimal conditions to anticipate and evaluate the person's ability to function in degraded system environments.

The problem is that introducing human flexibility into the testing equation makes it impossibly complex. It is difficult enough to design tests of a system where it is assumed that the human is optimally trained and functioning. If one now adds the complexity of potential human cognitive adaptations to the situation, the problem of designing adequate evaluations becomes formidable. One solution to this problem may be to abandon the attempt to probe every possible approach the human may take in operating a system, and rather to concentrate on the *skills necessary for successful performance under any reasonable range of system conditions.* In other words, what cognitive (and psychomotor) skills must the person have to successfully operate the system under any expected conditions?

To answer that question, it will be necessary to develop a way to match the cognitive capabilities of the human to a variety of system demands. Necessarily, this will demand an innovative type of human cognitive testing as well as a technique or model that allows the measured capabilities to be matched to the various system demands. Ultimately, this means there is a need to introduce more ecologically valid techniques into the testing situation so that the limits of human cognitive abilities are factored into the arguably nonlinear demands of the real-world environment.

This article presents a summary of 6 years of efforts devoted to investigating how these demands can be met. Several interrelated steps are involved in the solution, and they are schematically illustrated in *Figure 1.* Each of these steps is then described briefly below. Finally, a brief discussion of how this approach might be applied in the test and evaluation environment is provided.

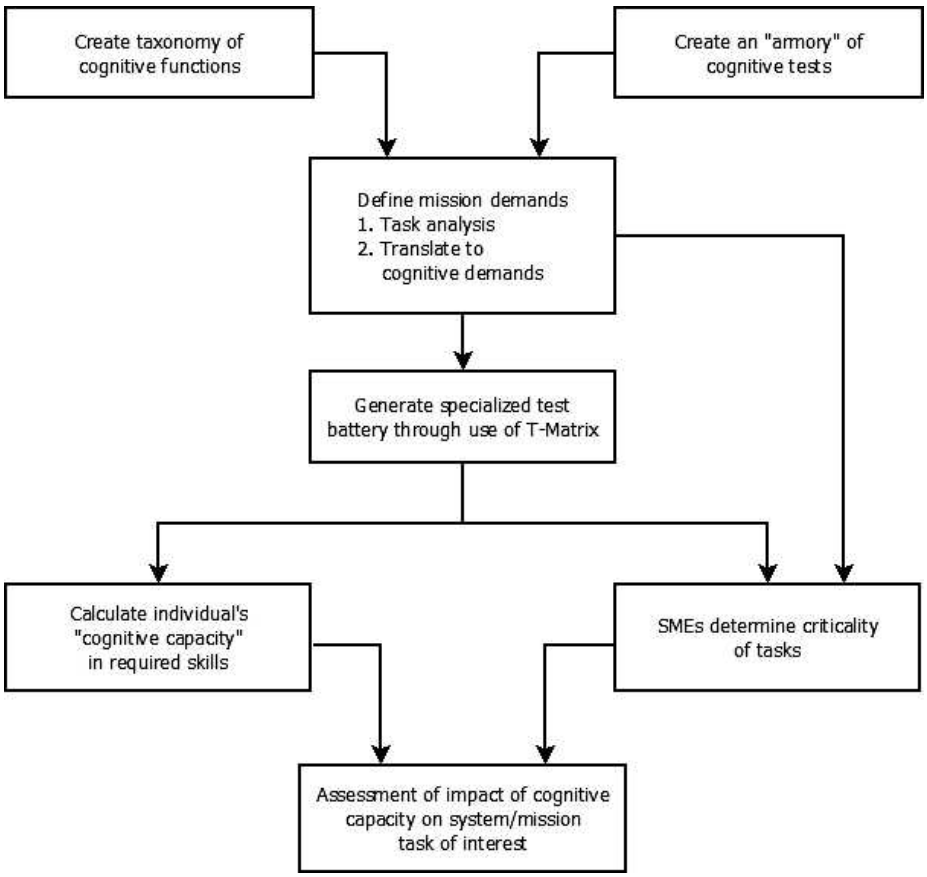


Figure 1. Overview of the approach for creating cognitive test batteries applicable to the evaluation of specific jobs, tasks, missions, or systems.

Overall concept

Stated most generally, the overall goal of these efforts has been to develop a cognitive assessment system that could be directly related to any specific job or mission. More specifically, the desire was to define the minimum set and level of cognitive abilities that would allow the person to successfully operate a given system in the context of a given mission and/or system condition.

Generically, these goals are not new to the testing environment. In one form or another, they have formed the basis for evaluating any system–human interaction. Many cognitive test batteries have been developed. In each of them, the basic concept was to utilize a fixed set of tests, usually administered in a fixed order, to assess a range of cognitive capabilities. The capabilities were typically defined rather broadly (mathematical functions, spatial functions, short- and long-term memory processes, etc.) or based on a theoretical orientation (working memory, declarative memory, episodic memory, etc.). On the testing side, each procedure was assumed primarily to probe one type of cognitive function. For example, one procedure developed by Sternberg (1966) was commonly referred to as a "memory retrieval" test. Although some recent test batteries allow the user to generically determine what cognitive skills are of interest in a particular application, this is typically an "armchair" exercise, not well quantified.

This evolution of cognitive testing has led to a desire to make the process more universal and quantified. While there is as yet no "grand unified theory" of cognition as envisioned by Newell (1994), experimental and theoretical developments have developed to the point where reasonably comprehensive and well-defined taxonomies of human cognitive abilities can be defined. These, in turn, led to a realization that what is actually measured by common test procedures is more complex than the simple unidimensional view that had typically been assumed. In fact, it became clear that any test procedure actually is dependent on a number of cognitive abilities for successful performance. The path to greater specification and quantification of cognitive testing therefore became clear. If tests probing the entire defined range of human cognitive abilities could be identified in terms of the

Table 1. The initial list of cognitive functions.

Sustained attention	Procedural memory	Spatial visualization
Divided attention	Time/velocity estimation	Math functioning
Selective attention	Language/semantics	Problem sensitivity
Directed attention	Decision making	Cognitive flexibility
Visual–motor control	Planning/problem solving	Situation awareness
Declarative memory	Task multiplexing	

degree to which they actually probed each ability, a foundation would be established for relating them to the cognitive demands of a task or job in a way that was much more detailed than had previously been done.

Development of taxonomy of cognitive functions

If one wishes to eventually develop a testing system that will be appropriate for a wide range of operational systems and missions, the first step is to define the range of cognitive abilities that will need to be tested. In an effort supported by Dr. Susan Chipman, of the U.S. Office of Naval Research, the authors attempted to synthesize a large amount of diverse literature in the field of current cognitive theory, such as reductionistic computer models, stage theories, brain imaging studies, and psychometric approaches. The goal of this extensive literature review was to integrate the diverse approaches to cognitive theory into a framework that would yield clues regarding how cognitive function could be probed in a comprehensive way. Based on these analyses, a suggested list of the cognitive skills was constructed (O'Donnell, Moise, and Schmidt 2005) *(Table 1)*.

It can be seen that a major criterion for selecting these particular categories was that they could be related to actual activities required in performing jobs. We were not only interested in identifying "pure" cognitive processes that might be studied in the laboratory, but also in those more complex cognitive abilities that might involve several more basic ones. For instance, "situation awareness," although poorly defined, probably involves elements of attention, working memory, and others. However, from a functional point of view, it is an identifiable cognitive process that is important to many jobs. Therefore, it can serve well as a separate cognitive function.

Admittedly, this synthesis is an initial effort. Subsequent research and theoretical developments could, and even should, add or subtract from this list, or even supplant it entirely in the light of new insights. As a starting point, however, this taxonomy illustrates the beginning of the process leading to a richer analysis of human cognitive capacities.

Creation of a test armory

In considering the tests that should be available for a general assessment system, it was desirable to avoid the traditional approach of having a fixed battery designed to be given in the same order every time. This was because it was envisioned that the assessments would be tailored to a wide range of applications. A concept first suggested by Hunt (1991) and Hunt and Pelligrino (2002)—the test "armory"—was therefore adopted. The concept was that rather than developing another cognitive test "battery," one should identify a large number of tests that would generally cover *all* of the cognitive functions described in the taxonomy. This would allow the user in the testing situation to generate specific batteries that would be tailored to the cognitive demands of a given job or mission.

Again, under Office of Naval Research sponsorship, the testing literature was surveyed to select a reasonably large number of probes of various cognitive functions. Ultimately, a total of 19 test procedures were selected and programmed, with several variations of some that bring the total to 24 tests. These now constitute the NTI Cognitive Assessment Armory *(Table 2)*. The tests include traditional techniques (e.g., reaction time, mathematical functioning, Sternberg memory search, etc.) as well as new techniques designed to probe more complex functions (e.g., rapid decision making, motion inference, directed attention, etc.). Tests are described in O'Donnell, Moise, and Schmidt (2005). A detailed

Table 2. Tests currently in the NTI cognitive assessment armory.

Continuous memory	Dichotic listening	Digit span
Manikin	Match to sample	Math processing
Motion inference	NovaScan	Precision timing
Peripheral processing	Rapid decision making	Reaction time
Relative motion	Sternberg memory search	Stroop test
Tower of Hanoi	Tracking	Visual vigilance
Wisconson Card Sorting		

user manual is presented in O'Donnell, Moise, and Schmidt (2004).

Selecting a test battery for specific applications

The existence of a taxonomy and series of tests is only the beginning of actually finding ways to generate specific batteries for specific jobs. It is necessary to develop a technique for *matching* the tests to the system or mission demands. If an objective and quantified way of doing this could be developed, it would provide not only an "audit trail" for assessing the relevance of the battery to the job, but also lay the foundation for modeling and prediction of human cognitive capacities within the system. As noted earlier, it was recognized early that no performance test is dependent on a single cognitive skill identified in the taxonomy. For instance, although a simple reaction time test is certainly dependent on visual–motor coordination, it obviously also requires some degree of attention allocation, sustained attention, focused attention, and other skills. Further, though important, these skills are not all equally critical to successful performance on the test. If some estimate of the *degree* to which each skill is probed by a particular test could be obtained, it could lay the foundation for relating tests to system demands in a more precise way.

The psychometric literature suggested a technique that could be modified to provide this more objective technique. DiBello, Stout, and Rousos (1995) described a "Q-matrix" approach for selecting questions in a test. Using their basic approach, a matrix (now called a "T-Matrix") was constructed in which the cognitive attributes described in the taxonomy (e.g., spatial visualization, working memory, etc.) constituted one dimension, and the tests in the armory (e.g., continuous memory, manikin test, etc.) constituted the other. Cognitive scientists familiar with the tests were then asked to arrive at a consensus opinion about whether a test does or does not require a given cognitive attribute from the taxonomy for successful performance. If it did, they were asked to estimate the *degree* (on a scale from 0 to 9) to which the test was dependent on that attribute for successful performance. A representative sample of the T-Matrix is presented in *Table 3*. The full 18 × 24 matrix is presented in O'Donnell, Moise, and Schmidt (2005).

As can be seen from *Table 3*, the matrix of values produced in this way provides a two-dimensional mathematical vector of the tests in the armory that specifies, at least on an ordinal scale, the array of cognitive functions that are probed by each test. To the extent that these vectors are accurate, they provide a "cognitive map" of each test. For instance, in this sample, it can be seen that the Sternberg test probes two major skills, working memory and situation awareness. The rapid decision-making test, on the other hand, samples a variety of cognitive skills to different degrees than any of the others. It is also noted that each cognitive skill is sampled by more than one test in the armory, opening up the opportunity for multidimensional analysis of each skill.

Table 3. Representative segment of the T-matrix cognitive skill.

TEST	Divided attention	Working memory	Decision making	Situation awareness	……etc.
NovaScan	5	6	0	8	X
Sternberg	0	9	0	3	X
Rapid decision	4	3	8	4	X
Unstable tracking	0	4	0	3	X
Etc.	X	X	X	X	X

The ultimate goal of using the T-Matrix is to select a set of cognitive tests that optimally probe the cognitive demands of a particular job, task, or mission. To do this, it is necessary to know what those demands are. In effect, we must ask, "What are the cognitive requirements for operating the system and, from a test and evaluation viewpoint, what are the boundary cognitive capabilities of the human that will allow the system to be operated successfully?"

In practice, this is a two-step process. In the first step, the job, mission, or system of interest must be decomposed into its component tasks. For instance, if the application of interest is the reentry and landing of the space shuttle, the starting point of the analysis should be an investigation of the individual tasks involved in that mission. Current data on those tasks could be obtained from existing task analyses, training manuals, or interviews with subject-matter experts (SMEs). The product of this step is a list of essential activities that the individual must perform to carry out the mission. These activities are phrased strictly in terms of actions (e.g., enter data, roll wings level, initiate remote hand controller inputs, etc.), without regard to the cognitive skills required by those actions.

Once the essential tasks are identified, the second step translates those actions into the cognitive skills demanded. This involves an analysis of each category of action identified. The standard descriptions of cognitive skills in the earlier taxonomy are used to identify those that are required for each task category. For instance, the category "monitor" clearly involves considerable working memory, sustained attention, situation awareness, and others. However, it probably does not involve a great deal of visual–motor control,

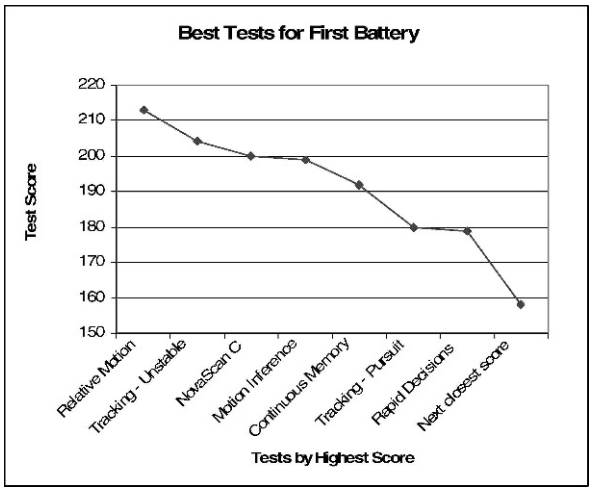


Figure 2. Relative power for the first eight tests from the NTI cognitive assessment armory in probing the cognitive demands of the reentry and landing activity of the space shuttle.

problem solving, or cognitive flexibility. In practice, it was a relatively easy exercise for a group of cognitive psychologists to arrive at consensus opinions regarding which skills were involved in each task category.

These two sets of data provide the information necessary for the T-Matrix to *select the optimum set of tests that best assess the cognitive requirements of the mission*. This can be carried out automatically through use of a simple optimization algorithm, the output of which presents a prioritized list of tests from the armory. An example of such an output for the reentry and landing activity of the space shuttle is presented in *Figure 2*. The list gives a quantified estimate of how well each test probes the demands of that specific mission, allowing the investigator to select the number of tests required based on the unique testing requirements (e.g., time available, subject characteristics, environment, etc.). Currently, once the cognitive demands of the activity are entered, the system automatically optimizes and configures the battery.

Although this approach is still based on subjective judgments, it represents a significantly more detailed and precise picture of what cognitive skills a test actually probes. More objective techniques such as factor analysis (which this approach mimics) could be used to add a degree of objectivity to the assessments, although the time and expense of doing this is virtually prohibitive.

Calculating cognitive capacity from test battery results

Use of the T-Matrix allows a user to select a set of cognitive performance tests that are optimized for a particular system and mission. The next step in the process is to estimate the individual's capacity in the required cognitive skills relative to his or her "normal" capacity. If the tests in a battery do indeed assess the kinds of skills necessary for a given job, it is a reasonable and common assumption that the "normal" level of functioning on those tests approximates the way the person typically does the job. If the well-trained person typically does the job successfully, motivated performance on the tests should reflect his or her usual cognitive "capacity" on the required skills. By implication, any decrement from that "normal" baseline constitutes an indication that the person's ability to perform the job might be impaired to some degree.

Determining the actual meaning of a decrement, however, is not a trivial task. Traditionally, one carries out a number of simple mathematical procedures, such as normalization of scores and calculation of standard deviations from the person's baseline, to determine degree of decrement. However, these linear mathematical operations simply provide more numbers that, in themselves, give no clue about their practical meaning. We believe that to proceed to the next level of analysis, a major assumption must be made. It appears reasonable to assume that there is a meaningful *nonlinear distribution* around the person's average performance on the tests. If so, a one-unit decrement from baseline does not indicate half the change in a person's capacity as a two-unit decrement. We do not know exactly what this distribution should be, but at our present level of sophistication, it would appear that a reasonable place to start is the traditional Gaussian, or "normal" distribution. For instance, using that approach, a decrement in performance of one standard deviation from a person's mean would indicate about a 34 percent decrement in that person's capacity, or that the person is operating at about 66 percent of capacity. Of course, other distributions might be employed (e.g., Poisson, chi-square, or individualized distributions that might skew in one direction or another). It would be a feasible, though nontrivial, task to determine the optimal distribution experimentally. For the present purposes, the critical point is that a performance score on a test is translated into a metric that begins quantitatively to evaluate the person's capacity in the various cognitive skills measured by the test.

This conversion of basic test scores to performance capacity measures can be done for each test in a battery. So, for each test, the person's performance relative to his or her "norm" can be multiplied by the degree to which the test measures a cognitive skill (taken from the T-Matrix). This yields an array of numbers for the entire battery. This array provides several assessments of the person's capacity, one for each of the cognitive skills demanded by the job in question. For instance, if a person's score on a spatial manipulation test indicates

that he or she is operating at 50% capacity, and if that test demands a great deal of spatial visualization, the fact that that test "loads" heavily on spatial visualization (has a high value in the T-Matrix) demands that it should be given great weight. On the other hand, if the same test requires only a minimum amount of attention allocation, the meaning of the test score, while not zero, would be much lower. In effect, the values calculated in the array constitute a multidimensional assessment of the person's capacity in each of the relevant cognitive skills. By amalgamating this multidimensional estimate into a single number, it is possible to arrive at an overall assessment of the person's capacity in each of the cognitive skills required by the job or mission.

It is important to note that this estimate is now based on composite performance of the person on the entire battery, not just on a single test. Among other things, we believe that this approach begins to account for various levels of interaction among cognitive skills that occur in complex performance situations, and therefore represents a more ecologically valid way to estimate a person's cognitive capacity than those based on the results of a single test. The establishment of a single measure of capacity in each cognitive skill lays the foundation for approaching the question of what this level of cognitive ability means to actual performance of the job or operation of a system.

Assessment of the mission impact of cognitive capacity

While objective assessment of cognitive capacity is a necessary step in assessing whether the human can operate a system, it only provides the raw data for doing so. Essentially, it provides one of two required sets of input data. What is still needed is more detailed data on the *criticality* of each cognitive demand of the system. "Criticality" here is operationally defined as the degree to which the mission would be compromised if the operator were not functioning at his or her normal level of cognitive proficiency—if the operator was cognitively degraded to any extent, how likely is it that the mission would fail, or the system would not operate? As an example, two cognitive skills might be required for successful performance. However, one might be absolutely critical to success in the sense that if it were not carried out successfully, a catastrophic result would ensue. The other might be necessary, but might have minimal impact on success if it was not performed perfectly. There might be "workarounds" for it, or it might simply degrade the degree of success in the mission. In the first case, if the person's capacity was minimally degraded, the result might be total mission failure. In the second case, the person might be significantly impaired, and the mission would still succeed at some lower level.

To generate these data, individuals intimately familiar with the system, at least with its design, must consider how "important" it is that an operator be functioning at his or her normal level of ability. This, of necessity, involves input from SMEs. Operationally, the SME is presented with the activity's task categories developed previously and asked to consider and rank the impact to the total mission if that task is not performed up to normal standards. The result of these rankings is a "criticality" dimension added to the task demands. Motivated SMEs are able to understand the context of the question and adapt to it. The output of this exercise is a set of "criticality" ratings by the SMEs that is used to make the measured capacity of the person unique to the demands of the specific mission.

If these criticality ratings are then integrated with the person's cognitive capacity for each cognitive skill involved, an overall picture of the individual's current capability to carry out the mission can be generated. This can be presented in a number of ways—as a probability of mission success, as a graphic display of the individual's present capability to carry out the mission, or as an assessment of the limiting or boundary conditions in which a system could be operated successfully. In any case, the final output of these analyses yields a measure of the overall capability of the person to carry out the mission.

Discussion and application

The process described previously constitutes a new way of assessing what has been called a person's "readiness for duty." Obviously, it is more complex than simply giving one test or acquiring one physiological measure because it attempts a much more detailed analysis of cognitive demands in any given activity. It goes far beyond approaches that simply assume that almost all jobs or missions can be evaluated by a single test. It represents perhaps the first attempt to actually integrate separate assessments of the person, the system, and the mission into a composite assessment of the cognitive performance demands. This microscopic analysis of the total performance environment provides a rich set of data upon which operational decisions or system assessments can be based.

Of course, this richness comes at a price. Customizing a test battery for every job, even when the battery is selected and configured automatically, requires considerable attention to detail. The cognitive demands of the job must be defined from task analyses or SME inputs, and this introduces significant subjectivity into the process. One benefit of this is that it grounds the test procedure in the real world and

provides a degree of face validity that should increase user acceptance. The SME input also allows this process to consider levels of criticality in the various cognitive demands, allowing for a more nuanced analysis than has been possible in previous testing. Finally, the analyses and interpretations involved in determining the mission impact of a given set of test results involve several assumptions, such as the use of the normal curve to estimate performance capacity. While these are not unreasonable assumptions, it would certainly be better if they could be verified experimentally. As has been pointed out, this might be possible in some cases. For the present, however, it is probably necessary to make the best assumptions possible, making sure to identify them and consider their appropriateness.

Given the previously discussed cautions, and certainly others that will appear as the new approach is used, it is worthwhile to explore areas of potential application for the methodology described here. Specifically, in the test and evaluation area, the need pointed out earlier to consider the capacity of the human to adapt to unusual system demands in unique ways can be addressed with this technology. This can be done by simply adjusting the "criticality" ratings used in the calculations. For instance if, in the nominal system, the "problem solving" skill is considered only minimally important (i.e., has a criticality rating of "2" or "3") that rating might be increased to "9" in the face of unexpected system problems. Each of the other required cognitive skills could similarly be adjusted based on various hypothesized system problems. By manipulating these values within the algorithm, it will be possible to define the minimum set and level of cognitive capacities required of the operator under any foreseeable degree of system malfunction. In effect, this permits the system evaluation to expand beyond the "nominal operator," and to factor in the adaptability of the human to the system and the mission. ❑

Dr. Robert O'Donnell received the doctoral degree from Fordham University, and completed postdoctoral studies at UCLA in the Brain Research Institute. He retired from the U.S. Air Force at the rank of Colonel, last serving as Chief of the Workload and Ergonomics Branch of the Aerospace Medical Research Laboratory. He currently is the chief scientist of NTI, Inc., Fairborn, Ohio. E-mail: ODNova@aol.com

Dr. Samuel Moise received the doctoral degree from Carnegie-Mellon University, and completed postdoctoral studies and served on the staff of the UCLA Brain Research Institute. He specializes in cognitive testing and test software development. He currently serves as vice president and principal scientist of NTI, Inc., Fairborn, Ohio. E-mail: majormoise@sbsglobal.com

Dr. Douglas Eddy received the doctoral degree from Carnegie-Mellon University, and served as chairman of the Psychology Department, Trinity University. He specializes in the effects of stressors on performance and test development. He currently serves as principal scientist and Southwest regional director of NTI, Inc., Fairborn, Ohio. E-mail: Douglas.eddy@upwardaccess.com

Dr. Regina Schmidt received the doctoral degree from Wright State University, and served as a research scientist with NTI, Inc., Fairborn, Ohio. She currently carries on research in cognitive testing in the Air Force Research Laboratory, Wright-Patterson AFB, Ohio. E-mail: regina.schmidt.ctr@WPAFB.mil

References

DiBello, L. V., L. A. Stout, and W. F. Roussos. 1995. Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques. In *Cognitively diagnostic assessment*, ed. P. D. Nichols and S. F. Chipman, 361–389. Hillsdale, NJ: Lawrence Erlbaum Associates.

Hunt, E. B. 1991. A modern arsenal for mental assessment. *Educational Psychologist* 25 (3/4): 223–241.

Hunt, E., and J. Pellegrino. 2002. Issues, examples, and challenges in formative assessment. In *New directions in teaching and learning: Using cognitive psychology & learning theories as a pedagogy for higher education*, ed. D. Halpern. San Francisco: Jossey-Bass.

Newell, A. 1994. *Unified theories of cognition*. Upper Harvard, MA: Harvard University Press. Reprint edition.

O'Donnell, R. D., S. L. Moise, and R. Schmidt. 2004. Comprehensive computerized cognitive assessment battery. Arlington, VA: Office of Naval Research; 2004. Contract No. N00014–01-C-0430. NTI, Inc.

O'Donnell, R. D., S. L. Moise, and R. Schmidt. 2005. Generating performance test batteries relevant to specific operational tasks. *Aviation, Space, and Environmental Medicine* 76 (7): C24–C30.

Sternberg, S. 1966. High speed scanning in human memory. *Science* 153: 652–654.

ITEA Journal 2010; 31: 217–231

Measuring Human Performance in a Mobile Ad Hoc Network (MANET)

Elizabeth K. Bowman, Ph.D.
Army Research Lab,
Computational and Information Sciences Directorate,
Aberdeen Proving Ground, Maryland
Randal Zimmerman, Ph.D.
Zimmerman Consulting Group, Leavenworth, Kansas

Tactical warfighter networks represent the final leg of the Network Centric Warfare space to mud continuum. The challenges associated with developing, evaluating, and fielding these networks are significant, as experience from field evaluations demonstrates. Even more critical is the capability to quantitatively measure the extent to which tactical networks serve the Warfighters who depend on them for data and information. Such analysis is constrained in two respects. The first is the lack of measures for the reliable correlation of human performance to network Quality of Service levels. The second is the lack of applied data collection methodologies for objective analysis of timeliness and accuracy of decision inputs in the context of integrated Command, Control, Communications, Computers, Intelligence, Surveillance, and Reconnaissance networks. This article reports on research applications addressing both issues. Our findings were developed over a 5-year period from experience in a tactical networked operations field test environment. We describe a methodology for the reliable collection of human situational awareness measures and report human performance findings in the context of network metrics. We suggest emerging linkages between human and network performance metrics. Our conclusions recommend future actions that will support user-centric test and evaluation of tactical networks, systems, and networked Command and Control.

Key words: Decision accuracy; decision timeliness; human performance; mobile ad hoc network; situational awareness; work load.

The first decade of the new millennium saw an avalanche of research in Network-Centric Warfare (NCW).[1] Much of the early research and lessons learned from current operations focused on the strategic and operational levels of command (Conner 2005), where the four tenets of NCW (Office of Force Transformation 2005) were somewhat easier to address compared with the more mobile tactical force:

1. A robustly networked force improves information sharing.
2. Information sharing enhances the quality of information and shared situational awareness.
3. Shared situational awareness enables collaboration and self-synchronization, and enhances sustainability and speed of command.
4. These, in turn, dramatically increase mission effectiveness.

The relative ease of analysis at the higher command echelons (e.g., Joint Operations, Coalition Air Operations Centers) is determined in most part from the stable networks that are utilized in these settings to link centralized and remote nodes that are stable in terms of location and satellite links. If viewed from the lens of the tactical echelon, however, the four tenets of NCW are less intuitive, highly dependent on the variable performance features of Mobile Ad Hoc Networks (MANETs), and require adaptive Command, Control, Communications, Computers, Intelligence, Surveillance, and Reconnaissance (C4ISR) human system interface capabilities that can mitigate the MANET drop-off/self-healing node design. For example, in his analysis of the Operation Iraqi

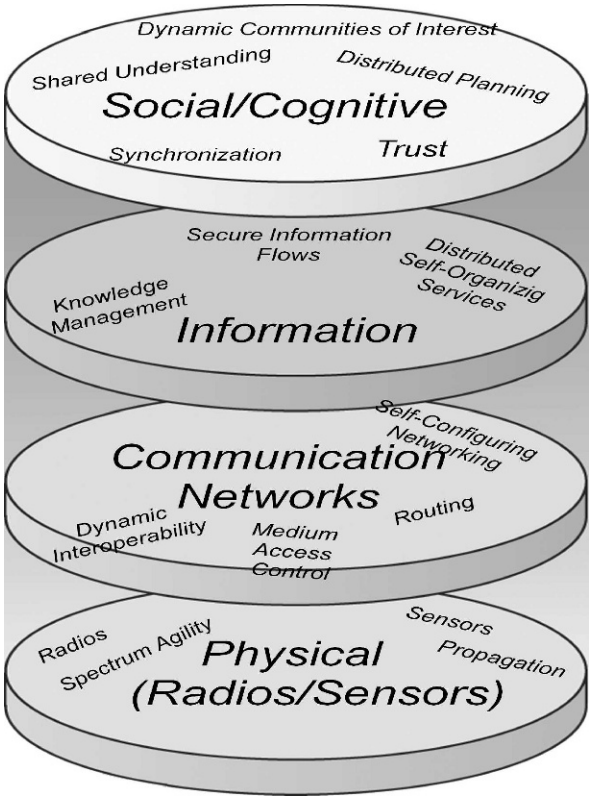


Figure 1. Network component levels (NRC 2005).

Freedom 2003 Thunder Run (where lead elements of the 2nd Brigade, 3rd Infantry Division attacked Baghdad from the southern outskirts, through the city and west to the airport [Conner 2005, p. 15]), Conner noted that carrying the "robust intelligence capability [through a common operating picture] forward to the tactical level would prove almost completely lacking" (p. 18). He characterized the existence of a "digital divide" between operational and tactical commands and suggested that the reasons for this divide were the great distances covered by tactical units and the vast amount of data they were attempting to share. Conner notes several examples in the early phase of Operation Iraqi Freedom where "the promise of technology providing near perfect situational awareness had failed the tactical commander" (p. 20).

Our central thesis is that productive research in the realm of tactical networking must focus on linking the physical, information, communications, and cognitive/social dimensions of the network. Intuitively, this is a sound assertion; human performance should always be investigated *in the context* of supporting technologies. Practically, this is a large challenge in terms of choosing metrics for comparison and in collecting data for analysis. This point is easily made by examining the representation of the network levels in *Figure 1* (National Research Council 2005). If we were to choose one metric from each level that could be expected to cluster, we might choose propagation (Physical), routing (Communication), secure information flows (Information), and shared understanding (Social/Cognitive). The first three metrics could be quantitatively measured, and we could develop correlations among these results. However, no quantitative metric is available for shared understanding suitable for correlation with the Quality of Service (QoS) metrics. In addition, MANET performance is designed to be dynamic, with nodes dropping off and self-healing due to terrain and weather conditions. This performance is invisible to the human eye; users must detect network anomalies from characteristics such as latency of communications, failed messages, or garbled radio speech. Also, unlike the QoS measures, human performance measures such as shared understanding will be degraded for several reasons; network performance is only one contributor. Experience, training, workload, and fatigue provide additional factors that contribute to shared understanding. Partitioning out the variance in this factor due to network performance adds complexity to the human-network analysis problem. As we address this issue, we briefly consider the network environment that serves as the setting for our research.

U.S. Army tactical networks will be MANETs, characterized by wireless radios with limited bandwidth and no fixed infrastructure support. Instead of fixed network nodes, the MANET nodes will be dynamic; they will enter and leave the network at any time due to mobility and terrain conditions (Chiang et al. 2008; Ikeda et al. 2009). In MANETs, network nodes will include vehicles, dismounted soldiers, and unattended sensors and unmanned vehicles. The challenges associated with developing information and communication systems that can operate on a MANET are far from trivial. Porche, Jamison, and Herbert (2004) used a high-resolution simulation to model variations of communication and network parameters to determine the impact on battle command displays. For example, these authors determined that if sensor inputs were limited to 50 percent of the run time, and Common Operational Picture (COP) update rates were no more than once per minute, message completion rates could be above 75 percent. Notably, no human-in-the-loop testing was used to determine the effect of these parameters on decision making. For example, COP update rates of once per minute would be acceptable for a static force, if that force were moving; however, such a rate would be insufficient for navigation from the COP.[2] Simulations are useful in network performance testing because of

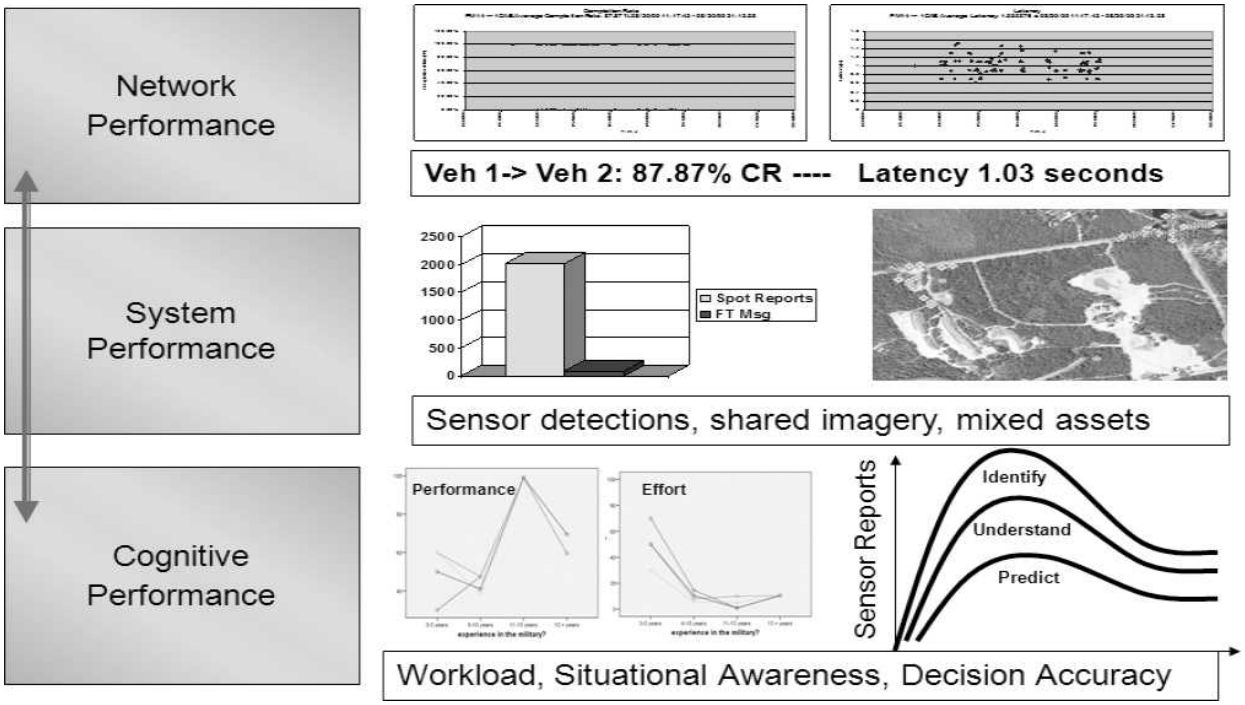


Figure 2. Schema for measuring cognitive performance in context of system and network performance.

the unpredictable and dynamic nature of MANETs in open terrain conditions. They have been used for a variety of performance testing, such as packet latency (Anna and Bassiouni 2006), wireless communication protocols (Gao and Boscardin 2006), and cross-layer routing (Iannone, Kabassanov, and Fdida 2007). These authors, respectively, systematically varied traffic rates for system load and data packet sizes (Anna and Bassiouni), network density and protocol performance (Gao and Boscardin 2006), and hop-count rate and transmission rate (Iannone, Kabassanov, and Fdida 2007). These evaluations provide a foundation for field experimentation in that they support the development of physical MANET capabilities and document simulated network performance bounds.

Field experimentation of human-in-the-loop experiments on MANETs is complex from several perspectives. First, the very nature of MANETs makes controlled variation of key factors difficult. MANET structure and performance (e.g., connectivity, node links, bandwidth, message completion rates) is dynamic by design and cannot be easily controlled. This inability to systematically vary performance parameters presents a challenge to hardware, software, and human evaluation efforts (Bowman and Kirin 2006; Porche, Jamison, and Herbert 2004). Relating human performance to specific network and system characteristics is a second challenge. For example, consider the example described above from the Porche, Jamison, and Herbert simulation. What quantitative impact would a 50 percent sensor feed, a 1-minute COP update, and a 75 percent message completion rate have on workload and situational awareness? Further, what impact would those levels of workload and Situational Awareness (SA) have on timeliness and accuracy of decision making? These questions imply a third challenge: measuring human performance in objective and quantifiable ways with tools and techniques that can be applied across a range of systems, experiments, and evaluations. Partial solutions to these challenges have been developed by the authors over a 5-year period. In that time we have produced meaningful conclusions in the realm of the cognitive impact of networked C4ISR technologies on battle command performance at the tactical level.

The schematic in *Figure 2* provides a high-level view of our efforts to integrate performance metrics across the network (Physical and Communications), the System (Information), and Cognitive (Social/Cognitive) domains. This diagram shows excerpts from our data collection methodology. Network performance shows Message Completion Rates (MCR) and latency of messages. System performance shows numbers of spot reports and free text messages sent during the mission, and shows a COP screen shot of sensor spot reports (the yellow clover leaves). Cognitive performance shows variable levels of workload in terms of performance satisfaction and effort expended and three

levels of SA over time. This schematic shows data for one record run when the network was performing quite well (high MCR and low latency). This high network performance is borne out by the high number of spot reports (over 2,000) and free text messages (87), substantiated by the COP screen shot (yellow icons fade out after 5 minutes). The resulting sensor icons appearing automatically on the display explains the spike in workload and the initially high levels of SA, followed by diminishing SA as the spot reports became too numerous to track. This example of the impact of network and system performance on human performance provides a view to our study objectives. In the remaining sections of the article, we delve into the network and cognitive levels in more detail. We do not address system performance in this article.

The remainder of this article is organized as follows. We provide a brief description of the venue in which our measures and methodology were developed. We then illustrate the data collection methodology that has proven useful for our purposes as detailed above. Next, we review the network and human performance measures used in our analysis. Finally, we offer conclusions that can extend our work for test and evaluation analyses of tactical networks and system-of-systems applications.

Historical perspective: C4ISR On The Move (OTM)

Since 2005, we have engaged resources at the annual Communications Electronics Research, Development, and Engineering Center (CERDEC) C4ISR OTM experiments to determine how tactical soldiers will benefit from an integrated sensor and communications suite to improve timeliness and accuracy of decision making. At this venue, soldiers use instrumented vehicles and various mounted and dismounted communications devices, interact with unmanned air and ground vehicles and unattended ground sensors, and view Battlespace entities on a COP that is an enhanced version of Force XXI Battle Command Brigade and Below (FBCB2).[3] Annually, a range of C4ISR technologies are integrated in a tactical network for soldiers to use against live Opposition Forces (OPFOR) (PM C4ISR OTM 2005; PM C4ISR OTM 2006; PM C4ISR OTM 2007; PM C4ISR OTM 2008; PM C4ISR OTM 2009). Soldiers operated in operationally relevant missions daily against a live, but scripted, enemy force. All enemy forces (vehicles and personnel) were also instrumented. This important capability is further described in the following section.

