INSTITUTE FOR DEFENSE ANALYSES

"How Much Testing is Enough?" 25 Years Later

Heather Wojton, Project Leader

Matthew R. Avery
James R. Simpson

OED Draft

December 2019

Approved for Public Release.

Distribution Unlimited.

IDA Document P-10994

Log: H 2019-000627

INSTITUTE FOR DEFENSE ANALYSES 4850
Mark Center Drive
Alexandria, Virginia 22311-1882

About This Publication
This work was conducted by the Institute for Defense Analyses (IDA) under contract HQ0034-19-D-0001, Task BD-09-2299, "Test Science Applications," for the Office of the Director, Operational Test and Evaluation.

The views, opinions, and findings should not be construed as representing the official position of either the Department of Defense or the sponsoring organization.

Acknowledgments
The IDA Technical Review Committee was chaired by Mr. Robert R. Soule and consisted of Dr. Arthur Fries, Dr. Daniel J. Porter, and Dr. Dean Thomas from the Operational Evaluation Division, and Dr. James M. Gilmore from the System Evaluation Division.

For more information:
Heather Wojton, Project Leader
hwojton@ida.org • (703) 845-6811

Robert R. Soule, Director, Operational Evaluation Division
rsoule@ida.org • (703) 845-2482

INSTITUTE FOR DEFENSE ANALYSES

IDA Document P-10994

"How Much Testing is Enough?" 25 Years Later

Heather Wojton, Project Leader

Matthew R. Avery
James R. Simpson

Executive Summary

The question of “How much testing is enough?” is persistent across Department of Defense (DoD) test and evaluation (T&E) endeavors. In 1994, the Military Operations Research Society (MORS) and the International Test and Evaluation Association (ITEA) attempted to answer this question, or at least focus future inquiry, with a three-day mini-symposium. Three of the organizers of the 1994 symposium edited the content from the sessions and keynotes to create a report summarizing the major points of discussion and providing some general recommendations. They organized this information into nine Discussion Areas (DAs). It has been 25 years since that symposium and, in some respects, the T&E community has experienced great progress in answering the question that inspired the symposium. This report summarizes progress since 1994 in answering the question, “How much testing is enough?”

Since 1994, T&E has undergone substantial change, some of which has been technologically driven, as new capabilities and new threats have required testers to adapt their techniques to the modern warfighting environment. Other changes have been driven by policy, as leaders within the T&E community have pushed for testing to become more efficient and provide more accurate information to warfighters and decision makers. Others have pushed for T&E to become more responsive to changing program and warfighter needs in order to field new capabilities more quickly.

These changes and improvements have affected some of the challenge areas identified at the 1994 symposium more than others. Table E.1 lists the nine DAs and provides a qualitative assessment of the progress that has been made in each area. Note that “minimal” does not mean “none,” and “substantial” does not indicate that all issues have been completely resolved. Also, these assessments refer specifically to progress in answering the question, “How much testing is enough?” as it relates to each DA. System models are substantially more sophisticated today than they were in 1994, but this doesn’t necessarily mean we’ve answered all the questions about how the use of modeling and simulation (M&S) affects the amount of live testing required.

The areas in which progress has been most substantial include the use of M&S, the use of statistical methods in T&E, and the pooling of data across developmental test (DT), operational test (OT), and M&S. One key similarity among these areas is their use of statistical techniques. The T&E workforce has more training options in these areas than ever before, and policy across the T&E community encourages the use of rigorous statistical methods. Case studies have shown substantial gains from data pooling (especially for reliability analyses), but not all programs take advantage of pooled data to the greatest extent possible. Although the use of M&S data for T&E is more widely accepted now than it was in 1994, quantitative verification, validation, and accreditation methods are not the universal standard.

Table E.1. Progress since 1994 in Each Discussion Area

Discussion Areas	Progress
Impact of ATD/ACTDs (now JCTD) on T&E	Minimal
Early coordination between DT and OT communities	Moderate
Using early T&E data to influence budgeting/acquisition strategy	Moderate
Integration of modeling and simulation with testing	Substantial
Lack of visibility of testing costs	Minimal
Statistical application in the T&E discipline	Substantial
Pooling/sharing data across DT, OT, and M&S	Substantial
Characterizing different types of risk	Moderate
Greater statutory/regulatory flexibility to address “the unexpected”	Moderate

Other areas have seen less progress. Most notable here is the visibility of test costs. Inconsistencies between Services, between programs within Services, and within single programs have left us without a standard accounting method for T&E costs. Although numerous case studies have documented benefits from T&E, these benefits are frequently qualitative, meaning that analysts can quantify neither the costs nor benefits to perform a cost/benefit analysis. Similarly, early technology demonstrations remain unincorporated with the rest of the test and evaluation process. The Middle Tier Acquisition pathways defined in the FY16 National Defense Authorization Act may provide a better mechanism for incorporating early technology demonstrations into T&E, as they were designed to provide program managers with greater flexibility. However, more flexible pathways do not ensure that the systems tested at early events closely resemble the versions that are eventually fielded, or that early technology demonstrations are planned in such a way as to provide value to the T&E process. It remains to be seen whether this new approach will bear fruit where previous ones did not.

In the remaining areas, the results are more mixed. This often indicates a lack of codified and enforced best practices. This report identifies not only multiple examples of programs using early DT data to change their acquisition strategy but also examples of programs that opted not to alter their strategy based on early results. The reasons for these decisions vary by case. Some parts of the T&E community have defined protocols and long histories of close coordination between DT and OT, with the Electronic Warfare community being a prime example. For most other programs, this collaboration is done on an ad hoc basis, making overall progress uneven. Similarly, some types of risk (such as statistical risk) are well defined and widely used in test planning. Other types of risk remain difficult to define and hard for acquisition chiefs to manage systematically.

Additionally, some areas that were not considered pressing or critical challenges to T&E in 1994 have emerged in the intervening years. Table E.2 introduces three modern challenge areas and offers a similar qualitative assessment of progress. While symposium participants were aware of the challenges in testing software in 1994, the scope of the challenge has increased by orders of magnitude since then. Cybersecurity testing wasn't mentioned in the 1994 symposium summary, but today nearly every system undergoes cybersecurity testing as part of both operational and developmental testing. Although cybersecurity testing is now more comprehensive and sophisticated than it was even 10 years ago, quantitative approaches for determining how much cybersecurity testing is necessary remain nascent. The use of artificial intelligence (AI) and autonomous systems present the T&E community with unique challenges to overcome. These challenges are all vital, and the T&E community must continue to develop approaches for determining the appropriate amount of testing required to address concerns associated with these challenge areas.

Table E.2. Progress in Modern Challenge Areas

Modern Challenge Areas	Progress
Software-intensive systems	Moderate
Cybersecurity	Minimal
Artificial intelligence & autonomous systems	Minimal

The editors of the MORS/ITEA symposium report noted that the three-day event may have brought up more questions than it answered. Given that we are still grappling with the question of "How much testing is enough?" this assessment appears prescient. The persistent challenge is that no single approach can answer this question overall. As a result, this report does not aim to answer the question set forth in 1994 but instead documents the progress made by the T&E community in answering those questions in the intervening years.

Contents

1. Introduction

In 1994, the Military Operations Research Society (MORS) and the International Test and Evaluation Association (ITEA) held a joint three-day symposium titled, "How Much Testing Is Enough?"[1] The symposium consisted of keynote addresses, presentations, and working group sessions. The goal of the symposium was to "provide a forum in which the military operations research and test and evaluation communities could identify key issues and develop novel and useful insights into more cost-effective test and evaluation." Attendees included the "principal decision makers and the doers of testing, analysis and acquisition communities," from government, industry, and academia. Included in this group were representatives from the office of the Director, Operational Test and Evaluation (DOT&E). The symposium identified challenges, made recommendations, and, according to the chairs, may have produced more questions than it answered.

Shortly after the symposium, organizers published a report consisting of a summary of findings, a review of the discussions from the various panels and working groups, a partial list of participants, and major recommendations. While many of the panels and working groups provided recommendations pertaining to their focus area, the executive summary contains three overarching recommendations (quoted here verbatim from the report):

- The T&E planning process should be subjected to fundamental Operations Research (OR) scrutiny, including explicit consideration of alternative T&E strategies, contingency planning, and cost/benefit tradeoffs.
- Each individual T&E program should empower a small, stable "integrated T&E team" (with some mix of contractor, developer, trainer, operational test and evaluation, user, and office of the secretary of defense representation) to manage design, evaluate, and implement issues; continually monitor T&E planning activities and review emerging results; and revise T&E plans as warranted. The team's dual emphasis should be on comprehensiveness and efficiency.
- More emphasis should be placed on ensuring that each individual test and evaluation activity is efficiently designed and analyzed. In particular, experimental design techniques and other established statistical approaches should be better exploited.

[1] Gherig, et al. (1994)

It has been 25 years since this symposium took place, and the question that inspired it remains a constant concern in the test and evaluation community. But that is not a sign that the state of the art has remained stagnant. We can look back at the concerns expressed at the symposium and see many areas in which the test and evaluation community has made progress, some areas where we have not, and others where the changing world forces us to reexamine old solutions and address new challenges.

The symposium report's Executive Summary highlights nine major Discussion Areas (DAs) that will frame the discussion of the progress made by the T&E community in the past 25 years in answering the question, "How much testing is enough?"

1. Impact of ATDs/ACTDs[2] on T&E
2. Early coordination between DT and OT communities
3. Using early T&E data to influence budgeting/acquisition strategy
4. Integration of M&S with testing
5. Lack of visibility of testing costs
6. Statistical application in the T&E discipline
7. Pooling/sharing data across DT, OT, and M&S
8. Characterizing different types of risk
9. Greater statutory and regulatory flexibility to address "the unexpected."