Our early efforts at documenting the cognitive impact of an integrated and networked sensor suite depended on a large force of human data collectors to observe soldiers at multiple node locations and to administer surveys at strategic points in the runs (PM C4ISR OTM 2005; PM C4ISR OTM 2006; PM C4ISR OTM 2007). Although we triangulated our data collection (Cresswell 1998) with multiple inputs (observation, survey, individual and group interviews), we determined that a field study of this type demanded novel methods that were less human-intensive and intrusive into the soldiers' experiences (Bowman and Kirin 2006). In subsequent years, we modified the Army SALUTE (Size, Activity, Location, Unit, Time, Equipment) report to include two subsequent fields: Assessment and Prediction. The subsequent SALUTE-AP survey was an effective tool to extract SA reports from soldiers, but it still required human intervention and Subject Matter Expert (SME) scoring of reports (Bowman and Thomas 2008, 2009). The SALUTE-AP tool also restricted SA findings to enemy-specific information; no awareness of own forces was included. And, while the SALUTE-AP represented an improvement, it was still a subjective report. We continued to search for objective measures that would serve to support the subjective reports. The Geospatial Environment for Command and Control Operations (GEC2O)[4] visualization tool finally provided the analysis medium for which we had been searching.

Methodology

GEC2O

The GEC2O tool was designed originally as a Rapidly Operational Virtual Reality (ROVR) system to provide a large-scale three-dimensional (3-D) immersive visualization environment. The system allows users to interact with terrain models from Google Earth as well as high-fidelity models created by the developers. The modeled terrain provides extremely accurate representations of wooded and urban training ranges at Fort Dix where missions were conducted. Map overlays, which can be easily printed for later use, are created using standard drawing tools and embedded 2525B military symbology. The ROVR supported both virtual walk-through and flyover navigation through its modeled terrain. It received and stored all FBCB2 tactical messages and unit position information transmitted over the tactical local area network, which was viewable in real time or later via playback. The ROVR was installed in the Tactical Operations Center and configured using two 6- by 4-foot rear projection screens for an overall footprint of about 12-feet wide by 7-feet high by 8-feet deep. The soldiers used this capability infrequently in early years due to the remote location of the large screen display. Typically, they used rock drills as shown in *Figure 3*.

Figure 3. Rock drill, Geospatial Environment for Command and Control Operations (GEC2O) display, and GEC2O mission planning.

Also in *Figure 3* are the large display and a slightly smaller 3-D version used for planning.

In 2008 and 2009, the GEC2O developers provided the visualization capability on a smaller flat screen and a tabletop display as shown in *Figure 4*. The overlays created in GEC2O were electronically uploaded into FBCB2 in the vehicles. With these technologies, the rock drills with sticks and pinecones became a phenomenon of the past.

A feature of GEC2O that enabled it to be used as an analysis tool is the ingestion of instrumented position reports, sensor spot reports, and free text messages for geospatial and time-stamped display. In 2008 and 2009, we used the GEC2O to provide objective analysis of subjective reports of SA, and discontinued the SALUTE-AP surveys (though we did continue to administer surveys to record workload, trust in network, and usability concerns for technologies). We also continued a short, four-question survey on SA in order to provide context to our objective SA findings.

Prior to the daily mission runs, the soldiers were instructed to send SALUTE reports through their FBCB2 free text messaging system whenever they learned of new enemy activity. Since these were captured and time-stamped in GEC2O, they provided us an archive of such reports. By viewing the map display with Red and Blue icons (vehicles and dismounts), we could substantiate the SALUTE reports to determine accuracy of Size, Location, and Time elements of the report. For Activity, Unit, and Equipment, we relied on the script for the enemy runs and for records of activity maintained by the lead analyst assigned to the enemy force. We were also able to assess timeliness of reports through the time-stamps in the GEC2O system. We compared the message received time with the XML time sent to calculate network travel time for the message. In addition, we were able to impute some system performance variables in the human decision-making calculus. For example, in GEC2O, sensor spot reports are displayed as yellow cloverleaves. In *Figure 5*, the display shows one sensor field providing many reports while another field (marked by the red circle in the middle of the map) provides none. Therefore, enemy vehicles approaching through the inactive sensor field would arrive undetected, whereas those arriving through the overactive field could be undetected due to the soldiers' inability to integrate a high number of redundant reports.

GEC2O analysis of decision-making accuracy and timeliness

Timeliness and accuracy of decision making were analyzed with one representative day's retrospective review of data from GEC2O playbacks. This was an

Figure 4. Geospatial Environment for Command and Control Operations tabletop display and 2-D flat-screen display.

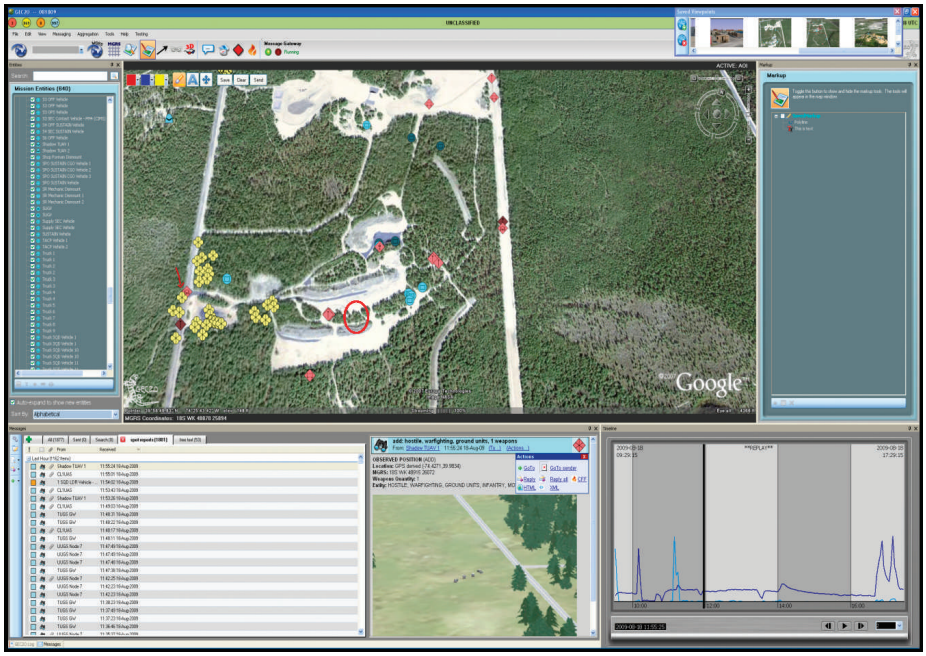


Figure 5. Geospatial Environment for Command and Control Operations screen shot.

exploratory methodology and was not undertaken for each day's record runs. The run from August 20, 2009, was chosen for analysis because it provided a good selection of network, system, and human performance data points. On this date, 12 separate messages were investigated. These messages are shown in *Table 1*. In this table, we have extracted the relevant features of the message contents as well as our analysis findings. The messages extracted for analysis were all free text messages sent from one soldier (usually the Squad Leader) to the group of soldiers, all located in stationary vehicles. The messages were generally uniform in composition and size and contained text that communicated that soldier's knowledge of enemy activity, location, and size. The messages were evaluated for timeliness and accuracy according to low, medium, and high scores. High scores for these measures reflect reports that contained activity descriptions that were very close to ground truth (accuracy) and took little time to process (timeliness). The time delays reported in this analysis reflect network transit time, not time delays due to human activity.

The latency rates were calculated in the following way. Each message received in the GEC2O system showed a "received" time and a "sent" time. The latter was available in the XML script showing the message properties. The analyst subtracted the received time from the sent time to calculate latency. Review of message latency times shows that latency was variable in the mission run; messages were delayed significantly in the early portion of the run but experienced very low latency throughout most of the remainder of the mission. Latency began to rise again at the end of the trial runs. This characterization of tactical network performance shows the influence of the airborne communications relay. On each day, the airship had to travel from the nearby airport to the test site. The early and late messages with high latency reflect message traffic that used ground, rather than airborne, relay mechanisms. Also, even when the airship was present, it sometimes experienced problems that affected the vehicle network. For example, if the airship was not at a high enough altitude to affect the communications relay, ground network pathways were used, resulting in network delays.

Decision accuracy

Decision accuracy was measured on a low-medium-high scale based on the soldier's description of enemy location and activity compared with ground truth. Ground truth was ascertained from the GEC2O playbacks of the instrumented OPFOR vehicles and personnel. Because these position reports were captured locally and manually collected after each mission, these reports are accurate on the display to within 10 meters. This position accuracy is compared with the BLUE position reports that are transmitted via the satellite network and are accurate to within 20 meters. The BLUE reports noted in *Table 1* contain information that the soldiers extracted from various sensors, including soldier visual detections, unmanned aircraft system (UAS) imagery, and simulated Unattended Ground Sensor (UGS) sensor reports. The UAS and UGS reports included a 10-digit grid location of the

Table 1. Decision accuracy and timeliness measures August 20, 2009.

Time of report	Time delay (s)	Description of report	SME interpretation of report	Score	
				Accuracy	Timeliness
1029	349	Reported 4 vehicles with grid location	Grid location within 14 meters of enemy vehicle	High	High
1040	131	Reported 4 moving vehicles	Only 1 was mobile, 3 were stationary	Medium	Medium
1045	103	1 Vehicle reported at grid location	Grid location was within 36 meters of vehicle	High	Medium
1117	20	Reported 11 dismounts at grid location (based on simulated spot report with grid ID)	Grid location was exact	High	Low
1125	17	Reported 1 vehicle and 1 dismount (based on simulated spot report with grid ID)	Grid location was exact	High	Low
1129	21	Reported 1 stationary vehicle (based on UUGS spot report)	Grid location was exact	High	Low
1135	19	Reported 1 stationary vehicle (based on UUGS spot report)	Grid location was exact	High	Low
1146	23	Reported 1 dismount walking south from grid location (based on UAS image)	Grid location was within 10.57 meters	High	Low
1151	21	Reported 1 vehicle driving (based on UUGS spot report)	Grid location was exact	High	Low
1201	21	Reported 1 stationary yellow sedan with grid location (based on DCGS-A image)	Grid location was exact	High	Low
1214	20	Reported 1 stationary vehicle and 1 dismount	Grid location was exact	High	Low
1215	80	Call for fire on yellow sedan (based on simulated spot report with grid ID)	Grid location was exact	High	Medium

SME, subject matter expert; ID, identification; UUGS, urban unattended ground sensors; UAS, unmanned aircraft system.

target. All of the UAS sensor reports and some UGS sensor reports included a simulated image (see *Figure 6* for examples of simulated images).

As might be expected, the soldier messages of enemy activity that were based on the simulated sensor reports were highly accurate. It is clear why the soldiers were so accurate using simulated sensors; the images are very clear, and the reports contain 10-digit grid locations of the OPFOR (e.g., in *Figure 6*, the image of the burning vehicle shows a grid location of 18 S WK 50105 24717).

Decision timeliness

Decision timeliness was also measured on a low-medium-high scale, with low ratings being the more preferred state. The timeliness ratings for each message are shown in the second column of *Table 1*. These results show a consistent trend in the message timeliness factor noticed by experiment observers over the 3 weeks of the experiment. The initial report required 349 seconds to reach the GEC2O system; this was followed by rapidly descending time periods until the last message, which began to increase in time required to transit the network. The curve of message latency is shown in *Figure 7*.

It is at this point that the difficulty in analyzing cross-domain measures comes into play. When we sought to understand the reasons behind message latency, we learned that the network engineers had a different perspective on that network performance metric. The latency measures recorded by the network engineers are presented in *Figure 8*. Comparison of the

Figure 6. Simulated sensor images (Unattended Ground Sensor on left and a Class IV UAS on right).

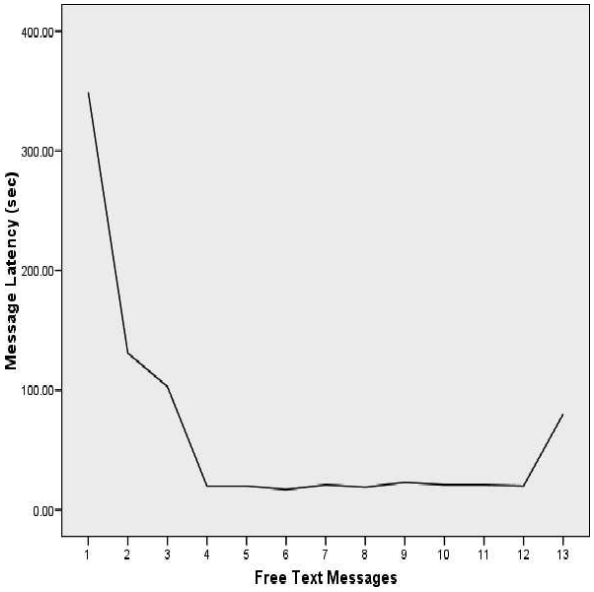


Figure 7. Visual display of message latency in seconds.

latency metrics shows a best (e.g., lowest) latency of 19 seconds in *Figure 7*, but a best latency of less than 1 second in *Figure 8*. We believe that this represents a difference between network level and application level measurement; however, this was not confirmed by the network engineers.

Figure 8 shows the average latency of messages as recorded by the automated network data collection tools. The one consistency between *Figures 7 and 8* and *Table 1* is the spike at 14:52:48 shown in *Figure 8* and the 349 second latency in *Table 1*, row 1. The network times were recorded in Greenwich Mean Time (GMT), which was a 4-hour additive time from the GEC2O recording format (Eastern Daylight Time). It appears that the network latency recorded a large spike at approximately the same time period as the message latency in the application layer of GEC2O. Though these latency figures are far from equal (349 seconds vs. 36 seconds), the same phenomena appear to affect both

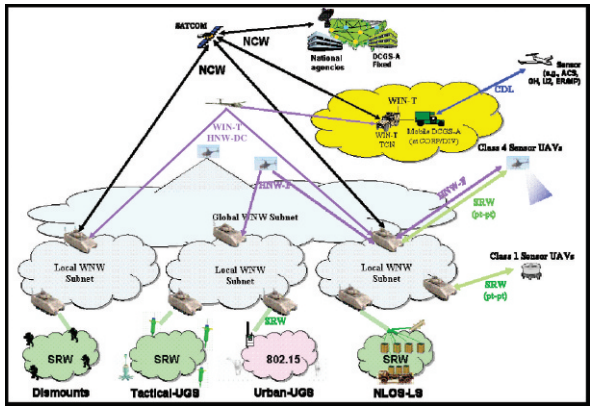


Figure 9. Future combat systems (FCS) multi-tiered transport architecture.

systems. We believe that the inverted bell-shape curve, displayed in *Figure 7*, is the result of several confounding issues. Some of the issues include network delays at the beginning and end of the record runs, message-processing delays on the GEC2O system because of volumes of simulated position reports, a delay in message arrivals due to GEC2O system re-booting during a mission, and differences in time synchronization between the GEC2O system time and the network time. Those differences are under investigation by a joint team of analysts from CERDEC Command and Control Directorate, PM C4ISR OTM and the Army Research Laboratory, and serve to point out the difficulties in this type and level of analysis. We will now shift our discussion to cover the cognitive and network measures used in our analysis.

Measures: cognitive and network performance

The emphasis on investigating the cognitive impact of an integrated C4ISR suite of technologies was driven by the understanding that warfare will always be a human activity. Technology must support human decision making, but it will never replace it. The drive

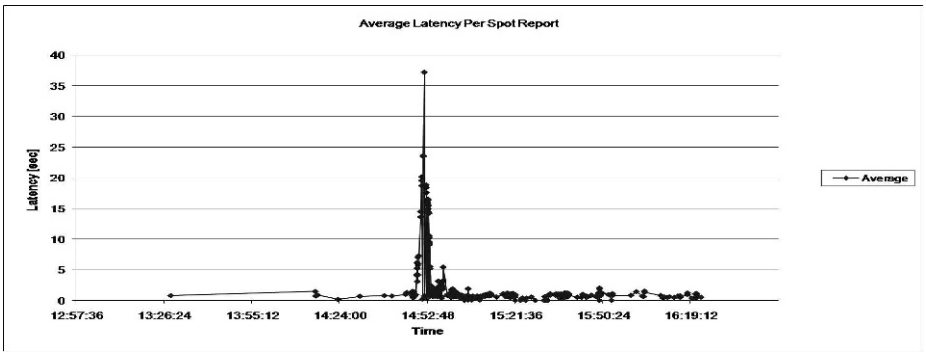


Figure 8. Average latency per report.

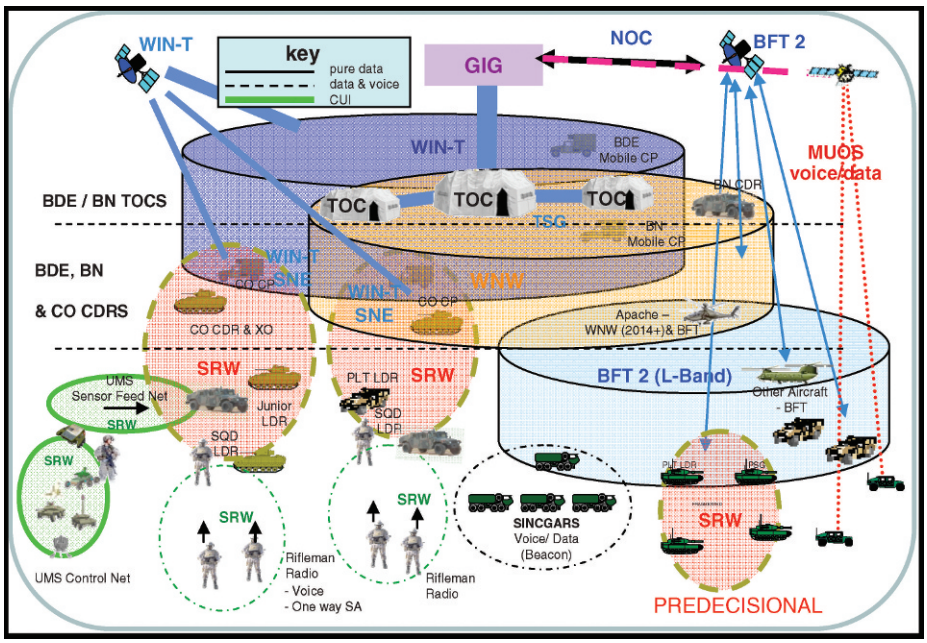


Figure 10. 2013 Modular brigade combat team architecture.

to develop sophisticated decision support systems and intuitive visual displays must include consideration for the ways in which soldiers will utilize these systems on the battlefield. Networked C2 places the human Warfighter at the center of a complex, dynamic, and uncertain web of information. This study measured the impact of networked human and sensor information on the cognitive performance of soldiers at the tactical level. This study provided a benchmark for future analysis of how valid the basic network centric warfare tenets may be for the tactical Warfighters. In the next section, we briefly introduce the network architecture that supported these capabilities and review network performance measures. We follow this discussion with cognitive measures.

Network architecture design

The architecture design elements of the 2009 study were drawn from several including the Future Combat Systems (FCS) Multi-Tiered Transport Architecture, the Unified Battle Command 120 Day study (Moore 2008) and the 2013 Modular Brigade Combat Team (MBCT) Architecture (Latham 2008). Each of the communications architectures employed a variety of systems depending upon the specific configuration under examination, as shown in *Figures 9 and 10*.

Within these architectures, a common element includes satellite communications provided by the NCW, which support an on-the-move satellite capability. NCW is a communications waveform that acts as the primary satellite mechanism for Increments 2 and 3, as well as being available for technology insertion at Increment 1. NCW is an Internet Protocol (IP)-based, bandwidth-on-demand protocol that supports block file data, voice, and video services, as well as other IP-based services to support disadvantaged terminals that may have smaller dish sizes, rendering them less capable. In addition, Soldier Radio Waveform (SRW) is common to both architectures. SRW is an IP-based, software-defined digital communications waveform that supports voice, data, and video services. SRW is designed to provide communications to dismounted soldiers, unattended ground sensors, intelligent munitions systems, non-line-of-sight launch systems, and unattended ground and air vehicles.

Network performance

Several tables are presented to illustrate the performance of the network during the C4ISR OTM 2013 MBCT platoon trial runs. *Table 2* represents the view of network performance from the Brigade Headquarters' perspective. The tables were derived from the daily data sets that were harvested after the completion of the runs. These data include any unicast data between C4ISR Information Management System (CIMS) and FBCB2s, as well as any multicast data that originated either from FBCB2 or CIMS. These multicast groups include the standard groups for CIMS Gateway messages, which include imagery notifications, imagery request, imagery transfers, and chat.

The primary focus of our attention in these network performance metrics is the lowest tactical leader—the

Table 2. Network performance – Brigade Headquarters view.

Radios	From the BDE HQ to	August 14, 2009 1337–1653		August 18, 2009 1311–1757		August 20, 2009 1313–1630		August 21, 2009 1717–2000	
		Completion rate	Average latency	Completion rate	Average latency	Completion rate	Average latency	Completion rate	Average latency
NCW	CO CDR	76.9%	0.60 s	77.5%	0.45 s	95.8%	0.60 s	89.4%	0.59 s
SRW CO Net	VPL1	12.7%	0.95 s	69.1%	1.31 s	87.3%	1.80 s	88.0%	1.32 s
SRW CO Net	VPL2	12.5%	0.95 s	69.1%	1.13 s	87.1%	1.80 s	89.2%	1.33 s
SRW CO Net	PL	66.3%	0.78 s	58.2%	0.73 s	76.9%	0.83 s	82.0%	0.85 s
SRW PLT Net	SL	65.3%	1.00 s	55.6%	0.96 s	75.4%	1.00 s	76.8%	1.00 s

BDE, brigade; HQ, headquarters; NCW, network-centric warfare; CO, company; CDR, commander; s, seconds; SRW, soldier radio waveform; VPL, virtual platoon leader; PL, platoon leader; PLT, platoon; SL, squad leader.

Platoon Leader (PL). These performance metrics presented in *Table 2* show that the MCRs for messages coming from the Brigade Headquarters to the PL ranged from 58.2 percent to 82 percent, with latency measures of less than 1 second.

The network performance metrics from the Company Commander to the PL on the same dates *(Table 3)* show better MCRs; the range of MCR was 60.7 percent to 97.1 percent. We note that only 1 day had a lower than 87 percent MCR, however. Again, latency was a non-issue, with scores in the less than 1/4-second range.

The PL perspective, shown in *Table 4*, shows fairly positive network metrics from the MCR and the latency categories. On August 14, 2009, the PL had low MCRs with the Virtual Platoon Leaders 1 and 2, but these scores rose on the remaining days. The completion rates from the PL to his Company and Brigade Headquarters ranged from the mid-80s to mid-90s range, with low latency in all cases.

A comparison of the network performance for these seven trials offers a mix of both expected and some unexpected results. The network, composed of NCW and SRW, yielded a completion rate of 74.8 percent. This result indicates that just under three quarters of the messages sent during the trials actually reached their intended recipient, yet the soldiers were generally able to complete their assigned missions. These results do not suggest that a 75 percent message completion rate is acceptable but rather illustrates how adaptive the soldiers were in accomplishing their missions despite network performance issues. For example, when the soldiers were not receiving the reports or messages that they were expecting, they would use the provided voice communications, either Single Channel Ground and Airborne Radio System (SINCGARS) or SRW, to send or request the needed information. If they were unable to reach the intended person on the radio, they would frequently rely on personal cell phones to send text messages and images of the OPFOR. Finally, when other means of communication were unavailable, the soldiers would simply drive to the other person's location and conduct a face-to-face meeting. These results do not imply that "alternate" means of communication were preferred by the assessment team; rather they demonstrate the ingenuity of the participating soldiers to work around the limitations of the experimental network. The next section will describe the impact of the technologies on soldier performance.

Table 3. Network performance – Company Commander view.

Radios	From the CO CDR to	August 14, 2009 1337–1653		August 18, 2009 1311–1757		August 20, 2009 1313–1630		August 21, 2009 1717–2000	
		Completion rate	Average latency	Completion rate	Average latency	Completion rate	Average latency	Completion rate	Average latency
NCW	BDE	90.0%	0.74 s	76.7%	0.72 s	75.9%	0.60 s	67.8%	1.37 s
SRW CO Net	VPL1	35.4%	0.44 s	71.6%	0.39 s	78.1%	0.76 s	88.4%	0.65 s
SRW CO Net	VPL2	36.3%	0.44 s	70.6%	0.32 s	77.7%	0.80 s	88.4%	0.65 s
SRW CO Net	PL	97.1%	0.18 s	87.0%	0.18 s	60.7%	0.21 s	94.2%	0.21 s
SRW PLT Net	SL	95.2%	0.39 s	78.3%	0.38 s	59.5%	0.44 s	90.7%	0.42 s

CO, company; CDR, commander; NCW, network-centric warfare; BDE, brigade; s, second; SRW, soldier radio waveform; VPL, virtual platoon leader; PL, platoon leader; PLT, platoon; SL, squad leader.

Table 4. Network performance – Platoon Leader view.

Radios	From the PL to	August 14, 2009 1337–1653		August 18, 2009 1311–1757		August 20, 2009 1313–1630		August 21, 2009 1717–2000	
		Completion rate	Average latency	Completion rate	Average latency	Completion rate	Average latency	Completion rate	Average latency
NCW	BDE HQ	81.0%	0.92 s	69.5%	1.09 s	83.4%	0.87 s	77.9%	1.50 s
SRW CO Net	CO CDR	85.2%	0.16 s	96.7%	0.22 s	82.4%	0.32 s	96.2%	0.24 s
SRW CO Net	VPL1	22.6%	0.32 s	72.4%	0.26 s	88.8%	0.91 s	82.4%	0.55 s
SRW CO Net	VPL2	20.6%	0.34 s	71.0%	0.29 s	87.7%	0.98 s	84.4%	0.55 s
SRW PLT Net	SL	98.2%	0.23 s	69.8%	0.22 s	96.9%	0.22 s	92.1%	0.23 s

PL, platoon leader; NCW, network-centric warfare; BDE, brigade; HQ, headquarters; s, second; SRW, soldier radio waveform; CO, company; CDR, commander; VPL, virtual platoon leader; PLT, platoon; SL, squad leader.

Table 5. Confidence in computer use.

Survey question	Mean	Standard deviation	Minimum	Maximum
I am confident in my ability to use computers in general.	4.8	.41	4.0	5.0
I am confident in my ability to use Army C2 digital systems.	4.0	.89	3.0	5.0
I am confident in my ability to learn to use new software quickly.	4.5	.55	4.0	5.0
I am confident in my ability to perform multiple tasks at the same time.	4.7	.52	4.0	5.0
I am confident in my ability to use personally owned computers.	4.8	.41	4.0	5.0

Cognitive performance measures

Demographics

Soldier participants in this study included five Noncommissioned Officers (NCOs) from the New Jersey Army National Guard (NJ ARNG) and one Officer from CERDEC. All study participants were male. These soldiers were assigned roles as the Company Commander, Platoon Leader, Platoon Sergeant, Squad Leader, and Robotics operators. The NJ ARNG soldiers had returned from a year-long combat deployment earlier in the summer. The rank of the soldiers was Major (1), Sergeant First Class (1), Staff Sergeant (2), and Sergeant (2). The soldiers represented a range of experience in the military. Due to the small number of participants, experience levels were divided into less than and more than 10 years of service.

We documented the soldiers' subjective perception of their ability to use computer systems because the Unified Battle Command study revolved around digital battle command displays. *Table 5* displays these data. The questions were measured on a five-point Likert scale from 1 (strongly disagree) to 5 (strongly agree). On average, the soldiers were extremely confident in their ability to use computers in general, to use personal computers, to perform multiple tasks at the same time, and to learn new software quickly. Slightly lower scores were recorded for the confidence in using Army C2 digital systems. The mean score of 4.0 (standard deviation [SD] = .89) represented two soldiers who rated this question as "neutral," two who "agreed," and two who "strongly agreed."

Workload

Within the network configuration described above, four record runs were achieved from August 18–21, 2009. A repeated measure Multivariate Analysis of Variance (MANOVA) was used to analyze these data. Though some minor differences in workload were noted in the two experience groups, none of these differences are significant at the $P = .05$ level. As can be seen in *Figure 11*, the less experienced soldiers consistently reported slightly higher workload levels than did their more experienced counterparts. However, the highest level of workload reported was achieved on day four of this MBCT trial, and that mean score was 38.04, SD = 10.36. Considering the scale of the workload measure (0–100), this was a low workload rating.

Situation awareness

A repeated measures MANOVA was used to analyze these data, which are graphically displayed in *Figure 12* by day. On each day, the less experienced soldiers reported that they experienced more difficulty than their more experienced counterparts in achieving Levels 1, 2, and 3 SA, but these differences were not significant. It should also be noted that this amount of difficulty, though relatively higher than the more experienced soldiers, is still rather low. That is, none of

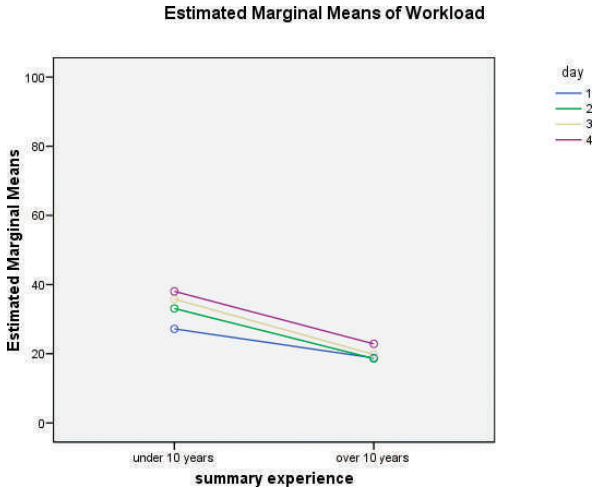


Figure 11. Average workload by experience level in modular brigade combat team.

the soldiers reported much difficulty in achieving SA at all levels.

Conclusions

This report documents a number of major contributions to tactical networked battle command engineering and soldier performance challenges. The vision of Unified Battle Command provides an evolutionary path for the integration of dissimilar applications, communication and sensor technologies, and forces in full spectrum operations. The C4ISR OTM Unified Battle Command Cognitive Impact Study (UBC-CIS) explored the major tenets of this vision. The findings enumerated in this report validate the fundamental assumptions of UBC and clarify technology and human challenges in realizing networked command and control at the lowest tactical echelon.

The network architecture developed for the UBC-CIS provides a firm foundation for continued test and evaluation of networked communication and sensor technologies for tactical elements. Specifically, the 2009 architecture confirmed that multiple pathways (air-satellite-ground) for data transmission are feasible with quantifiable QoS consequences for each route. These QoS impacts (latency, message completion rates) were correlated with human situational awareness and decision making. Knowledge of lesser collaboration pathway consequences can be used by commanders to select, in advance, pathways based on mission requirements and technology availability.

This study also examined technology and human performance factors to investigate the feasibility of extending the network to the lower tactical elements. From a technology perspective, challenges were documented in maintaining connectivity with mobile and static vehicle configurations in forested and open terrain. Soldiers derived adaptive ways of sharing information with less advantaged members of the unit. Redundant communication modes were essential to this capability.

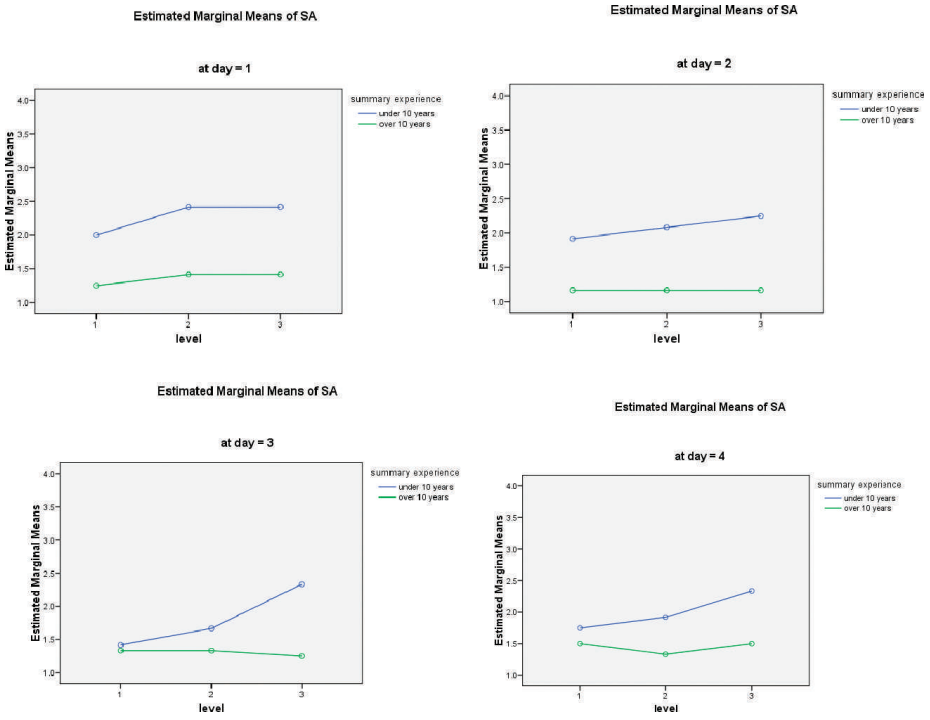


Figure 12. Reported levels of situational awareness of 4 days of trials.

The nature of networked communication and sensor technologies required the tactical leader to be aware of the unique contribution of each system to the mission. He also needed to understand the dependencies between systems and between those systems and the network. Given the combination of air and ground sensors and communication types, this was not an insignificant task. This experiment clearly demonstrated the need for the tactical leader in a networked unit to manage the network architecture for optimal system and soldier performance. Such a requirement cannot be managed by static tactics, techniques, and procedures (TTPs) alone given the dynamic nature of system and network performance in the context of terrain and unit configurations. For soldiers to gain maximum benefit from networked technologies, these systems need to be deployed in an optimal configuration *in the context of the mission*. The role of Network Manager NCO would provide services to the force such as field interpretation of system specifications for optimal use. For example, unmanned and unattended sensors have unique requirements for network support. When planning for placement of unattended ground sensors (UGS) or routes for UAS/UGVs, network connectivity is a primary factor. Also, perception of network health is critical for manned and unmanned teaming and selection of communication mode. For example, in low bandwidth conditions, text communications may be favored over voice or sharing imagery. System integration features are also a necessary consideration for successful network-enabled performance at the tactical level. Questions a Network NCO might consider include: How are systems connected? What are individual and composite system capabilities and limitations? How does the use of one system impact another system? What is the link status of network nodes? How is node drop-off or re-connectivity signaled? How can network problems be diagnosed and repaired? How can errors be diagnosed as human error (and subject to training solutions) or technology failures? These issues will be the subject of a future research effort as we attempt to harness network architectures and performance for tactical soldiers' benefits.

The UBC vision recognizes a need for dissimilar units to interoperate with organic battle command systems and share common geospatial data. The UBC-CIS study examined the ability of soldiers to interact with three different battle command systems for ISR missions against an adaptive and resilient enemy force. Though the analysis procedures were limited by the small number of subjects, this investigation provides a good insight into soldier performance results and analysis methodologies for exploring the cognitive impact of new technologies in a mobile ad hoc network environment. In addition, the advanced capabilities demonstrated by the M&S team allowed the live participants to interact with data simulating two Brigade Combat Teams, thus expanding the tactical perspective.

The workload measures document that the soldiers generally had low workload scores. Though the less experienced soldiers did report slightly higher difficulty with these tasks, they reported, on average, that this was not a particularly demanding task. This is due, in large part, to the effective use of unmanned technologies provided to the soldiers for this experiment. The coordinated use of air and ground vehicles and unattended ground sensors allowed the soldiers to maintain an awareness of enemy activity in the area of operations. This integration of manned and unmanned teaming for ISR highlights the major contribution of the Cognitive Impact Study.

This study represents the authors' efforts to plan, execute, and analyze complex instrumented and human performance data in the context of the system-of-systems demonstrated in this experiment: Network, Systems, and Cognitive. This analysis capability was made possible by the focused contributions of the entire PM C4ISR OTM experimentation and data collection staff, and required a complex set of instrumentation capabilities, data reduction methodologies, human data collection staff (to observe and administer human performance surveys), and senior researchers on site to match network, system, and cognitive performance metrics to observed behavior. This triangulation of data analysis methods provided a quantification of human performance in the context of network and system performance. These insights are valuable as mobile ad hoc networks become available for tactical units. These networks are designed to be accommodating to network disruptions and will have self-healing capabilities. Engineering these networks to optimize human performance in decision-making areas has been shown in this experiment to be possible under less than 100 percent bandwidth conditions. That the soldiers could perform well in significantly degraded video conditions suggests that limited bandwidth can accommodate decision making for certain task scenarios.

The availability of a technology such as GEC2O is the critical underpinning of this analysis activity. This technology allowed after action playback/play-forward of mission runs with archival review of messages, spot reports, and sensor images. Thus, the analysis team was able to retrieve SALUTE reports and objectively contrast the content of each message with actual instrumented activity. This capability significantly

reduced the need to record detailed activity notes on the part of observers and solved the problem inherent with field experiments of this type and duration: remembering and distinguishing between events of one day and another.

The final contribution of this study was to document the ability of soldiers to conduct planning, execution, and analysis tasks in ISR scenarios with a range of unmanned systems, communications technologies, and battle command systems. That these disparate battle command systems could be engineered in a mobile ad hoc network was a feat of merit in itself; the additional fact that the soldiers could adaptively work between these systems was of particular interest. The soldiers combined their experience with personal and Army computers with a range of classroom and field training in preparation for the field experiment and performed with low workload and high situational awareness against an adaptive enemy force.