In addition to the above DAs, new technologies have created challenges that did not exist or were not viewed as critical in 1994:

- Software-intensive systems and automated software T&E
- Cybersecurity testing
- Artificial intelligence and autonomous systems

In each of these areas, the T&E community has been asked to develop new approaches and adapt what exists from industry to provide timely, decision-quality information to leadership and warfighters. While there has been some success, these areas remain challenges for the T&E community.

The goal of this report is to assess the progress the T&E community has made since the 1994 symposium. As the organizers of the symposium noted at the time, the three-day event may have produced more questions than it answered, and emerging technologies

[2] Advanced Technology Demonstrations/Advanced Capability Technology Demonstrations

have added even more. The T&E community has answered some of those questions in the intervening time, but others remain unaddressed.

Given the breadth of the subject, no report can comprehensively cover the past 25 years. Since 1994, there have been eight confirmed Secretaries of Defense, and the organizational structure by which DoD and the Services do test and evaluation has changed multiple times. Rather than offering a detailed retelling of this history, this report focuses on the top-level progress in the nine DAs from the 1994 symposium report, as well as emerging issues driven by new technology. The selected case studies used to highlight progress are especially interesting examples, rather than a comprehensive look at all programs since 1994.

The remainder of this report is organized as follows: Chapter 2 addresses each of these DAs, starting with a brief summary of the major points coming out of the 1994 symposium, before identifying progress that has been made in that area. After this brief summary, we discuss in detail examples of progress, such as the use of (1) formal policies and guidance and (2) case studies, to highlight what has changed (and what has not changed) since 1994. Chapter 3 discusses new challenges in answering the question, "How much testing is enough?" that have emerged since 1994. Chapter 4 gives a few concluding remarks.

2. Progress to Date

A. DA 1 - Impact of ATD/ACTDs on T&E

1. Notes from the MORS/ITEA symposium

At the 1994 symposium, several presenters commented on the need for better incorporation of T&E activities with Advanced Technology Demonstrations (ATDs) or Advanced Capability Technology Demonstrations (ACTDs). The lack of attention to proper test planning for ATD/ACTDs was the theme in many focus group discussions. One commenter went so far as to question what ATD/ACTDs did to contribute to T&E in acquisition. Another commenter suggested that T&E for ATD/ACTDs could be part of early acquisition and be included as an early Developmental Test (DT) phase. An additional presenter suggested that more rigorously designed ATDs could provide useful data for evaluating any resulting program of record, reducing the overall cost of T&E. While many participants believed that rigorous testing (vice "experimenting") in ATD/ACTDs could help reduce testing down the line, some discussants believed that new policy (i.e., DoDi 5000.02) on testing for ATD/ACTDs was a prerequisite for fixing the process.

2. Progress since 1994

Little progress has been made in addressing most of the concerns outlined above. However, the way that DoD does prototyping and technology demonstrations has undergone many rounds of change. ATDs and ACTDs no longer exist. DoD and the Services have attempted to use a variety of structures to more rapidly incorporate new technology with the active force, but many of the issues with ATDs outlined in 1994 persist in similar early technology demonstration events that occur today.

a. ACTD to JCTD

The acronym and business model changed in 2006 from ACTD to the Joint Capability Technology Demonstration (JCTD), and a presentation from the Deputy Undersecretary of Defense for Advanced Systems and Concepts (DUSA(AS&C)) discussed the rationale behind the change. Following is a summary of that presentation: ACTDs were intended to field mature technologies to joint warfighters by helping to tailor technology and develop tactics, techniques, and procedures. From 1995 to 2007, more than 150 ACTDs were initiated, and 76 percent of all ACTDs transitioned at least one product into a warfighting capability. However, ACTDs often failed to solve immediate military problems, often failed to transition to program of record quickly, and weren't adequately addressing joint problems. The JCTD program was created in 2006 to address these weaknesses. The goal of JCTDs was to tailor solutions to Combatant Commander needs, yield faster starts and deliveries, and provide a mechanism for joint science and technology focused on capabilities from concept to production. In terms of T&E investment, ACTDs were originally tied to a specific military exercise, while JCTDs were allowed more flexibility in order to provide more opportunities to show operational utility.[3]

b. Other pathways

Other novel acquisition paths designed to rapidly field emerging technology continue to present testing challenges. Multiple pathways for rapidly fielding technology now exist beyond the JCTD. Quick Reaction Assessments, Early Fielding Reports, and most recently the Middle Tier Acquisition Pathways provide the Department with tools for quickly deploying new technology and forgoing at least some of the traditional steps in the acquisition process.

The Operational Test (OT) community has been involved in multiple rapid acquisition efforts, supporting urgent operational needs (UONs) and joint urgent operational needs (JUONs). DOT&E has consistently supported rapid fielding through early involvement of testing, with the Mine-Resistant, Ambush Protected (MRAP) vehicle[4] and the MQ-1C Gray Eagle[5] being two prominent examples. The Gray Eagle is an example of a program of record (rather than a technology demonstrator) that successfully underwent testing and was rapidly fielding.

Early testing does not necessarily lead to the success of a program, but it can identify important deficiencies. One common challenge is system integration. For example, a new system (e.g., the Large Aircraft Infrared Countermeasures (LAIRCM) system) needed for rapid fielding may integrate well on one platform but not well on another. While LAIRCM was rapidly fielded and successfully integrated onto the AH-64E,[6] there were considerably more challenges integrating it onto the MV-22.[7]

3. Summary

Many of the problems initially described to justify the transition away from ACTDs to JCTDs persist with modern technology demonstration efforts. While technology

[3] Peterson (2006)

[4] DOT&E (2010, 1), DOT&E (2010, 2)

[5] DOT&E (2009)

[6] DOT&E (2015)

[7] DOT&E (2018)

demonstrations may help to identify promising capabilities, they rarely feature systems that are close to deployable. The environments of a technology demonstration are rarely similar to the conditions in which warfighters might employ a system with the demonstrated capabilities. Examples of successful rapid fielding efforts include traditional programs of record. Middle Tier Acquisition pathways provide yet another tool for rapidly fielding technology, but it is too soon to know how successful they will be.

B. DA 2 - Early coordination between DT and OT communities

1. Notes from the MORS/ITEA symposium

Acquisition programs can benefit from early and continued coordination between developmental and operational test teams. A major theme of the 1994 symposium was the general lack of coordination and communication between the two teams. Presenters noted cases of developmental testers not releasing data to the OT team, and several cases in which the OT team expressed that they would not accept "tainted" or operationally unrealistic data from DT. Members of the symposium stressed the need to go beyond coordination to collaboration, and work together early in the test planning stages in a more integrated fashion. They stressed the need for the program office, the contractors, and the test groups to all work together. They suggested that DT and OT testers identify opportunities to share data between test phases and determine synergistic test objectives. They also stressed that proper test design leads to data usable for multiple purposes and across a larger operational envelope.

2. Progress since 1994

The DA concerning early collaboration between test communities is a relative success story when compared to the other DAs. There are many notable examples of successful DT/OT coordination since 1994, including the electronic warfare (EW) and rotary-wing aviation communities. While changes to policy and encouragement from executive leadership have been critical, collaboration is enhanced when individuals in program offices and DT/OT test teams have good working relationships. Better policy, more encouragement from leadership, and sound processes would move DoD closer to routine and effective collaboration throughout all phases of acquisition

a. OSD, DOT&E, and Operational Test Agency (OTA) policies

Since 2000, the Office of the Secretary of Defense (OSD) and DOT&E have directed that DT and OT collaborate early to improve the likelihood of system acquisition success.

A 2000 OSD memo[8] detailed the need for developmental, operational, and deployment testers to collaborate in efforts to remedy interoperability problems across the Services.

Likely the most influential and direct memos from DOT&E and the Deputy Under Secretary of Defense for Acquisition and Technology (DUSD(AT&L)) came in December

[8] AT&L, DOT&E, ASD C3I & Joint Staff (2000)

2007[9] and April 2008,[10] mandating and defining the use of integrated testing in acquisition. The 2007 memo co-signed by DOT&E and USD(AT&L) encourages teaming between the OT and DT communities. The integrated testing definition from the April 2008 memo reads:

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 particularly the developmental (both contractor and government) and operational test and evaluation communities.

The Defense Science Board (DSB) also highlighted the need for collaboration between developmental and operational testers in its 2008 report on developmental test and evaluation. The DSB recommended policy mandating integrated test planning throughout the program cycle, including data sharing and integrated test events (where practical) designed to satisfy both OT and DT requirements.[11]

In May 2019, the OTAs agreed upon six core test principles that were codified in a memorandum:[12] Early OT Involvement; Tailor to the Situation; Continuous and Cumulative Feedback; Streamline Processes and Products; Integrated and Combined Collection/Test; and Adaptive. Many of these six principles emphasize areas in which OT and DT can work together early on. While the concepts described in that memo have been used in the acquisition community for some time, the memo places renewed focus on the benefits of early collaboration between the DT and OT communities.

b. Examples of DT/OT collaboration

The Marine Corps H-1 upgrade program[13] is an example of a successful DT/OT collaboration. The H-1 Upgrades Test and Evaluation Master Plan (TEMP) Rev B[14] describes the Continuous DT Assist concept. To implement Continuous DT Assist, the Navy's Commander, Operational Test and Evaluation Force[15] (COTF) stationed a detachment of Marine pilots and maintainers (H-1 operational test team (HOTT)) to Pax River to monitor DT every day, from the start of DT until the end of OT. With this much visibility of DT, the OT pilots were able to participate in early DT flights before conducting two operational assessments, both of which provided useful information. At the first OA,

[9] DOT&E & AT&L (2007)

[10] DOT&E & AT&L (2008)

[11] DSB (2008)

[12] DoD Operational Test Agencies (2019)

[13] Crabtree, et al. (2007)

[14] Program Manager PMA-276 (2005)

[15] COTF is the Navy's OTA and the lead test agency for the H-1 Upgrades OT.

the OT pilots discovered how poorly the Targeting Sight Sensor (TSS) had been integrated onto the AH-1Z attack variant.[16] In the second OA, the OT pilots found other problems[17] with both aircraft and resolved an issue that had been raised during DT. During DT, the Navy discovered that the four-bladed composite rotor system on each aircraft flexed under a load much more than the designers expected. The Navy then imposed a G-limit to maneuvers that was below the requirement. In OT-IIB, the OT pilots discovered no lack of maneuverability in spite of the new, lower G-limits. So, although the rotor system did not meet specs, it met all reasonable expectations for operational maneuverability. Their close integration with the developmental testers meant that the HOTT was familiar with this issue prior to the OA and was able to successfully resolve it.

The B61 program used an integrated 52-shot matrix to generate the live test data required for both DT and OT. Early and continuous collaboration took place among the contractor, the program office, Sandia National Laboratory, and the developmental and operational test units. The program office made integrated experimental design one of the critical statements of objectives in the contract. The program office further motivated collaboration by holding a series of investigatory test working groups to examine the outcome of an experimental campaign to characterize the weapon performance using data from integrated flight simulation. Because the operational testers and other stakeholders were routinely involved in this investigatory group, they were more willing to consult and collaborate with the program office and the contractor in designing the operational shots. This collaboration led to trust among the participants. The result was that positive early test results and the detailed understanding that the OT team had of those results led to the program office canceling 4 of the 26 government developmental test shots, and declaring the system ready for dedicated IOT&E. This reduction saved a minimum of $8 million dollars of direct savings and shortened the delivery timeline by three months.[18] By working together, the test teams reduced the total number of shots required.

Another successful example is Small Diameter Bomb (SDB) II.[19] For SDB II, key tools for collaboration were sequential testing phases, design of experiments (DOE) planning for each phase, and contractor/DT/OT collaboration. The series of tests included contractor DT serving as the foundation for government DT, leading to Air Force IOT&E, then extending to Navy DT and OT.[20]

[16] Comfort, et al. (2003)

[17] Aldridge, et al. (2005)

[18] Huffman, et al. (2019)

[19] Ortiz (2019)

[20] Dailey (2016)

c. Integrated testing includes the contractor

Participants in the 1994 symposium emphasized the need for a more integrated approach to testing involving the contractors, DT, and OT, stating that all three entities should employ statistical test design. The idea was that if contractor testing is sufficiently rigorous and data are made available to the government testers, some test points may not need to be repeated in subsequent developmental testing. This approach hasn't become common practice in the intervening years, but some systems have attempted to use contractor test data for OT evaluation.

In the case of the B61, the contractor worked with government testers throughout system development. The contractor, developmental test team, and operational test team collaborated successfully to use Robust Product Design. The B61 OT evaluation incorporated contractor knowledge of how the system was physically constructed and data from early developmental testing.[21] A comparable evaluation done using OT data exclusively would have required additional time and test resources.

Unfortunately, many development contracts do not cover the release of data, and in such cases, the contractor may not be willing to provide the data to the government testers. In the 1990s, a common cost savings tactic for the government was to not purchase the intellectual property (IP) rights of system models and software. While this may have saved money in the short run, it meant that when systems were updated or testers wanted to save costs by using data from M&S, the government had to pay every time the model was used. These problems persist today, where it was recently revealed that the ambiguity in IP rights ownership for a key Joint Strike Fighter (JSF) model caused a delay in JSF testing.[22]

d. DT/OT collaboration in the EW test process

In 1994, in response to congressional direction (and later revised in 1996), OSD published a DoD T&E process for EW systems.[23] This process was a "best practice" of how to do T&E on EW systems to reduce occurrences of incomplete early testing and poor performance at IOT&E. Inherent in the EW test process is substantial use of DT information by OT, so a great deal of collective planning is required to guarantee the efficient use of time and test resources. DT data from chamber tests and M&S help to determine the most useful data points to explore in OT. Data from OT feed back into the M&S platforms to improve their accuracy and validity. In recent years, there has also been a transition from EW testing that relied on subject matter expertise (e.g., from former Electronic Warfare Officers (EWOs)) to more M&S-based EW testing, which is relevant

[21] Ortiz (2019)

[22] Albon (2019)

[23] DTSE&E (1996)

to DA 4 from the MORS/ITEA symposium. Since 1996, when the process was published, the scope of EW systems employed by the U.S. military has increased dramatically. An update to the process could address newer technologies that rely on the electromagnetic spectrum but that are not classic EW systems (e.g., radars; radios; Position, Navigation, and Timing).

e. Validation as standard T&E practice

There can be great value in reserving some OT test points for validating the primary findings from DT&E. In a recent Targeting Pods test, the test team found unexpectedly good results using an atypical target tracking tactic. The DT team encouraged the OT group at Nellis Air Force Base (59th and 422nd Test and Evaluation Squadrons) to repeat the results with different atmospheres, targets, pods, tails, and pilots, and obtained comparable results, which validated the initial finding. The general test design and execution practice should ensure that a sequential test plan incorporates a validation step to replicate and confirm findings (positive or negative) from DT.

f. What still hampers collaboration

Early collaboration is clearly in the best interests of both the OT and DT communities. The DT community benefits from having the operational view of the risk areas employing the new system. The OT community gets an early look at performance to inform the design of any operational assessments or operational utility evaluations, as well as the design of dedicated IOT&E. Early coordination is hampered, however, by three primary ongoing issues:

1. DT and OT have different objectives, which makes designing tests that are useful to both communities challenging. Experimental design can help bridge this gap by providing a framework through which both sides can describe their goals. Developmental testers may need to assess certain system specifications, while operational testers may need to measure system performance under a variety of different operational conditions. By using a common framework, finding points of overlap and potential efficiencies may be easier. Currently, many offices lack the relevant technical expertise (e.g., a senior operations analyst/statistician), making a dialogue centered around DOE challenging.
2. Program office representatives often don't believe they will benefit from more integrated testing. Unless the number of OT test events is reduced, there is little reason for DT to go through the processes of coordinating with the operational side for developmental test events. Conversely, the operational testers are concerned that unforeseen test limitations will mean that they will not get the required data from DT, and are therefore unwilling to give up OT events. A "wait and see" approach, by which OT observes what developmental testers do and

harvests any data they can for their operational evaluation, is often the suboptimal result.

3. For many programs, more than a decade will pass between late technology development and the dedicated IOT&E. With the turnover of government personnel, especially in the military, up to three generations of personnel might be involved with the program from beginning to end. Such turnover makes it difficult to sustain commitments made by earlier program participants. There are no simple ways to resolve this challenge, and it is likely to persist for the foreseeable future.

g. Ad hoc collaboration processes

Although gains have been and will continue to be made by programs recognizing early the need for DT and OT to work together for the common goal of gaining maximum information from test and evaluation, the current process for collaboration is inconsistent across Services and programs. While these ad hoc efforts are laudable, a standardized approach may not be possible because of the different goals and requirements of DT, OT, and live fire testing. Although a variety of statistical methods to plan for the analyses of the combined DT and OT data have been developed (see “DA 6 – Statistical application in the T&E discipline” for more details), these are not substitutes for the process for early collaboration between DT and OT.

h. Learn from history

A 2013 paper in the journal *Quality Engineering* discusses ways to reduce duplication of effort during T&E through improved efforts to archive and share data among communities. According to this paper, data from one organization may not be easily found by people in other organizations, or even by other people from the same organization. Test planners should leverage information from defense information sites, such as the Defense Technical Information Center and the Defense Systems Information Analysis Centers; they could then summarize any useful results for reference in the subsequent test plan, and include how the relevant historical data might be incorporated into the analysis to follow the test. Improving the way reports and data are indexed, stored, and archived could make these tasks easier. [24]

3. Summary

Since 1994, there have been many examples of programs reaping the benefits of early and continued DT and OT collaboration, including those for rotary-wing aircraft and air-to-ground munitions. These are not comprehensive, however; success stories exist across

[24] Simpson, et al. (2013)

Services and products. The EW community has a long history of using DT data for OT evaluation, in many cases by necessity. Additionally, DOT&E, DUSD(AT&L), and the Service OTAs have published policy encouraging this collaboration, and leadership from such organizations remains critical to continued success.

Impediments to successful collaboration remain. Processes for DT/OT collaboration are still ad hoc. DT and OT have different objectives, which makes collaboration a perpetual challenge. Turnover of uniformed personnel make long-term “handshake” agreements between DT and OT difficult to commit to, and the value of making such commitments is not always readily apparent to either the OT or DT testers. Despite these challenges, the progress since 1994 and the examples of success noted above are encouraging.

C. DA 3 - Using early T&E data to influence budgeting/acquisition strategy

1. Notes from the MORS/ITEA symposium

New programs should use current knowledge about existing or similar systems to shape their acquisition strategies early on in system development. More specifically, programs should develop alternative acquisition strategies and use "prior information" about systems to determine the most cost-effective strategy. This knowledge can be used to anticipate challenges and devise relevant contingency plans, which should in turn be discussed in the TEMP.

TEMPs present rigid schedules and sequences of test events, but participants at the symposium recommended that the TEMPs instead include multiple approaches so that the acquisition strategy could be informed by early test results. Rather than the "success-oriented" approach favored in TEMPs, more realistic test strategies would include slack time and consideration for what should occur when problems inevitably arise. Finally, T&E planning should be subjected to operations research scrutiny, including contingency planning and cost/benefit analysis.