This analysis investigated the strength of the theoretical underpinnings of the UBC vision. The methodological approach of an integrated analysis of a system-of-systems MANET represents a pioneering first step toward the goal of network optimization for human understanding in a mission-relevant environment. This method showed promise in mapping performance metrics on behavior across the network levels. Future efforts will focus on the continued resolution of quantifiable quality of service metrics, system performance characteristics, and human decision making. ❑

ELIZABETH K. BOWMAN, PH.D., is an operations research analyst with a behavioral science background and serves currently as the Social Network Analysis team leader for the Army Research Lab, Computational and Information Sciences Directorate. She has experience in applied field and laboratory experiments that examine distributed collaborative environments at the operational and tactical levels of command. She is the chairperson of The Technical Cooperation Program (TTCP), C3I Technical Panel 2, Command Information Interfaces, where she interfaces collaboratively with defense scientists from Australia, Canada, United Kingdom, and the U.S. Service research laboratories. These alliances provide Dr. Bowman with a varied perspective on challenges associated with networked communications for C2, to include coalition, operational, and tactical echelons. E-mail: liz.bowman@us.army.mil

RANDY ZIMMERMAN, PH.D., recently retired from his first career as a distinguished U.S. Army Infantry Officer who served in a wide variety of leadership and command positions. During his time in the Army, he led teams as small as 3–4 to more than 600 people. Since 1996, Randy has focused his operations management talent on solving complex problems for the federal government. He has a solid record of performance in all aspects of operations research from requirements definition to data postprocessing and presenting results, as well as a proven record of identifying core issues and achieving results under pressure, on time, and under budget. While serving in the Army, Randy worked on a wide variety of problems ranging from logistics issues to information management systems. His dissertation supported the Defense Logistics Agency in identifying the least efficient supply depots around the world and determining what measures were necessary to improve system efficiency. Since retiring from the Army in 2005 and forming the Zimmerman Consulting Group, he continues leading large groups of people focused on solving complex problems. E-mail: r.zimmerman@zcgrp.com

Endnotes

[1]Network-Centric Warfare is also commonly referred to as Network-centric operations or Network-enabled operations. For ease of discussion, we will use the NCW term in this article to generically refer to the concept that the network is providing information communications technologies to Warfighters.

[2]Forces use the FBCB2 Blue Force Tracker capability to navigate when mobile. A 1-minute COP update rate would have the following impact: a vehicle traveling 20 mph would cover .5 miles in 1 minute. This could be an acceptable update rate or not, depending upon the nature of the navigated terrain. Such a rate would clearly be inadequate for city or village navigation.

[3]CERDEC enhancements to the FBCB2 display included windows to provide chat, sensor image links, and a view of network health status of all vehicles in the network. These enhancements were made by linking additional programs to the display.

[4]GEC2O was developed by Mechdyne, Future Skys, and JB Management at the direction of the U.S. Army CERDEC C2 Directorate. GEC2O was used in this study as a tool for replaying the missions in order to effectively understand the sequence and timing of events, which assisted in scoring SA measures.

References

Anna, K., and M. Bassiouni. 2006. Evaluation of packet latency in single and multi-hop WiFi wireless networks. In *Proceedings of SPIE: Wireless sensing and processing,* April 2006, Orlando, Florida, ed. Raghuveer M. Rao, Sohail A. Dianat, and Michael David Zoltowski, Vol. 6248. Bellingham, WA: SPIE.

Bowman, E., and J. A. Thomas. 2008. C2 of unmanned systems in distributed ISR operations. In *Proceedings of the 13th International Command and Control Research Technology Symposium,* June 17–19, 2008, Seattle, Washington, Article ID 083. Washington, D.C.: CCRP.

Bowman, E., and J. A. Thomas. 2008. Cognitive impact of a C4ISR tactical network. In *Proceedings of*

the 14th International Command and Control Research Technology Symposium, June 15–17, 2009, Washington, D.C., Article ID 069. Washington, D.C.: CCRP.

Bowman, E., and S. Kirin. 2006. The impact of unmanned systems on platoon leader situation awareness. *Proceedings of the 2006 Command and Control Research Technology Symposium,* June 20–22, 2006, San Diego, CA. Washington, D.C.: CCRP.

Chiang, C. J., R. Chadha, S. Newman, R. Orlando, K. Jakubowski, Y. Kumar, and R. Lo. 2008. *Building a versatile testbed for supporting testing and evaluation of tactical network management tools and their interoperability*. Piscataway, NJ: IEEE.

Conner, W. D. 2005. *Understanding first in the contemporary operational environment*. Ft. Leavenworth, KS: School of Advanced Military Studies, U.S. Army Command and General Staff College.

Cresswell, J. W. 1998. *Qualitative inquiry and research design: Choosing among five traditions*. Thousand Oaks, CA: Sage.

Gao, R. X., and T. Boscardin. 2006. Design of a test bed for performance evaluation in a wireless sensor network. In *IMTC: Proceedings of Instrumentation and Measurement Technology Conference,* April 24–27, 2006, Sorrento, Italy. San Ramon, CA: IMTC.

Iannone, L., K. Kabassanov, and S. Fdida. 2007. Evaluation of cross-layer-aware routing in a wireless mesh network test bed. *Journal on Wireless Communications and Networking*, Vol 2007, Article ID 86510.

Ikeda, M., L. Barolli, M. Hiyama, T. Yang, G. De Marco, and A. Durresi. 2009. Performance evaluation of a MANET testbed for different topologies. In *IEEE: Proceedings of 2009 International Conference on Network-Based Information Systems,* August 19–21, 2009, Indianapolis, IN, 327–34. Piscataway, NJ: IEEE.

Latham, J. 13 November 2008. JTRS Basis of Issue (BOI) for the Current Force. Washington, D.C.: Army Science Board.

Moore, D. 2008. Command and Control Roadmap Briefing. Track 7, Session 3, Acquiring, Fielding, Sustaining LandWarNet.

National Research Council (NRC) of the National Academies. 2005. *Network Science*. Washington, D.C. http://www.nap.edu (accessed March 24, 2009).

Office of Force Transformation. 2005. The Implementation of Network-Centric Warfare. Washington, D.C.: Office of the Secretary of Defense.

PM C4ISR On The Move. 2005. C4ISR On-The-Move Testbed 2005 Experiment. Ft. Monmouth, NJ: Author.

PM C4ISR On The Move. 2006. 2006 Capstone Experiment Final Report. Ft. Monmouth, NJ: Author.

PM C4ISR On The Move. 2007. C4ISR On-The-Move Event 07 Final Report. Ft. Monmouth, NJ: Author. https://www.us.army.mil/suite/doc/9627504 (accessed November 12, 2009).

PM C4ISR On The Move. 2008. C4ISR On-The-Move Event 07 Final Report. Ft. Monmouth, NJ: Author.

PM C4ISR On The Move and Army Research Laboratory. In press. Unified Battle Command Cognitive Impact Study. Ft. Monmouth, NJ: PM C4ISR OT.

Porche, I., L. Jamison, and T. Herbert. 2004. *Framework for measuring the impact of C4ISR technologies and concepts on Warfighter effectiveness using high resolution simulation*. Santa Monica, CA: RAND. http://oai.dtic.mil/oai/oai?&verb=getRecord&metadataPrefix=html&identifier=ADA466098 (accessed March 24, 2009).

ITEA Journal 2010; 31: 232–239

The Cognitive Performance Component in Networked System of Systems Evaluation

B. Diane Eberly
U.S. Army Evaluation Center, Aberdeen Proving Ground, Maryland

The Army's Future Force requirements contain ample descriptions of the physical architecture for manned and unmanned systems, and a multilayer network to which the Army will eventually migrate. Requirements specifications that will allow those entities to seamlessly function as an interoperable and integrated entity also exist. However, few descriptions exist of the cognitive performance requirements that will be essential for the individual, team, and units to perform command and control function because they increasingly rely on a networked system of systems. Even more elusive are the methods for evaluating, in an operational environment, whether the cognitive performance requirements have been met. Increased task complexity, uncertainty, workload, distributed command and control, battlefield visualization, and situational understanding are but a few of the areas that a future networked system of systems design is required to address.

The volume of communications and information exchange within and between layers of command call for simplification in tools and processes provided to warfighters. Networked-enabled command and control must allow warfighters to manage these increasing demands at operational tempos that drive proactive versus reactive maneuvers against a highly adaptive threat. This article describes a number of factors that affect the networked system of systems' ability to enhance cognitive performance and support the levels of coordination and collaboration required for distributed command and control in a complex battle space. Considering these factors in the evaluation of a networked system of systems is important given the increased levels of higher order cognitive processing necessary to operate in such an environment.

Key words: Cognitive performance; networked system of systems.

The 2009 Army Posture Statement characterizes the global security environment as "more ambiguous and unpredictable than in the past." Recent conflicts have demonstrated greatly increased complexity in planning and executing the range of military operations required for engaging in irregular, nontraditional, disruptive, and sometimes catastrophic warfare. As amplified in the Capstone Concept for Joint Operations (DoD, 2009), rapidly changing conditions in the joint operations area are posing significant challenges for commanders and are creating taxing demands on their ability to generate courses of action (COAs). These COAs are often in collaboration with other service, multinational, and interagency contributors to combat highly adaptive adversaries utilizing unconventional tactics (Department of the Army, 2007). These challenges can only be met with human and technological capabilities that enable greater levels of collaboration and information sharing, which then lead to self-synchronization. Aberts and Hayes (2003) identified these COAs as critical capabilities for the "next revolution in military affairs," more prominently known as network centric warfare, described by Cebrowski and Garstka (1998).

Unarguably, the greatest capability demand for achieving victory in the range of military operations is the development of technologies, techniques, tactics, and procedures that support the generation of proactive COAs. Such capability demands require effective employment of sensors, systems, and decision aids designed to deliver human and sensor inputs, fuse them

into information sources supplied to increase situational awareness, and support making resolute decisions that can increase speed of command. Clearly, these demands cannot be met through the application of human abilities alone. Several types of technologies are required to accomplish this feat. Further, these technologies cannot operate in isolation. Implementing network centric warfare requires assurance that a multitude of networked systems can communicate and exchange information, hence a networked system of systems (SoS); there must be assurance that they provide the qualities and characteristics that complement the way warfighters execute missions, versus providing ones that may ultimately burden them. Kevin P. Byrnes, General, United States Army, Commanding, said the following (Department of the Army, 2005):

> *"Technological advances alone will not constitute transformation. Our most critical asset is not technology, but the critical thinking of our Soldiers and leaders." (Department of the Army, 2005; Kevin P. Byrnes—General, United States Army, Commanding)*

A networked SoS naturally implies using computer technology to manage the transmission of information between two or more systems, or between devices attached to them. Conley (2009) further describes these systems to include weapons and vehicle platforms as nodes within this SoS and that these platforms as network nodes must be equipped with communications capabilities such that a network "sees" them no differently from any other networked system. These capabilities are usually generated from efforts within the discipline of computer networking, which can be considered a subdiscipline of telecommunications, computer science, information technology, or computer engineering. In practice, however, it's the integration of capabilities from these subdisciplines that synergistically provide the capability to move information, in a variety of formats, from one location to another. To some this capability seems quite mysterious; to others it's the simplistic application of modern technology, but there's likely little argument that it becomes a very complex issue to all when you also consider requirements for this capability to provide the exact information an individual needs to make decisions and operate systems anytime he or she needs that information, and from any location around the globe.

From an engineering point of view, developing a networked SoS is only limited by the available technology, but from the users' point of view, the requirements go far beyond technological feats. The most technologically advanced networked SoS could become a hindrance rather than an aid to the user if it doesn't support the way he or she needs to use it. Of course, people can usually modify their behaviors to adapt to technology, but there are potential risks in adopting an attitude that users *should* do that instead of providing technologies designed to better coincide with their behaviors.

Within the context of the Army's concept for unified battle command (UBC), the remainder of this article focuses on the factors that shape the way people use networks and networked SoS, influence how effectively they are able use the embedded capabilities, and determine how much they are willing to rely on them to execute their tasks. In addition, a number of other factors that shape not only individual performance but also collectively shape the performance of teams and organizations planning and executing net-centric operations.

Unified battle command

The UBC concept identifies a strategy that utilizes a SoS approach to federate several battle command applications. This federation delivers a suite of integrated battle command functions applicable to all Army echelons. The training and doctrine command capabilities manager for battle command identified the battle command essential capabilities (BCEC) that are essential for commanders to execute battle command in the realm of full spectrum operations. The integrated suite of applications is designed to provide a "robust" network, seamless sharing, and displays of "relevant" geospatial information, and a "standard" collaboration capability within and across all command levels, to include extension to the individual soldier. These are considered to be the critical components of the BCEC, supported by the UBC concept, and are further described in the following discussion.

Battle Command Essential Capabilities (BCEC)

An integrated capabilities development team for battle command identified 10 essential capabilities to implement the UBC concept. Those capabilities are (Department of the Army, 2008):

1. A robust network capability. The force must possess a commander-centric, secure, integrated, and adaptable communications network consisting of line-of-sight and beyond-line-of-sight means.
2. Execute tactical network operations. Commanders need the ability to have effective tactical network operations (network management) conducted and provide guidance to allow allocation of network resources to maximize performance through all phases of the joint phasing model.

3. Display and share relevant information. The Army's battle command system must enable the receipt and dissemination of essential information for display on the common operational picture from dismounted soldier through army-level command posts. This includes symbols, graphic control measures, friendly and enemy information, civil considerations, and the operational environment from disparate information systems.
4. A standard and sharable geospatial foundation. Commanders and leaders need common geospatial information to enable all battle command essential information requirements, create a common map foundation, and display and share this information on a tailorable and interoperable common operational picture.
5. Enable collaboration. Commanders and leaders need a common suite of collaborative tools to allow establishment of a collaborative environment to achieve shared understanding and ensure unity of effort in both high and low bandwidths.
6. Create and disseminate orders. The Army's battle command system must be able to create, change, and distribute mission orders (both voice and written) to include attached graphics between command post, platforms, and leaders.
7. Battle command on-the-move. The commander must have the ability to maintain situational awareness, make timely and informed decisions, and position himself at the decisive point during the battle.
8. Execute a running estimate. The Army's battle command system must be able to support running estimates by continuously gathering and tracking information to support tactical decision making by providing a continuous assessment of current and future operations, including conclusions and recommendations.
9. Joint, interagency, intergovernmental, and multinational interoperability. The Army's battle command system must be able to exchange relevant operational information with joint, interagency, intergovernmental, multinational partners; nongovernmental organizations; and contractors.
10. Rehearsal and training support. The Army's battle command system users must be capable of preparing for operations using embedded rehearsal and training tools that accurately represent the spectrum of missions and environments.

In addition to the BCEC, UBC calls for implementation of the battle command framework of the Training and Doctrine Command 2007 as shown in *Figure 1*. This framework identifies 10 functional battle command concepts. Each of the BCEC and functional battle command concepts has a technological solution within the networked SoS, but there are a number of factors that need to be addressed to support the cognitive performance of users of the networked SoS.

Significant advances have been made in technologies to acquire and move abundant levels of information across a networked SoS and deliver it to the appropriate user. However, human brains cannot consume and make sense of the sheer volumes of information presented within a timeframe to make it usable for planning and conducting high operational tempo (OPTEMPO) missions. The remainder of this article provides a comprehensive, though not exhaustive, list of factors that affect cognitive performance and how some of those impacts degrade effective use of a networked SoS to execute battle command processes.

Factors affecting cognitive performance in net centric operations

While the BCEC and functional battle command concepts support requirements definition for a networked SoS designed to support the battle command framework, those requirements ultimately serve to optimize collaboration, situational understanding, visualization, and information sharing capabilities. The end state for optimizing these capabilities is optimized command decision making (CDM), both in speed and effectiveness. It is the speed at which commanders receive the information they need to develop the necessary levels of situational understanding; visualize the battlefield; collaborate with other services, nations, or agencies; and make effective decisions to achieve desired end states that support achieving the ultimate end state of thwarting or defeating intentions of the adversary. *Figure 2* provides a graphic illustration of the dependent relationships between these capabilities that support CDM. This premise serves as the foundation for the discussion that follows regarding the factors that contribute to, or detract from the development of these capabilities.

There are a number of factors that contribute to the development of capabilities that affect speed and effectiveness of CDM. These factors are both internal and external to the decision maker. Internal factors relate to those that *form* the inherent abilities of the decision maker, while external factors are those that *affect* the inherent abilities of the decision maker. *Figure 3* outlines a number of the internal and external factors that impact cognitive performance, and hence contribute to developing the primary capabilities that

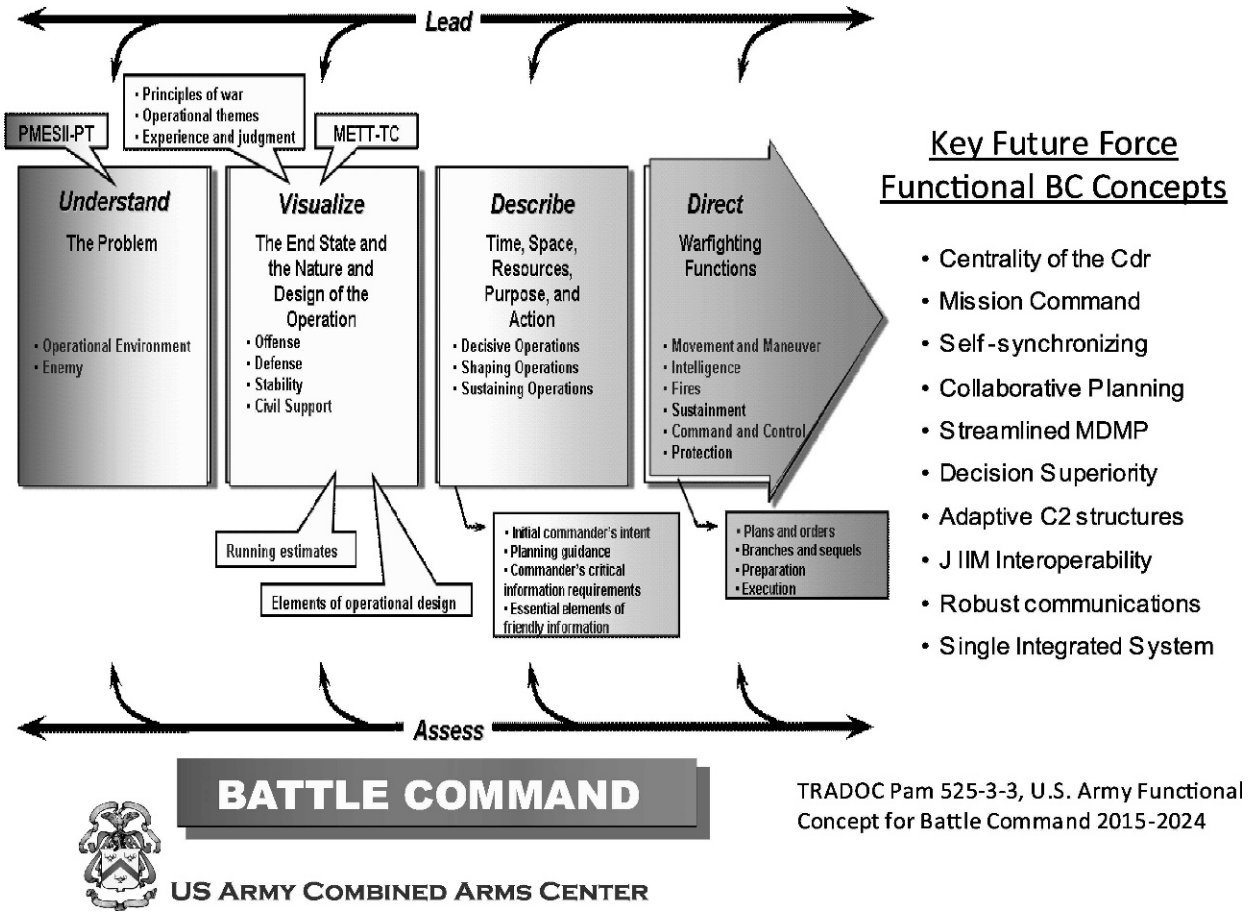


Figure 1. The battle command framework.

support CDM (see *Figure 2*). Internal factors are grouped into the categories of ability, disposition, and state, and external factors are grouped into automation, environment, and team. This is certainly not an exhaustive list, and the intent of this article is not to delve into all aspects of the effect of these factors on cognitive performance but to identify some of those that can be observed and/or measured for their impact on the primary capabilities supporting CDM. The following section is dedicated to identifying these factors, providing a brief description of each, and some of their impacts on cognitive performance, and why it is important to consider them in the test and evaluation of a networked SoS.

Internal factors

Shaping factors that affect cognitive performance and are internal to an individual are those he or she brings to the situation. They are a culmination of the things that are inherent to the individual's genetic makeup, learning experiences he or she has been exposed to, references committed to memory, and physiological condition(s) that determine his or her ability to cope with a set of circumstances in an operational setting. The author has grouped these factors into three areas: (1) ability, (2) disposition, and (3) state.

1. Ability. Shaping factors in this area primarily center on formal education, occupational training, and repeated exposure to events from which an intuitive reaction or thought process is developed and repeated when a like event is presented again. These are similar to learned behaviors such as testing how hot a cup of coffee is before taking a drink because a burned tongue had resulted sometime in the past when the temperature was unknown.

Professional military education provides the foundation for knowledge that warfighters need to plan and conduct military operations. Training provides warfighters the opportunities to use the knowledge gained from professional military education. Skills and abilities are the by-products of education and training. They are what has been gained by the individual as a

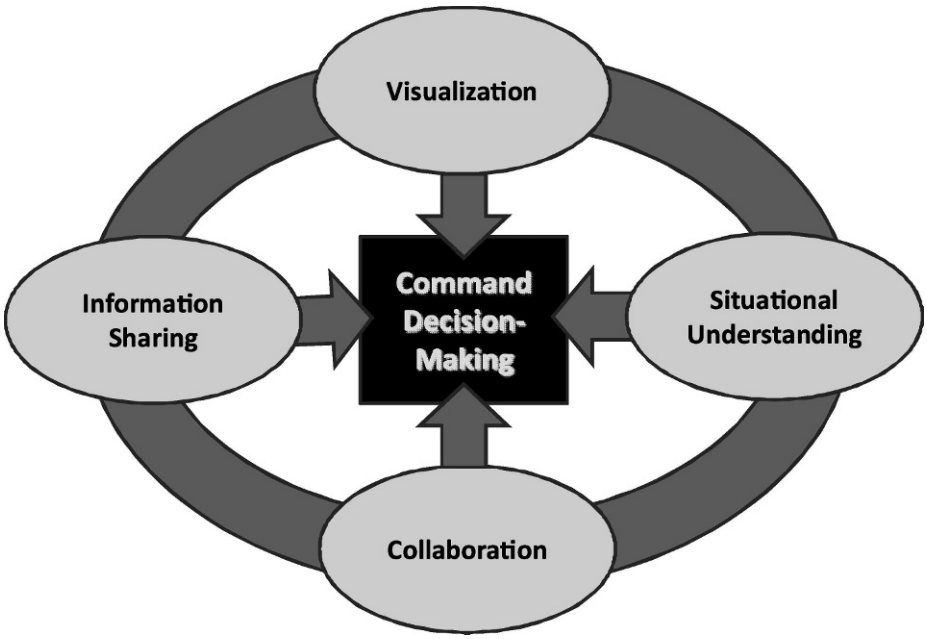


Figure 2. Primary capabilities supporting command decision making.

result of prior education and training. In a situation requiring a decision, education, training and the resulting skills and abilities obtained influence the quality of the decision made. It is the repetition and richness of experiences, however, that provide decision makers the opportunity to store in memory an intuitive response to a situation. Those stored memories contribute to developing a mental model that can be used in a similar situation and thus increase the speed at which the decision is made. Klein et al. (1993) refer to this as recognition-primed decision making, which has since become the accepted theory for how individuals make decisions in complex, time-constrained circumstances, provided the mental model for that situation has been developed.

With respect to the cognitive performance of an individual engaged in planning and executing military operations using a networked SoS, the abilities of the individual must be considered when evaluating the performance of the networked SoS. Because the capabilities of the networked SoS must support a decision maker in developing the most effective and rapid decisions he or she can, the abilities of that person must be considered before evaluating whether the networked SoS has met the prescribed requirements.

Internal Factors

- **Ability**
  - Education
  - Training
  - Experience
  - Skills/Competencies
  - Mental Models
- **Disposition**
  - Personality
  - Culture
  - Motivation
  - Need for Cognitive Structure
  - Risk Aversity
  - Uncertainty Tolerance
- **State**
  - Workload (physical/mental)
  - Fatigue
  - Nutrition
  - Awareness
  - Trust
  - Uncertainty

External Factors

- **Automation**
  - Networks
  - Decision Aids
  - Collaboration Tools
  - Information Filters/Fusion
  - Presentation (visual/auditory, other)
- **Environment**
  - Temperature
  - Austerity
  - Motion
  - Threat/Safety
  - Facilities
- **Team**
  - Shared Awareness
  - Unity
  - Back-up
  - Communication
  - Role Identity

Figure 3. Shaping factors that affect cognitive performance and the development of primary capabilities supporting CDM.

2. Disposition. Everyone comes into the world with inborn traits. In addition, people develop additional traits as a result of experiences throughout their lives. Regardless of their source, these traits influence the ways in which individuals use technology. Traits attributed to personality and culture factors tend to influence how people feel about technology, while factors such as motivation, need for cognitive structure, risk adversity, and uncertainty tolerance tend to influence how and how much people use technology. For instance, the more adventuresome and adaptive an individual, the more likely he or she will explore the potential for technology to support accomplishment of tasks. Conversely, the more risk adverse or lack of tolerance for uncertainty, the less likely he or she will seek help from technology.

Motivation to use technology can stem from a number of sources; however, necessity can often rule over desire depending on the perceived or real usefulness and/or ease of use of the technology. For the most part, the networked SoS supporting planning and execution of military operations leads to necessity for using it, unless it is possible for the user to develop a work-around that he or she finds more satisfying than relying on the technology available to accomplish a task.

An individual's need for cognitive structure refers to how much ambiguity in the information he or she is obtaining can be tolerated before it leads to dissonance, stress, uncertainty, or confusion. According to Roney and Sorrentino (1987), people are either certainty or uncertainty oriented. Certainty-oriented people tend to bin information as valid or invalid, and ignore information that is either inconsistent or ambiguous. Uncertainty-oriented people, on the other hand, have a greater ability to deal with the same ambiguous or inconsistent information by binning it as such and allowing the possibility for its usefulness. Therefore, when validating whether a networked SoS meets the users' requirement through the use of feedback from the user, it is necessary to know how much uncertainty users can tolerate before evaluating whether the prescribed requirements have been met.

Risk adversity and tolerance for uncertainty also determine how much an individual is willing to trust whether technologies provided will actually do what the individual is told they will do. If trust is high, willingness to use (without doubt) the technology will be higher. If trust is low, the individual's willingness to rely on the technology decreases. The factor of trust is further examined in the next section on state.

3. State. Shaping factors in this area are the physiological conditions that affect an individual's normal cognitive performance abilities. Considerable research has been conducted in assessing the effect of state factors on cognitive performance. Assessing cognitive performance in a net-centric environment is more critical to the evaluation of system or SoS performance because there are greater cognitive demands on individuals in a net-centric information-driven environment. Wesensten, Belenky, and Balkin (2005) explain that the ability to integrate information, anticipate, and plan depends on the brain's prefrontal cortex to execute. Various physiological stressors (or performance shaping state factors), such as high workload (physical and mental), fatigue, and poor nutrition degrade the functioning of the prefrontal cortex, and by extension degrades cognitive performance in general.

Lack of trust, too much uncertainty, and poor situational awareness, while not direct stressors to the prefrontal cortex, can also degrade cognitive performance. When trust in the information provided from technological devices becomes low, individuals feel uncertain that they have the necessary information to make a decision, which leads to increases in decision making time. If they learn to distrust the technology, or in this case the networked SoS, they can fall into patterns of ignoring the information produced when they shouldn't. If they become overconfident that the networked SoS is flawless and begin to overtrust the information, never questioning its validity, equally bad decisions can be made with equally bad outcomes.

Situational awareness is in itself a key capability supporting CDM and has received the recognition as being the primary enabler for decisive victory in planning and executing military operations. If the information provided from a networked SoS is not sufficient, poorly represented or formatted, and/or lacks the salient cues to adequately result in an accurate, current, and relevant level of situation awareness, decision makers are at risk for making poor decisions, not because of their inherent abilities, but because the networked SoS has not provided the right capabilities to ensure that the decision maker is equipped to make optimal decisions.

Understanding an individual's state while engaged in the use of a networked SoS is important for evaluating whether it meets the user's requirement for it to support execution of military operations.

External factors

Shaping factors that affect cognitive performance and are external to an individual are those that the

situation and surroundings impose on him or her. Just as with internal factors, they are a culmination of things that determine his or her ability to cope with a set of circumstances in an operational setting. The author has grouped these factors into three areas: (1) automation, (2) environment, and (3) team.

1. Automation. Shaping factors in this area primarily center on the capabilities within the networked SoS. They include the networks themselves, and all the tools and aids that collect, reduce, organize, present, and transmit information for use in the planning and execution of operations. Executing and planning the full range of military operations in the current joint operations area requires unprecedented levels of information that can be trusted and acted on immediately. It is critical that the networked SoS be fully operational nearly 100% of the time so that decision makers have all of the information they need (without an overload of unnecessary information) in a timely manner to support the development and execution of proactive COAs designed to interrupt and/or outpace the adversary's decision cycle.

Evaluation strategies must include decision makers' cognitive performance in this kind of net-centric environment. Assurance that the networked SoS is providing the opportunity to keep a full level of situational understanding, an accurate battle space visualization through a current common operational picture, the ability to share that information with mission collaborators through effective and efficient collaboration tools, and the timely delivery of commander's intent are among the more critical capabilities for the networked SoS. However, the technological solution to providing those capabilities is not the only consideration; they must support the rapid OPTEMPO and unique ways in which warfighters wish to use the networked SoS to plan and execute missions. Evaluating these capabilities needs to occur in the same mission context and OPTEMPO as real-world operations to determine if the networked SoS is suitable for the users' needs.

2. Environment. In the section on internal factors, it was noted that an individual's state can affect the functioning of the prefrontal cortex, which is responsible for higher level cognitive functions (Wesensten, Belenky, and Balkin 2005). They expound on that by stating environmental conditions can cause sufficient degradation in an individual's physiological state to result in impaired functioning of the prefrontal cortex, and thus his or her ability to successfully execute tasks requiring higher levels of cognitive functioning, such as decision making.

With that understanding, it is important to evaluate the environmental conditions of a test environment for its potential to affect cognitive functioning because degrading that ability can cause the user of a networked SoS to improperly use it, perform poorly while using it, or revert to a more habitual method of executing a task that bypasses using the technology altogether.

3. Team. The Army is acquiring more complex manned and unmanned systems, of which many require more than one person to set up, calibrate, operate, monitor, and/or interact. The complexity of the systems may require one individual to attend to the system nearly 100% of the time, thus requiring other individuals to monitor the environment, send and receive sensor information, or conduct a variety of other tasks depending on the unit and mission. The ability of the team, which may be colocated or distributed, to collaborate and synchronize tasks requires the team to have complete understanding of what every other member of the team needs to accomplish and where each person is in the process (Cooke et al. 2000). Additionally, when the team understands the commander's intent, and they share a good mental model with the commander, team processes improve (Serfaty, Entin, and Johnston 1998). Members of the team are able to back one another up, anticipate what actions another member is about to take, and interpret cues that might indicate excessive overload or stress on a team member.

Requirements for systems, and especially networked SoS, typically do not identify the team processes that are necessary for employment. Evaluators, therefore, should be cognizant of this and establish derived measures to evaluate the ability of the systems or networked SoS to support not only individual collaboration, visualization, situational understanding, and information sharing needs, but also those of the team.

Summary

The intent of this article is not to prescribe specific test and evaluation strategies and measures, but to make the readers cognizant of the factors that have the potential to influence how users will use a networked SoS and how well the networked SoS can support the users' needs. While the Department of Defense does not directly procure warfighters, it does invest heavily in training, educating, and otherwise "making ready" warfighters to conduct missions to defend our country

and promote peace around the world. As such, when we consider the capabilities we choose to procure for their use, we must consider not just the best technological solution, but the solution that best supports optimizing warfighters' performance, reducing demands on their already heavily taxed physical and cognitive abilities. Admittedly, designing test plans to examine these factors can be difficult, and certainly not all can be incorporated into an evaluation strategy. However, those systems or SoS that are being acquired to support higher order cognitive processes such as developing situational awareness, supporting collaborative processes, and visualizing the battle space should consider factors that shape the warfighters' abilities to do so. ❑

DIANE EBERLY is currently employed as an operations research systems analyst at the Army Test and Evaluation Command, Army Evaluation Center at Aberdeen Proving Ground, Maryland. She has also had the opportunity to work at the Joint Forces Command Experimentation Directorate (J9) in Suffolk, Virginia, as the field element chief for the Army Research Laboratory, Human Research and Engineering Directorate. Additionally, Diane was employed by the Commander, Navy Submarine Force in Norfolk, Virginia, as the Branch Head for Plans and Programs, and for Warfare Assessments. She received her master of science degree in information systems and operations analysis from the University of Maryland Baltimore County in 1997. Diane has over 25 years of research, development, test and evaluation experience, primarily focused on evaluating human systems integration for weapons and information systems in the defense acquisition lifecycle. E-mail: Diane.Eberly@us.army.mil

References

Alberts, David S., and Richard E. Hayes. 2003. *Power to the Edge,* DoD Command and Control Research Program; June; ISBN 1-893723-13-5; http://www.dodccrp.org/publications/pdf/Alberts_Power.pdf (accessed September 3, 2009).

Cebrowski, A. K. and J. J. Gartska, 1998. Network-centric warfare: Its origin and future. *Navy Proceedings* (January) 28–36.

Conley, S. F. 2009. Test and evaluation strategies for network-enabled systems. *ITEA Journal* 30 (1): 111–116.

Cooke, N. J., E. Salas, J. A. Cannon-Bowers and R. J. Stout. 2000. *Human Factors*. 42, 151–173.

Department of the Army. 2005. Training and Doctrine Command (TRADOC) PAM 525-3-0, v2.0 (2005). *Army*. http://www.tradoc.army.mil/tpubs/pams/p525-3-0.pdf (accessed November 25, 2009).

Department of the Army. 2007. *Training and doctrine command (TRADOC) PAM 525-3-3,* v1.0. *The United States Army functional concept for battle command 2015–2024*. http://www.tradoc.army.mil/tpubs/pams/p525-3-0.pdf (accessed November 25, 2009).

Department of the Army. 2008. *Training and Doctrine Command (TRADOC) capabilities manager—Battle command (TCM-BC). Battle command essential capabilities whitepaper*. http://www.tradoc.army.mil/tpubs/pams/p525-3-0.pdf (accessed November 25, 2009).

DoD (Department of Defense). 2009. Capstone concept for joint operations Version 3.0 (CCJO v3.0), January. *Future force capstone concept 2015–2024*. www.dtic.mil/futurejointwarfare/concepts/approved_ccjov3.pdf (accessed November 25, 2009).

Klein, G., J. Orasanu, R. Calderwood, and C. E. Zsambok. 1993. *Decision making in action: Models and methods*. Norwood, NJ: Ablex Publishing.

Roney, C. J. R., and R. M. Sorrentino, 1987. Uncertainty orientation and person perception: Individual differences in categorization. *Social Cognition* 5, 369–382.

Serfaty, D., E. E. Entin, and J. Johnston. 1998. Team coordination training. In *Decision making under stress: implications for training and simulation*, ed. J. A. Cannon-Bowers and E. Salas. Washington, D.C.: APA Press. pp. 221–245.

Wesensten, N. J., G. Belenky, and T. J. Balkin. 2005. Cognitive readiness in network-centric operations. *Parameters,* 35, 94–105.

ITEA Journal 2010; 31: 240–246

Integrating Situation Awareness Assessment Into Test and Evaluation

Cheryl A. Bolstad, Ph.D. and Haydee M. Cuevas, Ph.D.
SA Technologies, Marietta, Georgia

To guarantee the success of network-centric operations, warfighters need the ability to extract and share critical task-relevant information to develop and maintain the situation awareness that is so critical for effective team performance. As such, the design of emerging technologies and systems must adopt a "user-centric" approach, with consideration for human information processing capabilities and limitations. In turn, to ensure that these technologies and systems are meeting their design objectives, test and evaluation must similarly be expanded to include metrics that assess how well system features and functions are supporting critical human cognitive processes such as situation awareness and decision-making. In this article, we address this issue, focusing specifically on situation awareness. We discuss how situation awareness assessment, at both the individual and team level, can be integrated into test and evaluation. We also cite examples from our own research to demonstrate the diagnosticity afforded by situation awareness assessment.

Key words: Decision-making, diagnostics, human cognition, information technology, network centric warfare, team performance.

Network centric warfare promises to provide revolutionary command, control, and communications capabilities. With this increased network-centricity, the state of current military operations is shifting from traditional large command and control centers to small groups working together in a distributed manner through the use of information technology. Although advances in information technology are enabling this drive toward network-centricity through the development of networked databases, greater bandwidths, and more sophisticated collaboration tools, the deciding factor is how human operators will be able to work collaboratively to capitalize on this enormous volume of available information (Lawlor 2005).

To guarantee the success of network-centric operations, warfighters need the ability to extract and share critical task-relevant information to develop and maintain the situation awareness (SA) that is so critical for effective team performance. As such, the design of emerging technologies and systems must adopt a *user-centric* approach, with consideration for human information processing capabilities and limitations. In turn, to ensure that these technologies and systems are meeting their design objectives, test and evaluation must similarly be expanded to include metrics that assess how well system features and functions are supporting critical human cognitive processes such as SA and decision-making. In this article, we address this issue, focusing specifically on SA. We begin with a brief overview of SA, defining this construct at both the individual and team level. We then discuss how SA assessment can be integrated into test and evaluation, citing examples from our own research.

SA defined

Although several different definitions of situation awareness have been put forth in the literature (Fracker 1991; Sarter and Woods 1991; Smith and Hancock 1995), in this article, we focus on Endsley's (1995b) theoretical model of SA, which defines this complex cognitive construct as "the perception of elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future" (Endsley 1995b, 36). As implied by this definition, SA involves being aware of what is happening around you to understand how information, events, and your own actions will affect your goals and objectives, both now and in the near future. Endsley's definition highlights three levels of SA: perception, comprehension, and projection.

Perception (Level 1 SA) involves an active process whereby individuals extract significant cues from their environment, selectively directing attention to important information, while disregarding nonrelevant items. *Comprehension* (Level 2 SA) involves integrating this information in working memory to understand how the information will influence the individual's goals and objectives. *Projection* (Level 3 SA) involves extrapolating this information forward in time to determine how it will affect future states of the operating environment. Consideration of these three levels of SA is useful for understanding the types of difficulties human operators face while performing their tasks and also for determining how best to mitigate these challenges.

At the team level, SA can be viewed in terms of both team SA and shared SA. *Team SA* can be defined as "the degree to which every team member possesses the SA required for his or her responsibilities" (Endsley 1995b, 39). Thus, to ensure successful performance, each team member needs to have superior SA on those factors that are relevant for his or her job. In contrast, *shared SA* can be defined as "the degree to which team members possess the same SA on shared SA requirements" (Endsley and Jones 2001, 48). A major part of teamwork involves understanding the SA requirements that are relevant across multiple team members. Successful team performance, therefore, is influenced by the degree to which team members share a common understanding of what is happening on these shared SA elements. In other words, team members must be able to access and similarly interpret important information on the shared SA requirements that are relevant across their different positions.