2. Progress since 1994

Many methods and strategies exist for (1) ensuring that data collected early in a program's life cycle can be incorporated into programmatic decisions, and (2) anticipating the ways in which early test results can influence program timelines. This section identifies examples showing how incorporating early data helped put programs on a path to succeed, as well as examples of programs opting not to change course when new information became available. Tools that allow programs to systematically use early data to inform programmatic decisions, such as reliability growth planning, are also discussed. One challenge in this area in which progress has not been made is the integration of early data into budgetary decisions at the Department level. While program managers may be able to make informed decisions and move funds around within their own budgets based on early results from T&E, they must still operate within the constraints of the DoD budgeting process.

a. Delaying further testing to fix problems discovered in early testing can save money

Budgeting time to fix future problems that are inevitably discovered in DT and OT creates realistic expectations for programs. Current strategies that fail to budget time for future fixes result in “delays” that more accurately reflect good programmatic decisions to fix problems before full-rate production and fielding.[25]

When time is scheduled for fixes to be made before further testing, programs have positive outcomes. In November 2017, the Army Apache had 23 mission-critical problems with the v6 software, which was scheduled to go to test.[26] The Army decided to postpone the FOT&E by a year to allow the Apache program time to fix these problems and integrate JAGM onto the v6 Apache. They then conducted the Apache FOT&E at the same time as the JAGM IOT&E.[27] By conducting the two tests together, the Army saved an estimated $2 million.[28]

b. Sequential test designs

Sequential testing uses multiple test phases, arranged such that the results from earlier test phases influence the runs in subsequent test phases.[29] A multi-phase test strategy that schedules the most critical test events and runs early can be advantageous, since it enables the test team to use those early results to inform decisions about later testing. If critical hurdles are passed early on, less testing may be required later on. If the system struggles in those critical events, a more thorough set of tests than originally planned may be necessary. Sequential test designs are challenging to use in DoD, since the number of test runs, the conditions for those runs, and the resources required to execute those runs are often decided early on and codified in the TEMP. Issues with sequential test designs can arise if system performance is assessed by many measures, each of which is affected differently by multiple factors (i.e., they vary over different conditions). Sequential designs can also prove challenging to implement when the time required to score individual test events takes longer than the scheduled time between tests, and when OT&E stakeholders have divergent assessments of test runs.

c. Sensitivity tests

In a sensitivity test, the objective is to find a setting of a factor that produces a good estimate of a specific desired performance outcome, such as the bullet velocity that results in personnel armor penetration 50 percent of the time (the V50). The test approach takes advantage of previous settings and outcomes to derive the next series of tests. For example, the number of test events may be reduced if success or failure can be determined with necessary precision from the first few data points. Alternatively, when early results are

[25] DOT&E (2014, 3)

[26] Joint Attack Munition System Project Office (2017), Meely, M. (2017)

[27] Hoecherl (2018)

[28] ASA(AL&T) (2018)

[29] Fisher (1952), Box (1993)

mixed, testing continues so that system performance can be assessed to a level of precision agreed upon initially by the stakeholders.[30]

DOT&E has advocated sensitivity test designs in the past. One example in which this approach proved successful was DOT&E's First Article Test protocols, which provided assurances of acceptable performance under multiple conditions without increasing the number of shots required.[31]

d. Subsystem off-ramps

A subsystem off-ramp is a planned alternative that can be used to continue to develop and deploy a system in case a particular subsystem doesn't develop on the same timeline as the rest of the system. When early testing shows poor performance of critical subsystems, programs can continue progress on the overall system and avoid major delays by developing a subsystem off-ramp.

A recent example is CVN 78, which experienced significant reliability issues with its new arresting gear, eventually resulting in a Nunn-McCurdy breach. The legacy equipment is a form and fit replacement for the new equipment, and multiple third parties had suggested that the Navy have an off-ramp to install the legacy equipment if reliability for the new arresting gear did not improve. The Navy opted not to employ this off-ramp because of the new arresting gear's advertised capability of recovering heavier and lighter aircraft than the legacy system can. As of the time of this publication, the ongoing reliability issues have caused the Navy to forgo certifying the new arresting gear for these new aircraft types.[32]

e. Early T&E data

Discovering system problems through early T&E allows programs more time to adjust and make fixes. By the time programs reach OT, there is typically no time left to resolve newly discovered issues. This makes it even more critical to discover system problems early. While certain issues may become apparent only once the system undergoes OT, programs should strive to uncover problems in DT wherever possible. Unfortunately, problems that could have been discovered through earlier testing are sometimes identified only in OT. The 2013 DOT&E Annual Report highlights 12 examples of programs that exhibited problems in OT that should have been discovered during DT.[33]

[30] Johnson, et al. (2014), Wu and Tian (2014)

[31] DOT&E (2013, 2)

[32] DOT&E (2019, 2)

[33] DOT&E (2014, 4)

f. Reliability growth

Both reliability test resource planning and reliability growth require early failure data, so early T&E data are critical for making the best use of reliability growth techniques. Reliability and reliability growth are often afterthoughts in T&E, but progress has been made in both areas since 2005. Multiple DoD offices have prioritized reliability, contributing to the creation of the Systems Engineering Directorate in research, development, and test and evaluation, as well as the publication of several documents highlighting the importance of reliability. These documents include the DoD Guide for Achieving Reliability, Availability and Maintainability (RAM), the 2008 Defense Science Board Report, the RAM-C Report Manual, the OSD/AT&L Directive Type Memorandum 11-003, and the R&M Engineering Guide.[34]

g. Critical chain program scheduling

Staffing at OT and DT organizations throughout DoD is limited,[35] and this needs to factor into the rate at which new test efforts are started. When the number of test events to plan strains the staff capacity, staff members invariably start multitasking (frequently switching back and forth between tasks), trying to make at least some progress everywhere in order to comply with the program office schedule. Multitasking can dramatically reduce productivity and negatively affect quality, reducing the overall capacity of our test infrastructure. A solution to the problem is to regulate the release of test projects so as not to overload resources and to drive focused work. Capacity-based multi-project scheduling approaches such as those employed in critical chain project management[36] could help reduce the need for multitasking. Critical chain[37] has been applied since the late 1990s to a wide variety of projects to pace the release of projects and provide clear task priorities to project workers.

3. Summary

Since the symposium in 1994, some DoD programs have been successful in conducting useful early testing, and used the resulting information to inform programmatic decisions. However, these efforts remain program-specific rather than systematic efforts across the DoD acquisition system. Most efforts in making programs more reactive to early test data focus on transforming the TEMP into a less rigid document, which may make programs more nimble. Suggestions for improving the TEMP include incorporating explicitly responsive test designs, such as sensitivity test designs; building slack into the

[34] DoD Handbook, (2011)

[35] Thomas, et al. (2017), Warner (2013)

[36] Leach (2014)

[37] Steyn (2001)

schedule; and allowing subsystem off-ramps. While these efforts may provide some flexibility, they do not address the ways in which programs receive funding and build their budgets. Early T&E data, especially data from tests intended to stress the system, are extremely beneficial to program success, and programs should strive to collect these data and incorporate them into programmatic decisions.

D. DA 4 - Integration of modeling and simulation with testing

1. Notes from the MORS/ITEA symposium

The contributors to the symposium expressed many divergent points of view, but the consensus was that live testing and M&S are both essential to the T&E domain. Some cautioned that while M&S can add value regarding knowledge of system and subsystem performance, it is not a surrogate or replacement for live testing, so it does not provide an answer to the high cost of T&E. The DoD Comptroller suggested that testing should be done to calibrate the M&S, and the M&S (once validated) should be used to make decisions. Another participant commented that "simulation is the only way we can really test some of the interesting systems," which applies today to scenarios such as the most challenging electronic warfare environments and many advanced threats. Live, virtual, and constructive (LVC) simulations were burgeoning concepts at the time, and many symposium contributors were optimistic about their eventual utility. One participant was more circumspect about the chances of M&S solving the problem of "How much testing is enough," stating: "Distributed interactive simulation, for example, has become a common topic of much discussion among defense T&E communities, but there are few examples to date of how it has been applied to OT&E in a cost-effective manner."

2. Progress since 1994

Simulation-based acquisition was initiated not long after the 1994 symposium, when the Director for Test Systems Engineering and Evaluation (DTSE&E) commissioned in 1995 a one-year study to assess the effectiveness of the use of M&S in weapon systems acquisition and support processes. The DTSE&E study developed an approach to acquisition that was named Simulation-Based Acquisition. DTSE&E was disestablished by the Secretary of Defense on June 7, 1999, and some functions were transferred to the Director, Operational Test and Evaluation (DOT&E). Each of the Services now has its own M&S agencies and offices, while DoD supports the Defense M&S Coordination Office (DMSCO).[38] In 2018, the Deputy Assistant Secretary of Defense for Acquisition and Sustainment (DASD(A&S)) published a Digital Engineering Strategy.[39] While M&S constitutes a large part of the T&E strategy for many systems, it has not dramatically reduced T&E costs, and live tests remain a vital component of T&E.

a. M&S is integral in T&E

Since 1994, the maturity and capabilities of M&S have grown tremendously. M&S is now integral in providing the needed evidence for requirements verification and for

[38] National Research Council (2002), O'Bryon (2006)

[39] DASD(A&S) (2018)

demonstrating and validating system performance. It is increasingly being used to provide evidence for capability in regions of the operational space where live testing is not conducted (Figure 1). LVC simulation, integrated high-fidelity digital simulations, and hardware-in-the-loop (HWIL) simulations are developed and used in nearly all phases of DoD acquisition. Examples of weapon systems using and depending on M&S include but are not limited to aircraft (e.g., F-22, F-35), ships, vehicles, torpedoes and surface-to-air missiles, directed energy weapons, satellites (e.g., Space-Based Infrared System), and ground-based radar systems (e.g., Aegis Combat System, Ballistic Missile Early Warning System). Other types of simulations used by DoD include force-on-force engagement (e.g., F-22 Air Combat Simulation), mission and campaign constructive simulations (Brawler, Thunder, Metric, Suppressor), Electronic Attack and Infrared (IR) countermeasure systems (e.g., Joint Mobile IR Countermeasures Testing System), air refueling systems, targeting (e.g., Army Virtual Targets), and communications systems (e.g., Tactical Communication Network simulation).[40] Validated M&S systems are crucial for understanding initial subsystem, system, and system-of-systems capabilities (early evaluations) of DoD programs. They are also depended upon for assessment of the more mature systems, as well as follow-on testing, including operational training and tactics development.[41]

Note that while M&S is critical for testing these systems, it has not always been as useful as desired. The JSF is a recent example of a program that planned to make extensive use of M&S for testing. Unfortunately, delays in developing the M&S suite have contributed to delays in the completion of IOT&E.[42] In other cases, such as the Aegis Combat System, DOT&E identified issues with using the existing M&S system for OT&E.[43] These examples highlight the risk of relying on undeveloped or unvalidated M&S for T&E.

Figure 1 illustrates one of the ways that M&S can be used to supplement live T&E. In cases where live testing can be done across the full range of operational conditions (left panel), M&S can supplement these data by filling in the spaces between live test points and providing supplementary runs. In cases where live testing is limited to only a portion of the battlespace (due to threat limitations, safety concerns, etc.), M&S can be used to explore the space where live testing is not possible. In both cases, live test data should be used to validate the M&S to ensure high fidelity of collection.

[40] Holt (2016)

[41] Aegis Ballistic Missile Defense Program Office (2015), PEO Integrated Warfare Systems (2015)

[42] Albon (2019)

[43] DOT&E (2013, 1), DOT&E (2013, 3)

Strategies for selecting the design space and points for live and simulation-based testing

- **Ideal** : the live design (white points) encompasses the operational and simulation design space so comparisons between them are interpolations, not extrapolations.
- **Limitation in Live Space**: often due to practical constraints that exist in live testing. Here the domain of the live testing should span the maximum possible domain of the simulation experiment and regions of extrapolation should be clearly identified in the validation limitations

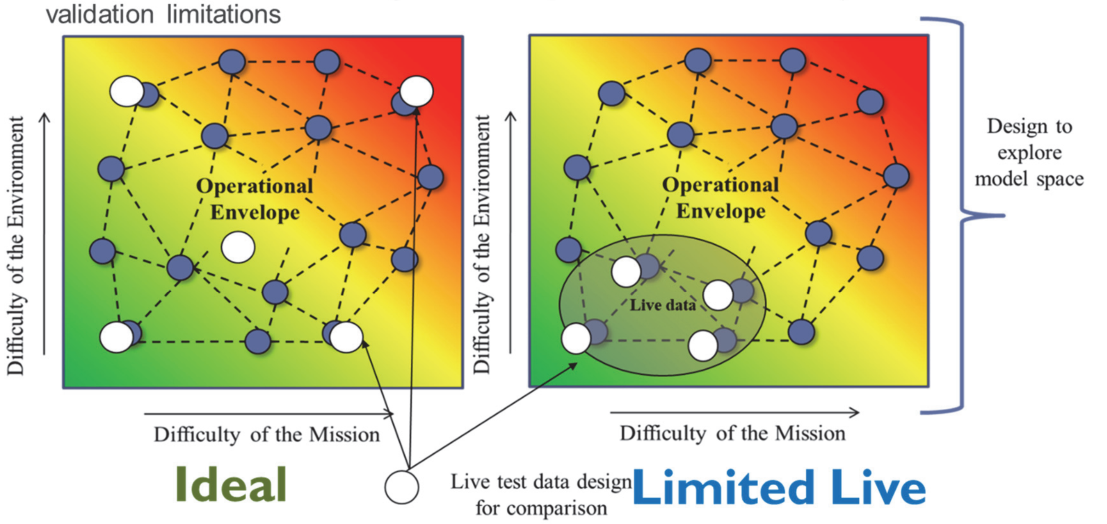
Difficulty of the Environment
Operational Envelope
Difficulty of the Mission
Ideal
Difficulty of the Environment
Operational Envelope
Live data
Difficulty of the Mission
Design to explore model space
Live test data design for comparison
Limited Live

Figure 1. The Challenges for Nearly All Current Weapon System Acquisition Programs Is That the Space Coverage Via Live Testing is Small Compared to the Full Operational Envelope. M&S is Often the Only Means of Gathering Evidence to Validate Capability In These Uncovered Regions.

b. M&S in EW

M&S has a long established history of use and is viewed as a necessary part of the T&E process for electronic warfare countermeasures systems. Firing actual munitions at manned aircraft is not an option because of safety concerns, and getting actual threats for use in testing is challenging. These circumstances make M&S vital for EW systems T&E.

Rigorous testing of modern electronic attack or electronic defense systems requires not only accurate representation of the system under test and relevant threats but also representation of complex employment environments. The complexity presents a major challenge for testers, whether the testers use live representations or fully digital simulations. Therefore, testing must be completed in a segmented fashion, combining actual hardware and software with simulated hardware and software. For example, the aircraft onto which the EW system under test is installed might be present for only a portion of the testing. For other parts of the testing, aspects that are absent may be represented by antenna pattern simulations, signature data, and flight path models.

Other aspects of EW testing likewise rely heavily on M&S. Most munitions (on both the red and blue sides) are represented via simulation. Enemy weapons that are not available for test must be modeled based on intelligence assessments and the modeler's best judgment, and EW techniques often cannot be radiated in free space because of security issues. To keep up with current and emerging weapon systems, the test community must perform more testing in anechoic chambers with direct injection and free-space radiation that is secure.[44]

c. Positive trends in the use of M&S

In the recent past, DoD often did not own the system-level models developed for each program, but this environment is rapidly changing for the better. M&S developments are becoming more like software development programs, and agile development operations are more commonplace. In some cases, the Services are investing the funds needed for processing cores to efficiently execute the simulations at rates that are often faster than those of the contractor. This capability enables the government to sometimes assist the contractor in simulation tasks, while at other times performing their own studies to gain additional insights. Some examples of this trend include the MV-22, JAGM, AMRAAM, CH-53K, SDB-II, B61 munition, and AIM-9X programs.[45]

d. Digital engineering

The Digital Engineering Initiative envisions a library of DoD-owned system-level models through which testers and analysts determine whether "off the shelf" models can be adequately adapted to suit the needs of the program.[46] Such a library would be consistent with the most optimistic visions of M&S expressed at the 1994 symposium.

DoD has attempted projects like this in the past. In the 1990s, the Defense Modeling and Simulation Office attempted to standardize the architecture used by DoD M&S systems. To do this, they developed the High Level Architecture (HLA) standard.[47] Although HLA didn't result in standardized, interoperable M&S across DoD, it was adopted by the Institute of Electrical and Electronics Engineers (IEEE) as a standard. HLA is still used as a standard today across a variety of domains beyond M&S.

e. DOT&E initiatives

DOT&E provided guidance and clarification, in two memos,[48] on validation methods and the use of live testing with M&S. These memos provided guidance to the T&E community, provided standard techniques for the use of M&S in T&E, and encouraged the

[44] Kyle (2017)

[45] COTF (2003), AFOTEC (2018)

[46] DASD(SE) (2018)

[47] Dahmann (1997)

[48] DOT&E (2016, 2), DOT&E (2017, 1)

collection of data for use in validation. The memos recognize that traditional approaches to document review, face validation, and subject matter expert (SME) evaluation are useful but not sufficient by themselves. These memos also stressed the need to incorporate the validation approach into test planning documents, such as the TEMP. The guidance also notes that a comprehensive strategy should include assessment of M&S output across the entire operational domain for which the M&S will be accredited, and that statistical analyses should be used to perform sensitivity analyses for SME review and assessment for consistency with reality.

DOT&E commissioned IDA to produce a handbook,[49] technical report,[50] and short course[51] on M&S validation. These products are available to the broader T&E community and are resources for continued use of M&S for T&E.

3. Summary

Substantial progress has been made in the use and integration of M&S in test and evaluation. This change can be attributed to the technological, computing, and capability enhancements of M&S, as well as changes in policy and emphasis by DoD leadership. The role of M&S in the T&E enterprise has grown as a result of the increased complexity of the battlefield, the increased demand for rapid acquisition, and the desire by program offices to know sooner whether systems under development will achieve the required performance and capabilities. Programs across the Services are requesting and building LVC simulations that incorporate their system into warfighting environments to better understand and work out integration issues, while generating the virtual operator–system interface to evaluate that aspect early in development. Simulations are now commonplace when dealing with highly technical and complex systems, which form the majority of systems today. Unfortunately, not all M&S applications are success stories. In some cases, contractual agreements prevent the government from obtaining their own copies of the digital simulation. In others, the cost of building and maintaining the simulation is greater than that of similar levels of live testing. But the successes outnumber the failures, and many programs have found great value by using M&S to better understand their system.

[49] Avery, et al. (2019)

[50] Avery, K., et al. (2017)

[51] Simpson, et al. (2019)

E. DA 5 - Lack of visibility of testing costs

1. Notes from the MORS/ITEA symposium

Participants in the 1994 symposium discussed cost from many perspectives, including the visibility of testing costs, ways to reduce the costs of testing, and how to best consider the cost of testing as part of the acquisition process. Approaches for minimizing costs, such as using M&S, combining DT and OT data, and using statistical techniques such as design of experiments, overlapped substantially with other discussion areas. One of the outcomes from the symposium was a project undertaken by DOT&E to initiate a study to establish a database of test and evaluation costs.