Role of SA in human performance

SA represents one of the most challenging aspects of human performance. In particular, in most complex tasks, effective decision-making largely depends upon the degree to which individuals have developed a good understanding of the situation, namely, their SA. SA is especially crucial in domains where information flow can be quite high and poor decisions may lead to serious consequences (e.g., piloting an airplane, functioning as a soldier, treating critically ill or injured patients). Indeed, SA has been recognized as a critical, yet often elusive, foundation for successful decision-making across a broad range of complex and dynamic systems, including aviation and air traffic control (Nullmeyer et al. 2005), emergency response and military command and control operations (Blandford and Wong 2004; Gorman, Cooke, and Winner 2006), and offshore oil and nuclear power plant management (Flin and O'Connor 2001). Lacking SA or having inadequate SA has been consistently identified as one of the primary factors in accidents attributed to human error (Hartel, Smith, and Prince 1991; Merket, Bergondy, and Cuevas-Mesa 1997; Nullmeyer et al. 2005). Yet, developing and maintaining SA imposes high cognitive demands upon human operators in terms of time, attention, and effort. Fortunately, the cognitive load associated with achieving high levels of SA can be mitigated through SA-oriented system design (see Endsley, Bolte, and Jones 2003) and SA-oriented training programs (Strater and Bolstad 2009). Hence, test and evaluation plays a major role in ensuring that these systems achieve their design objectives, a topic we turn to next.

SA assessment

An SA-oriented approach to test and evaluation goes beyond simply assessing a system's functional capabilities to also include how well the system's design supports human operators' critical cognitive processes underlying SA and decision-making. At the team-level, this also includes evaluating the system's effectiveness in supporting the team's ability to assess and track coordination, communication, collaboration, and information-sharing activities. In general, methodologies to assess SA vary in terms of direct measurement (e.g., objective real-time probes or subjective questionnaires assessing perceived SA) or indirect methods (e.g., process indices, trained observer ratings) that infer SA based on operator physiological state, behavior, or performance. Direct measures are typically considered to be "product-oriented" in that these techniques assess an SA outcome; indirect measures are considered to be "process-oriented," focusing on the underlying processes or mechanisms required to achieve SA (Graham and Matthews 2000). Selecting which methodology to use depends upon the researcher's objectives and what data collection facilities or setup is available. Examples of each of these SA measurement approaches will be further described next.

Process indices

Process indices, such as psycho-physiological measures, examine how individuals process information in their environment (Wilson 2000). Such measures include electroencephalography (EEG), event-related potentials (ERP), event-related desynchronization (ERD), heart rate variability (HRV), electrodermal activity (EDA), eye blinks, and eye tracking. Tracking eye movements, in particular, is one of the more common psycho-physiological approaches for providing insight into perception and comprehension. Eye-tracking devices can be used to monitor where

operators are directing their attention, and thereby, determine whether the saliency of important cues is sufficient or if nonessential cues are drawing away the operator's attention. Analyzing communications can also serve as process indices of operator SA. For example, verbalizations made by operators during a task can be analyzed to determine how well information is being acquired from a system designed to support this task.

Process indices are advantageous in that these offer objective assessment of operator SA and provide an indication of information access and utilization. However, process indices create large amounts of data to analyze and are difficult to implement in the real-world environment (e.g., eye-tracking devices, if head-mounted, can be cumbersome and intrusive). Further, process indices do not directly assess SA but rather can only be used to infer SA. In other words, these measures do not indicate what is actually done with the information acquired (processing) or whether the information is registered correctly or what is retained in memory. Instead, these measures simply indicate that the operator looked at the information. Given these limitations, process indices are more suitable for investigating specific research questions of information acquisition and for examining the processes underlying SA (e.g., perception, attention) rather than the final product.

Subjective measures

Subjective measures directly assess SA by asking individuals (or experienced observers) to rate their SA on an anchored scale (for a detailed review, see Jones 2000). These ratings can be collected during task performance or following task completion. Subjective measures of SA are attractive in that they are relatively straightforward, inexpensive, and easy to administer. However, several important limitations should be noted. Individuals making subjective assessments of their own SA are often unaware of information they do not know. Further, self-ratings may be tainted by performance outcomes. Subjective measures also tend to be global in nature and, as such, do not fully exploit the multivariate nature of SA to provide the detailed diagnostics available with objective measures. Nevertheless, self-ratings may be useful in that they can provide an assessment of operators' degree of confidence in their SA.

Subjective estimates of an individual's SA may also be made by experienced observers (e.g., supervisors, trained external experts). These observer ratings may be somewhat superior to self-ratings of SA because more information about the true state of the environment is usually available to the observer than to the operator, who may be focused on performing the task (i.e., trained observers may have more complete knowledge of the situation). However, observers have only limited knowledge about the operator's concept of the situation and cannot have complete insight into the mental state of the individual being evaluated. Thus, observers are forced to rely more on operators' observable actions and verbalizations in order to infer their level of SA. In this case, such actions and verbalizations are best assessed using performance and behavioral measures of SA, as described next.

Performance and behavioral measures

Performance measures infer SA from the end result (i.e., task performance outcomes) based on the assumption that better performance indicates better SA. Common performance metrics include quantity of output or productivity level, time to perform the task or respond to an event, and the accuracy of the response or, conversely, the number of errors committed. The main advantage of performance measures is that these can be collected objectively and without disrupting task performance. However, although evidence exists to suggest a positive relation between SA and performance, this connection is probabilistic and not always direct and unequivocal (Endsley 1995b). In other words, good SA does not always lead to good performance, and poor SA does not always lead to poor performance (Endsley 1990). Thus, performance measures should be used in conjunction with others measures of SA that directly assess this construct.

Behavioral measures also infer SA from the actions that individuals choose to take, based on the assumption that good actions will follow from good SA and vice versa. Behavioral measures rely primarily on observer ratings and are thus somewhat subjective in nature. To address this limitation, observers can be asked to evaluate the degree to which individuals are carrying out actions and exhibiting behaviors that would be expected to promote the achievement of higher levels of SA. This approach removes some of the subjectivity associated with making judgments about an individual's internal state of knowledge by allowing them to make judgments about SA indicators that are more readily observable.

Objective measures

Objective measures directly assess SA by comparing an individual's perceptions of the situation or environment with some "ground truth" reality. Specifically, objective measures can be used to collect data from operators' perceptions of the situation and compare this with what is actually happening at a given moment in

time. Thus, this type of assessment provides a direct measure of SA and does not require operators or experimenters to make judgments about situational knowledge on the basis of incomplete information. Objective measures can be gathered in one of three ways: during an interruption in task performance (e.g., queries), real time as the task is completed (e.g., probes), or posttest following completion of the task.

One common approach to directly and objectively measure SA is the Situation Awareness Global Assessment Technique (SAGAT) (Endsley 1995a). SAGAT utilizes a concurrent memory probe technique that presents queries related to the current task environment. Administration of the SAGAT involves freezing a simulation exercise at randomly selected times and hiding task information sources (e.g., blanking visual displays) while individuals quickly answer randomly ordered questions about their current perceptions of the situation. These responses are then compared with "ground truth" (i.e., actual data on the real situation) to assess the accuracy of the individuals' SA. However, because it involves interrupting task performance, SAGAT is better suited for assessing SA in simulation exercises and may not be practical for real-time measurement of SA.

For settings in which disruptions to task performance are not practical or desirable, real-time probes (e.g., open-ended questions embedded as verbal communications during the task) can be administered to naturally and unobtrusively assess operator SA (Jones and Endsley 2000). Real-time probes are similar to SAGAT in that they query operators on their knowledge of key task-relevant information in the environment; however, this methodology differs from the SAGAT in that task performance is not disrupted (i.e., the simulation or task is not stopped) but rather the queries are incorporated as a natural part of the task.

Modeling SA

SA modeling approaches can be used to objectively predict SA based on readily observable verbal and nonverbal communications. Specifically, team communications (particularly verbal communications) support the knowledge building and information processing that lead to SA construction (Endsley and Jones 2001). Thus, since SA may be distributed via communication, computational linguistics and machine learning techniques can be combined with natural language analytical techniques (e.g., Latent Semantic Analysis) to create models that draw on the verbal expressions of the team to predict SA and task performance (Bolstad et al. 2005b, 2007). For example, the Automated Communication and Situation Awareness (ACASA) tool offers near real-time, nonintrusive, quantitative assessment of SA by analyzing communication exchanges among team members (Foltz et al. 2008). Since the communication data are collected using either Automatic Speech Recognition (ASR) software or transcriptions of speech recordings, this methodology does not interrupt activities or affect performance. Thus, SA modeling approaches, such as the ACASA tool, are appropriate for use in both simulations and real-world environments. Further, this methodology can provide diagnostic information regarding current SA. For example, when coupled with ASR software, the ACASA tool can be used to quickly identify whether or not immediate action needs to be taken to address poor SA among team members.

Although evidence exists to support the utility of communication analysis for predicting team SA (Foltz et al. 2008), time constraints and technological limitations (e.g., cost and availability of speech recording systems and speech-to-text translation software) may make this approach more time consuming in terms of up-front investment. In addition, the models generated using this approach are domain- and task-specific; thus, unique models must be created for each environment or application. Last, this measure is only effective for measuring SA in a team environment and would not be suitable for situations in which a single operator is being evaluated.

Applying SA assessment to teams

Not surprising, assessing team and shared SA is more complex than assessing SA at the individual level. Some methodologies are inherently more readily applicable for team-level assessment. For example, the ACASA tool described earlier is specifically designed to be applied in a team context; thus it can be used to evaluate information flow during task performance in terms of how well team members are sharing the SA information requirements necessary for building and maintaining both team and shared SA. Similarly, behavioral measures can be used to support assessment of the types of overt team behaviors and communications that are indicative of SA.

Comparison of individual responses to objective measures of SA (e.g., SAGAT queries or real-time probes) across different team members can be used to ascertain the degree to which they have developed a common and accurate understanding of the situation or task environment (i.e., shared SA). Thus, this approach can provide the degree of diagnosticity needed to fully evaluate team performance. The simplest analysis involves comparing performance between two team members. For example, when analyzing two team members' responses to a SAGAT

query, one of four possible outcomes can occur: both individuals are correct; one individual is correct and the other is incorrect; both individuals are incorrect and they have the same response; or both individuals are incorrect but they have different responses (Endsley and Jones 2001). The latter three outcomes highlight different problems with the team members' shared SA, which in turn can provide insights on how to address this potential breakdown in team performance.

Lessons learned in SA assessment

Our work on assessing SA in team operations has demonstrated that using multiple metrics provides the greatest utility in terms of understanding how and why teams perform. For example, in a brigade-level simulated military exercise, we utilized the SAGAT methodology to evaluate a possible new unit formation (Bolstad and Endsley, In press). While the overall exercise was deemed a success, analysis of our SA assessment results indicated that placing the Deputy Brigade Commander away from the Commander hindered his ability to develop the same level of SA as the Commander. In another military exercise, we evaluated using cross-training as a method to improve team SA and performance (Bolstad et al. 2005a, 2005b). In addition to administering an objective measure of SA (i.e., SAGAT queries), we also included a subjective measure of team communication that specifically asked participants to rank order other team members based on their frequency of communication with them during the scenario; this measure was used to calculate social network distance, that is, the frequency with which team members communicated with each other. While results showed that cross-training, particularly in a leadership role, did lead to improved SA, analysis of the communication data also provided some insights on potential factors influencing team SA and performance. Specifically, physical distance (i.e., whether participants were co-located or distributed during the scenario) was found to be a significant predictor of both shared SA and social network distance. This finding supports the view that direct information exchange may be used as an input for building a team member's individual SA (Endsley and Jones 2001; Milham, Barnett, and Oser 2000).

One important lesson learned from these research studies is that increasing the sensitivity and diagnosticity of test and evaluation involves adopting a multi-faceted approach to assessment. Rather than rely on a single approach or metric, valid and reliable assessment should utilize a battery of distinct yet related measures that complement each other; this approach capitalizes on the strengths of each measure while minimizing the limitations inherent in each. Combining multiple measures together can provide valuable information with regard to factors influencing team SA, decision-making, and performance, such as the effect of team organization, distribution, and communication patterns.

Conclusion

Assessment of SA provides a degree of diagnosticity that is especially useful in the test and evaluation of new technologies and systems. SA assessment can be used to identify the source of the problem as well as to establish a baseline for comparison of the effects of different design concepts. More specifically, integrating SA assessment into test and evaluation can allow researchers to determine if a new technology or system is helping or hindering human operators' ability to perceive critical information (Level 1 SA), comprehend the relevance of this information to their task (Level 2 SA), and use this information to predict what will happen next (Level 3 SA); as well as to evaluate how these effects on operator SA influence decision-making and, ultimately, safety and performance. At the team-level, SA assessment can be used to determine the degree to which information is being exchanged among team members to support both team and shared SA.

In both cases, determining the best SA measures for inclusion in Mission-Based Test and Evaluation events depends upon multiple factors, such as the study's objectives, team size, other variables being assessed, and the ability to integrate the selected measures into the experimental test plan. However, whenever possible, a multi-faceted approach to SA assessment is desirable to ensure a higher level of diagnosticity in the overall assessment. ❑

Dr. Cheryl A. Bolstad is a principle research associate at SA Technologies, headquartered in Marietta, Georgia. She received her doctor of philosophy degree in psychology, specializing in cognition and aging, from North Carolina State University (Raleigh, North Carolina). Dr. Bolstad has 20 years of experience as a human factors engineer and has worked on a wide variety of projects including team performance analysis and measurement, training program design and evaluation, collaboration tool design, and cognitive readiness assessment. More recently, she has worked extensively in SA research including user interface design, training, and measurement. E-mail: cheryl@satechnologies.com

Dr. Haydee M. Cuevas is a research associate II at SA Technologies, headquartered in Marietta, Georgia. She received her doctor of philosophy degree in applied experimental and human factors psychology from the University of

Central Florida (Orlando, Florida). Dr. Cuevas has over 10 years of experience as a human factors researcher and has worked on projects funded by the National Science Foundation, Air Force Office of Scientific Research, Army Research Laboratory, Office of Naval Research, and Office of the Secretary of Defense. Her recent research has primarily focused on supporting human-automation team performance in complex operational environments. E-mail: haydee.cuevas@satechnologies.com

References

Blandford, A., and W. Wong. 2004. Situation awareness in emergency medical dispatch. *International Journal of Human–Computer Studies*. 61: 421–452.

Bolstad, C. A., and M. Endsley. In press. Measuring shared and team situation awareness in the U.S. Army's Future Objective Force. *Military Psychology*.

Bolstad, C. A., H. M. Cuevas, A. M. Costello, and J. Rousey. 2005a. "Improving situation awareness through cross-training." In *Proceedings of the 49th Annual Meeting of the Human Factors and Ergonomics Society*, September 26–30, Orlando, Florida, 2159–2163. Santa Monica, CA: Human Factors and Ergonomic Society.

Bolstad, C. A., H. M. Cuevas, C. Gonzalez, and M. Schneider. 2005b. Modeling shared situation awareness. In *Proceedings of the 14th Conference on Behavior Representation in Modeling and Simulation*, May 16–19, Los Angeles, California.

Bolstad, C. A., P. Foltz, M. Franzke, H. M. Cuevas, M. Rosenstein, and A. M. Costello. 2007. Predicting situation awareness from team communications. In *Proceedings of the Human Factors and Ergonomics Society 51st Annual Meeting*, October 1–5, Santa Monica, CA: Human Factors and Ergonomics Society.

Endsley, M. R. 1990. Predictive utility of an objective measure of situation awareness. In *Proceedings of the Human Factors Society 34th Annual Meeting*, Santa Monica, CA, 41–45. Santa Monica, CA: Human Factors and Ergonomics Society.

Endsley, M. R. 1995a. Measurement of situation awareness in dynamic systems. *Human Factors*. 37 (1): 65–84.

Endsley, M. R. 1995b. Toward a theory of situation awareness in dynamic systems. *Human Factors*. 37 (1): 32–64.

Endsley, M. R., and W. M. Jones. 2001. A model of inter- and intrateam situation awareness: Implications for design, training and measurement. In *New trends in cooperative activities: Understanding system dynamics in complex environments*. Edited by M. McNeese, E. Salas, and M. R. Endsley, 46–67. Santa Monica, CA: Human Factors and Ergonomics Society.

Endsley, M. R., B. Bolte, and D. G. Jones. 2003. *Designing for situation awareness: An approach to human-centered design*. London: Taylor & Francis.

Flin, R., and P. O'Connor. 2001. Applying crew resource management in offshore oil platforms. In *Improving teamwork in organization: Applications of resource management training*. Edited by E. Salas, C. A. Bowers, and E. Edens, 217–233. Hillsdale, NJ: Erlbaum.

Foltz, P. W., C. A. Bolstad, H. M. Cuevas, M. Franzke, M. Rosenstein, and A. M. Costello. 2008. Measuring situation awareness through automated communication analysis. In *Macrocognition in teams*. Edited by M. P. Letsky, N. W. Warner, S. M. Fiore, and C. A. P. Smith, 259–275. Aldershot, England: Ashgate.

Fracker, M. L. 1991. *Measures of situation awareness: Review and future directions* (Report No. AL-TR-1991-0128). Wright-Patterson Air Force Base, OH: Armstrong Laboratories.

Gorman, J. C., N. J. Cooke, and J. L. Winner. 2006. Measuring team situation awareness in decentralized command and control environments. *Ergonomics*. 49 (12–13): 1312–1325.

Graham, S. E., and M. D. Matthews. 2000. Modeling and measuring situation awareness. In *Workshop on assessing and measuring training performance effectiveness* (Tech. Rep. 1116). Edited by J. H. Hiller, and R. L. Wampler, 14–24. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Hartel, C. E. J., K. Smith, and C. Prince. 1991. Defining aircrew coordination: Searching mishaps for meaning. Paper presented at the 6th International Symposium on Aviation Psychology, April 29–May 2. Columbus, Ohio.

Jones, D. G. 2000. Subjective measures of situation awareness. In *Situation awareness analysis and measurement*. Edited by M. R. Endsley and D. J. Garland, 113–128. Mahwah, NJ: Lawrence Erlbaum.

Jones, D. G., and M. R. Endsley. 2000. Examining the validity of real-time probes as a metric of situation awareness. In *Proceedings of the 14th Triennial Congress of the International Ergonomics Association and the 44th Annual Meeting of the Human Factors and Ergonomics Society*. July 30–August 4, San Diego, CA. Santa Monica, CA: Human Factors and Ergonomics Society.

Lawlor, M. 2005. Researchers investigate cognitive collaboration. *Signal*. 59 (9): 30–34.

Merket, D. C., M. Bergondy, and H. Cuevas-Mesa. 1997. Making sense out of teamwork errors in complex environments. Paper presented at the 18th Annual Industrial/Organizational–Organizational Behavior

Graduate Student Conference, March. Roanoke, Virginia.

Milham, L. M., J. S. Barnett, and R. L. Oser. 2000. Application of an event-based situation awareness methodology: Measuring situation awareness in an operational context. In *Proceedings of the XIVth Triennial Congress of the International Ergonomics Association and 44th Annual Meeting of the Human Factors and Ergonomics Society*, July 20–August 4, San Diego, California, 2, 432–426. Santa Monica, CA: Human Factors and Ergonomics Society.

Nullmeyer, R. T., D. Stella, G. A. Montijo, and S. W. Harden. 2005. Human factors in Air Force flight mishaps: Implications for change. In *Proceedings of the 27th Annual Interservice/Industry Training, Simulation, and Education Conference* (Paper no. 2260), Arlington, VA: National Training Systems Association.

Sarter, N. B., and D. D. Woods. 1991. Situation awareness: A critical but ill-defined phenomenon. *International Journal of Aviation Psychology*. 1: 45–57.

Smith, K., and P. A. Hancock. 1995. Situation awareness is adaptive, externally directed consciousness. *Human Factors*. 37 (1): 137–148.

Strater, L. D., and C. A. Bolstad. 2009. Situation awareness in simulations. In *Human factors in simulation and training*. Edited by D. A. Vincenzi, J. A. Wise, M. Mustapha, and P. A. Hancock, 129–148. New York, NY: CRC Press, Taylor and Francis Group.

Wilson, G. F. 2000. Strategies for psychophysiological assessment of situation awareness. In *Situation awareness analysis and measurement*. Edited by M. R. Endsley and D. J. Garland, 175–188. Mahwah, NJ: Lawrence Erlbaum Associates.

ITEA Journal 2010; 31: 247–263

Development of an Autodiagnostic Adaptive Precision Trainer for Decision Making (ADAPT-DM)

Meredith Carroll, Ph.D., Sven Fuchs, Angela Carpenter, and Kelly Hale
Design Interactive, Inc., Oviedo, Florida
Robert G. Abbott
Sandia National Laboratories, Albuquerque, New Mexico
Amy Bolton, Ph.D.
Office of Naval Research, Arlington, Virginia

The Autodiagnostic Adaptive Precision Trainer for Decision Making (ADAPT-DM) is a framework for adaptive training of decision making skills. The training challenge is that decision making behavior is mostly unobservable with traditional behavioral measures, which generally only give access to outcome performance. This article describes the ADAPT-DM framework, which utilizes physiological sensors, specifically electroencephalography and eye tracking, to detect indicators of implicit cognitive processing relevant to decision making and accomplish the granularity required to pinpoint and remediate process level issues. Using these advanced measures, the trainee's performance on these cognitive processes can be assessed in real time and used to drive smart adaptations that individualize training. As a proof of concept, the ADAPT-DM framework was conceptually applied to the contact evaluation task in submarine navigation. Simulated data from 75 students, grouped into three levels of expertise (novice, intermediate, and expert), were used for principal component analysis to identify skill dimensions that reflect proficiency levels. Then ADAPT-DM's composite diagnosis was performed, which uses an expertise model that integrates automated expert modeling for automated student evaluation machine learning models with eye tracking and electroencephalography data to assess which proficiency level the simulated students actions were most similar to. Based on additional assessments, the diagnostic engine is able to determine whether the student (a) performs to criterion, in which case training could be accelerated, (b) is in an optimal learning state, or (c) is in a nonoptimal learning state for which remediation or mitigation are needed. Using root cause analysis techniques, the ADAPT-DM process level measures then allow instructors to pinpoint where in the decision making process breakdowns occur, so that optimal training adaptations can be implemented.

Key words: Adaptive training; decision making skills; expertise modeling; learning state.

In highly dynamic work situations, such as a submarine crew environment, individuals are required to function with high levels of decision making (DM) skill proficiency while in an environment marked by unforeseen threats, complex data streams, and high levels of uncertainty. The time typically available for training such DM skills is limited; therefore, there is a need for systems that can accelerate skill development, bringing trainees up to speed more quickly. Yet, existing training systems lack the capability to provide real-time adaptive training that can ensure effective and efficient training. An opportunity exists to precisely assess trainee performance and adapt the training experience to accelerate the learning process by (a) identifying and mitigating times when a trainee is in a nonoptimal learning state and time is being wasted, (b) identifying the root cause of performance deficiencies to allow feedback to be tailored to trainee-specific decrements, and (c) adapting training with increasing levels of trainee expertise to ensure efficient utilization of training time. The challenge with respect to assessing

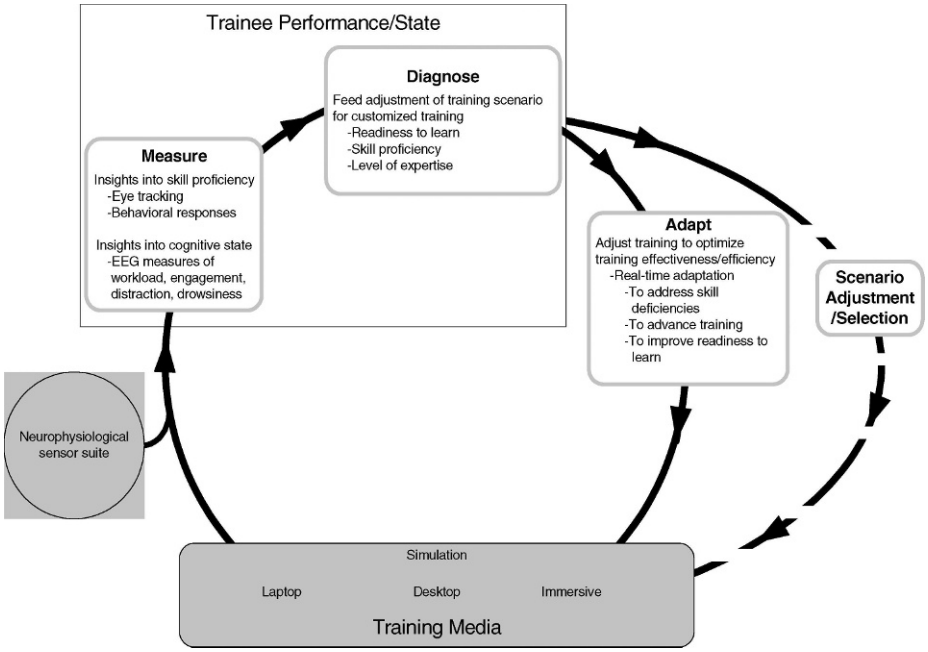


Figure 1. The Autodiagnostic Adaptive Precision Trainer for Decision Making (ADAPT-DM) framework.

the DM process during training, specifically, is that much of DM behavior is unobservable and thus difficult to measure with traditional behavioral measures, which generally only give access to outcome performance (Klein 1998). Outcome measures, such as decision outcomes, do not give the granularity needed to pinpoint and remediate process level issues. Implicit indicators are needed, such as visual scan patterns (i.e., how a decision maker is collecting information and what information is being considered), key cues entering into the decision, sources of distraction or confusion, or changes in cognitive processing that affect readiness to learn (e.g., fatigue, disengagement) (Klein and Hoffman 1992; Macklin et al. 2002). To increase assessment granularity for cognitive processes, we must (a) capture and evaluate perceptual and cognitive processes relevant to DM, (b) analyze the trainee's performance on these cognitive processes *in real time*, and (c) use these data to drive smart adaptations that are grounded in training science. As such there is a need for physiological-sensor–based real-time adaptive training.

The Autodiagnostic Adaptive Precision Trainer for Decision Making (ADAPT-DM) is a framework that aims to address this training gap. The framework is composed of three components necessary to ensure precision training: measurement, diagnosis, and adaptation *(Figure 1)*.

- The measurement component allows for the incorporation of a broad range of data collection tools, such as system collected, self-report, instructor assessment, behavioral, physiological, and neurophysiological measurement to gain a comprehensive understanding of trainee performance and state.
- By incorporating diagnosis methods, such as root cause analysis, expert comparison, and error pattern analysis, the diagnosis component analyzes these data to direct remediation and facilitate real-time training.
- Based on the diagnosis, the adaptation component triggers adaptations strategies designed to address performance and state issues through real-time adaptations, after-action feedback, and selection of future training content.

ADAPT-DM theoretical foundation

> *"Expertise is the key factor in decision making in natural environments." (Lipshitz et al. 2001)*

Two key models serve as the theoretical foundation for ADAPT-DM: the Stimulus- Hypothesis-Option-Response (SHOR) model (Wohl 1981) and the Skills-Rules-Knowledge (SRK) model (Rasmussen 1983). Similar to other contemporary models relevant to tactical DM, such as Endsley's (1995) situation awareness model and Klein's recognition primed decision-making model (Lipshitz et al. 2001), the SHOR model dissects the DM process into four distinct steps.

- *Stimulus:* In this step a decision maker gathers, recalls, filters, and aggregates information.

Table 1. SRK types of performance.

Type of performance	Level of cognitive control	Description of performance	Expertise typically associated
Skill-based	No conscious, cognitive control, highly automated	Routine activities conducted automatically that do not require conscious allocation of attention	High level of expertise
Rule-based	Low level conscious cognitive control	Activities controlled by a set of stored rules or procedures	Medium level of expertise
Knowledge-based	High level of conscious cognitive control	Novel situations are presented for which a plan must be developed to solve a problem	Low level of expertise

- *Hypothesis:* Here, the decision maker creates and evaluates hypotheses about the environment around them and selects the most plausible hypothesis.
- *Option:* The decision maker creates and evaluates decision options for how he or she should respond based on the hypothesis selected and potential positive and negative outcomes.
- *Response:* The decision maker plans, organizes, and executes the response selected.

This DM process becomes abridged as a decision maker develops expertise. According to Rasmussen's (1983) SRK model *(Table 1)*, as expertise develops a performer can successfully complete the decision task with greater levels of automaticity and hence lower levels of cognitive control.

Taken together, these models (Rasmussen 1983; Wohl 1981) suggest that as performers build expertise, they move from purely knowledge-based performance to skill-based performance *(Figure 2)*. For novices, situations are generally novel, and they have to perform the entire DM process, analyzing the environment and creating a hypothesis of what the pattern of cues means for the situation, then generating and evaluating potential responses. As expertise develops with experience base, the trainee starts to develop the ability to recognize patterns of cues, which can be successfully associated with existing mental models of a situation, so that known response rules associated with these familiar situations can be triggered. Thus, the DM process becomes abbreviated as the trainee quickly recognizes a situation and applies a preprogrammed rule. With high levels of expertise, the DM process becomes almost automated, wherein an expert reacts to familiar cues with an almost "wired response" based on almost immediate (and possibly parallel) recognition and evaluation of the situation.

These models provide a framework for evaluating at a very granular level where in the DM process breakdowns are occurring and at what level of expertise the decision maker is operating. Expertise is the key factor in DM in natural environments (Lipshitz et al. 2001), and the ability to identify level of expertise will allow a more comprehensive understanding of DM performance, including why performance breakdowns occur and what kind of scenario adaptations are most useful to address performance problems.

ADAPT-DM measurement component

For the first component of the ADAPT-DM framework—the measurement component—the essential question is what to measure. Within the natural-

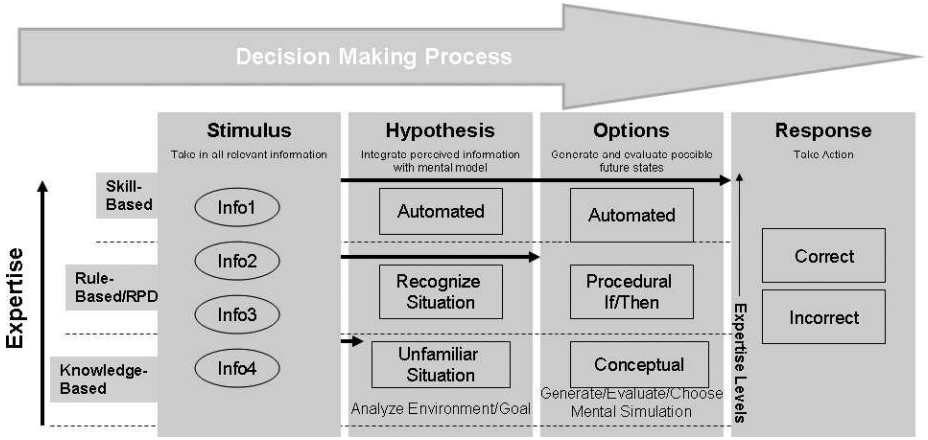


Figure 2. Adaptive DM model.

Table 2. Cognitive readiness problem states.

Problem state	Rationale/literature support
Workload	When workload is low and trainees are bored, they pay less attention, resulting in lower retention and decreased ability to apply information (Small, Dodge, and Jiang 1996).
	When workload is high, divided attention results, which is associated with large reductions in memory performance and small increases in reaction time during encoding, and small or no reductions in memory during recall, but comparatively larger increases in reaction time (Craik et al. 1996).
Engagement	Low levels of engagement indicate that a trainee is not actively engaged with some aspect of the training environment (Dorneich et al. 2004).
Distraction	Even if distraction does not decrease the overall level of learning, it can result in the acquisition of knowledge that can be applied less flexibly in new situations (Foerde, Knowlton, and Poldrack 2006).
Drowsiness	Drowsiness can causes lapses in attention and performance, as well as microsleeps (Neri et al. 2007).

istic decision making (Lipshitz et al. 2001) literature, some researchers have attempted to identify more granular measures of DM skills than performance time and accuracy by considering such measures as number of options considered (Klein and Peio 1989); however, few have considered how to operationalize real-time DM performance measurement and diagnosis. For example, Elliot et al. (2007) presented four metric categories linked to perceptual and cognitive skills associated with natural decision making, including speed (e.g., reaction time, response time), accuracy (e.g., accuracy of response), efficiency (e.g., shortest path to success), and planning (e.g., proactive actions taken). Although these measures provide some level of assessment of the DM process, they are not sufficiently granular to pinpoint where breakdowns in DM performance occur to provide real-time adaptations to target these deficiencies. This is the goal of the ADAPT-DM framework. One specific limitation of behavioral measures is that they are limited in their ability to discriminate performance within the "good" or "bad" performance categories for decision making. For example, an expert and a journeyman may both reach a good decision; however, the amount of effort (e.g., speed and flexibility) required for this level of achievement might differ significantly (Klein and Hoffman 1992). Time measures can typically capture a portion of this; however, they do not gauge internal states, such as workload, that might be critical factors when performing in novel or stressful situations. An expert who is not only performing well but has reached a certain level of ease and automaticity will be more prepared than a journeyman who is performing well but is using every available cognitive resource to achieve this level of performance. The journeyman may need more practice to maintain high performance under high stress levels in the field. It is thus necessary to understand the underlying cognitive states of the trainee, which both affect learning and are indicators of learning effectiveness, to comprehensively diagnose DM expertise and performance.

With the emergence of neurophysiological and physiological measurement technology that allows for real-time assessment of perceptual and cognitive processing, these unobservable processes become accessible. Specifically, some cognitive states that are measurable via electroencephalography (EEG), including workload and engagement, can provide neurophysiological measures of the unobservable aspects of DM skill development (Dorneich et al. 2007; Levonian 1972). *Table 2* outlines specific cognitive states that generally negatively affect the readiness for training by reducing attentional resources that facilitate learning and retention. Thus, it may be possible to utilize certain neurophysiological cognitive state metrics to detect issues with readiness to learn during DM performance:

- *Workload:* High cognitive workload is expected when performing in a knowledge-based control mode because no automaticity guides the process (Berka et al. 2007; Klein and Hoffman 1992). In rule-based control mode, rules are consciously retrieved from memory and applied to gathered information, also causing increased cognitive processing demands. Experts using skill-based DM, however, employ automated routines that require fewer cognitive resources. Thus, it is expected that the assessment of cognitive workload can contribute to the identification of the trainee's control mode.
- *Engagement:* Because of high task demands, novice and journeyman trainees are expected to exhibit higher levels of engagement than expert trainees because studies have shown a trend for decreasing EEG engagement with increasing task proficiency (Berka et al. 2007; Stevens, Galloway, and Berka 2007).
- *Distraction:* Distraction is a state characterized by a lack of clear and orderly thought and behavior, where a trainee becomes involved somewhere other than the cognitive tasks of interest

(Poythress et al. 2006). Expert performers have an exhaustive mental model of the task or situation so that very few situations cause distraction. Confusion is one element of distraction. In rule-based decision makers, confusion may stem from the conscious selection of rules and difficulties in applying them to the situation at hand. Naïve trainees are expected to show relatively high levels of confusion because their mental models are more likely to be incorrect or insufficient so that new situations may cause a mismatch.

- *Drowsiness:* Sleep disorders are common and can have deleterious effects on performance (Berka et al. 2004, 2005; Neri et al. 2007). In fact, loss of sleep can accumulate over time and result in a "sleep debt," which can lead to impairments in alertness, memory, and decision making. Individuals with chronic accumulation of fatigue are often unaware of the impact on their performance.

Eye tracking metrics provide a physiological measure with the granularity necessary to understand why DM-related performance failures occur to effectively adapt training. In particular, eye tracking offers an additional set of behavioral-based metrics to aid in assessing the information processing of individuals as it relates to perception. Toward this level of assessment, the following eye tracking metrics have been validated as providing information on perceptual processes (Hyönä, Radach, and Deubel 2003):

- *Number of overall fixations:* Inversely correlated with search efficiency.
- *Gaze percent on Areas of Interests (AOIs):* Longer gazes are equated with importance or difficulty of information extraction.
- *Mean fixation duration:* Longer fixations are equated with difficulty of extracting information.
- *Number of fixations on AOIs:* Reflects the importance of each area.

Thus, beyond traditional DM performance-based metrics, neurophysiological and physiological metrics can be used to provide an assessment of the unobservable aspects of DM skills development.

ADAPT-DM diagnosis component

The next component of the ADAPT-DM framework is the diagnosis component. ADAPT-DM diagnoses root causes in performance deficiencies and inefficiencies based on three important factors associated with DM skill development:

1. *DM performance:* The diagnosis component can use performance outcome (e.g., speed, accuracy, efficiency, and planning; Elliot et al. 2007) and eye tracking (e.g., number of overall fixations, gaze percentage on AOIs, mean fixation duration, number of fixations on AOIs; Hyona, Radach, and Deubel 2003) data to assess whether a trainee is collecting appropriate information, considering and understanding information appropriately, selecting good decision options, and appropriately executing these options.
2. *Learning state:* To ensure feedback and facilitate effective performance improvements, it is essential to ensure that trainees are operating in an effective learning state. The diagnosis component can use EEG-based metrics (e.g., workload, engagement, distraction, drowsiness; Dorneich et al. 2007; Levonian 1972) to ensure that the trainee's learning state remains at adequate levels to promote learning.
3. *Expertise:* Performance may not provide sufficient granularity to drive precise adaptations. A trainee can perform well but be using every spare resource, have inefficient performance, and substantial room for improvement in terms of strategies used. Additionally, performers operating at different expertise levels commit errors for different reasons. Thus, the diagnosis component assesses expertise to allow for more precise adaptations to be made.

Expertise is the most challenging of these skills to diagnose. To truly understand why trainees are performing as they are, one must take into account expertise level. Reason (1990) identified typical performance characteristics and failure modes related to the SRK levels (Rasmussen 1983) of cognitive control associated with varying expertise. These characteristics and failure modes *(Table 3)* can be used to diagnose deficiencies with respect to expertise level and select effective adaptations. However, given the multifaceted nature of expertise, it cannot be diagnosed by merely looking at a small subset of performance measures. Instead, it is necessary (though challenging) to consider several aspects of performance and cognitive state. The Automated Expert Modeling for Automated Student Evaluation (AEMASE) process can be used to support such diagnosis (Abbott 2006).

AEMASE is a process for subject matter experts to rapidly create and update their own models of normative behavior (Abbott 2006). First, examples of task behavior are recorded in a training simulator. The examples may be either good or bad behavior performed by either students or subject matter experts, but the examples must be accurately graded by a subject matter expert. Second, machine learning algorithms are

Table 3. Typical performance characteristics and failure modes related to the SRK (Reason 1990).

Expertise level	Typical control mode	Performance characteristics	Failure modes
Expert	Skill based	Errors occur during routine action	*Inattention*
		Attention during errors is not directed at task at hand	Double-capture slips
		Errors occur while applying known schemata	Omissions following interruptions
		Errors are "strong but wrong" and predicable	Reduced intentionality
		Error numbers may be high, but error/opportunity ratio is small	Perceptual confusions
		Low to moderate influence of (mostly intrinsic) factors	Interference errors
		Error detection is usually fairly rapid and effective	*Overattention*
		Knowledge of change is not accessed at proper time	Omissions
			Repetitions
			Reversals
Journeyman	Rule based	Errors occur during problem-solving activities	*Misapplication of good* rules
		Attention during errors is directed at problem-related issues	First exceptions
		Errors occur while employing stored rules	Countersigns and nonsigns
		Errors are "strong but wrong" and predicable	Informational overload
		Error numbers may be high, but error/opportunity ratio is small	Rule strength
		Low to moderate influence of (mostly intrinsic) factors	General rules
		Error detection is difficult and often requires external intervention	Redundancy
		Changes in the environment are anticipated but when and how is not known	Rigidity
			Application of bad rules
			Encoding deficiencies
			Action deficiencies
			Wrong rules
			Inelegant rules
			Inadvisable rules
• Novice	Knowledge-based	Errors occur during problem-solving activities	Selectivity
		Attention during errors is directed at problem-related issues	Workspace limitations
		Errors occur while employing limited, conscious processes	Out of sight out of mind
		Errors occur with variable predictability	Confirmation bias
		Error numbers are small, but high error/opportunity ratio	Overconfidence
		Influence of extrinsic situational factors on errors is high	Biased reviewing
		Error detection is difficult and often requires external intervention	Illusory correlation
		Changes in the environment are not prepared for and not anticipated	Halo effects
			Problems with causality
			Problems with complexity
			Problems with delayed feedback
			Insufficient consideration of processes in time
			Thematic vagabonding

applied to create a behavior model. Creating the model requires selecting the data fields that best distinguish between good and bad behavior (feature selection) and applying an algorithm to generalize assessments of observed behavior to assessments of new (potentially novel) student behavior. An appropriate algorithm must be selected for each student performance metric, depending on the type and amount of example data available. Third, student behavior is assessed using the behavior model. As each student executes a simulation-based training scenario, his or her behavior is compared with the model for each performance metric to identify and target training to individual deficiencies. The model determines whether student behavior is more similar to good or bad behavior from its knowledge base. Initially, the knowledge base is sparse, and incorrect assessments may be common. However, an instructor may override incorrect assessments. AE-MASE learns from this interaction, so the model improves over time. Real-time student assessment can be implemented by continuously reevaluating the model throughout a scenario to support dynamic scenario adaptation. In a previous pilot study, AE-MASE achieved a high degree of agreement with a

human grader (89%) in assessing tactical air engagement scenarios. In a subsequent study of E2 Naval Flight Officer tasks, AEMASE achieved 80%–95% agreement with a human grader on a range of metrics (Stevens et al. 2009). AEMASE is useful when data collection for a metric can be automated, but the metric is difficult to assess (i.e., grade performance) because the desired value for the metric depends on what is happening in the scenario, or there are several equally valid values. AEMASE can support real-time assessment and scenario adaptation by operationalizing complex or "fuzzy" assessments.

Based on a combination of relevant performance and state metrics, AEMASE can thus be used to determine the level of expertise to which a trainee's overall performance and state are most similar. This comparison can be made in near real time, thereby feeding the resulting categorization back to the ADAPT-DM diagnostic component.

ADAPT-DM adaptation component

The final component of the ADAPT-DM framework is the adaptation component, which precisely adapts training to support individualized DM skill development, based on the outcome of the diagnostic component. It uses a hierarchical adaptation strategy to adapt training without disrupting learning. Specifically, Bruner's (1973) constructivist theory can be formulated into a hierarchical adaptation strategy by applying the following principles:

- First, consider the student's willingness and ability to learn (i.e., cognitive readiness, as assessed via EEG-based cognitive state metrics). This adaptation stage should aim to enhance learning state to ensure learning can occur and mitigate any negative learning states, such as drowsiness and distraction.
- Second, structure training so that concepts can be easily grasped by trainees and skills deficiencies can be addressed (i.e., spiral organization). This adaption stage should aim to improve knowledge and skills to allow development of skilled performance and prevent trainees from practicing bad habits or perpetuating incorrect performance or error patterns.
- Third, once performance is at target performance levels, design difficult cases that facilitate extrapolation and fill any gaps in training (i.e., encourage trainees to go beyond the information given). This adaptation stage should aim to increase expertise levels to boost efficiency and effectiveness of performance by providing trainees with practice opportunities and instruction designed to move them up the expertise continuum to skilled performance *(Figure 3)*.

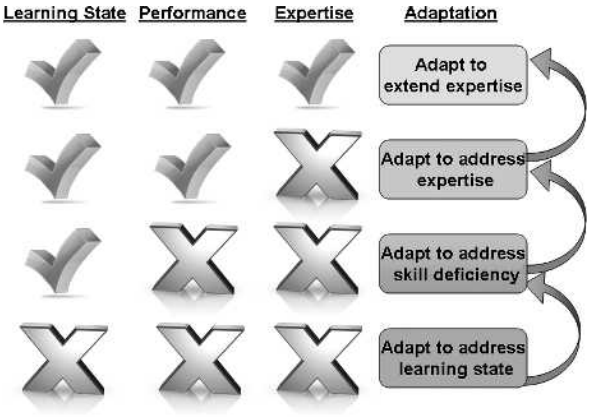


Figure 3. Adaptation goals with respect to diagnosed problem areas.

A generalizable adaptation matrix was constructed detailing adaptation strategies that can be used to address each stage in the hierarchical adaptation strategy *(Table 4)*.

Case study: Submarine navigation, contact evaluation task

As a proof-of-concept, the ADAPT-DM framework was conceptually applied to submarine navigation, particularly the contact evaluation task, which is a critical decision point in navigation. Based on a task analysis, it was determined that the contact evaluation task *(Figure 4)* entails the following perceptual, cognitive, and response components. *Perceptual components:* (1) scan the radar display for contacts; (2) detect contacts; (3) scan for other relevant cues to assess the contact. *Cognitive components:* (4) assess contact relationship to own ship; (5) use tools to aid in assessing contact relationship; (6) decide whether contact is of enough concern to monitor. *Response components:* (7) hook and monitor contact; (8) communicate contact information to the Contact Coordinator (CC).

Based on the task analysis, behavioral performance metrics (including eye tracking metrics) were identified for all tasks within the task flow *(Table 5)*. In addition, EEG-based cognitive state metrics were identified to assess trainee state *(Table 2)*.

Based on the performance metrics identified, the next step is diagnosing the adequacy of DM performance. While many of the metrics have straightforward thresholds, which divide good and poor performance (e.g., relevant contact hooked or not), several of the metrics have complex performance thresholds (e.g., scan data). It was determined that AEMASE machine

Table 4. Adaptation strategies.

Performance	Expertise	Diagnosis	Real time adaptation	Future adaptation
Good	Expert	Criterion	Increase difficulty	Once criterion met for highest level of difficulty, move on to new training objective
	Expert	Optimal learning state	None	Continue practice at this level of difficulty
	Journeyman	Optimal learning state	None	Continue practice at this level of difficulty
	Journeyman	Nonoptimal learning: drowsy	Increase pace of training	Give trainee a break, encourage to get up and walk around
			Novel situation to challenge	Increase difficulty of next scenario
	Journeyman	Nonoptimal learning: distracted	Auditory cue to bring back into focus	Increase difficulty of next event
	Novice	Nonoptimal learning: drowsy	Give positive feedback until not drowsy: "You are scanning relevant areas, keep up the good work!"	Give trainee a break, encourage to get up and walk around
				Continue practice at this level of difficulty
	Novice	Nonoptimal learning: distracted	Auditory cue to bring back into focus	Continue practice at this level of difficulty
Bad	Journeyman	Skill deficiency	Hints to abbreviate process or increase efficiency of performance	Decrease difficulty of next event
			Correction of error patterns/bad rules/ misapplication of good rules	
	Journeyman	Nonoptimal learning: drowsy	Cue to wake them up	Give trainee a break, encourage to get up and walk around
			Increase volume of auditory cues	Continue practice at this level of difficulty
			Increase intensity of visual cues	
	Journeyman	Nonoptimal learning: distracted	Auditory cue to bring back into focus—feedback relevant to performance decrements	Continue practice at this level of difficulty
	Novice	Skill deficiency	Scaffolding to assist in building rules (training wheels, faded feedback, etc.)	Decrease difficulty of next event
			Feedback to deal with typical failure modes	
	Novice	Nonoptimal learning: drowsy	Give feedback on errors until not drowsy: "You are spending too much time on irrelevant areas."	Give trainee a break, encourage to get up and walk around
				Decrease difficulty of next event
	Novice	Nonoptimal learning: distracted	Auditory cue to bring back into focus—feedback relevant to performance decrements	Decrease difficulty of next event

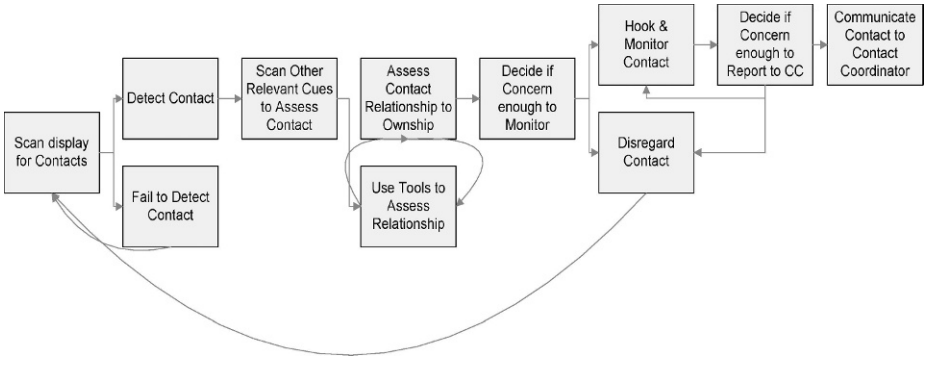


Figure 4. Contact evaluation task.

Table 5. Behavioral performance metrics for the contact evaluation task.

Task	Metrics
Scan radar screen for contacts	Appropriate view/scale of Field of View (FOV)
	% of relevant areas scanned
	% of areas scanned that were relevant
	Time until each/all relevant areas scanned
	Overall fixation duration on individual AOIs and screen
	Average fixation duration (on relevant and irrelevant)
	No. of times scan pattern changes directions (and moves significant length)
Detect contact	Target fixated (yes/no)
	Time until first target fixation
	No. of target fixations
	Duration of target fixations (average duration, total duration)
Scan relevant cues needed to assess contact	% of areas scanned that are relevant (cues and contact)
	Appropriate view/scale of FOV
	% of relevant areas scanned
	% of areas scanned that were relevant
	Time until each/all relevant areas scanned
	Overall fixation duration on individual AOIs and screen
	Average fixation duration (on relevant and irrelevant)
	No. of times scan pattern changes directions (and moves significant length)—fixation pattern on contact, on cue, on contact, on cue
Assess contact relationship to ship	No. of target fixations
	Duration of target fixations (average duration, total duration)
Use tools to assess contact relationship to ship (e.g., threat rings)	Appropriate tool use (occurrence and duration of use)
	No. of fixations on tools
Decide whether contact is of concern enough to monitor	Reaction time (time from detection/fixation until response)
	No. of target fixations
	Duration of target fixations (average duration, total duration)
Decide whether contact is of concern enough to report to CC	Reaction time (time from detection/fixation until response)
	No. of target fixations
	Duration of target fixations (average duration, total duration)
Hook contact/not	Response accuracy: contact hooked or not
	Response time (time from start to completion of response)
Communicate contact to CC/not	Response accuracy: Occurrence of communication to CC (either measured via instructor event-based checklist or voice recognition/Sandi software) and whether contact relevant
	Response time (time from start to completion of response)

learning models (Abbott 2006) could be used to compare performance on these metrics to expert and novice models to effectively assess performance. Each metric was thus defined by the behavioral or physiological variables for expert or novice comparison, the contextual variables that determine appropriate behavior or expected physiological response, and the algorithm proposed for modeling expected behavior from the context *(Table 6)*.

Most of the proposed metrics deal with the allocation of attention over time. These metrics can be implemented with occupancy grids. An occupancy grid is a two-dimensional histogram that accumulates the amount of time spent in each cell of a grid. It is weighted to reflect the recent past using a decay function. The visualization of an occupancy grid is similar to heat maps used in eye tracking studies. However, the purpose of the occupancy grid is not mainly to produce a visualization; rather it is to create a quantifiable similarity metric for expert versus trainee attention allocation. The relevance of a context is determined by a similarity metric over contextual variables, such as the positions of a submarine and contacts, and by ocean currents, etc. The similarity between expert and trainee actions is the cross product (or area of overlap) between the expert and trainee occupancy grids.

In the example occupancy grid in *Figure 5*, a trainee student (S, Left) is navigating toward a port in the presence of other surface vessels. The knowledge base (1-3, Right) contains recordings of previous expert scenario executions. The knowledge base is searched for relevant contexts (1 and 2, highlighted in green), defined by similar positioning of the submarine and other vessels, currents, etc.

After selecting relevant contexts 1 and 2 *(Figure 5)*, AEMASE determines whether the trainee's actions are similar to any performed by an expert. The red areas

Table 6. Metrics proposed for AEMASE evaluation.

Metric	Description	Context	Algorithm
Metrics collected from the simulation			
Field of view and zoom scale of radar operator interface	Radar operators control display settings specifying area and scale. Maintaining overall situational awareness requires adjusting the settings to maintain the "big picture" while frequently zooming in to view important detail.	Position of the submarine in the port, presence of tracks, and distracters.	Occupancy grid.
Reaction time for appearance of new contact	Radar operators must maintain situational awareness to react promptly to new radar returns. A delayed reaction reduces the amount of time to take measures in response to the new contact.	The position of the new contact relative to the carrier. Other contacts or navigation by own ship may also influence the allowable reaction time.	One-sided Gaussian distribution of expert reaction times, which captures the proportion of experts requiring at least *x* seconds to respond.
Metrics collected from eye tracking			
Percentage of relevant areas scanned	This metric quantifies whether the student is monitoring all areas that an expert would monitor. It requires correlating the view area (determined by radar scope settings) with the onscreen gaze position.	The relevance of areas is conditioned on the terrain (contour of the ocean floor or inlet). Relevance also depends on entities in the scenario, including their locations, attributes, and actions.	Using the occupancy grid, this is the area of the overlap between student and expert scan areas, divided by the expert's total scan area.
Percentage of areas scanned that were relevant	This metric quantifies whether the student is spending an inordinate amount of time and effort monitoring areas that are unlikely to be salient. The hypothesis is that experts know which cues in the environment are most salient, while novices' patterns of attention allocation are more randomized.	The relevance of areas is determined as before, by retrieving examples of expert attention allocation in similar contexts.	Using the occupancy grid, this is the area of the overlap between student and expert scan areas, divided by the student's total scan area.

show where the trainee student (S, Left) or experts (1,2 Center) have been looking recently. S*1 and S*2 are the dot product (or overlap) of trainee student attention with expert attention 1 and 2, respectively. S*1 (highlighted in green) has the larger area. However, S*1 covers only a portion of 1, so the trainee is neglecting some important areas.

The composite diagnosis is driven by an expertise model that integrates the AEMASE metrics with eye tracking and EEG data to assess trainee proficiency. The first step in this data integration process was to identify a minimal set of skills necessary to characterize trainee performance and expertise. Because trainees learn a progression of skills throughout their training, metrics that are appropriate for novices might be irrelevant for experts (and vice versa). Through the skills identification process, relevant metrics can be identified for trainees at each level in the training progression. Then Principal Component Analysis (PCA) can then be used to identify skill dimensions that reflect each proficiency level.

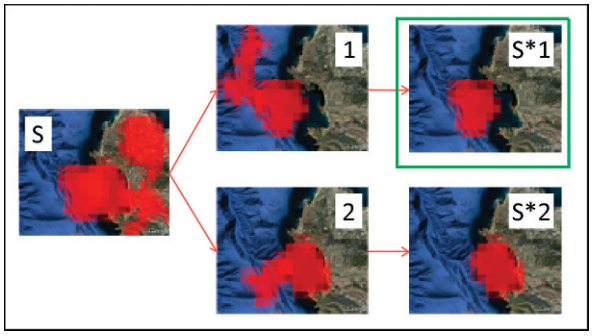


Figure 5. Comparing expert versus student actions with occupancy grids.

Table 7 shows hypothetical data as an example. In the example, three metrics have been applied to four trainees. The metrics include: ScanRelevance, which is the overlap between expert and trainee occupancy grids from eye tracking data; RadarZoom, which is the overlap between expert and trainee occupancy grids from radar center of view/zoom settings, and ResponseTime, which is the number of seconds from the appearance of a new track until it is hooked by the trainee. *Figure 6* shows a scatter plot for each pairing of two variables with the hypothetical data.

The values for RadarZoom and ScanRelevance are strongly correlated; they lie nearly on a straight line. This means either can be accurately predicted from the

Table 7. Hypothetical metric data.

Trainee	ScanRelevance	RadarView	ResponseTime
1	.80	.75	8
2	.50	.55	4
3	.49	.40	7
4	.74	.81	3

other, so there is no need for both. Thus—in this hypothetical sample—trainees who correctly select radar settings also tend to focus visual attention on the most important areas. ResponseTime, in contrast, is not strongly correlated with either of the other metrics. From these data, PCA would identify two dominant dimensions: The first would correspond closely with both ScanRelevance and RadarView, and the second with ResponseTime.[1]

The second step of the expertise model assesses general expertise. For this aspect of the diagnosis, an instructor assesses the general expertise of each trainee by watching the trainee execute a task scenario. A model of the instructor's assessment is trained using multiple linear regression and the trainee's skill ratings as predictors. Models for different expertise levels (i.e., novice, journeyman, expert) use different skills (Klein and Hoffman 1992), so the expertise model is particular to each skill level. The model also reveals the importance of each skill in the instructor's general assessment of expertise. The model is intended to yield several insights:

- The system simulates the instructor's assessment of general expertise of trainees in the future.
- If a skill does not contribute significantly to overall expertise, it might be because the skill is not very important. Alternately, it might be that the selected task scenarios do not exercise the skill, and additional scenario development is needed.
- If the model does not fit the instructor assessments very well, it may be that the set of metrics (and physiological metrics) is insufficient, and new metrics should be added. Or, overall expertise might be a nonlinear function of the skills. In this case nonlinear models (e.g., neural networks, support vector machines, etc.) could be explored. Alternately, the instructor's assessments might simply be subjective and unreliable.
- Creating models for several instructors would allow for determination of whether instructors are consistent with each other in assessing expertise and placing value on particular skills.

The expertise model was explored by prototyping the algorithms for the model. The prototype was implemented using synthetic data, so the associated results (such as figures showing the contribution of specific metrics to the expertise model) are notional and serve only to illustrate the expertise model concept. In developing the prototype, we simulated a subject population of 75 students grouped into three levels of expertise (novice, intermediate, and expert) for the set of metrics presented in *Table 8*, which lists the population mean and standard deviation for each metric broken down by level of expertise. The units for each metric in the synthetic data set are not specified (e.g., negative values have no special significance).

In the prototype, PCA was performed on the data for each level of expertise independently to explore the hypothesis that different skills are developed at each level of expertise. *Figure 7* shows a "scree plot" for components of variance (skills) for intermediate-level students. This plot shows that most of the variance from the 14 original metrics is explained by only the first 2 principal components (52%), and the first 4 capture 80%, while the first 6 metrics capture 90% of the metrics. Thus it is possible to construct new composite metrics to simplify trainee assessment.

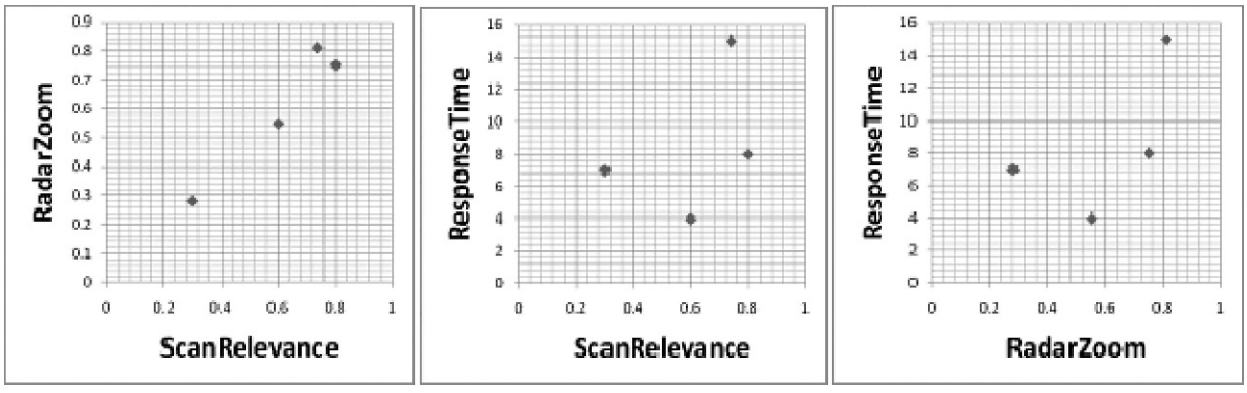


Figure 6. Scatter plot for each pairing of two variables with the hypothetical data.

Table 8. Synthetic data.

	Mean			Standard deviation		
	Novice	**Intermediate**	**Expert**	**Novice**	**Intermediate**	**Expert**
RADARView	1.05	4.63	7.87	2.06	1.67	1.72
ReactionTime	3.14	4.92	7.14	1.18	1.73	1.89
ResponseTime	6.04	8.01	10.19	2.42	2.15	2.19
Workload	10.60	−1.17	−16.71	2.31	2.08	2.90
Engagement	2.01	2.12	2.05	0.49	0.39	0.62
Distraction	0.83	0.90	1.25	1.27	0.82	1.07
Drowsiness	3.55	4.27	3.48	1.85	1.96	1.38
GazeCoverage	13.30	26.76	61.92	3.85	4.35	3.84
GazeRelevance	15.71	41.79	96.01	4.52	4.11	4.67
GazeTargetTime	19.21	57.07	65.92	4.34	3.11	4.47
GazeTargetDuration	1.46	0.72	−7.45	1.67	1.30	1.47
GazeToolFixations	11.42	−16.52	−15.48	4.47	2.94	3.41
BlinkRate	10.93	10.05	12.20	3.06	2.54	2.54
PupilSize	5.05	5.38	4.85	1.26	1.21	1.39

As such, composite metrics were extracted. Each of the principal components is a composite metric, which is a combination of the 14 original metrics. But in most of the composite metrics, only a few of the original metrics have significant influence. For the intermediate trainee in the synthetic data set, most of the weight in the first principal component is assigned to Distraction. Metrics that do not contribute significantly to the composite metrics may be discarded entirely. *Figure 8* shows the original 14 metrics projected onto the three first principal components, which reveals which of the original metrics best align with the principal components. This information is used to derive meaningful names for the composite metrics.

Based on the aforementioned three-tier diagnoses (DM performance, learning state, and expertise), it was then necessary to identify how these streams of data would be integrated to identify adaptation trigger points. First, the diagnosis engine would continuously assess cognitive state based on neurophysiological measures, including levels of workload, engagement, distraction, and drowsiness. These assessments would be based on predefined thresholds and evaluate adequacy of cognitive learning state. Second, the diagnosis engine would assess predefined behavioral and physiological (i.e., eye tracking) performance metrics associated with each step in the DM process (see description of the SHOR DM model; Wohl 1981). Third, the diagnostic engine would identify the level of expertise the trainee's performance and state that most closely matches based on a combination of all relevant performance and state metrics. Based on outputs from these two steps, the diagnostic engine would place the trainee within one of three categories:

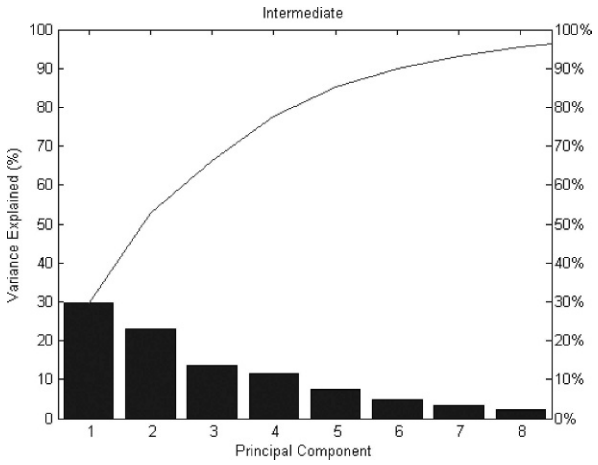


Figure 7. Scree plot for intermediate level of expertise.

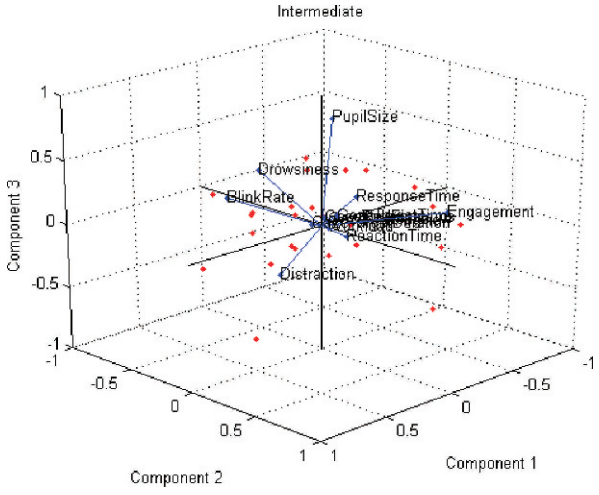


Figure 8. The original 14 metrics projected onto the three first principal components, which correspond roughly with engagement, drowsiness, and radar view settings.

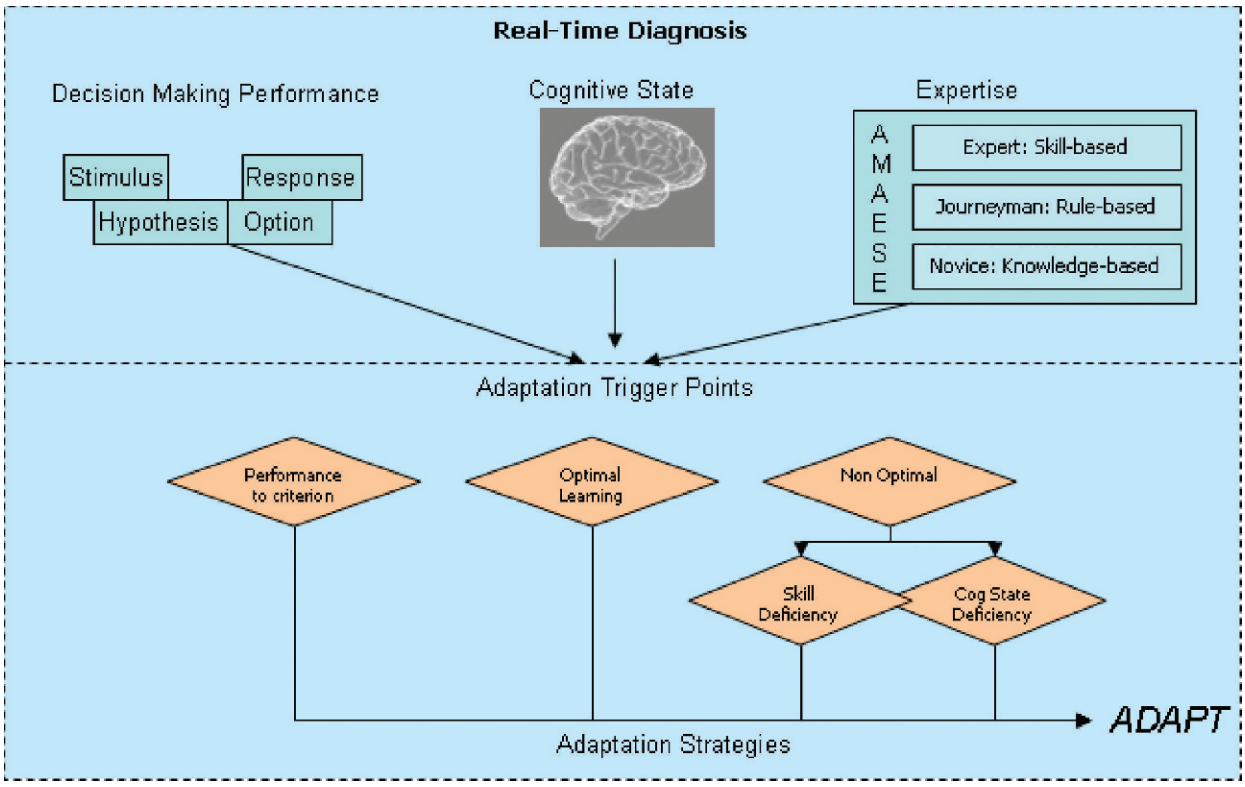


Figure 9. ADAPT-DM real-time diagnosis concept.

1. *Performance to criterion* in which the trainee's performance is effective and efficient across a broad range of situations,
2. *Optimal learning state* in which a trainee's performance is effective; however, practice is necessary to increase efficiency and build experience base,
3. *Nonoptimal learning state* in which the trainee is having performance or state issues that need remediation or cognitive state issues that need mitigation.

Students in the last category would be further categorized based on performance and state indicators to pinpoint the root cause of nonoptimal learning state, specifically identifying whether there was a skill deficiency or a cognitive state deficiency of drowsiness or distraction. Based on these categorizations and the context-specific performance measures, appropriate adaptations would be triggered *(Figure 9)*.

Table 9 presents the generalizable diagnosis matrix that shows precisely how the streams of data will be combined and resulting diagnoses.

Conclusions

This effort has resulted in conceptualization of the ADAPT-DM framework for supporting precision training, which is adaptive to trainees' differing needs, skill proficiency levels, learning states, and expertise levels. Implementation of this framework into a training system should accelerate DM skill development by

- Developing a comprehensive picture of a trainee's knowledge, skills, and cognitive state through continuous performance and state measurement.
- Using sophisticated models of expert and novice performance to evaluate expertise, along with performance and learning state, to understand key deficiencies and opportunities to accelerate learning.
- Ensuring an optimal mix of experiences and instruction (such as real-time feedback, real-time scenario modification, and automated cueing and scaffolding strategies) to rapidly develop robust and effective DM skills.

Through root cause analysis based on physiological and neurophysiological data, ADAPT-DM goes beyond simply assessing whether trainees made good decisions. Process level measures become feasible, enabling instructors to pinpoint where in the DM process breakdowns occurred. The expected benefits of a system based on the ADAPT-DM framework are

- Training is compressed and accelerated because the system detects and adapts to the acquisition of specific skills, learning state, and expertise.

Table 9. General diagnoses.

Performance measures						
Performance	**Workload/difficulty**	**Engagement**	**Distraction**	**Drowsiness**	**Expertise**	**Diagnosis**
Good	Low	High	Low	Low	Expert	Criterion
					Journeyman	Optimal learning state
					Novice	
		Low	Low	High	Expert	Criterion
					Journeyman	Nonoptimal learning: drowsy
					Novice	
		Low	High	Low	Expert	Criterion
					Journeyman	Nonoptimal learning: distracted
					Novice	
	High	High	Low	Low	Expert	Optimal learning state
					Journeyman	Optimal learning state
					Novice	
		Low	Low	High	Expert	
					Journeyman	Nonoptimal learning: drowsy
					Novice	Nonoptimal learning: drowsy
		Low	High	Low	Expert	
					Journeyman	Nonoptimal learning: distracted
					Novice	Nonoptimal learning: distracted
Bad	Low	High	Low	Low	Expert	
					Journeyman	Skill deficiency
					Novice	Skill deficiency
		Low	Low	High	Expert	
					Journeyman	Nonoptimal learning: drowsy
					Novice	Nonoptimal learning: drowsy
		Low	High	Low	Expert	
					Journeyman	Nonoptimal learning: distracted
					Novice	Nonoptimal learning: distracted
	High	High	Low	Low	Expert	
					Journeyman	Skill deficiency
					Novice	Skill deficiency
		Low	Low	High	Expert	
					Journeyman	Nonoptimal learning: drowsy
					Novice	Nonoptimal learning: drowsy
		Low	High	Low	Expert	
					Journeyman	Nonoptimal learning: distracted
					Novice	Nonoptimal learning: distracted

- Trainees are better prepared for live training and operations by ensuring an optimal experience base.
- Seamless integration with existing DM trainers. ❑

Dr. MEREDITH CARROLL is a senior research associate at Design Interactive, Inc., and is currently supporting design, development, and evaluation of performance assessment tools and virtual training systems for the office of Naval Research's Human Performance, Training and Education (HPT&E) Program. Her work focuses primarily on Individual and Team Performance Assessment, including physiological and behavioral measurement, performance diagnosis, and training remediation through feedback and training adaptation. She has also performed extensive work conducting Task Analyses, designing Virtual Training Environments and Training Management tools, and conducting training effectiveness evaluations. Her research has focused on human–team performance and training in complex systems in aviation and military domains, with the focus on perceptual skills. She received her bachelor of science degree in aerospace engineering from the University of Virginia, her master of science degree in aviation science from Florida Institute of Technology, and doctor of philosophy degree in human factors and experimental psychology from the University of Central Florida. E-mail: Meredith@designinteractive.net

SVEN FUCHS is a senior research associate at Design Interactive, Inc., where he is currently involved with the development of adaptive training frameworks that are enhanced with physiological sensors. He previously sup-

ported several Augmented Cognition research projects sponsored by DARPA, IARPA, and ONR. Sven has authored and coauthored over 10 papers and 3 book chapters on Augmented Cognition. In 2006, he was named an "Augmented Cognition Ambassador" and received the Augmented Cognition International Society's "Foundations of Augmented Cognition" award. Other interest areas include usability evaluation, multimodal interface design, and innovative human–system interface technologies. He holds an undergraduate degree in computer science for media from the Flensburg University of Applied Science in Germany and was a Fulbright scholar at DePaul University, Chicago, where he earned a master of science degree in human-computer interaction. E-mail: Sven@designinteractive.net

Angela Carpenter is a research associate in Design Interactive, Inc.'s Human Systems Integration (HSI) division with 4 years of HSI experience. Her work has focused on use of multimodal design science to optimize operator situational awareness and workload in C4ISR environments, development of neurophysiological metrics to assess signal detection, and assessment of astronaut cognitive state via a handheld game. She received her master of science degree in human factors and systems at Embry-Riddle Aeronautical University with a focus on human–systems integration and systems engineering, where she was a graduate teaching assistant involved in pilot training studies. She earned a bachelor of arts degree from Flagler College in psychology and Spanish. E-mail: Angela@designinteractive.net

Dr. Kelly Hale is the human–systems integration director at Design Interactive, Inc., and has over 10 years' experience in human-computer interaction, training transfer, usability evaluation methods of advanced technologies, and cognitive processes within multimodal systems and virtual environments. Kelly has been involved with DARPA's AugCog program and ONR's VIRTE program, and has gained extensive experience in evaluating various devices using a variety of empirical and nonempirical usability evaluation methods. She has been principal investigator of multiple Phase I and II SBIR efforts funded by ONR, DARPA, and NASA, and was research and technical lead of innovative augmented cognition technology development funded by IARPA. Kelly directed Multimodal Information Design Support product development, which is currently being transitioned into the IMPRINT modeling software. Kelly holds a bachelor of science degree in kinesiology/ergonomics from the University of Waterloo, Canada, and master of science and doctor of philosophy degrees in industrial engineering from the University of Central Florida. E-mail: Kelly@designinteractive.net

Dr. Robert G. Abbott is a principal member of the technical staff in the Cognitive and Exploratory Systems group at Sandia National Laboratories, where his team develops software for automated behavior modeling. He holds a doctor of philosophy degree in computer science from the University of New Mexico. He has been a member of the technical staff at Sandia since 1999. His current research focuses on automating the creation of human behavior models with the objectives of reduced cost and rapid development. Applications include trainable software agents to assume the roles of friendly and opposing forces, and automated student assessment for distributed virtual training environments. This line of research is supported primarily by the U.S. Navy and includes validation experiments with human subjects to assess the impact of new training technologies. Other research interests include distributed systems, security-related data mining, and computer vision. E-mail: rgabbot@sandia.gov

Dr. Amy Bolton is a program officer at the Office of Naval Research where she manages applied research and advanced technology demonstration projects as part of the Code 34, Capable Manpower Future Naval Capability program. Dr. Bolton's portfolio includes manpower, personnel, training and human system design projects. The aim of the projects is to increase human performance by advancing training methodologies and technologies, devising better personnel selection tools, and designing better systems to take the operator's capabilities and limitations into consideration. Prior to joining the Office of Naval Research, Dr. Bolton was a research psychologist at the Naval Air Warfare Center Training Systems Division in Orlando, Florida where she served in the role of principal investigator on a variety of research efforts. E-mail: amy.bolton@navy.mil

Endnotes

[1]It would also identify a third dimension but with a very small eigenvalue, indicating that the third dimension is negligible.

References

Abbott, R. G. 2006. Automated expert modeling for automated student evaluation. *Intelligent Tutoring Systems* 4053: 1–10.

Berka, C., D. J. Levendowski, M. Cvetinovic, M. M. Petrovic, G. F. Davis, M. N. Lumicao, M. V. Popovic, V. T. Zivkovic, and R. E. Olmstead. 2004. Real-time analysis of EEG indices of alertness, cognition and memory acquired with a wireless EEG headset. *International Journal of Human-Computer Interaction* 17 (2): 151–170.

Berka, C., D. J. Levendowski, M. N. Lumicao, A. Yau, G. Davis, V. T. Zivkovic, R. E. Olmstead, P. D. Tremoulet, and P. L. Craven 2007. EEG correlates of task engagement and mental workload

in vigilance, learning, and memory tasks. *Aviation Space and Environmental Medicine* 78 (5, Suppl.): B231–B244.

Berka, C., D. J. Levendowski, P. Westbrook, G. Davis, M. N. Lumicao, R. E. Olmsted, et al. 2005. EEG quantification of alertness: Methods for early identification of individuals most susceptible to sleep deprivation. In *Proceedings of the SPIE Defense and Security Symposium, Biomonitoring for Physiological and Cognitive Performance during Military Operations,* ed. J. A. Caldwell and N. J. Wesensten, vol. 5797, 78–89. Orlando, FL: SPIE: The International Society for Optical Engineering.