The symposium discussants addressed the challenges of using cost/benefit analysis and balancing decision risk against testing costs. Although they provided general outlines for what would constitute a "cost" and a "benefit," they were short on the practicalities of how to perform such a study with any degree of accuracy.

2. Progress since 1994

Cost remains a major concern for the T&E community. Regarding the question of "How much testing is enough?", test costs are a critical consideration. Although participants in the 1994 symposium noted that cost transparency was challenging, the proceedings treated it as a challenge that could be solved relatively simply. The working group, when discussing cost/benefit analysis, mentioned the difficulties of defining what the costs were and how to assign them to specific programs or test events, but did not anticipate how much of a challenge this would pose to the efforts that followed the symposium. The T&E cost database envisioned by DOT&E was never completed, and subsequent efforts to establish the "cost of testing" have not resulted in a consensus. One consistent challenge with estimating the cost of testing is the lack of standardized terminology and accounting practices across the Services and OTAs. Another is that although documents such as Selected Action Reports and TEMPs typically include estimated testing costs, actual test costs are tracked inconsistently.

a. Quantifying the cost of testing

A 2013 IDA study[52] commissioned by DOT&E attempted to discern the cost of operational test and evaluation through case studies. The objectives of this study were to:

1. Develop a taxonomy of OT&E resource and cost elements
2. Collect resource and cost data on different commodity groups

[52] Dominy, et al. (2013)

3. Research and document Service rules on financing, budgeting, and accounting for OT&E costs
4. Quantify OT&E costs relative to other program costs (acquisition, production, etc.).

This study found large variations in OT&E cost estimates among the Services, suggesting that finding objective ways to define the cost of testing is difficult. The study further found large variations in reported test costs *within individual programs* in various documents such as the TEMP, Test Resource Plan, and Test and Evaluation Exhibit, reporting considerably different cost estimates. This sample paragraph discusses the Miniature Air-Launched Decoy (MALD) program:

> "For MALD/MALD-J, OT&E cost was $10 million based on the TEMP, $36.2 million based on the Program Office estimate, $49.52 million based on the Program Office estimate when direct support cost was included, and $54.8 million based on the IDA taxonomy (which includes direct support cost). The relative [OT&E] cost for MALD/MALD-J, compared to system acquisition cost, ranged from 0.6 percent based on the TEMP to 3.0 percent based on the IDA taxonomy."

b. Challenges in evaluating the benefits of T&E

On the other side of the cost/benefit equation, there have been many attempts to identify the benefits of T&E, but few of these link monetary costs to identified benefits. At the request of DOT&E, IDA collected case studies illustrating direct value from operational test and evaluation.[53] The value identified in these case studies is typically qualitative (e.g., identification of issues with system effectiveness, evaluation of system performance in operationally realistic environments not otherwise captured by specification testing, etc.) rather than monetary. These examples demonstrate that quantifying the cost side of the equation is not the only challenge would-be cost/benefit analysts must confront.

c. Benefits of M&S

While the use of M&S for T&E has obvious benefits, it is not prima facie obvious that using M&S in lieu of live tests is cheaper. Simulations are generally orders of magnitude cheaper to run than a live test event, are somewhat repeatable, and allow for hundreds to hundreds of thousands of runs to be performed over modest time horizons. M&S is also beneficial in its ability to extend test conditions into areas that cannot be reasonably repeated in developmental or operational environments. However, models must be validated and accredited as adequate for the chosen application. The requirement

[53] DOT&E (2016, 3), DOT&E (2011, 1), DOT&E (2014, 3), DOT&E (2011, 2), Avery M., et al. (2017)

identified in the MORS Verification, Validation, and Accreditation (VV&A) Workshop[54] was that "experimental data" be used as the preferred validation source. For example, when radar cross-section data are required in a model, the source should be physical radar measurements. Also, collecting adequate data to validate a model so that it might be accredited for use in T&E could end up costing as much or more than using live runs for T&E. The cost challenges can be compounded if the government does not own the rights to the model once it has been developed. DA 3 includes a more complete discussion of both the costs and benefits of using M&S for T&E.

3. Summary

Since 1994, little has changed when it comes to evaluating the cost of testing. Attempts to quantify the costs of T&E have failed to overcome the challenges posed by the varying accounting practices used across different Services and programs. Existing examples show the utility of test and evaluation but do not attempt to translate those benefits into monetary values. The lack of progress in this area underlines the difficulty of this challenge as it was described 25 years ago.

[54] Williams, et al. (2002)

F. DA 6 – Statistical application in the T&E discipline

1. Notes from the MORS/ITEA symposium

Perhaps because the symposium was co-sponsored by an operations research society with an optimization and statistical methods mindset, more than half of the pages in the proceedings used some variant of the word "statistical." Although this is only one of the nine discussion areas, operations research and statistics were at the heart of all three of the 1994 symposium recommendations. Specific comments regarding the role of statistics in T&E stand out:

- "Statistical techniques are tools to increase the efficiency of T&E."
- "There is a compelling need to upgrade and maintain the level of statistical interest, skills, sophistication, and appreciation in T&E."
- "T&E team members should be trained in tools like design of experiments (DOE)."
- "[The] use of regression models and especially response surface methodology seems natural in the operational testing environment."

Beyond making these strong statements, symposium participants also wanted to encourage the use of graphical methods over tables of data, use statistical methods to analyze early test data to inform OT planning and analysis, and couple data management with statistical analysis to more efficiently and effectively process data. However, it was clearly emphasized in the symposium that the question of "How much testing is enough?" is best answered by addressing most or all of the nine discussion areas. Statistical methods are most useful in efficiently sizing specific test events with known objectives and limitations. In the macro sense, though, these tools – combined with DT/OT collaboration and sharing, early test data, M&S, cost/benefit analyses, risk analysis and more flexibility – can answer the question throughout the full life of the program.

2. Progress since 1994

Statistical methods, including graphical exploratory analyses, descriptive statistics, and more formal statistical modeling, have been used to good effect in hundreds of programs over the past decade. Benefits include improved understanding of system performance and more efficient testing. Experimental test design methods are routinely applied to identify and systematically vary relevant factors to adequately cover the operational test space. Resulting data are analyzed to identify the key factor effects and interactions that influence the measures of performance and measures of effectiveness tied directly to the test objectives. Leadership from DOT&E and the Service OTAs has made sure that gains from using statistical methods will continue to accrue across DoD in the coming years.

a. DOE in OT since 2009

The application of statistical methods and designed experiments practiced today throughout the OT community is due in large part to the initiatives and memos authored by DOT&E starting in 2009. DOT&E commissioned a report in 2010 on the use of DOE for planning operational test events.[55] The report found that a few programs used DOE, but not many. Leadership did not require the inclusion of uncertainty estimates in reports, and the use of tools such as operating characteristic curves to size tests for evaluating reliability was inconsistent. Many times, the "factor of three" rule of thumb[56] was used to size reliability testing rather than criteria based on producer and consumer risk, or uncertainty about the reliability estimates.[57] The report also found that although DOT&E provided guidance on how to design a test, this guidance offered little detail and few explicit steps.

DOT&E issued guidance memos throughout the early 2010s encouraging the use of DOE in operational test design. These memos identified multiple statistical measures that provided stakeholders with objective criteria to consider when comparing test designs. These measures help articulate an analytical trade-space in which to evaluate different tests, which helped shift the focus from whether one more run was needed to the amount of information that would be gained from an additional run, and how that information would help decision makers.[58]

b. Cost savings from DOE

There are many examples of cost savings from using DOE, but one of particular interest is the F-35 AMRAAM integration vibration testing. The government required the contractor to employ statistically based test design and analysis techniques. Government experts assisted in developing an efficient split-plot design with high statistical power for modeling vibration intensity across all combinations of F-35 variants and aircraft missile store stations. Combining this test design with novel methods for analyzing vibration profiles enabled the contractor team to identify potential causes for faults. The contractor chief engineer estimated that this approach saved $15 million.[59]

[55] Thomas, et al. (2010)

[56] "The test length should be at least three times the system's mean time to failure requirement."

[57] The origin of this rule of thumb is actually based in producer and consumer risk, albeit very high ones. However, these risks are not fully understood or consciously chosen by most of the testers who use the rule of thumb.

[58] DOT&E (2010, 2); DOT&E (2013, 4); DOT&E (2013, 5)

[59] Hutto, et al. (2018)

c. Acquiring skills

The Naval Surface Warfare Center created an initiative in 2018 called the T&E Renaissance, with forums in 2018 and 2019, stressing the increased application of test science and statistically based methods in T&E. Also, as a major part of this initiative, there is a strong push toward accelerating the hiring of the best and brightest new STEM talent into their T&E workforce.[60]

d. Skills and training in statistics and DOE

DoD took the 1994 call to "upgrade and maintain the level of statistical interest, skills, sophistication, and appreciation in T&E" seriously and devoted substantial time and resources to training, coaching, and mentoring the T&E workforce in the use of statistical methods, DOE,[61] and reliability. Since 2007, DoD has made short courses and trainings in statistics available to the T&E workforce. Conservative estimates show more than 500 courses offered, with over 7,300 students participating. Members of the armed forces of our international allies, including the Great Britain Air Force, Australian Air Force, Taiwanese Air Force, and Brazilian Army, have also taken advantage of these opportunities. The most in-demand courses cover DOE, reliability growth, reliability estimation, and statistical modeling, but more advanced courses, including "Statistical Methods for M&S Validation," "Predictive Analytics," "Split-Plot Design and Analysis," and "Repeated Measures Design and Analysis," are also included. In part due to this additional training, the part of the T&E workforce with statistical skills has increased over the past 10 years.[62]

Table 1. DoD Statistics and Design of Experiments Training Courses (2007-2019)