Bruner, J. 1973. *Going beyond the information given*. New York: Norton.

Craik, F. I., R. Govoni, M. Naveh-Benjamin, and N. D. Anderson 1996. The effects of divided attention on encoding and retrieval processes in human memory. *Journal of Experimental Psychology General* 152 (2): 159–180.

Dorneich, M. C., S. D. Whitlow, S. Mathan, P. M. Ververs, D. Erdogmus, A. Adami, M. Pavel, and T. Lan. 2007. Supporting real-time cognitive state classification on a mobile individual. *Journal of Cognitive Engineering & Decision Making* 1 (3): 240–270. (Special Issue on Augmented Cognition: Past, Present, and Future).

Elliot, T., M. Welch, T. Nettlebeck, and V. Mills. 2007. Investigating naturalistic decision making in a simulated microworld: What questions should we ask? *Behavior Research Methods* 39 (4): 901–910.

Endsley, M. R. 1995. Toward a theory of situation awareness in dynamic systems. *Human Factors* 37 (1): 32–64.

Foerde, K., B. J. Knowlton, and R. A. Poldrack. 2006. Modulation of competing memory systems by distraction. *Proceedings of the National Academy of Sciences of the United States of America* 103 (31): 11778–11783.

Hyönä, J., R. Radach, and H. Deubel, eds. *The mind's eye: Cognitive and applied aspects of eye movement research*. Amsterdam, The Netherlands: North-Holland.

Klein, G. A. 1998. *Sources of power: How people make decisions*. Cambridge, MA: MIT Press.

Klein, G. A., and R. R. Hoffman. 1992. Seeing the invisible: Perceptual-cognitive aspects of expertise. In *Cognitive science foundations of instruction,* ed. M. Rabinowitz, 203–226, Mahwah, NJ: Erlbaum.

Klein, G. A., and K. J. Peio. 1989. Use of a prediction paradigm to evaluate proficient decision making. *The American Journal of Psychology* 102 (3): 321–331.

Levonian, E. 1972. Retention over time in relation to arousal during learning: An explanation of discrepant results. *Acta Psychological* 36 (4): 290–321.

Lipshitz, R., G. Klein, J. Orasanu, and E. Salas. 2001. Taking stock of naturalistic decision making. *Journal of Behavioral Decision Making* 14 (5): 331–352.

Macklin, C., M. J. Cook, C. S. Angus, C. S. G. Adams, S. Cook, and R. Cooper. 2002. Qualitative analysis of visualisation requirements for improved campaign assessment and decision making in command and control. In *21st European Conference on Human Decision Making and Control,* ed. C. W. Johnson. Glasgow, Scotland: Department of Computing Science, University Of Glasgow, GIST Technical Report G2002-1. http://www.dcs.gla.ac.uk/~johnson/eam2002/EAM_2002.pdf (accessed February 21, 2010).

Neri, D. F., D. F. Dinges, and M. R. Rosekind. 1997. *Sustained carrier operations: Sleep loss, performance, and fatigue countermeasures*. Moffett Field, CA: NASA Ames Research Center, http://human-factors.arc.nasa.gov/zteam/PDF_pubs/Nimitz1997.pdf (accessed July 24, 2009.

Poythress, M., C. Russell, S. Siegel, P. D. Tremoulet, P. Craven, C. Berka, and D. J. Levendowski. 2006. Correlation between expected workload and EEG indices of cognitive workload and task engagement. In *Augmented cognition: past, present and future,* ed. D. Schmorrow, K. Stanney, and L. Reeves, 32–44. Arlington, VA: Strategic Analysis, Inc.

Rasmussen, J. 1983. Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models. *IEEE Transactions on Systems, Man, and Cybernetics* 13 (3): 257–266.

Reason, J. 1990. *Human error*. Cambridge, UK: Cambridge University Press.

Small, R. V., B. J. Dodge, and X. Jiang 1996. Dimensions of interest and boredom in instructional situations. In *Proceedings of the 1996 Conference of the Association for Educational Communications and Technology*. Washington, D.C.: AECT Publications.

Stevens, S. M., J. C. Forsythe, R. G. Abbott, and C. J. Gieseler. 2009. Experimental assessment of accuracy of automated knowledge capture. In *Foundations of Augmented Cognition, HCII 2009,* San Diego, CA. Berlin, Germany: Springer.

Stevens, R., T. Galloway, and C. Berka. 2007. Allocation of time, EEG-engagement and EEG-workload resources as scientific problem solving skills are acquired in the classroom. In *Proceedings of 3rd Augmented Cognition International, held in conjunction with HCI International 2007*, Beijing, China, July 22–27, 2007. Heidelburg, Germany: Springer.

Wohl, J. G. 1981. Force management decision requirements for Air Force tactical command and control. *IEEE Transactions on Systems, Man and Cybernetics* 11 (9): 618–639.

Acknowledgments

This material is based upon work supported in part by the Office of Naval Research under STTR contract N00014-09-M-0326. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views or the endorsement of the Office of Naval Research. We are grateful to Wendi Van Buskirk and her team at the Naval Air Warfare Center Training Systems Division and the Navy subject matter experts who helped guide the content and design of the ADAPT-DM framework.

ITEA Journal 2010; 31: 264–274

Process Instrumentation Systems for Training and Operational Test Needs With Case Study of Use at JEFX 09

Jennifer Ockerman, Ph.D., F. T. Case, Nathan Koterba, and Greg Williams
Johns Hopkins University Applied Physics Laboratory, Laurel, Maryland
Glenn Conrad
MITRE, Pensacola, Florida
Susi McKee
U.S. Air Force 505th Command and Control Wing, Hurlburt Field, Florida
Oscar Garcia
Air Force Research Laboratory,
Warfighter Readiness Research Division, Mesa, Arizona
James Welshans, Ed.D.
Teledyne CollaborX, Navarre, Florida

Process instrumentation provides trainers and testers with deeper insight into the activities, both human and technology, that affect system and warfighter performance and mission effectiveness. To this end, the authors have designed and developed several process instrumentation systems, working with both U.S. Air Force warfighter trainers and testers. Both trainers and testers need to, in operationally relevant environments, reconstruct events to meet teaching needs or testing requirements. These similar reconstruction activities drive the need for a method to capture human, as well as technology, activities. This article describes process instrumentation systems and their benefits for trainers and testers and then describes, with examples and a case study, two process instrumentation systems.

Key words: Chat systems; communication; human activity; instrumentation; JEFX-09 experiment; reconstruction; trainee decisions; training; work process.

In training and test environments, the instrumentation of user–operator work processes and communications, in addition to the instrumentation of technology (e.g., computer systems and networks), can provide new and needed insights into mission effectiveness. In training environments, trainers can use the information collected and displayed by process instrumentation to reconstruct training events and augment after-action reviews, highlighting key trainee decisions. Similar information and reconstructions can assist operational testers in determining a system-under-test's impact on total performance and mission effectiveness.

During both training and test events, trainers and testers develop scenarios using Master Scenario Event List (MSEL) inputs. Both use these MSELs to stimulate desired outcomes for focused observation, analysis, and debrief and reporting. Training after-action reviews rely on reconstruction technology and human activities to determine student performance and reinforce teaching points. Test analysis of MSEL test inputs involves the reconstruction of technology and human activities to determine the efficacy of the system or network under test.

Over the past 5 years, the Johns Hopkins University Applied Physics Laboratory (JHU/APL) has developed several work process and personnel communication instrumentation systems. These systems have been used by the Time Sensitive Targeting/Dynamic Targeting trainers at the U.S. Air Force's Combined Air and Space Operation Center—Nellis (CAOC-N) and researchers at the Air Force Research Laboratory—711th Human Performance Wing for several training research exercises focused on dynamic effects, as well as assessors of the Warplan-to-Warfighter Forwarder (WWF) Spiral II initiative during the Air

Force's Joint Expeditionary Force Exercise (JEFX) 09-3. In addition, the U.S. Joint Forces Command has incorporated several JHU/APL instrumentation capabilities into their Joint After-Action Review Resource Library.

This article provides an overview of the benefits of work process and personnel communication instrumentation in both training and test environments, a description of the instrumentation systems developed by JHU/APL with examples of their use by CAOC-N trainers, and a case study of WWF assessors during JEFX 09.

Benefits of work process instrumentation

Process instrumentation systems automatically collect, organize, and archive large amounts of work process and personnel communications data, as well as pertinent technology data. During events, instead of trainers and testers manually searching and recording the information they will later use for reconstructions and analyses, process instrumentation systems automatically, and with minimal intrusiveness, collect work process, communications, and system data. Collected data are organized and archived in databases, making it easily accessible to performance analysis and assessment systems, as well as searchable with user-specified criteria. Such instrumentation significantly improves the ability of trainers and testers to evaluate overall mission effectiveness, assess human-centric issues, such as effectiveness of Command and Control (C2), and contribute to Tactics, Techniques, and Procedures (TTP) development.

In addition to relieving trainers and testers of the burden of manually collecting data in real time during an event and the danger of missing key information because of trainer or tester task overload or data inundation, the persistent archive of data by process instrumentation systems creates a historical record of events and activities. Nowadays, all too often, the work process, communications, and system data during training and test events is lost after the event ends or is captured in nondigital, not easily accessible Information Technology (IT) system formats, like Microsoft Office documents or proprietary, stove-piped data sources. Without historical data to establish process and mission effectiveness baselines, the benefits of training programs and new systems under test become difficult to determine and are often subjectively assessed.

The ability to archive data and refer to it later also enables trend analysis. This can be particularly useful in the training environment to gauge students' progress in the training process. Trend analysis can also provide insights into effectiveness of training procedures as well as improvements over time of a technology being tested—given analysis is conducted to isolate contributions to effectiveness of specific material systems, processes, or learning.

Process instrumentation systems also provide the capability to reconstruct both technology and human activities, using empirically based "truth" data, that is, data from IT systems used by humans to complete an activity or process, as opposed to notes or memories of the participants, trainers, and testers. A reconstruction can take many forms but is most useful if it contains both technology actions and human actions that have been correlated and can be examined through specific events (e.g., MSEL items) in a larger mission context, often called a thread. With the activities correlated and focused on a single event thread, testers and trainers can better understand the timing and process actions to evaluate the performance impact of the MSEL input on the overall mission. Additional context from other threads provide insights into how C2 (e.g., leadership, team coordination, and decision making) and TTP positively (or negatively) influenced performance outcomes. With the technology and human activities correlated, attention is focused on total performance and mission effectiveness, as opposed to just warfighters' opinions.

Based on extensive JHU/APL observations, event reconstruction is often performed manually, consuming much time and labor, and limiting the amount and depth of analysis. Automating the collection and analysis of user–operator process data and communications results in a reduction in time and labor as well as improved quality in reconstruction and analysis. Existing instrumentation systems, such as the Air Combat Maneuvering Instrumentation and the Nellis Air Combat Training System used for tactical aircrew training, support this position.

A distinction needs to be made between IT system instrumentation, which focuses on data and system operation validation, and process instrumentation. Process instrumentation brings the human aspects of a system, such as workflows, decision making, and collaboration and communication, into focus for the training and testing communities and allows a full accounting of total system performance and mission effectiveness.

Process instrumentation systems: CPAS 1.2 and 1.3

To research and exploit the benefits of process instrumentation, JHU/APL developed two systems, the CAOC Performance Assessment System (CPAS) v1.2 and CPAS v1.3 (or eCPAS for Enhanced CPAS).

JHU/APL built CPAS v1.2 for the CAOC-N trainers. This version of CPAS collects work process

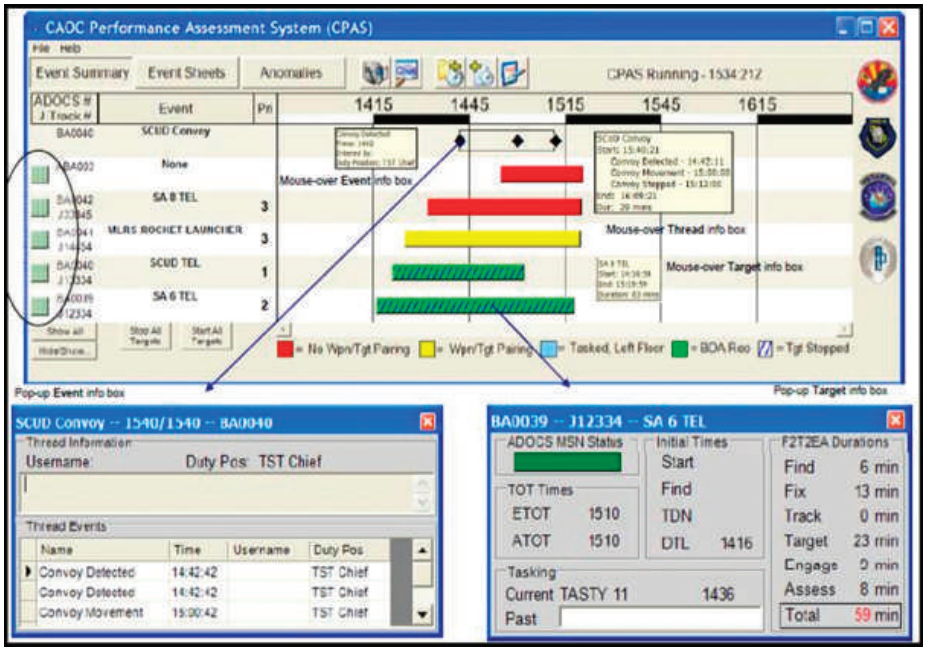


Figure 1. CPAS composite and drill-down views.

data from the Joint Automated Deep Operations Coordination System (JADOCS) mission collaboration technology and communications data from several text chat collaboration systems. CPAS stores the collected data in a relational database and displays it visually using timelines. Other capabilities include filtering and searching the collected communication messages, correlating key process events with communication messages, and detecting process anomalies using user-configured parameters.

A later version of CPAS, Enhanced CPAS (eCPAS), was customized to meet testing requirements from the Air Force's 605th Test and Evaluation Squadron. eCPAS added the capabilities to collect Link-16 messages, computer screen displays, and voice communications, as well as transcribe voice communications into text for improved search and correlation. Multiple data sources, including voice, Link-16, chat, and JADOCS remarks, may be searched simultaneously and displayed in user-customizable timelines.

CPAS version 1.2

CPAS was delivered to the Air Force in 2006. The primary purpose of the CPAS instrumentation system was to support individual and team training of the Air Force's Air Operations Center's dynamic targeting (DT) processes. Operational trainers use CPAS to monitor in near–real time individual and collective trainee performance. They compare observed performance against expectations, especially in relation to the lesson objectives, MSEL events, and other stimuli. By monitoring these activities in near–real time, trainers can flag critical learning opportunities for later analysis and debriefing, collaborate with others, and frame their initial understanding of individual and collective trainee performance. Using CPAS performance feedback, the trainers and event managers can also monitor and manipulate the event timeline to assess progress toward achieving lesson or exercise objectives. This includes monitoring the CPAS-generated JADOCS timeline (e.g., workflow manager playbacks) and warfighter chat systems for correlated activity. Finally, trainers use CPAS to support training debriefs after the training events and reconstruct student events and activities.

CPAS uses a combination of composite activity views (see top of *Figure 1*) and drill-down focus views (see bottom of *Figure 1*). The timescale is located along the top of the view, with each major event thread indicated by the track number, event name, and priority. In Figure 1, the trainer is exploring the DT Chief's SCUD convoy chat activity alongside the SA-6 TEL targeting process. Note the amount of detail CPAS displays (see *Figure 2*), mapping actual events onto the U.S. military's joint kill chain process steps: Find-Fix-Track-Target-Engage-Assess (F2T2EA).

Through interactions with Air Force personnel, JHU/APL realized that different training groups and warfighting regions apply region- or mission-specific approaches to their process steps. As a result, CPAS allows users to tailor each of the major process views. *Figure 2* shows the basic F2T2EA process with discrete color-coded steps, starting and ending times

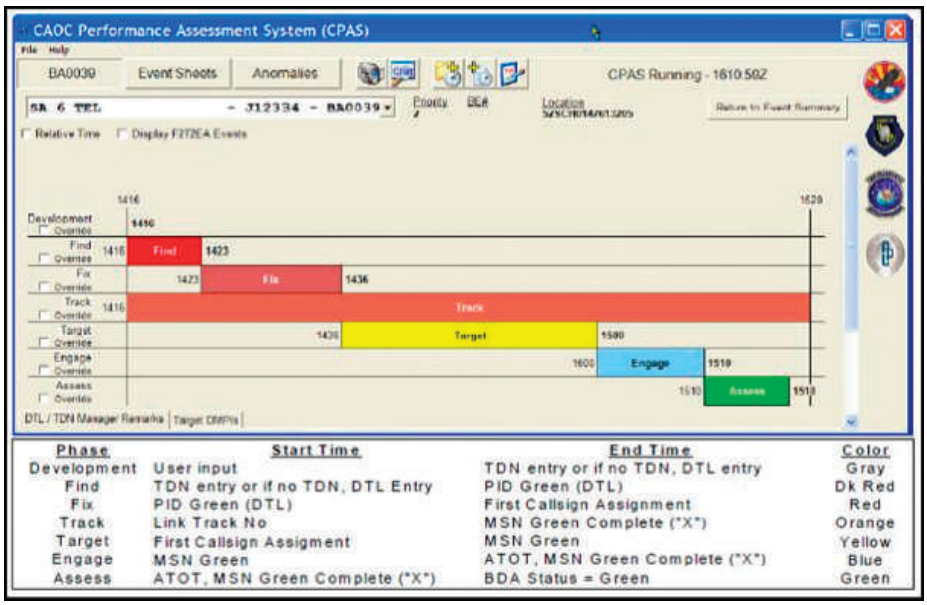


Figure 2. CPAS F2T2EA depiction of single mission.

defined as part of the CPAS configuration file for this unique training event.

CPAS was designed for realistic, dynamic scenarios and active instructor involvement in the training process. The system allows instructors to add their own notes, based on observations, trainee performance issues, or other potential learning opportunities, to the CPAS database.

Based on CAOC-N trainer requirements, CPAS exports relevant data fields into a Microsoft Excel spreadsheet for further analysis and development of debrief handouts.

The major analysis function in CPAS is a work-process anomaly detector (see *Figure 3*). Based on "school solution" process activities and timelines, trainers can compare individual and team trainee performance against the norm. In *Figure 3*, an intuitive display enables the trainer to recognize deviations from the expected work processes, providing instruction points for trainers to use during feedback sessions to the training audience.

CPAS augments trainer analysis and trainee performance feedback with a comprehensive JADOCS workflow manager playback feature (see bottom of

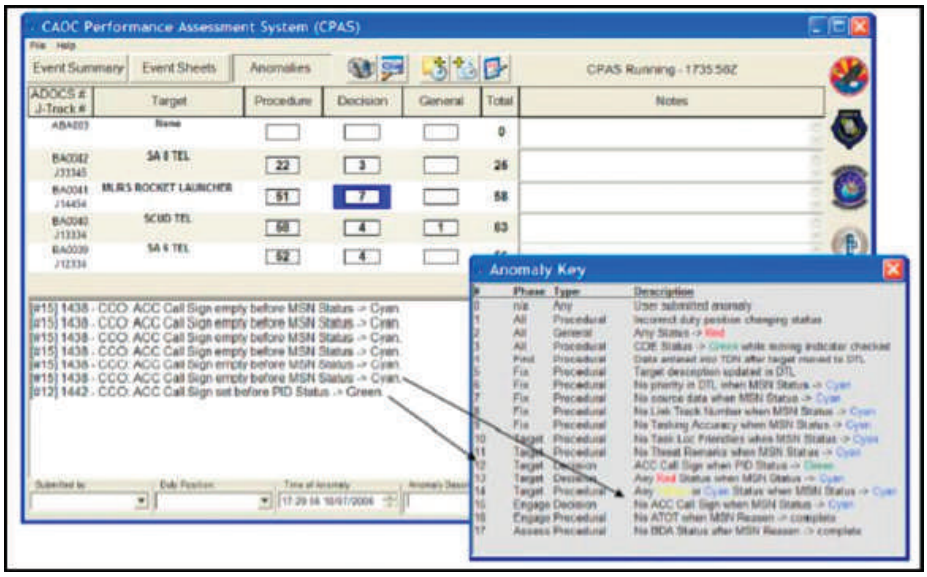


Figure 3. CPAS process anomaly display.

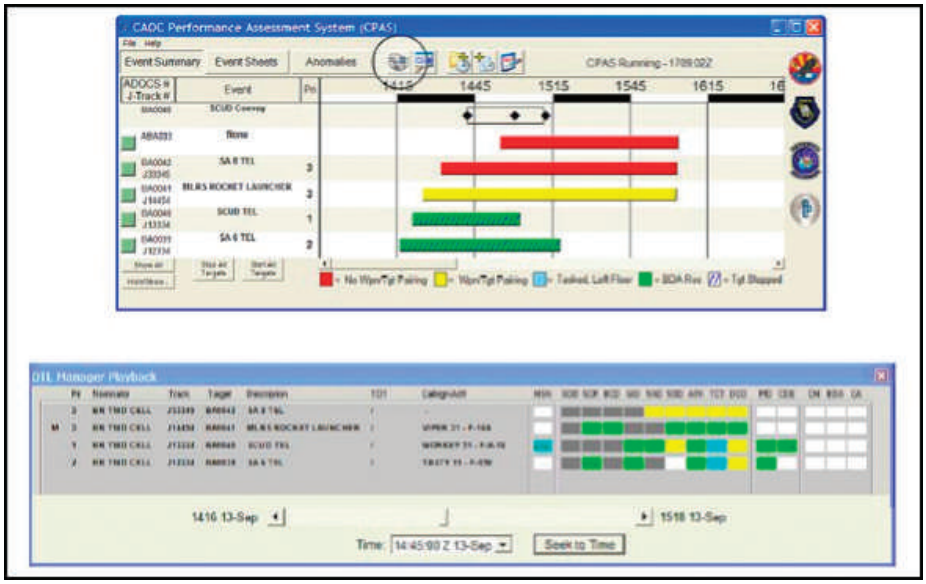


Figure 4. CPAS JADOCS manager playback display.

Figure 4). Using the slider bar at the bottom of the figure, the user can "playback" and observe activity and state changes in the respective JADOCS managers over time.

Based on JHU/APL's analysis of warfighter DT processes and TTP documentation, CPAS 1.2 focused on the chat environment as the primary context variable. This gave the instructor an increased understanding of "what's going on" among DT team members during a mission. *Figure 5* shows the wide variety of search and filter functions available for the trainer to analyze chat communications.

By collecting JADOCS and chat transactions, CPAS provided trainers with the ability to easily correlate these two disparate data sources for their trainee performance analysis and debriefing. *Figure 6* illustrates CPAS's correlation capability.

The capabilities CPAS provides to Air and Space Operations Center (AOC) DT trainers is enabling a transformation in the way AOC DT training is accomplished. CPAS reduces the subjective assessments of trainers and students, and creates more reliable, objectively recorded and measured facts. Already, CPAS functions as a training-force multiplier.

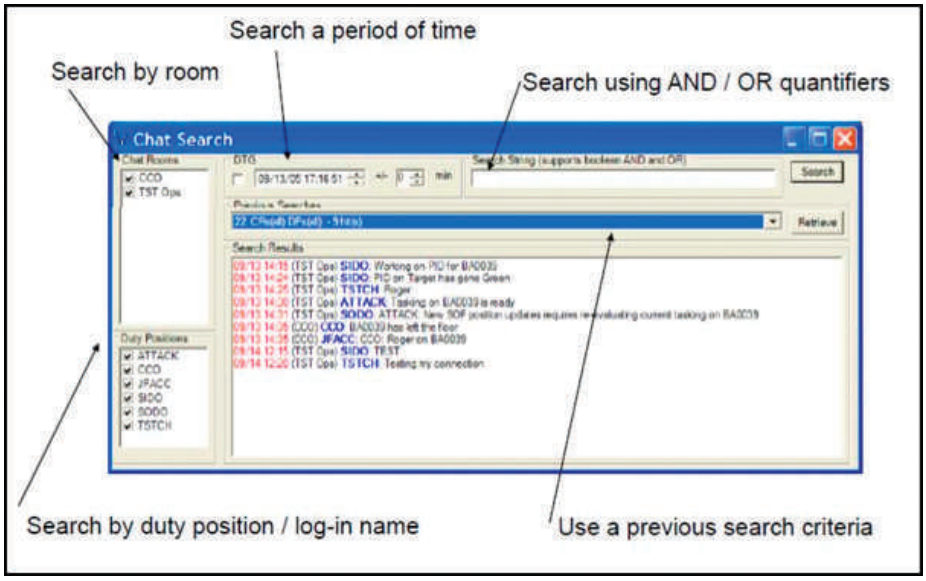


Figure 5. CPAS chat search interface.

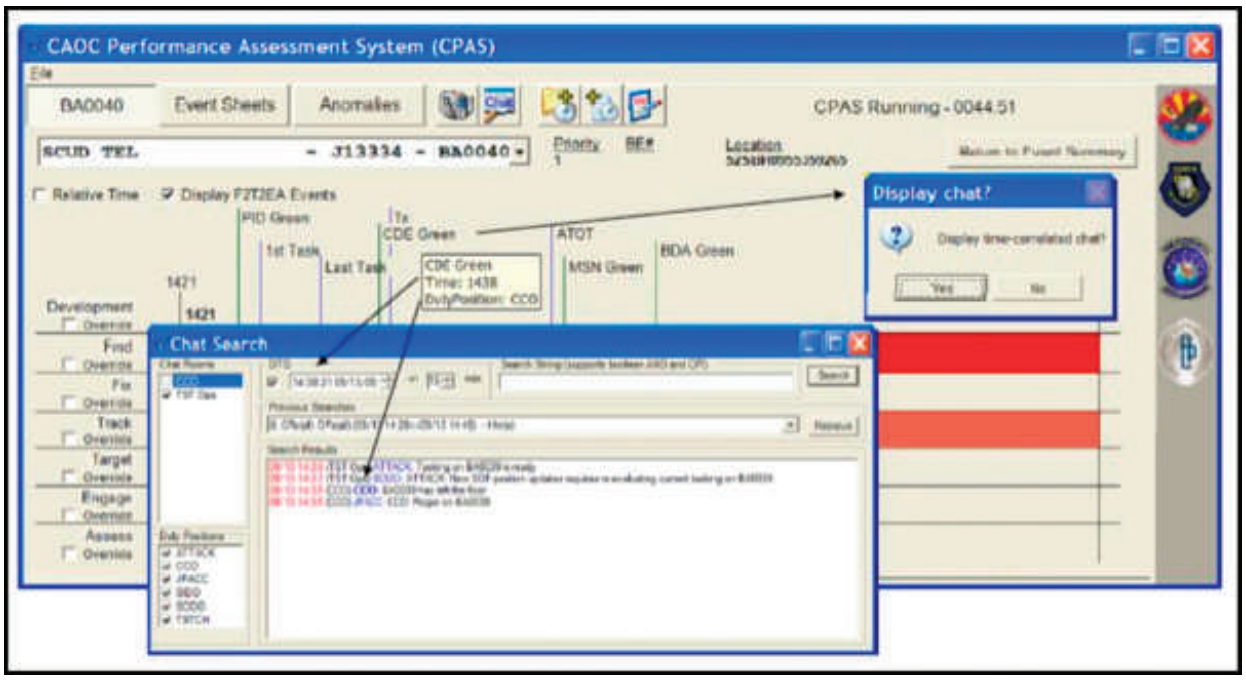


Figure 6. CPAS timeline-chat correlation displays.

No longer are trainers required to spend valuable training time manually collecting data for reconstruction. Instead, trainers have more time to observe, interact, and assess. Although not formally evaluated, the trainers have estimated that the use of CPAS has provided a 75% reduction in the amount of time required to prepare a debrief. Furthermore, CPAS has enabled the training debrief process to evolve from filling in key event times in a Microsoft Excel spreadsheet to spending more time analyzing the training threads and understanding what happened, when, and why.

CPAS version 1.3

CPAS 1.3, or eCPAS, was delivered to the Air Force in 2009. The primary purpose of the eCPAS instrumentation system was to support operational testing, as well as continue support for individual and team training of several of the AOC's combat operations division processes. eCPAS enables operational testers to monitor mission thread activities under test in near–real time. They can compare observed activities and events against expectations outlined in the test plan. By monitoring these activities in near real-time, they can flag critical observations for later analysis, collaborate with others, and frame their initial understanding of tested system performance in the context of the mission conditions, processes, and systems used.

Building on the successes of CPAS, eCPAS added several valuable capabilities. eCPAS retained the JADOCS timelines and workflow manager playbacks, as well as the warfighter chat systems for correlated activity. Based on operational tester requirements, eCPAS added the collection of additional chat systems, voice on C2 audio channels, and selected Link-16 messages. Testers can also identify significant observations via eCPAS by inserting flags directly onto the user interface. Flags can include both comments and screen captures to provide primary source data for anomaly analysis. The test team can take an instantaneous screen capture or set up periodic capture for the duration of the test event.

eCPAS has a "unified search" to support its correlation and analysis capability. *Figure 7* shows the increased fidelity of data capture that distinguishes eCPAS from CPAS. Testers can selectively search chat systems, rooms, and users; data link sources; and voice data sources. Testers can also query JADOCS data to view specific transactions from JADOCS' mission history files. In addition, testers can filter by keywords and times across all the selected communications simultaneously. Once the user-defined search data are returned, three different keywords can be highlighted throughout the data to show relevant trends and pick out threads. Finally, user-defined search data can be sent to a new timeline for display and analysis.

eCPAS improves on the CPAS framework for the JADOCS mission manager playbacks (see *Figure 8*) by making it easier for testers to configure and update the manager displays in real time.

Testers can inject events into the eCPAS database by generating process "flags" at any time during test event planning or execution (see *Figure 9*). Flags allow the testers to integrate their notes and observations with the

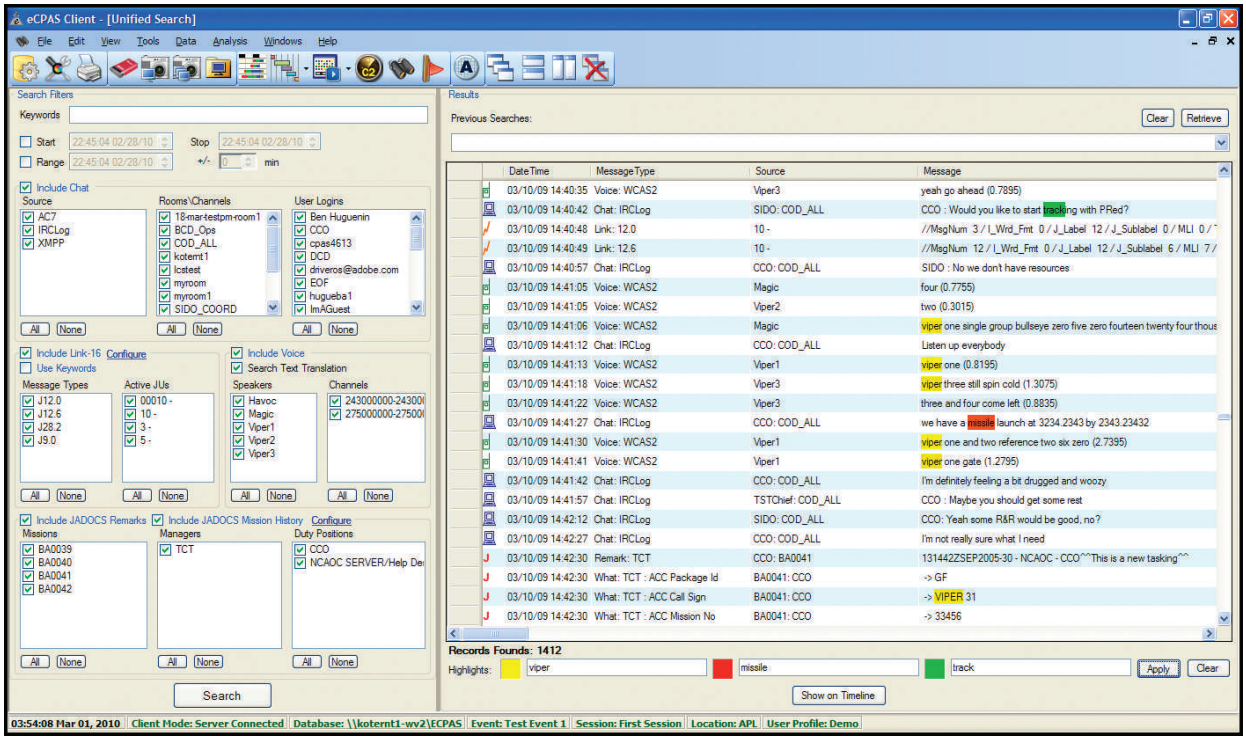


Figure 7. eCPAS unified search control display.

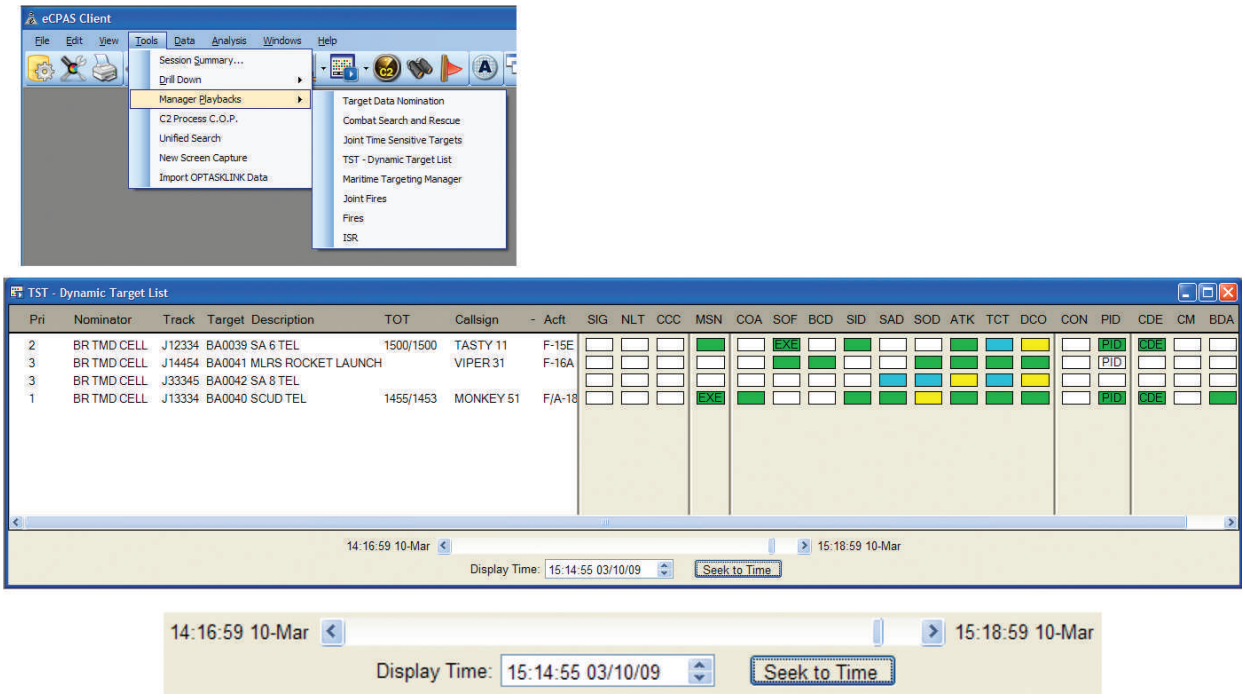


Figure 8. eCPAS JADOCS manager playback display.

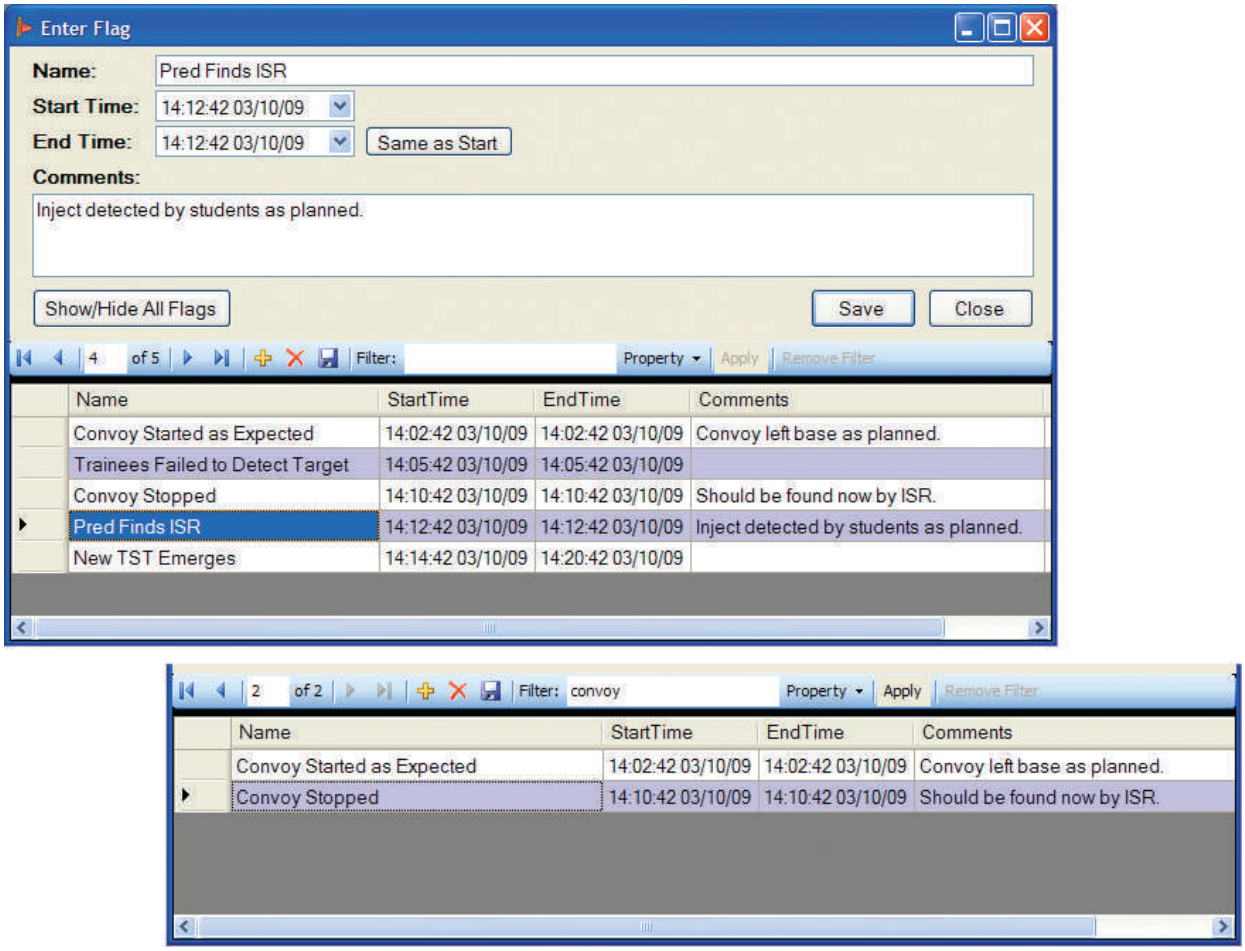


Figure 9. eCPAS "flag" entry.

rest of the data being collected by eCPAS. Multiple testers can create and edit these flags, which are viewable by the entire test team. Flags can be filtered (see bottom of *Figure 9*) for quick access and editing.

Another capability added in eCPAS allows testers to view timelines of data alongside one another. This new view, the Command and Control Process Common Operating Picture, or "C2 Process COP," is shown in *Figure 10*. The aggregate activity in JADOCS, chat, Link-16, and observer flags in area (1) shows the correlation of events in these different domains over time. This view would logically trigger the tester to conduct further analysis in this time slice and also provide the "at a glance" overview for the test team leadership. Conversely, the proliferation of observation flags (2) over a greater time slice might cause the tester to question why no activities have been recorded in the other (e.g., JADOCS, chat, Link-16) domains.

Table 1 provides an overview of the CPAS and eCPAS data sources, and the information they provide. In addition, it provides a one-sentence summary of each version's targeted use.

JEFX-09 case study

The goal of the Warplan-to-Warfighter Forwarder Spiral II (WWF II) system was to enable JADOCS, Target Package Generator, and Network Enabled Weapons Control Interface Manager to send machine-to-machine targeting data from the AOC directly to the airborne C2 aircraft, combat aircraft, or Network-Enabled Weapons (NEW) using Link-16 standards (MIL-STD-6016D) and allow automated status updates in a dynamic targeting environment. The task being performed by the team assessing WWF II in JEFX 09-3 was to analyze the timeliness, accuracy, and completeness of all the J-series messages that were transmitted on the network. Although not a formal operational test event, the JEFX 09-3 experiment was a suitable surrogate that included real operators and aircrew conducting operations in a live, virtual, and constructive environment. Live aircraft (F-15Es, F-16CMs, a B-2, and an E-3 AWACS) flew missions in the Nellis ranges; simulated aircraft, both virtual (E-8 JSTARS) and constructive (F-15E, F-16C), were located at Eglin Air Force Base (AFB), FL;

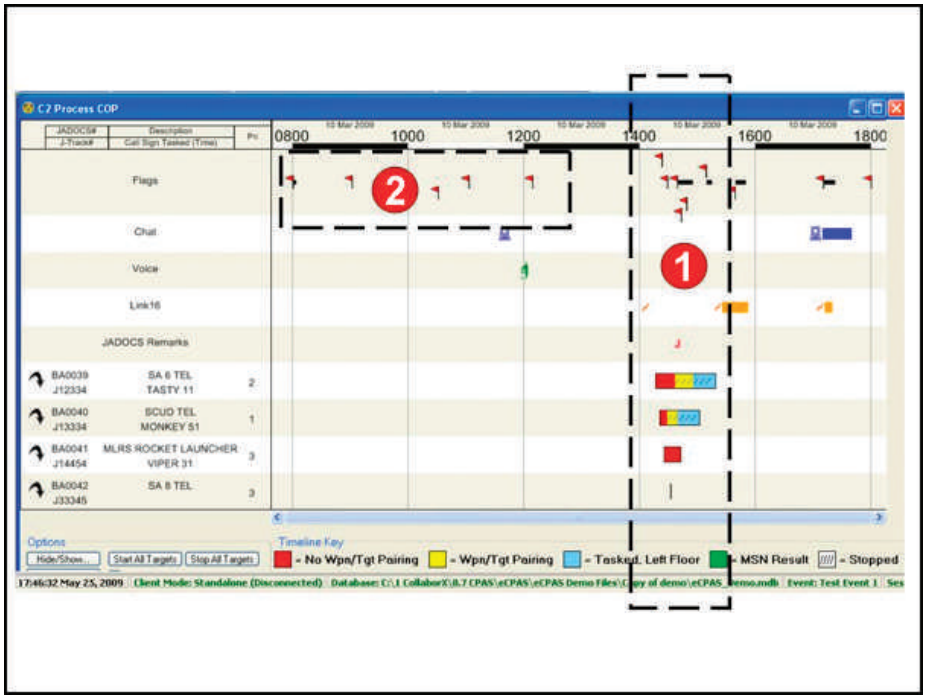


Figure 10. eCPAS C2 Process COP for visual correlation and event overview.

and a live AOC manned by real operators was in place at Nellis.

The primary technology used by one member of the assessment team to monitor, in near–real time, the WWF II system during JEFX 09-3 was JADOCS. The JADOCS workflow managers (Joint Time Sensitive Target Manager and the Intra-AOC Manager) together with the Coordination and Target Data tabs in each mission folder provided most of the data needed to maintain an understanding of the mission information flow. Other technologies used to build and maintain situational awareness during live-fly operations included the CPAS instrumentation displays, the XMPP chat client (transverse), the multilink translational and display system, the tactical view command and control, the joint windows warfare assessment model, and audible voice communications.

During both live-fly and modeling and simulation operations in the experiment, the event summary timeline display of CPAS version 1.2.5.3 was used to maintain insight into the state of each of the 15 to 20 missions that had been tasked. Multiple missions were prosecuted simultaneously; therefore, the event summary timeline feature helped the assessor to keep track of the missions and key events as they occurred. When required, further details of a specific mission were available by drilling down to the detailed mission phase's display from the event summary window. These drill-down views allowed analysis of key events in the context of the phases defined by WWF II assessors.

Another useful CPAS feature was the JADOCS workflow manager's playback displays. Using the Intra-AOC Manager playback, an assessor could quickly find when the critical columns changed status by viewing changes in the color and/or three-letter status code of the manager's display lights, known as "chicklets." Many times during the exercise, the JADOCS Intra-AOC Manager chicklets (which are incredibly small on an extremely crowded computer monitor) changed very quickly and an assessor had to be observing the specific chicklet at the exact moment to verify the transitory status. With the playback manager in CPAS, however, the assessor could easily go back and step through the scenario to verify which actual transactions occurred at which time. The CPAS chat search also proved useful by allowing different chat rooms and roles to be selected and then searched around a specific timeframe or for specific keywords. Thus, CPAS helped in tracking the various human and machine actions and chat communications, helping assessors reconstruct and understand the exact interactions among the systems under test.

Although eCPAS version 1.3.0.14 wasn't fully functional during JEFX 09-3, it did demonstrate its potential value with certain features, such as screen capture technology that helped the assessor to record and replay an operator's screens during the experiment. Other useful features, such as capturing Link-16 messages and displaying JADOCS remarks and mission history information in a timetable with the unified search function, proved their worth during detailed data

Table 1. CPAS/eCPAS capability summary.

	Data sources	Information provided
CPAS 1.2*	JADOCS	Dynamic targeting mission histories (e.g., who, what, when, where, how)
	IWS	Dynamic chat content (e.g., who, when, what)
	Jabber (XMPP)	
	mIRC	
CPAS 1.3†	JADOCS	Dynamic targeting mission histories (e.g., who, what, when, where, how)
	IWS	Dynamic chat content (e.g., who, when, what)
	Jabber (XMPP)	
	mIRC	
	Adobe Connect	
	Link-16	Dynamic data link content (e.g., C2 system messages between AOC and tactical units)
	Voice	Recorded .wav files of tactical communications
	Voice text	Transcribed files of tactical voice communications
	Video capture	Single view and/or sequenced screen shots of selected user work station displays

IWS = Ezenia InfoWorkspace, XMPP = eXtensible messaging and presence protocol, mIRC = Madam-Bey Internet relay chat.

*CPAS 1.2 supports individual and team training. The system correlates AOC dynamic targeting activities with chat, exposing activities in context to exploit learning opportunities.

†CPAS 1.3 supports individual and team training and operational testing. The system correlates AOC dynamic targeting activities with multiple interactive chat systems, data link messages, and voice transmissions exposing activities in a rich context to exploit learning opportunities and generate empirical assessments of systems under test.

analysis (using eCPAS version 1.3.0.27) after JEFX 09-3. The scalable timetable feature also helped assessors zoom into critical timeframes, such as when messages are sent from the AOC to identify potential problems when mission scenarios do not proceed as planned.

In summary, the CPAS process instrumentation systems proved useful as both experiment monitoring tools and analysis augmentation tools for WWF II assessors. These CPAS instantiations still require additional functionality to serve as the primary analysis tools. Specifically, feedback from the WWF II assessors indicates the need to expand the Link-16 data collection to other J-messages, as well as to pursue the ability to collect and transcribe voice communications with better accuracy.

Conclusion

The authors believe process instrumentation, which focuses on collecting information about human activities, provides great benefits to the training and test communities. Future efforts to improve and enhance the current CPAS capabilities will more fully realize the goal of process instrumentation and analysis. JHU/APL currently has, or is, applying the concept to Air Force Air Support Operations Center operations, U.S. Navy Maritime Operations Center operations, and the Joint Forces Command Joint After-Action Review Resource Library. In addition, JHU/APL is working on the next iteration of the C2 process instrumentation concept, the Operational Command and Control Instrumentation System (OC2IS, pronounced Oh-sis). OC2IS is an FY10 Resource Enhancement Program project sponsored by the Office of the Secretary of Defense's Test Resource Management Center and managed by the Air Force 505th Command and Control Wing. Although requirements for OC2IS are not yet finalized, potential enhancements include distributed data collection, remote data collector control, Web-based user interfaces, a service-based architecture, more robust voice transcription, automated analyses capabilities, workflow tracking, improved user configurations, more powerful reconstruction tools, and collectors for new data sources. By capturing data on the human view and perspective during process execution and then correlating that data to other data to determine the overall impact of the human-in-the-loop, OC2IS will contribute to the assessment of how well systems under test are meeting mission effectiveness requirements.

Along with developing technology to collect new types of data, JHU/APL is also researching how to best portray these data to provide meaningful insight into the processes and the human component of total system performance. By providing a better understanding of human activities and their effects, JHU/APL wants to develop capabilities to evaluate the impacts to C2 and TTP of new technologies and training methodologies. This includes tracing threads of human and technology activities, as well as evaluating human aspects, such as workload, situation awareness, and decision making for individuals, and the collaboration and coordination of teams. ❑

Dr. Jennifer J. Ockerman is employed as a senior cognitive systems engineer at JHU/APL, where she applies

cognitive engineering techniques to military C2 projects. Dr. Ockerman received her bachelor of science degree in industrial engineering and operations research from Virginia Tech and her master's degree and doctor of philosophy degree in industrial and systems engineering, with a specialty in human integrated systems and a minor in cognitive science, from Georgia Tech. She has over 50 publications and is a reviewer for several conferences and journals. She is a member of the Human Factors and Ergonomics Society (HFES), the Association of Computing Machines (ACM), and ITEA. E-mail: jennifer.ockerman@jhuapl.edu

F. T. Case joined JHU/APL in 1997 and serves as program manager of Operational C2. Mr. Case directs, manages, and provides technical contributions to several technology development programs focused on operational C2. He is a coinventor of the CPAS technology. Mr. Case is a former DARPA PM where he managed and directed the activities of multiple advanced technology C2 and modeling and simulation programs. He holds a master's degree in systems technology from the Naval Postgraduate School. He is a rated senior navigator with over 1,000 flying hours in fighter aircraft. Mr. Case is a member of the Military Operations Research Society, the Project Management Institute, and ITEA. E-mail: ft.case@jhuapl.edu

Nathan Koterba has 6 years' experience designing, developing, and implementing process instrumentation systems for U.S. military sponsors at JHU/APL. Using his bachelor of science degree in computer engineering from the University of Maryland and his master of science degree in computer science from Johns Hopkins University, he builds technology to objectively evaluate individual, team, and IT system performance through process instrumentation. In particular, he's exploring ways to improve understanding of processes and their performance using visualization technologies. Mr. Koterba also researches user-interaction techniques for trainers, testers, and assessors to rapidly and intuitively conduct process forensics and reconstruction. As an implementer of instrumentation systems, he's a strong advocate for open, exposed, and accessible data sources. E-mail: nathan.koterba@jhuapl.edu

Greg Williams is a systems engineer at JHU/APL after completing a 20-year Air Force career that he began as an electronic warfare officer in the RF-4C and F-4G. After graduating from the USAF Test Pilot School as a test navigator in 1997, he conducted developmental testing in the F-16CJ, B-1B, F-15E, and RQ-4A Global Hawk. His final two tours were at the Pentagon, HQ Air Force T&E, and Patuxent River, Maryland, with the V-22. He has a bachelor of science degree in physics from the U.S. Naval Academy, 1988, and a master of science degree in physics from the Air Force Institute of Technology, 1993. E-mail: gregory.williams@jhuapl.edu

Glenn Conrad is a lead communications engineer with MITRE supporting U.S. Air Force C2 test activities at the 46 Test Squadron on Eglin AFB and the 505 Command and Control Wing, Hurlburt Field, Florida. Previously, Mr. Conrad worked at the JHU/APL and SAIC Corp on various test and training projects. His background is real time and distributed systems, control systems, and communications. He has master of science and bachelor of science degrees in systems analysis from the University of West Florida. He is a member of the IEEE and International Test and Evaluation Association (ITEA). E-mail: glenn.conrad@mitre.org

Susi McKee has more than 23 years of Air Force test and evaluation experience. Ms. McKee is 505th Command and Control Wing's (CCW) chief of capabilities development, enhancement, and integration at Hurlburt Field. Previously, Ms. McKee was director of test for 605th Test and Evaluation Squadron where she oversaw C4ISR operational testing efforts. At Eglin AFB, she served as chief of analysis for 53d Wing's Air-to Ground Weapon System Evaluation Program and was an analyst for operational flight program upgrades. In addition, while at the 46th Test Squadron, she conducted safe escape and ballistics analyses. E-mail: susana.mckee@hurlburt.af.mil

Oscar A. Garcia is the team lead for Air and Space Operations Center (AOC) training research for the Air Force Research Laboratory's Warfighter Readiness Research Division in Mesa, Arizona. The AOC Training Research Team is focused on improving readiness for AOC personnel. Mr. Garcia received a bachelor of science degree in human factors engineering from the U.S. Air Force Academy in 2001 and an master of business administration from St. Mary's University in 2005. Mr. Garcia is a member of the Simulation Interoperability Standards Organization and an Interservice/Industry Training, Simulation and Education Conference subcommittee member for Human Performance. E-mail: oscar.garcia@mesa.afmc.af.mil

Dr. James S. Welshans is a former active-duty U.S. Air Force fighter pilot, instructor, and war planner. Currently an independent scholar, he is employed as a senior requirements engineer with Teledyne CollaborX, advising the U.S. Air Force Research Laboratory on military C2 projects and technology transition. A founding member of the U.S. Air Force Operational Command Training Program, Dr. Welshans taught strategy and operational assessment to military officers worldwide during major command and control exercises. Dr. Welshans received a bachelor of science degree in international affairs and history from the U.S. Air Force Academy (1977), a master of science in management (1985), and a doctor of education in curriculum and instruction and educational leadership (2008). He serves on the executive board of the Society for Phenomenology in the Human Sciences. E-mail: lectricsix@mediacombb.net

ITEA Journal 2010; 31: 275–279

Augmenting Test and Evaluation Assessments Using Eye-Tracking and Electroencephalography

Anthony Ries, Ph.D., and Jean Vettel, Ph.D.
U.S. Army Research Laboratory,
Human Research and Engineering Directorate,
Translational Neuroscience Branch, Aberdeen Proving Ground, Maryland

Tools that provide continuous, objective measurements of human-system interactions can augment measures obtained through subjective assessments and/or expert observation by providing near–real time performance metrics. Two tools for the Test and Evaluation (T&E) community will be discussed: eye-tracking applications that are viable for use in T&E today and electroencephalography-based metrics that hold promise for the future.

Key words: Continuous measurements; Electroencephalography (EEG); eye-tracking; mental state; subjective assessments.

System evaluation would be much easier if testers could recruit the ideal operator: someone who never gets fatigued, who maintains a consistent level of concentration on the task, and who can accurately recall their moment-by-moment experience of task difficulty during the testing session. This would ensure that any performance decrements during testing were not a result of the operator being tired, not concentrating, or simply not remembering what happened at a particular point in time. Unfortunately, these operators are hard to find! Instead, evaluators must try and measure the operator's mental state in order to assess how that state affects performance during system interaction, or alternatively, how the system interaction influences mental state, which in turn affects performance. Currently, to evaluate operator mental state, evaluators must rely on self-assessment questionnaires, such as the National Aeronautics and Space Administration Task Load Index (NASA-TLX) (Hart and Staveland 1988), that interrupt the operator at discrete times throughout the testing session to provide an introspective assessment. Not only does the interruption break mental concentration on the task, but self-reports are not sensitive to fluctuations of cognitive state within a task; rather they provide an average subjective estimate over a length of time.

Self-assessment measures are not ideal because they lack objective means to measure particular mental state changes, what influenced the state change, and what happened because of the state change. Consider a scenario where the operator must perform several tasks on a new system, and a self-assessment questionnaire is administered at different intervals throughout the study (i.e., discrete measurements). The operator's fatigue level may fluctuate over the course of the study; however, since individuals must rely on memory to recall past events and are not always accurate in their self-assessment reports, discrete measurements can lead to inaccuracies when trying to capture dynamic mental state fluctuations during the test. The lack of a continuous measure can lead to errors in system evaluation, attributing the decrement in user performance to system design rather than attributing it to changes in the operator's mental state (i.e., level of fatigue). One solution to the problem of tracking performance changes over time is to simply take measurements more frequently; however, this comes with the serious disadvantages of breaking continuity of the task, not to mention the introduction of additional task complexity created by the demands of completing multiple surveys as well as performing the task itself. A more efficient solution to this problem would be the use of tools that permit continuous measurement of task performance, eliminating interruptions for self-assessments.

This article discusses how two measurement tools, eye-tracking and Electroencephalography (EEG), provide continuous measurements related to operator performance without creating task disruption. Eye-tracking provides numerous measures of user eye

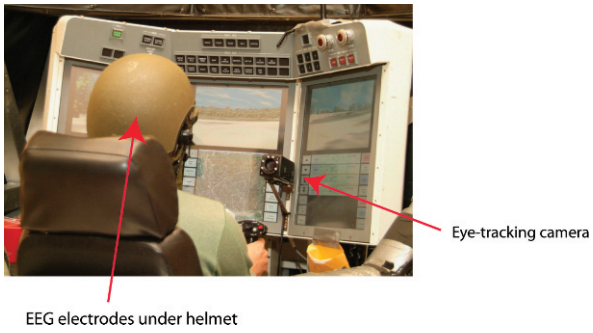


Figure 1. Integration of eye-tracking and Electroencephalography (EEG) tools during operator-system interaction. Eye-tracking and EEG provide continuous, objective measurements of operator performance.

movements and visual scanning patterns, while EEG provides a measure of electrical activity in the brain that can be linked to many complex behaviors and operator mental states. These two tools complement one another and may eventually be used effectively together in T&E environments. For example, *Figure 1* shows an operator in a futuristic crew station, wearing a helmet containing EEG electrodes to record brain activity. In addition, a camera mounted within the driving simulation continuously records operator eye measurements. Both of these tools can capture dynamic changes in the operator's mental state throughout the testing session, providing information regarding how the operator interacts with the system interface, without interrupting operator task performance. In this article, potential applications of these two tools to T&E environments will be discussed, with an emphasis on eye-tracking applications that are ready for use today, and EEG applications that provide promise for the future.

Eye-tracking in T&E

Eye-tracking is the process of measuring either the point of gaze (where the operator is looking), or the motion of the eye relative to the head. There are a number of methods for measuring eye movement, including the use of video images from which the eye position is extracted. Advances in computer and video technology have lead to the development of eye-tracking systems that are portable and simple to use (Babcock, Lipps, and Pelz 2002). Eye-tracking offers the evaluator an objective and unobtrusive means to continuously measure human performance in a diverse set of environments and field settings. After a quick calibration, eye-tracking can provide several continuous user measures that can be linked to operator mental state. These measures include blink rate, frequency of eye movement, pupil dilation, the amount of eyelid closure over time (described as Percentage Eye Closure or PERCLOS), and the length of time spent looking at a particular location. Although all of these continuous measures offer advantages, we will highlight just a few of them here to demonstrate how these measures can augment a variety of T&E scenarios today.

Figure 2 shows an example of how an eye-tracker could be used to plot the gaze path of an operator using a crew station computer interface with multiple display screens. In this example, the operator scanned several displays showing urban environments and system status, searching for images of people who could pose a threat to the security of their vehicle. The operator used the touch-screen interface on the center console to complete a threat report anytime a threatening person was seen. The red circles indicate where the operator fixed their gaze on the screen, with the numbers inside the circles indicating the order in which the operator's eyes traveled across the multiple display screens. The size of the red circles describes the relative amount of the time spent looking at each location, with larger circles representing longer gaze fixation times. As can be seen, the operator searched from left to right, starting first with the exit points on the corner building at the left-most display, continuing to the right across the center, and ending at a point between the right-most display and the status display. We can see that the operator spent time looking at the threat report console (circle 3) before scanning the other two buildings (circles 4 and 8). These continuous measures show operator scanning patterns used to perform the task, as well as which system interface features are used most frequently.

Eye-tracking measures have been used successfully to reveal differences between novice and expert operators. Ottati, Hickox, and Richter (1999) compared eye movement patterns of novice and experienced pilots in a flight simulator that required the use of electronic maps for navigation. Traditionally, pilots use printed maps for this task. However, in Ottati's task, the pilots were required to identify critical terrain navigation landmarks from electronic instrumentation. The eye-tracking data revealed search path differences between novices and experts. Unlike the experienced pilots, novice pilots gazed longer and more frequently outside the cockpit instead of at the cockpit instrumentation. In addition, when the novice pilots did look at the mapping instrumentation, they gazed much longer than the experts, suggesting that they struggled to identify the critical landmarks on the electronic map. These differences, revealed by eye-tracking measurements, can be useful for developing training procedures to teach novice operators to use expert-like strategies in reading electronic maps.

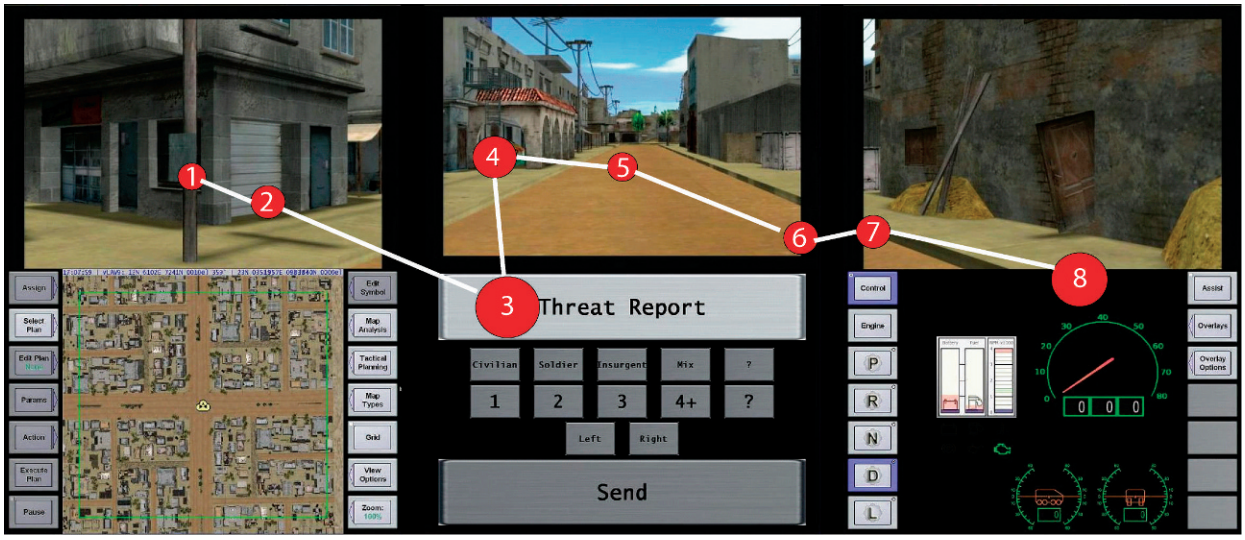


Figure 2. Eye movements superimposed on a crew station computer interface. Numbers indicate the order of eye fixations, and the size of each circle represents the total fixation duration.

Eye-tracking measures such as PERCLOS can help characterize dynamic changes in the operator's mental state owing to factors such as fatigue. In a study by Dinges et al. (1998), operators were intentionally sleep deprived for 42 hours and were then required to monitor a system for target events and to report when the events occurred. The operators completed questionnaires to monitor how sleepy they felt, and an eye-tracking system was used to monitor their visual search. Results of the study indicated that PERCLOS reliably predicted when operators were fatigued more effectively than the questionnaires, and the measure revealed points in time in which operators experienced changes in their level of fatigue. Performance decrements can be linked to the operator's state by capturing these dynamic changes.

This short review only touches the surface of the current-day capabilities of incorporating eye-tracking measures into T&E assessments. The examples presented here highlight potential applications of continuous measures to assess and/or compare how operators interact with system interfaces, as well as monitor dynamic changes in the operator's fatigue level. However, many other applications have been identified in several commercial applications, including the automobile and aviation industries. With their portability and ease-of-use, eye-tracking systems can easily be incorporated into T&E environments to identify additional assessment capabilities.

EEG in T&E

EEG provides a measurement of electrical activity in the brain using recordings from electrodes on the scalp. These measurements have been linked to dynamic changes in behavior and factors related to an operator's mental state. Traditional EEG systems have been very bulky, entailing set-up time to attach electrodes to the scalp with gel, and requiring the operator to minimize any head or body movement while physically tethered to the EEG recording devices. Traditional EEG systems are also highly susceptible to electrical artifacts from nearby equipment and other non-brain sources of electrical activity. System calibration must be performed for each operator before each testing session because the day-to-day variability of operator EEG measurements can be high (East, Bauer, and Lanning 2002). These attributes make traditional EEG systems impractical for current T&E environments; however, in the past few years, EEG systems have been developed that minimize these limitations. These newer systems are light-weight, often incorporating the electrodes into a hat or helmet, and are designed for use in real-world, operational environments. They contain advanced amplification and wireless transmission technology to minimize the impact of electrical interference. One such system is shown in *Figure 3*. As technology develops, these newer EEG systems will likely evolve to be as portable and easy-to-use as the eye-tracking systems in use today.

Results from EEG-based measurements collected in controlled environments show potential for application in operational settings. For example, Pope, Bogart, and Bartolome (1995) utilized a real-time index of operator task engagement, based on the power of EEG spectra, in an adaptive system to mitigate the effects of fatigue. EEG was collected while operators performed several

Figure 3. An example of an electroencephalography system designed to be worn under a helmet. A wireless base station receives and decodes the neural signals from several meters away, allowing the operator to be fully mobile. While this device shows promise in controlled environments, there is still work to be done to increase reliability in operational settings.

tasks including monitoring, resource management, and compensatory tracking. The tracking task was either automated by the system or manually controlled by the operator, and the system dynamically switched between the two modes based on the changes in the EEG-based index in order to minimize the effects of fatigue.

Similarly, Wilson, Lambert, and Russell (2000) also used an EEG-based measure in an automated system that adapted its functioning based upon changes in operator mental state. As in the Pope, Bogart, and Bartolome (1995) study discussed above, operators were required to perform multiple tasks. When the operator could not complete all assigned tasks, the system assisted with the monitoring task, requiring operators to solely focus on the remaining two tasks—resource management and tracking. The system switched into this assist mode only when the EEG-based measure of operator state indicated that they were overloaded and struggling to complete all tasks simultaneously (referred to as "high workload" by the authors). Of importance, by using the EEG measure to dynamically adapt the system, the number of errors decreased significantly in both the resource management task (33 percent) and in the tracking task (44 percent).

Other EEG findings obtained in controlled, laboratory settings show promise for future application in T&E environments. For instance, EEG-based measures have revealed neural signatures that are sensitive to error detection. The error-related negativity (ERN) produces a distinct pattern of brain activity when an individual makes an incorrect response or error. Monitoring operator errors in near–real time would allow testers the ability to identify when certain errors occur during system interaction as well as the ability to adapt system components based on errors committed. The ERN is not only sensitive to when an operator makes a known error, but it is also seen when an operator witnesses someone else making an error (van Schie et al. 2004). Preliminary applications of the ERN are currently being employed during the testing of brain-computer interfaces (Ferrez and Millan 2008). The ability to detect operator errors in real time could substantially augment T&E assessment capabilities.

Another laboratory-based finding demonstrates the utility of EEG for detecting predefined targets or identifying sudden changes in the environment without relying on overt responses from the operator. Using EEG in this way may provide testers with the ability to detect anomalies or unexpected changes that occur during system interaction. One instance of using this approach comes from image classification where an operator must detect a target of interest embedded in a series of rapidly presented images. Systems have shown their ability to accurately identify particular images for further analysis based on EEG measurements alone even though the images were shown for 100 milliseconds (Gerson, Parra, and Sajda 2006). This approach proved to be highly accurate in discriminating between target and nontarget images, and much faster than simply relying on self-reports from the operator performing the target-scanning task. The potential of EEG to rapidly identify what information is critical to the task at hand could greatly enhance how operator performance and system evaluation are analyzed in T&E environments.

These examples highlight the potential of EEG-based measures to index factors related to operator mental state. The examples described here are just a subset of the laboratory-based EEG findings that may have direct relevant applications to T&E. EEG-based measures can greatly augment the types of continuous operator assessments currently available with eye-tracking, and the combined use of both eye-tracking and EEG measurements may improve real-time assessment of operator mental states across many tasks, systems, and environments above and beyond using each method alone.

Conclusions

Eye-tracking and EEG are two tools that provide continuous measurements of operator performance, and both provide powerful analysis tools to the T&E community. Eye-tracking systems are simple to use, portable, and provide measurements related to operator mental state, such as blink rate and PERCLOS, that can be applied in a T&E setting today. Although they are not ready for use in all field settings, EEG systems

have the potential to provide additional measures related to operator mental state that may be of use to testers and evaluators. Studies using EEG have successfully adapted systems and improved operator performance, and other more preliminary studies using laboratory-based EEG measures such as the ERN show how real-time assessment of operator performance can be augmented in applied settings. As additional research is conducted, EEG-based measures obtained in controlled settings can be applied to the complex and dynamic environments of T&E. Future research should explore the combined use of eye-tracking and EEG systems to create a broad-based measure of changes in operator performance, expanding assessment capabilities to a diverse set of tasks and environments. ❑

Dr. Anthony Ries is a research psychologist at Aberdeen Proving Ground in Aberdeen, Maryland. He is currently serving as deputy manager of the "High-Definition Cognition in Operational Environments" U.S. Army Technology Objective (Research), which is focused on enabling assessment of soldier performance in operationally relevant environments using behavioral and physiological measures. Dr. Ries received a bachelor of science degree in psychology from Northwest Missouri State University in 2000 and master of arts and doctor of philosophy degrees in cognitive psychology from the University of North Carolina at Chapel Hill in 2003 and 2007, respectively. E-mail: Anthony.ries@us.army.mil

Dr. Jean Vettel is employed by the U.S. Army Research Laboratory in the Translational Neuroscience Branch at Aberdeen Proving Ground in Aberdeen, Maryland. Her research focuses on improving soldier–system performance by conducting neuroscience research in complex, operational settings. She received a bachelor of science degree from Carnegie Mellon University in 2000 and master of science and doctor of philosophy degrees from Brown University in 2008 and 2009, respectively. E-mail: Jean.vettel@us.army.mil

References

Babcock, J., M. Lipps, and J. B. Pelz. 2002. How people look at pictures before, during, and after image capture: Buswell revisited. In *Proceedings of SPIE, Human Vision and Electronic Imaging*, 4662:34–47. Bellingham, WA: SPIE.

Dinges, D. F., M. Mallis, G. Maislin, and J. W. Powell. 1998. Evaluation of techniques for ocular measurement as an index of fatigue and the basis for alertness management. Department of Transportation Highway Safety, pub. 808 762.

East, J. A., K. W. Bauer, and J. W. Lanning. 2002. Feature selection for predicting pilot mental workload: A feasibility study. *International Journal of Smart Engineering System Design* 4: 183–193.

Ferrez, P. W., and J. del R. Millan. 2008. Error-related EEG potentials generated during simulated brain-computer interaction. IEEE Trans Biomedical Engineering, March, 923–929.

Gerson, A. D., L. C. Parra, and P. Sajda. 2006. Cortically-coupled computer vision for rapid image search. IEEE Trans Neural Systems and Rehabilitation Engineering, June 1–6.

Hart, S. G., and L. E. Staveland. 1988. Development of a multi-dimensional workload rating scale: Results of empirical and theoretical research. In *Human mental workload*, ed. P. A. Hancock and N. Meshkati. Amsterdam, The Netherlands: Elsevier.

Ottati, W. L., J. C. Hickox, and J. Richter. 1999. Eye scan patterns of experienced and novice pilots during Visual Flight Rules (VFR) navigation. In *Proceedings of the Human Factors and Ergonomics Society, 43rd Annual Meeting*, Houston, Texas. Santa Monica, CA: Human Factors and Ergonomic Society.

Pope, A. T., E. H. Bogart, and D. Bartolome. 1995. Biocybernetic system evaluates indices of operator engagement. *Biological Psychology* 40: 187–196.

van Schie, H., R. Mars, M. Coles, and H. Bekkering. 2004. Modulation of activity in medial frontal and motor cortices during error observation. *Nature Neuroscience* 7: 549–554.

ITEA Journal 2010; 31: 280–283

Leader Development by Design

Robert A. Cassella
Booz Allen Hamilton, Leavenworth, Kansas

U.S. Army doctrine has recently embraced a "methodology for applying critical and creative thinking" to develop approaches to solve complex, ill-structured problems prior to the initiation of detailed planning. The just-published Army Capstone Concept advocates "a mindset based on flexibility of thought" by leaders who "have a tolerance for ambiguity, and possess the ability and willingness to make rapid adjustments according to the situation." Who are these people, and how will they be developed? While much has been done to produce the doctrine, as much thought must be given to how the prototypical leader of the twenty-first century will be developed and assessed in some set of "vital few" professional competencies.

Key words: Cognition; flexibility of thought; operational adaptability; visualization.

It seems that everyone agrees—the operational environment of the recent past was undeniably complex, today's operational environment is even more complex, and the forecast for the future is no less complex. In the postulated future, the U.S. military will face a daunting array of state and nonstate adversaries armed with everything from antiquated Kalashnikovs to horrific weapons of mass destruction, and the locus of conflict will range from subterranean lairs to outer space and cyberspace.

Military operations will continue to encompass everything from peacetime engagement to humanitarian assistance to irregular and general war. In today's and tomorrow's missions, the salient difference will be the simultaneous nature of operations and the frequency of transitions from operation to operation. This modern phenomenon is what former Marine Commandant, General Charles Krulak,[1] referred to as "three-block war."

> *"In one moment in time, our Service members will be feeding and clothing displaced refugees, providing humanitarian assistance. In the next moment, they will be holding two warring tribes apart—conducting peacekeeping operations—and, finally, they will be fighting a highly lethal mid-intensity battle—all on the same day…all within three city blocks."*

In recognition of this reality, General Martin Dempsey, the current commander of the U.S. Army Training and Doctrine Command, prefaced the most recent Army Capstone Concept with these words: "…[t]he Army must…achieve…operational adaptability…a mindset based on flexibility of thought calling for leaders at all levels who are comfortable with collaborative planning and decentralized execution, have a tolerance for ambiguity, and possess the ability and willingness to make rapid adjustments according to the situation…" General Dempsey went on to emphasize that "[t]he training and education of our entire force must aim to develop the mindset and requisite knowledge, skills, and abilities required to operate effectively under conditions of uncertainty and complexity."[2]

Correspondingly, the 2010 Department of Defense quadrennial defense review states that "…part of our commitment [is] to ensure that tomorrow's leaders are prepared for the difficult missions they will be asked to execute…"[3] Other sources postulate what capabilities will be needed for twenty-first century leaders: innovativeness, creativity, agility, adaptability, critical thinking, versatility—without advancing a method to produce what former Army Chief of Staff General Peter Schoomaker repeatedly described as "soldier-pentathletes."[4]

Recently, Army doctrine has embraced an approach to solve complex, ill-structured problems prior to the initiation of detailed planning. This approach requires practitioners who will endeavor to understand environments populated by diverse actors to define the problem and, ultimately, design an approach to resolve it.[5] Other doctrinal literature prescribes the preferred method of battle command—through decentralized execution—based on orders that emphasize the results to be attained, not how they are to be accomplished.[6]

Table 1. The 2008 Battle Command Battle Lab study of cognition and visualization identified 19 KSAEs grouped into four competency areas.

Domain knowledge	Decision making	Communications	Adaptability
Doctrine	Red teaming	Oral expression	Metacognition
Tactics/operational art	Critical thinking	Written expression	Tolerance for ambiguity/uncertainty
Systems and processes	Fusion	Instructor	Bias recognition
System capabilities/limitations	Assessment/analysis	Interpersonal communication	Attention management
Observer/controller	Second-in-command		Operations and training

This doctrine of mission command demands operators whose appreciation of the situation guides the adaptive employment of forces. Decentralized execution also requires leaders who are willing to act in the absence of orders when existing orders no longer fit the situation or when unforeseen opportunities or threats arise. Although the discussion of decentralized execution usually implies a cogent choice, in the reality of a degraded command and control environment, adaptive leaders will be forced to self-synchronize their actions within the commander's intent and the mission variables to achieve their objective.

All of these requirements for the development of future leaders will certainly necessitate changes to how they are accessed and matured from precommissioning through the gamut of professional military education. Anyone who has been involved in managing the curricula of Service schools recognizes that there is no lack of good ideas, but there is a paucity of available time—first, the amount of time parceled out for leaders to spend away from the operating force and, second, the "zero-sum-game" nature of scheduling leader development opportunities. In this environment, how can leaders receive the education they need to refine their strengths and overcome shortcomings in some number of crucial knowledge, skill, and ability sets?

And it's not just about military professionals—during UNIFIED QUEST 2009, participants generally agreed that, rather than emphasize processes, "…the professional development of civil-military designers and planners should be the critical focus." The UNIFIED QUEST 2009 final analysis report goes on to say that because the development of a single multiagency methodology or process is unlikely and the common ground between government agencies is people, a common education for civilian and military designers and planners would provide a unifying factor between government agencies and the Department of Defense.[7]

One effort to identify requisite knowledge, skills, abilities, and experiences for future full spectrum operators was the 2008 study commissioned by the U.S. Army Battle Command Battle Lab at Fort Leavenworth. The purpose of this study was to identify the cognitive skills required to visualize full spectrum operations; it was conducted in conjunction with the annual OMNI FUSION experiment to inform the evolution of battle command doctrine and the development of future leaders.