Course Type	Courses	Students
2 day	188	2703
1 Week	315	4610
Total	503	7313

The Air Force Institute of Technology (AFIT) provides graduate education in statistics for future T&E practitioners. Since 2009, AFIT has administered a 15-month, 5-course T&E graduate certificate program annually to 20-25 students. Students are full-time military and government civilian employees taking one course per quarter via a distance learning program. The curriculum for the program includes probability and

[60] Lawrence (2019)

[61] Freeman, et al. (2013)

[62] Thomas (2017)

statistics, linear regression, operational experimentation, reliability engineering, and a capstone course focusing on time series analysis and response surface methods.

e. DATAWorks

DOT&E, in collaboration with NASA, co-sponsors an annual 2-day workshop called DATAWorks (Defense and Aerospace Test and Analysis Workshop). This popular workshop draws participants from across DoD and other government agencies – from practitioners to top leadership – all of whom come together to share the latest in rigorous T&E methods and applications. DATAWorks includes day-long courses and mini-tutorials in topics related to T&E, including statistics and experimental design.[63]

f. Methods adapted to DoD T&E

DoD engineers and operations analysts have often taken existing statistical methods and adapted or enhanced them as needed. When existing methods do not exist, they create new methods tailored to the specific problems (referred to as statistical engineering).[64] Analysts share these new tools across similar programs, or programs with comparable methodological challenges. Progress in collaboration, communication, and networking needs to continue. Examples of methods[65] adapted include:

1. Lognormal and binomial parametric survival analysis for known skewed, binary, or censored data, including range to detect or time to detect a threat
2. Use of a three-phase sequential design for sensitivity tests involving binary response variables (e.g., determine bullet exit velocity penetrating armor 50 percent of the time, or V_{50})
3. Creating interactive apps that use Monte Carlo and parametric methods to estimate statistical power for non-normal (binary or logistic, exponential, and Poisson) performance statistics situations
4. Parametric survival modeling for decile regression, especially for tail percentiles (e.g., 90 percent) of performance measures
5. General fixed geometric blocks, crossed with classical or optimal modal blocks, for sensor designs
6. Factor-covering arrays and space-filling designs for deterministic responses, such as software defect detection and location

[63] DATAWorks Archive (https://testscience.org/archive/)

[64] Anderson-Cook and Lu (2012), Hoerl and Snee (2010), Hare (2019)

[65] Freeman (2018)

7. Routine use of optimal constrained design regions and/or disallowed combinations
8. Characterization of existing agile test cases, then augmentation of them to develop an agile test design strategy
9. Optimal augmentation of historical designs with new variables added
10. Creative combining of multiple test designs, each with different test objectives and factors, into a single test plan matrix to be executed simultaneously
11. Augmenting of existing, historical, or proposed space-filling designs with D-optimal designs for polynomial modeling of high-dimensional M&S designs (i.e., 20-60 variables)
12. Sample size tool and statistical power analysis for planning a test, regardless of the underlying probability distribution of the performance measures
13. Application of generalized linear models to test data with non-normal distributions
14. Planning tools for cyber T&E using adaptive covering arrays and designed experiments
15. Split-plot and split-split plot designs and analyses
16. Decision analysis-based designs
17. Linear programming-optimal designs
18. Simulation-based designs using a hybrid of maximum projection space-filling designs augmented with 20 percent I-optimal points to improve statistical modeling properties.

Citations are provided for a small sampling of peer-reviewed published journal articles describing and demonstrating DoD-centric design of experiments and statistical modeling of new methods or unique applications.[66] Additional references are given for a series of National Research Council (NRC) reports that address various statistical challenges in DoD T&E. These reports, spanning from 1994 to 2015, were sponsored by DOT&E and AT&L.[67]

[66] Roth, et al. (2010); Tucker, et al. (2010); Johnson, et al. (2012); Gilmore (2013); Kass (2015); Landman, et al. (2007); Zessin, et al. (2017); Simpson and Wisnowski (2001)

[67] National Research Council (2002); NRC (2004); NRC (2012, 1); NRC (2012, 2); NRC (2014); NRC (2015); Nair and Cohen (2006); Rolf and Steffey (1998); Cohen, et al. (1998)

g. Statistical Methods Center of Excellence

In 2012, the Deputy Assistant Secretary of Defense for Developmental Test and Evaluation (DASD(DT&E)) formed the Scientific Test and Analysis Techniques (STAT) Center of Excellence (COE) to provide independent Ph.D.-level statistics skills directly to DoD major acquisition programs in order to increase T&E efficiency and effectiveness. STAT COE experts work directly with program managers in supporting their efforts by planning, designing, and devising rigorous T&E strategies and test events. The STAT COE also develops case studies and captures STAT Best Practices for wider dissemination across the acquisition community. In addition, the STAT COE supported smaller special programs advancing the practice of STAT in DT&E across the DoD T&E infrastructure. To date, the STAT COE has engaged with 71 programs and projects, including 55 Acquisition Category I major defense acquisition programs and major defense information systems, and has saved DoD more than $197 million in unrealized or avoided costs.[68] This is in addition to delivering more efficient test designs, increasing knowledge, and providing direct technical T&E support.

This support was directly called out in the DOT&E 2016 and 2017 Annual Reports.[69] In 2016, DOT&E noted that the STAT COE provides program managers with the scientific and statistical expertise to plan efficient tests, ensuring that programs obtain valuable information from each test event. The report also states that programs that engage with the STAT COE early on have better structured test programs for generating valuable information, and that the STAT COE provides direct access to experts in test science methods that otherwise would be unavailable. DOT&E also noted that smaller programs with limited budgets do not have access to strong statistical help in their test programs and cannot afford to hire a full-time PhD-level statistician to aid their developmental test program. Having access to these capabilities in the STAT COE on an as-needed basis is one means to enable these programs to plan and execute more statistically robust developmental tests. The 2017 report highlighted STAT COE expertise in software and cybersecurity testing.

h. New Navy policy

OPNAV has drafted a proposed Secretary of the Navy Instruction for Department of the Navy (DON) Acquisition T&E Policy and Guidance that requires programs to integrate STAT into DT/OT processes to support efficient and effective test planning and execution.[70] This instruction stipulates the following:

[68] STAT COE (2019)

[69] DOT&E (2017, 2); DO&E (2018, 3)

[70] Secretary of the Navy (2019)

- STAT should be used to support TEMP descriptions of efficient and rigorous DT/OT/LFT&E planning and cost estimates.
- T&E leads shall coordinate with DON T&E/N94 to secure the support of the STAT Center of Excellence SMEs to support test planning.
- STAT shall be used to support informed program decisions and at each TEMP update.

3. Summary

Tremendous progress has been made in the application of statistical methods for planning, designing, and analyzing the results from T&E since 1994, particularly for OT&E. The cause of success in integrating statistical methods is multifaceted, but involved the necessary aspects of being anchored by policy, prioritized by leadership, proven by projects, and sustained by practitioner training/coaching.

Notable contributions have come from both small (unit level) and large (center and agency level) organizations where statistical test design and statistical analyses are standard default practice. Specific contributors to the positive momentum have been the deep efforts to communicate the benefits to leadership; the initiatives taken by leaders, managers, and practitioners; the education process; and the sharing of expertise and lessons learned. More advancements, better adaptation of methods to solve real problems, and a more routine and universal application throughout the spectrum of DoD T&E would continue to increase system knowledge and contribute to efficiency and effectiveness.

G. DA 7 - Pooling/sharing data across DT, OT, and M&S

1. Notes from the MORS/ITEA symposium

The major theme of this discussion area was to encourage the gathering, pooling, and sharing of test-related information: (1) data from prior phases of the current program's test (e.g., DT, OT, and M&S data), and (2) related test data from older systems. Despite agreement that pooling could "shorten the acquisition process, reduce test time, and get technology into the field faster," the conditions under which pooling is appropriate were debated. Some thought it more important to facilitate the sharing of data between DT and OT before considering how to pool it. By planning to pool data early on in the acquisition process, testing can be accomplished more efficiently.

2. Progress since 1994

This discussion area has much in common with DA 2, which addressed early coordination between the DT and OT communities. Many of the comments on DA 2 also apply here to DA 7. The discussions below focus exclusively on pooling and sharing data, and attempt to repeat as little as possible from DA 2.

The majority of the efforts to pool or share data since 1994 have been grassroots initiatives within individual programs, or by individual OT or DT organizations. For the recognized challenge of how best to combine M&S results with test data, usually for M&S validation, approaches vary greatly. Some programs simply execute the simulation and ask for SME opinion, while others use more data- and analysis-driven approaches.

a. Prerequisites for pooling data

Testers must take care when pooling data. Before combining data, analysts should consider how the different venues, system configurations, system operators, and operational conditions could affect the data. In 2004, a report from the National Resource Council commissioned by the Army Test and Evaluation Center discussed methods for combining data for OT&E, noting that "both formal and informal methods [for combining data] require the judicious selection and confirmation of underlying assumptions as well as a careful and open process by which various types of information, some of which involve subjective judgment, are gathered and combined."[71]

Prior to pooling data across test events, phases, or M&S, analysts should make explicit what assumptions are required and evaluate to what extent those assumptions hold for their data. Formal approaches (including statistical techniques, such as Bayesian methods) provide a framework for this step.