The exercise of battle command comprises three interdependent components: (a) understanding, visualizing, and describing a situation; (b) directing and leading forces; and (c) assessing operations—all so as to impose one's will upon an adversary. These component parts of battle command are fundamental for all leaders at all levels in all operations and are generalizable across the depth and breadth of military and civilian leadership.

The results of this research into cognition and visualization coalesced 19 knowledge, skills, abilities, and experiences (KSAE) into four interconnected competencies of domain knowledge, decision making, communications, and adaptability *(Table 1)*.

The study projected that some level of expertise across the identified competencies could produce results that contribute to visualization, shared understanding, collaborative teaming, self-synchronization, self-awareness, sense making, creativity, insight, and agility.

To complete the articulation of these core competencies, the study proposed novice, journeyman, and master level behaviors associated with the competencies that can be used for assessment *(Table 2)*. Because the output, or results, of a competency (adaptability, for example) may be observed and judged, a person could be assessed to be performing well or not so well with respect to a given competency. It is not reasonable, however, to generalize that the person has mastered the depth and breadth of any given competency. Thus, evidence of expertise in a competency is demonstrated by actions (behaviors) that can be observed and assessed to distinguish the outstanding performers from the average.

The study concluded that mastery of these competencies would allow leaders to artfully exploit the capabilities of the human and technological subsystems

Table 2. The behaviors for adaptability illustrate the difference between novices, journeymen, and masters with regard to their tolerance for ambiguity or uncertainty.

Competency = adaptability	Novice behaviors	Journeyman behaviors	Master behaviors
Tolerance for ambiguity/uncertainty	Tolerates risk and uncertainty.	Handles risk and uncertainty comfortably.	Rises to the challenge, thrives on situations involving risk and uncertainty.
	Decides and acts without having a sense of the total picture.	Uses ingenuity to compensate without having a sense of the total picture.	Uses ingenuity in dealing with ambiguous situations, and guides others to cope effectively.
	Copes with change and transitions when necessary.	Effectively copes with change and transitions comfortably.	Anticipates impact of change; plans how to transition and guides others in transitions.

of the command and control system to achieve operational results.[8]

There are obstacles to producing the prototypical "pentathlete"—most notably, the cultural resistance to developing innovative, critically thinking devil's advocates who challenge conventional thinking. Like any significant cultural change in any military service, acceptance of an innovative approach to leader development will require doctrine—the engine of change—and a four-star champion for the leader development programmatics that are necessary to implement the idea of operational adaptation.

The development of tactical, operational, and strategic leaders who can realize and exploit the idea of operational adaptability cannot be deferred until—or contained within—command and staff college level education. The process of growing the generation of leaders who embody the "right stuff" must begin at entry level and continue in and out of professional military education opportunities.

Another challenge would be to overcome the cultural artifact that abhors the idea that producing better warfighters means longer incubation in an educational venue. If future uses of military force will demand more than just warfighting skills, there should be a means to classify requirements for an individual's development within the institutional environment and through their self-development.

There are a number of assessment instruments currently in use throughout the military—primarily used for specialized applications like special operations selection—like the 16 Personality Factor, Wonderlic Personnel Intelligence Test, the Minnesota Multifacet Personality Inventory, and the Individual Adaptability Measure. These tools can be exploited for *classifying*—not *selecting*—entry-level leaders in areas that contribute to the development of the cognitive capacities that enable acquisition of the knowledge and skills, as well as the refinement of innate abilities, required of the objective twenty-first century leader. The results of this classification effort, which would be akin to an Armed Services Vocational Aptitude Battery for leaders, could facilitate personalized development. Emphasis on the individual avoids the tendency to pursue a "sheep dipping" approach to professional growth in which every person receives the same treatment regardless of their strengths and weaknesses.

By systemically determining desired results (such as visualization, collaborative teaming, self-synchronization, self-awareness, sense making, creativity, and agility), we could derive sets of requisite knowledge, skills, abilities, and experiences that contribute to their attainment. Once the required knowledge, skills, and abilities of the objective twenty-first century leader are determined, based on future operational demands, behaviorally based performance assessment tools can be developed and exploited that indicate some level of proficiency (novice, journeyman, master).

The cost of producing a generation of full spectrum warriors will be resource-intensive—the least of which is not time—but these costs can be mitigated by personnel policy changes that require altering the current concept of a "normal" career in content and duration. These changes might include extending time-in-grade gates and redefining the minimum retirement threshold to 25 or 30 years.

Other necessary policy revisions would have to be made to separate developmental assessment and feedback and to the winnowing process that determines which officers will be promoted and selected for positions of greater responsibility. The current Army evaluation reporting system, contained in Army Regulation 623-3, is designed to "select and develop tomorrow's leaders." This bifurcated purpose is at odds with itself—the evaluation narrative that supports selection for schooling, promotion, and command will probably not be developmental in nature and any suggestions for an officers' personal or professional development probably won't get them selected for career advancement.

This conflict of purpose might be mitigated by a solution similar to the Multi-Source Assessment and

Feedback program that provides "…a 360-degree approach…*without ties to Army personnel management processes or systems* [emphasis added]." This leadership assessment tool specifically addresses nine leader competencies such as leads by example, develops leaders, and creates a positive environment. As designed and executed, the Multi-Source Assessment and Feedback program provides feedback to the assessed individual and the chain of command and to the institutional education system, albeit anonymously and in aggregate as formative data to provide command climate insights in units and for the refinement of curricula in service schools.[9]

Another area where improvements must be made is in the delivery of professional military education; leaders must be educated to cope with unstructured problems rather than being given both the problem and all of the attendant information. It is not practical to teach the rudiments of critical thinking in one classroom, for example, and then eschew critical examination of assumptions and the thought underlying these assumptions in an adjacent classroom—and expect students to understand how to solve complex problems or to thrive in an environment that demands decentralized execution.

Depending upon an individual learners' developmental needs, some will need abstract, theoretical explanations while others need concrete, procedural explanations; most will probably need a combination of both approaches. Without a definitive assessment of these needs, we will be forced into applying the same old solution to novel situations and thus defer to the default position rather than leader development by design. ❑

ROBERT A. CASSELLA is a consultant with Booz Allen Hamilton in Leavenworth, Kansas, currently supporting the U.S. Army Training and Doctrine Command Analysis Center at Fort Leavenworth. He is the author of Project Management Skills for Kids *(Trafford, 2005). E-mail: cassella_robert@bah.com*

Endnotes

[1]Charles Krulak, 1997. *The Three Block War: Fighting in Urban Areas, an Address to the National Press Club*, October 10, 1997.

[2]U.S. Army Training and Doctrine Command, "The Army Capstone Concept: Operational Adaptability—Operating under Conditions of Uncertainty and Complexity in an Era of Persistent Conflict," December 21, 2009, i.

[3]Department of Defense, "Quadrennial Defense Review Report," February 2010, 54.

[4]Peter Schoomaker, "Farewell Message," April 6, 2007.

[5]Department of the Army, "The Operations Process," February 2010, Chapter 3.

[6]Department of the Army, "Operations," February 2008, 3–6.

[7]U.S. Army Training and Doctrine Command Analysis Center at Fort Leavenworth, "UNIFIED QUEST 2009 Final Analysis Report," July 31, 2009.

[8]U.S. Army Battle Command Battle Lab at Fort Leavenworth, "Cognition and Visualization Study Report," September 26, 2008.

[9]Multi-Source Assessment and Feedback Web site, downloaded from https://msaf.army.mil/Default.aspx (accessed on February 21, 2010).

ITEA Journal 2010; 31: 284–290

Electromagnetic Spectrum Test and Evaluation Process

Marcus Shellman, Jr.
Defense Information Systems Agency,
Defense Spectrum Organization, Annapolis, Maryland

With today's challenges of evolutionary acquisitions compounded by the dictates of balancing mission needs, Joint interoperability concerns, and decreasing budgets, the combat materiel developer faces increasingly complex acquisition decisions and is besieged with a multitude of requirements when developing and fielding systems. These factors often overshadow the need to address spectrum supportability and electromagnetic environmental effects control when procuring many of our military systems and during test and evaluation. However, by not assessing spectrum supportability and electromagnetic environmental effects during test and evaluation, the probability of systems experiencing electromagnetic interference (EMI), safety hazards, and/or denied operation/deployments increase dramatically. This article outlines the Department of Defense's approach to address and mitigate electromagnetic spectrum concerns throughout the system's acquisition life cycle.

Key words: Electromagnetic Environmental Effects (E3), electromagnetic spectrum, Spectrum-Dependent (S-D), Spectrum Supportability (SS), Spectrum Supportability Risk Assessment (SSRA).

Over the past several decades, the military has documented hundreds of electromagnetic interference (EMI) problems between blue forces that have resulted in diminished mission effectiveness, system failure, and even loss of life. Significant investments have been forfeited or lost due to a failure to address electromagnetic environmental effects (E3) control and spectrum management (SM) during test and evaluation (T&E). In addition, many fielded systems operate with limited capabilities and mission constraints due to vulnerabilities that would have been discovered if spectrum supportability (SS) and E3 controls had been addressed early during acquisition, as recently reported in General Accounting Office Report, GAO-03-617R, "Defense Spectrum Management."

In addition, the demand for electromagnetic (EM) spectrum, both nationally and internationally, coupled with increased worldwide implementation of emerging spectrum technologies, has resulted in new and challenging operational problems not previously encountered by our military. Our military now must compete for the use of the EM spectrum in an environment primarily driven by economic factors of the commercial marketplace. To overcome these challenges and reduce the potential for EMI and other ills associated with noncompliance, SS and E3 control needs to be designated as a mandatory critical operational issue (COI) during developmental and operational test and evaluation (DT&E/OT&E) processes.

Background

The operation of the Defense Acquisition System (DAS) is delineated in Department of Defense (DoD) Instruction 5000.02. The operation of the Joint Capabilities Integration and Development System (JCIDS) process is established by CJCSM 3170.01C. Procedures for certifying JCIDS programs are established in CJCSI 6212.01E. SS and E3 control requirements are required throughout the DAS beginning with the preparation of JCIDS documentation and validated through DT&E and OT&E. The relationship between the JCIDS, DAS, E3, and SS processes is depicted in *Figure 1*.

To ensure that these major concerns are addressed, DoD has issued the following policies:

- *DoD Instruction 4650.01 (Policy for Management and Use of the Electromagnetic Spectrum)*. This instruction establishes policy for management

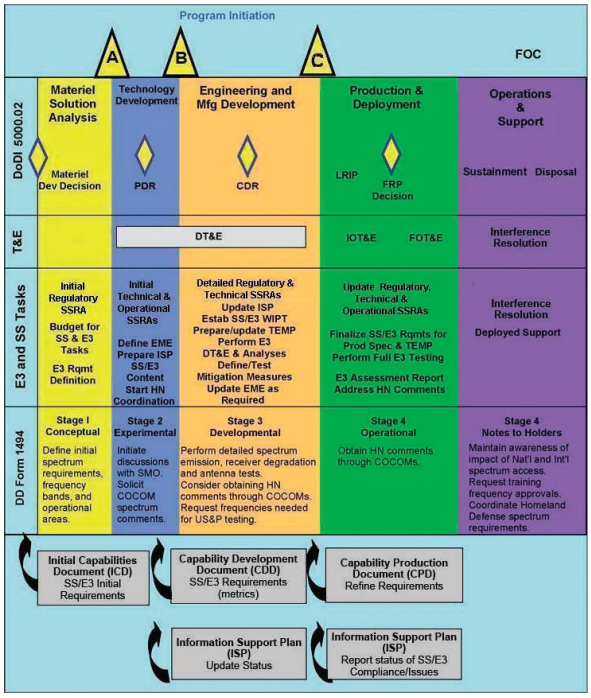


Figure 1. Electromagnetic environmental effects / Spectrum supportability implementation in the DoDI 5000.02 acquisition process.

and use of the EM spectrum within DoD and requires the DoD Components developing or acquiring spectrum-dependent (S-D) equipment or systems to perform a series of spectrum supportability risk assessments (SSRAs). SM is the planning, coordinating, and managing Joint use of the EM spectrum through operational, engineering, and administrative procedures, with the objective of enabling electronic systems to perform their functions in the intended environment without causing or suffering unacceptable interference.

- *DoD Directive 3222.3 (DoD E3 Program)*. This directive requires all electrical and electronic systems, subsystems, and equipment, including ordnance containing electrically initiated devices, to be mutually compatible in their intended electromagnetic environment (EME) without causing or suffering unacceptable mission degradation due to E3. E3 is defined as the impact of the EME on the operational capability of military forces; equipment; systems; and platforms and encompasses the disciplines of electromagnetic compatibility (EMC); EMI; electromagnetic vulnerability (EMV); electromagnetic pulse (EMP); electronic protection; electrostatic discharge (ESD); hazards of electromagnetic radiation to personnel (HERP), ordnance (HERO), and fuels or volatile materials (HERF); lightning; and precipitation static (p-static). E3 also addresses the impact from directed energy weapons and high-powered microwave devices.

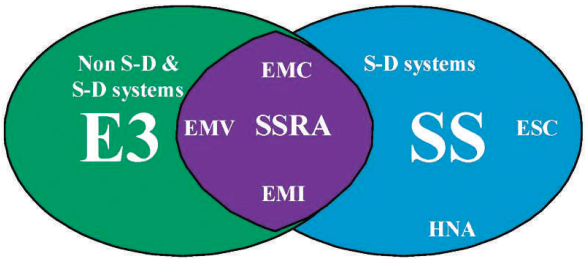


Figure 2. Interrelationship between electromagnetic environmental effects and spectrum supportability.

Together, these policies establish requirements for implementing SS and E3 control throughout the acquisition life cycle including design, development, T&E, and ultimately deployment and sustainment of military platforms, systems, equipment, and forces. These requirements must be addressed early in the program and enforced during milestone reviews by Milestone Decision Authorities. Operational impact assessments of SS and E3 control must be accomplished during both DT&E and OT&E. Doing so has proven to be cost-effective and greatly reduces risks associated with system deployment and supportability. The interrelationship between E3 and SS is depicted in *Figure 2*. The primary overlap occurs during the mutual concern for achieving EMC and preventing EMI for S-D systems and equipment.

Defense Spectrum Organization (DSO)

The DSO is situated strategically in the Defense Information Systems Agency (DISA) to provide leadership in addressing EM spectrum challenges facing the DoD. DSO comprises the Joint Spectrum Center (JSC), the Strategic Planning Office, the Global Electromagnetic Spectrum Information System Program Management Office, and the Business Management Office. Among these divisions, DSO promotes efficient, compatible use of the EM spectrum among our military forces and Allies. DSO's primary missions are to promote effective and efficient use of the EM spectrum to ensure interoperability, reliability, and survivability of military platforms, systems and equipment, and to ensure that system limitations and vulnerabilities are mitigated and documented for the warfighter. The DSO concept of operations also includes provisions to provide support to the Director, Operational Test and Evaluation (DOT&E) and the Services' Operational Test Authorities (OTAs).

Integrated approach for implementing SS and E3 T&E tasks

The following guidance was developed by DSO for program managers (PMs), Materiel Developers (MATDEVs), and OTAs for assessing E3 and SS during the T&E process:

1. Determine the spectrum required to support the mission and define the intended EME in which the system will operate.
2. Ensure E3 control and SS requirements are addressed in acquisition and procurement documentation including JCIDS documents such as the Initial Capabilities Document (ICD), the Capability Development Document (CDD), the Capability Production Document (CPD), Information Support Plan (ISP), and acquisition documents including the Test and Evaluation Master Plan (TEMP), Requests-for-Proposals, Contract Specifications, and other pertinent documents. Additional guidance for implementation of E3/SS during acquisition is provided in MIL-HDBK-237 and the Defense Acquisition Guidebook (DAG).
3. Apply interface standards such as MIL-STD-464 and MIL-STD-461 to ensure that the system and its subsystems and equipment will operate compatibly in the mission EME. The system must meet its performance requirements when exposed to the operational EME.
4. Define E3/SS test objectives in the TEMP and allocate sufficient resources to conduct test objectives.
5. Verify and document SS and E3 control issues during DT&E and OT&E.
6. Conduct early E3 and SS operational assessments that consider the intended mission including single Service, Joint, and international deployments.
7. Provide E3 assessments during operational test readiness reviews. Report the operational impact, system limitations, and vulnerabilities from unresolved E3 and SS problems.

Defining the EME

Fundamental to the process is defining the intended operational EME. MIL-STD-464 establishes maximum external EMEs for shipboard operations, space and launch vehicles, ground systems, fixed wing and rotary wing aircraft, and ordnance. MIL-HDBK-235 provides the assumptions, scenarios, and rationale used to derive the levels in MIL-STD-464. The following steps are provided to further refine and tailor these EMEs based on specific mission scenarios:

- Step 1. Identify the mission scenarios in which the system or equipment is targeted and the associated platforms and systems supporting the missions.
- Step 2. Determine the major geographic regions and countries in which the system or equipment is expected to operate.
- Step 3. Conduct engineering analyses to identify EMI source/victim pairs with the proposed system or equipment during these missions.
- Step 4. Run Joint E3 Evaluation Tool (JEET)[1] analysis based on mission profile to identify all systems contributing to the operational EME.
- Step 5. Use MIL-HDBK-235 to verify spectral characteristics of systems and equipment identified in the above steps.

The EME should include intentional and unintentional electromagnetic radiation (EMR) from DoD systems, as well as from civil and foreign systems. Specific mission-oriented EME profiles are defined in MIL-HDBK-235 and are composed of a combination of measured and calculated data.

Equipment Spectrum Certification (ESC)

ESC is required in accordance with Office of Management & Budget (OMB) Circular A-11 and DoD Instructions 5000.02 and 4650.01 for all S-D systems and equipment. OMB Circular A-11 requires ESC by the National Telecommunications and Information Administration (NTIA) prior to submitting budget estimates for program development. Furthermore, all military S-D systems must conform to the spectrum regulations delineated in the "NTIA Manual of Regulations and Procedures for Federal Radio Frequency Management." ESC requests must be submitted by the PM, MATDEV, or other acquisition authority via the appropriate Service frequency management office (FMO) using procedures in DoDI 4650.01 and the NTIA Manual. As indicated in *Figure 1*, ESC is required at each phase of the acquisition process. Prior to operating S-D systems during DT&E and/or OT&E, the PM/MATDEV must obtain a frequency allocation and, in most cases, a frequency assignment to radiate.

Spectrum Supportability Risk Assessment (SSRA)

The SSRA is used to identify and assess regulatory, technical, and operational EM spectrum and E3 issues with the potential to affect the required operational performance of the overall system. As shown in *Figure 1*, SSRAs are required throughout the acquisition process with the level of detail in the SSRAs increasing as the item's design matures. Specifically

- Initial SSRAs evaluate the system's spectrum needs versus national and international spectrum regulatory requirements, availability of spectrum, and the potential for E3 problems:
    - The Initial Regulatory SSRA addresses the relative regulatory status of the candidate system with respect to host nation spectrum policy governing projected deployments and operational frequencies.
    - The Initial Technical SSRA focuses on candidate technologies and required technical parameters, such as system type, platform type, bandwidth requirements, etc. Preliminary EMC analyses are appropriate at this point to identify potential interactions that will require further study.
    - The Initial Operational SSRA takes into account the full complement of S-D systems anticipated to be in the operational environment and requires a more extensive EMC analysis to identify in operational terms (e.g., frequency-distance separations, steps that may be needed to preclude interference).
- Detailed Regulatory and Technical SSRAs, performed prior to Milestone C, provide increased specifics based on the findings of the initial assessments as the program matures. Developmental data are reviewed for impact to systems operation, and potential risks and mitigation measures are discussed.
- Updated SSRAs in each area are required prior to Production and Deployment, with mature Spectrum and EMC sections. Operational environments should be refined and spectrum and EMC risks reduced to acceptable levels through mitigation measures and/or tactical procedures. At this point the system is ready for deployment.

When evaluating SS, operational restrictions, availability of frequencies, host nation approval (HNA), and known incidents of EMI need to be considered. S-D systems and equipment cannot be operated legally until they have been granted ESC by National and DoD authorities; in addition, a frequency assignment must be obtained from the appropriate area frequency manager. For systems that will operate outside the United States and Possessions, an HNA also is required prior to operation in each foreign country designated for use.

Developers of S-D systems and equipment shall identify and mitigate regulatory, technical, and operational SS risks using the suggested tasks in DoDI 4650.01. System developers shall increase the detail of these risk assessments as the item's design matures. Developers shall assess the risk for harmful EMI with other S-D systems and manage it with other developmental risks. SSRAs should be initiated concurrently with the appropriate stage of certification of spectrum support. Complex "family of systems" (FoS) or "system-of-systems" (SoS) acquisition programs may require more than one SSRA.

DT&E E3 considerations

DT&E will demonstrate that the system design sufficiently mitigates E3 risks and that the system is in compliance with its contractual E3 specifications, based on tailored military standards or commercial standards. Developmental testing (DT) usually is conducted in a test laboratory or open area test site. These tests include production acceptance tests and evaluation and first article E3 testing after an item has been approved for full-rate production. Compliance with E3 control requirements provides a high degree of confidence in achieving platform/system compatibility upon integration but does not guarantee it. However, it is known that noncompliance often leads to operational EMI problems; the greater the noncompliance, the higher the probability that an operational EMI problem will occur.

Equipment and subsystem E3 design requirements must be specified early in the program to avoid costly fixes and ensure mission effectiveness. MIL-STD-461 provides detailed performance and verification requirements for emissions and susceptibility characteristics of equipment and subsystems. MIL-STD-464 provides system-level E3 requirements for airborne, sea, space, and ground platforms and systems, including associated ordnance. The design characteristics, as well as the intended mission, installation, shielding integrity, choice of components, and use of filtering should be considered when performing developmental tests.

OT&E E3 considerations

OT&E assessments are required to validate unresolved E3 problems and to document mission limitations and/or vulnerabilities. During OT&E, E3 testing should be structured to identify and resolve issues that impact mission effectiveness. The assessment should evaluate the impact to other key performance parameters described in the TEMP. These evaluations, which may include both tests and analyses, also may be used to formulate operational procedures and tactics for the item. OT&E assessments should be accomplished in as realistic an operational EME as possible. It is important that resources and assets required for

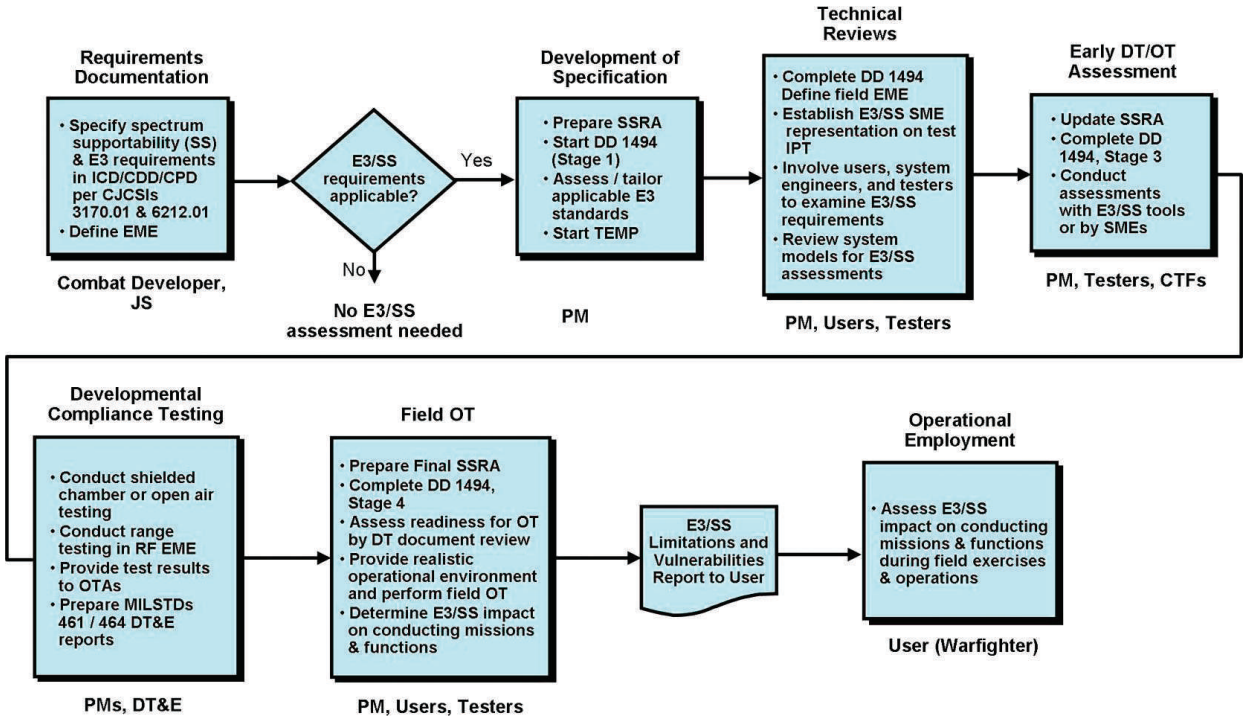


Figure 3. Spectrum supportability and electromagnetic environmental effects assessment process.

verification of E3 requirements be identified early in the program to ensure their availability when needed.

During operational testing (OT), potential EMI source versus victim pairs corresponding to the susceptibilities observed during DT should be identified and systematically evaluated by exercising the subsystem and equipment onboard the platform or system through the various modes and functions while monitoring the remaining items on the platform or system for degradation. Both "one source versus one victim" and "multiple sources versus one victim" conditions should be evaluated. The most common approach is to monitor performance through visual and aural displays and outputs. The need to evaluate antenna-connected receivers across their operating frequency ranges is important for proper assessment. In addition, detection of undesired responses during EMI testing may necessitate an EMV analysis during OT&E to determine the impact on operational performance. EMV analyses require identification of both friendly and hostile emitters that the item may encounter during its life cycle and a determination of the likelihood that the source system will be encountered during operation.

Assessment process for SS and E3 control

Figure 3 depicts the assessment process for SS and E3 control during acquisition. It highlights key objectives from the initial development of requirements to operational fielding. After requirements have been validated by the Joint Requirements Oversight Council, a decision must be made by the PM/MATDEV to determine whether the materiel solution will require SS and E3 control (i.e., is it electrical/electronic and/or S-D). If either SS or E3 control is required, then the subsequent technical reviews, assessments, and testing are mandatory throughout the remainder of the acquisition process. Early involvement from testers and the user community is recommended. Test events should be planned and resourced appropriately to achieve test objectives. SS and E3 control tests should be incorporated into the TEMP. Once DT and OT are completed, a system limitations and vulnerabilities report should be produced and updated periodically if SS and E3 control issues are discovered during operations.

The SSRA is discussed in detail above. The E3 assessment should document and examine compliance with tailored E3 requirements based on the mission needs defined by the combat developer and/or Joint Staff and required by the ICD, CDD, and CPD. In addition, the PM/MATDEV should ensure that the TEMP outlines the specific COIs aimed at verifying EMC. Any additional problems uncovered by verification testing need to be documented and the mission and design of the system may need to be reevaluated. Once all E3 concerns have been identified, an E3

Table 1. Data requirements for spectrum supportability and electromagnetic environmental effects assessments.

Objective: To identify, to the best extent possible, the E3 and SS limitations and vulnerabilities of the subject system.	
Information as appropriate to program development	**Responsibility**
1. DD Form 1494 submitted to the Service Frequency Management Office (FPO)	PM
2. Spectrum Supportability Risk Assessment (SSRA)	PM
a. Regulatory SSRA b. Technical SSRA c. Operational SSRA	
3. Description of operational electromagnetic environment (EME) (e.g., operational environment, theater, mission in the OPLAN)	PM
4. Latest program documentation (e.g., ICD, CDD, CPD, ISP, TISP, Specification)	PM
5. TEMP, which contains	PM
a. E3 within the scope of a critical operational issue (COI) b. list of tests and analyses used to determine the equipment effectiveness/suitability/survivability performance in the operational EME	
6. Copy of the following analyses and/or test and evaluation data	PM
a. intra-platform/system analyses (1) antenna coupling and blockage analyses and/or test data (2) subsystem/equipment EMC analyses and/or test data (3) CI/NDI/GFE EMC analyses and/or test data b. inter-platform/systems EMC analyses and/or test date for spectrum-dependent and non-spectrum-dependent equipment c. special E3 analyses and/or test data (i.e., HERO, HERP, HERF, EMP, Lightning, and P-Static), if required by the CDD, CPD, or TEMP	
7. E3 and SS impact assessments that identify and define operational limitations and vulnerabilities (i.e., lessons learned)	PM
8. DT&E Test Plans and results/reports	PM
9. OT&E Test Plan and results	OTA
10. User-initiated test results	OTA

E3, electromagnetic environmental effects; SS, spectrum supportability; PM, program manager; OPLAN, operation plan; ICD, initial capabilities document; CDD, capability development document; CPD, capability production document; ISP. information support plan; TISP, tailored information support plan; TEMP, test and evaluation master plan; EMC, electromagnetic compatibility; CI, commercial item; NDI, non-developmental item; GFE, government furnished equipment; HERO, hazards of electromagnetic radiation to ordnance; HERP, hazards of electromagnetic radiation to personnel; HERF, hazards of electromagnetic radiation to fuel; EMP, electromagnetic pulse; DT&E, developmental test and evaluation; OT&E, operational test and evaluation; OTA, operational test authority.

Assessment Report stating any anticipated operational issues can be prepared and incorporated into the SSRA, where applicable. In cases where these concerns were not identified during DT, it will be necessary to conduct further assessments as part of field OT prior to preparation of the final E3 Assessment Report and final regulatory, technical, and operational SSRAs.

Supporting documentation

Documentation including DD Form 1494, HN agreements, EMC control plans, EMI test plans and reports, etc., is the foundation for developing E3 and SS test events during OT&E. DT&E test data must be captured and documented. The core elements of the T&E effort are the test procedures and data collection. Faithful execution of the test procedures and explicit data collection will lead to meaningful evaluations during the assessment process. Test reports should summarize the results into viable conclusions and recommendations, thus finalizing the process. To aid in this process, the data item descriptions associated with MIL-STD-461 and MIL-STD-464 should be invoked through the contract specification by the PM/MATDEV.

SS/E3 assessment data requirements checklist

Table 1 presents the data requirements checklist to be used as a guide for the information needed by an SS/E3 assessor. All items except items 9 and 10 should be provided by the PM or MATDEV; items 9 and 10 should be provided by the OTA.

Summary

To overcome the difficult challenges discussed in this article, verification of SS and E3 control during T&E should be mandatory for DoD procurements. System limitations and vulnerabilities must be identified, documented, and provided to the warfighter.

Compliance must be enforced by the DOT&E, milestone decision authority (MDA), PM/MATDEV, and the various Service OTAs. Experience has shown that addressing and mitigating SS and E3 issues early during the acquisition process and verifying that these critical issues are achieved through the T&E process increases both cost and mission effectiveness. In support, DSO can provide the necessary T&E tools that allow for an acquisition to have a successful life cycle from cradle to grave. ❑

MARCUS SHELLMAN, JR., is a senior engineer for the DSO Research, Development, and Acquisition Division of the Defense Information Systems Agency. He received his bachelor of science degree in electrical engineering from the University of Maryland, College Park, Maryland, in 1984. He is the chairman of the American National Standards Institute Accredited Standards Committee on EMC, C63,® Subcommittee 2, EMC Terminology and Definitions and a National Association of Radio and Telecommunications Engineers certified engineer. He has held several engineering positions in DoD including positions at the Space and Naval Warfare Systems Command and the Naval Ordnance Station/Indian Head. He also holds Defense Acquisition Workforce Level III certifications in both program management and systems planning, research, development, and engineering. Mr. Shellman manages the Under Secretary of Defense for Acquisition, Technology, and Logistics Defense Standards Program Electromagnetic Compatibility Standardization area, overseeing the development and maintenance of approximating 40 military standards, specifications, and handbooks. He also has participated in preparing several DoD and Joint Staff directives, instructions, and manuals. E-mail: marcus.shellman@jsc.mil

Endnote

[1]Joint E3 Evaluation Tool (JEET). Request a copy from j5@jsc.mil.

References

Defense Standards Program. MIL-HDBK-235, "Operational Radiated Electromagnetic Environment Guidance for Specification and Evaluation of Military Platforms, Systems, and Equipment." http://assist.daps.dla.mil/quicksearch/ (accessed December 15, 2009).

Defense Standards Program. MIL-HDBK-237, "Electromagnetic Environmental Effects and Spectrum Supportability Guidance for the Acquisition Process." http://assist.daps.dla.mil/quicksearch/ (accessed December 15, 2009).

Defense Standards Program. MIL-STD-461, "Requirements for the Control of Electromagnetic Interference Characteristics of Subsystems and Equipment." http://assist.daps.dla.mil/quicksearch/ (accessed December 15, 2009).

Defense Standards Program. MIL-STD-464, "Electromagnetic Environmental Effects Requirements for Systems." http://assist.daps.dla.mil/quicksearch/ (accessed December 15, 2009).

Department of Defense (DoD). 2004. DoD Directive 3222.3, "DoD Electromagnetic Environmental Effects Program." http://www.dtic.mil/whs/directives/ (accessed December 15, 2009).

Department of Defense (DoD). 2009. DoD Instruction 4650.01, "Policy and Procedures for Management and Use of the Electromagnetic Spectrum." http://www.dtic.mil/whs/directives/ (accessed December 15, 2009).

Department of Defense (DoD). 2009. Interim Defense Acquisition Guidebook, Chapter 7.6 (Electromagnetic Spectrum). https://acc.dau.mil/dag (accessed December 15, 2009).

General Accounting Office. 2003. GAO-03-617R, "Spectrum Management in Defense Acquisitions." http://www.gao.gov/new.items/d03617r.pdf (accessed December 15, 2009).

Joint Staff. 2008. Chairman of the Joint Chiefs of Staff Instruction, CJCSI 6212.01E, "Interoperability and Supportability of Information Technology and National Security Systems." http://www.dtic.mil/cjcs_directives/cdata/unlimit/6212_01.pdf (accessed December 15, 2009).

Joint Staff. 2008. Chairman of the Joint Chiefs of Staff Manual, CJCSM 3170.01C, "Operation Joint Capabilities Integration and Development System." http://www.dtic.mil/cjcs_directives/cdata/unlimit/m317001.pdf (accessed December 15, 2009).

Joint Staff. 2009. Chairman of the Joint Chiefs of Staff Instruction, CJCSI 3170.01G, "Joint Capabilities Integration and Development System." http://www.dtic.mil/futurejointwarfare/strategic/cjcsi3170_01g.pdf (accessed December 15, 2009).

Acknowledgments

The author extends his appreciation to Mr. Jose Blanco (DISA/DSO) and to Mr. Stephen Caine, Mr. Michael Duncanson, and Mr. Joseph Snyder from the staff of EG&G Technical Services, a division of United Research Services (URS), for their support during the preparation of this article.

2010 ITEA Journal *Themes*

The ITEA Publications Committee has established themes for the 2010 issues of *The ITEA Journal* and invites articles in the following areas:

The Role of T&E in Systems Engineering (March issue). Systems engineering is the engineering of complex systems and is intrinsically multi-disciplinary just as test and evaluation (T&E) are. Systems engineers, with their broad view of a program, are in a unique position to diagnose problems in the event of system failure. In T&E the operational requirements of a system must be decomposed to technical requirements and a strategy developed for measuring parameters that can lead back to assessment of mission performance. Systems engineering supplies the process and tools and along with integrated testing is the foundation for future T&E. Integrated testing is the collaborative planning and collaborative execution of test phases and events to provide shared data in support of independent analysis, evaluation, and reporting by all stakeholders. The purpose of integrated testing is to identify system deficiencies early, comply with accelerated schedules, and reduce cost. Design of experiments enables an efficient test design considering all key factors and conditions affecting performance. This issue examines all aspects of systems engineering as well as design of experiments; T&E workforce and training; verification, validation, and accreditation; standards, metrics, data, analysis, and more. ***(Manuscript deadline: December 1, 2009)***

User-Centric Systems (June issue). Systems of systems and network-centric systems are viewed as force multipliers deriving from the mutual connectedness of the elements and the perceived value of timely, critical information. Yet the right information provided to the right person at the right time does not guarantee success. The human is the key and in emerging complex systems of systems the human is more than an operator, and instead is part of the system, a node in the network. Testing a system includes objectively testing the user and requires characterizing human performance. Systems can become so complex that training the user to operate them is no longer possible; rather the systems must be designed with human limitations in mind. This issue examines cognitive performance and measures in addition to traditional form, fit, and function; instrumentation; personal protective equipment; human-machine integration; and situation awareness. ***(Manuscript deadline: March 1, 2010)***

Simulation – Where is T&E Today? (September issue). Simulation is not new and is known by terms such as modeling & simulation and live-virtual-constructive simulation. In one form or another it has been around T&E for more than 20 years and spawned simulation-based acquisition in the Department of Defense, simulation-based design in industry, and a host of other initiatives and hopes. Yet the predictions and expectations have not been realized and the capabilities are often oversold. Where is simulation today in the business of T&E? What has prevented realizing the full power of simulation, what needs to change, or have we arrived already? This issue looks at simulation past, present, and future in T&E and addresses technology, policy, history, success stories, and lessons learned; as well as simulation experience in operational testing, training, design, and other applications. ***(Manuscript deadline: June 1, 2010)***

Cyberspace Test and Evaluation (December issue). Cyberspace is the fifth combat domain, beyond air, land, sea, and space and is the realm of computers, networks, and software. The terrain of cyberspace is not physical but is virtual and ever in flux as network topology and system connectivity dynamically change. In the Department of Defense (DoD) the importance was recognized by creation of cyber commands. Beyond the DoD, cyberspace encompasses commercial networks, the communications industry, power distribution, commerce, transportation, and nearly everything that touches our lives and business today. Systems and networks are subject to continual attacks including spam, phishing, viruses, Trojan horses, worms, root kits, spyware and other malware, and distributed denial of service. Cyber-crime is multi-jurisdictional and spam is being replaced by scam. This issue looks at cyber-infrastructure, data-driven security, information assurance, information operations, electronic warfare, network electronic attack, and other cyber-threats and defenses. ***(Manuscript deadline: September 1, 2010)***

In addition: 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.; **"TechNotes"** 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**: should submit contributions to the **ITEA Publications Committee Chairman (itea@itea.org, attn.: Dr. J. Michael Barton)**. Detailed Manuscript Guidelines can be found at **www.itea.org** under the ITEA Publications tab.