[71] National Research Council (2004)

b. Sharing simulation data

Although DT is now generally willing to share data with OT, government DT personnel may not have the authority to share data from proprietary or contractor simulations. In many cases, the government does not own the simulation, and may be paying for the data "by the run." To ensure that data are sharable across organizations and that the new data can be generated more easily, the government should ensure that it can access the contractor's integrated simulation, or own the simulation outright. See Section 2.B.2.c and 2.D.2.a for further discussion.

c. Sharing between DT and OT

The DT and OT teams for the AIM-9X missile shared data and experienced some success.[72] DT data from simulation runs and captive-carry flight tests were made available to the OT team. Because the OT community had access to these data, they were able to characterize the driving variables for system performance, yielding a more efficient and informative test design for the live shots. The limiting factor was that collaboration on flight test design was missing from the contract. The hardware contractor had little incentive to collaborate and plan with government testers. As a result, direct cooperation between individuals on the contractor, DT, and OT sides, rather than a broader policy umbrella or planning dating back to the contracting period, were the drivers of success.

d. Pooling of data for reliability estimation

DoD has provided substantial guidance in the area of T&E for reliability. Pooling data is essential for using techniques such as reliability growth planning. Properly executed reliability growth planning requires early integration and evaluation of DT data. These DT data feed the Reliability Growth Tracking curve, which, when compared to the Reliability Growth Planning curve, can alert a program early if it is not on track to meet its reliability growth goal. The Army Materiel Systems Analysis Activity (AMSAA) has published templates for programs to build curves for the growth, tracking, and projection phases of their reliability programs. In addition, a number of DoD reliability guidance documents have been published since 2002:

- *Reliability Issues for DoD Systems: Report of a Workshop*, National Academies Press, 2002
- *DoD Guide for Achieving Reliability, Availability, and Maintainability*, DOT&E and AT&L, 2005

[72] 53rd Wing USAF (2009)

- *Defense Science Board Report on Developmental Test & Evaluation*, Defense Science Board, 2008
- *DoD Reliability, Availability, Maintainability, and Cost Rationale Report Manual*, OSD, 2009
- *Next Steps to Improve System Reliability*, DOT&E Memo, 2009
- *Handbook: Reliability Growth Management*. MIL-HDBK-189C, 2011
- *Army Materiel Systems Analysis Activity Design for Reliability Handbook*, AMSAA, 2011
- *DTM 11-003 – Reliability Analysis, Planning, Tracking, and Reporting*, OSD AT&L, 2011
- *Reliability Growth: Enhancing Defense System Reliability*, National Research Council, 2015
- *DoD Reliability and Maintainability Engineering Management Guide*, DASD SE, 2016
- *DOT&E TEMP Guidebook 3.1*, DOT&E, 2017

e. Statistical methods for combining data

Combining data from multiple sources has proven to be valuable in T&E. Data from legacy systems can help provide a baseline understanding of a new (but similar) system, or provide a basis for comparison. Similarly, when systems receive updates, data from the older version can be used to inform assessments of the new version. Less formal exploratory and graphical techniques are encouraged for initial analyses, while more formal statistical modeling and Bayesian methods[73] can be used for quantitative assessments. For example, data collected in DT on a less mature version of the system can often provide substantial value for OT assessments.[74]

f. Case studies illustrating the utility of pooling data

Case studies have documented how these approaches can provide value for a variety of programs, particularly in the area of reliability.

The Stryker Family of Vehicles includes 10 separate systems. Traditional approaches would treat each system separately, resulting in large uncertainty bounds. Some precision

[73] Bayesian methods are standard statistical techniques used in a variety of applications. They are well-suited for combining data from different sources, particularly when the analyst believes the sources differ in quality or credibility. For more information, see Freeman, et al. (2015).

[74] Freeman, et al. (2015)

can be gained by carefully pooling DT and OT data to produce tighter confidence intervals. Pooling data across the different systems produces even more precise estimates while allowing evaluators to estimate performance for variants that experienced no failures during testing.[75]

The Joint Light Tactical Vehicle (JLTV) program included multiple system variants. Some vehicles had four seats while others had only two, allowing more room for storage. The variants each came with different mission packages. As a result, while all JLTVs had many components and subsystems in common, the different variants and different mission packages meant there were also substantial differences. IDA demonstrated that pooling data across all systems could result in more precise estimates of system reliability while still accounting for differences across system variants and configurations.[76]

Pooling data across multiple test events can sometimes be the only way to estimate overall system performance because of test limitations. End-to-end tests of the Ballistic Missile Defense System is challenging because of the complexity of the systems involved and overriding safety concerns. As a result, analysts are typically faced with a smattering of partial tests, results from simulation, and a small number of end-to-end live tests. By combining data from each of these sources, analysts gain a better understanding of the full picture than any single source alone could provide.[77]

g. Understanding how the data were collected

One challenge when combining data sets is knowing how older or earlier data sets were generated. Without understanding the testing and data collection processes, analysts will struggle to understand the appropriate amount of weight to give data collected in DT. Reliability data are often available from DT, but if those data were not scored using the same criteria (e.g., a Failure Definition Scoring Criteria (FDSC)), then operational testers will not know how the frequency of failures observed in DT will translate. A stronger focus on thorough and complete metadata that includes information on data pedigree will help facilitate the pooling of data and make testing more efficient. The decision about which data are relevant to an assessment is always going to require expert knowledge and documentation from an analyst.

The JLTV's vendor down-selection process during early OT illustrates the importance of understanding how well results from DT will translate to OT. Three vendors competed for the JLTV contract throughout multiple DT phases, culminating in an OT event, which informed the down-select decision. Only one JLTV vendor produced vehicles

[75] Dickinson, et al. (2013), Dickinson, et al. (2015)

[76] Fronczyk, et al. (2015)

[77] Avery, M., et al. (2017)

that met the desired reliability growth goal at the time of OT. This vendor had applied the FDSC that would be used in the OT event when self-assessing performance in DT, allowing that vendor to identify the most critical failure modes to fix prior to OT.[78]

h. M&S sharing with test

Early DT data can also be used in the development of models. The model can then be used to characterize system performance across a wider set of conditions than might be possible during DT. This can in turn identify important areas where live DT testing may reveal useful results for system developers, which can then be fed back into the M&S, creating a virtuous cycle. Pooling or combining DT data with M&S development and M&S runs for system characterization can best inform M&S validation, along with the comparison of live test results vs. M&S post-test reconstruction runs.[79]

i. VV&A and test

Data from M&S should be subjected to Verification, Validation, and Accreditation (VV&A) before being used in T&E. The VV&A process defined in the 2002 MORS VV&A workshop and approved by the Services and OSD was intended to ensure that M&S used to support T&E was "adequate for the intended application or uses." Models developed under this process can be evaluated to the degree necessary to ensure that they represent the real world. This validation was intended to be based on the comparison of model results to experimental data.[80]

j. Other ways to pool data

While the symposium discussion focused on sharing data across DT, OT, and M&S, it is important to note that there are other reasons for pooling data in T&E. Fleet data, such as flight hours, repair rates, and unit/system availability, can be more useful than data from follow-on testing for evaluating how well systems are performing or for evaluating the success of deployed upgrades or system changes. Programs designed to upgrade existing systems or new-start programs replacing legacy systems can use data from the field, T&E, and M&S to help set requirements. All of these sources can be valuable inputs to the test design process.

[78] Freeman, et al. (2016)

[79] Reese, et al. (2004)

[80] Williams and Sikora (2002)

3. Summary

While there are many successful examples of pooling and sharing DT, OT, and M&S data, these efforts were not united, nor were they based on a new or better process. Unless the new way of doing business is well communicated, leader driven, and practiced as a default, only pockets of excellence will exist. Unless proper planning occurs, the chances of using data for anything but the sole purpose intended by the organization that originally created the data are slim. Policies proposed to resolve this problem date back to 2000, but many in the T&E community may not be aware of these efforts.

One area in which some gains have taken place is reliability/reliability growth, where data are most scarce and pooling can be extremely beneficial. Some programs facilitate the use of DT data for OT by maintaining consistent data collection procedures throughout DT and using FDSCs, but others do not. While examples exist of programs that build their M&S capabilities alongside their live testing and use a rigorous approach to verification and validation, this is not the norm. A common challenge is matching the conditions depicted in the M&S to live conditions under which data were collected. To make the best use of both DT data and M&S capabilities, careful planning is required to ensure that live data support the M&S effort, and vice versa.

H. DA 8 - Characterizing different types of risk

1. Notes from the MORS/ITEA symposium

Symposium participants suggested several approaches for using test and evaluation to better understand risk. Some identified specific areas where better quantification of risk would be valuable: (a) program failure, (b) technology obsolescence, and (c) life and limb. But it was not clear to the presenters how T&E should characterize each of these risks. Quantifying actual risk levels and defining acceptable levels of risk is difficult for program managers, which makes it challenging to plan tests around levels of risk. Risk reduction and mitigation are a separate challenges from risk characterization, though the symposium participants did spend some time discussing both. One symposium attendee suggested that one way to account for risk is to build time into program schedules for surprises. This suggestion is consistent with some of the discussion in DA 3 regarding more realistic schedule planning in TEMPs. In contrast with the three types of risk mentioned above, it was pointed out that statistical risk is well defined, but alone is inadequate for making program decisions.

2. Progress since 1994

The symposium report offered little specific guidance on ways to improve how the T&E community characterizes risk, though they emphasized that a better understanding of risk would be helpful when deciding how much testing was necessary. Since the 1994 symposium, there has not been much progress in formalizing the ways that T&E organizations and specific programs account for different types of risk. There is not even a consensus that the three risk categories proffered in 1994 are the right types of risk to focus on in T&E.

a. Policy on quantifying risk

The initial policies for OT&E of information assurance[81] and for OT&E of software-intensive systems[82] are examples of policies written in the same time frame as the 1994 symposium. They illustrate the desire at the time across DoD to use risk when determining test requirements. These policies prescribe a four-step process in which risk is assessed after each step. The guidelines for OT&E of software-intensive systems recommended that the OTA prepare a risk assessment as the first step in the process.

[81] DOT&E (1999)

[82] DOT&E (1996)

More recently, the National Institute for Standards and Technology published and subsequently updated a Risk Management Framework for identifying, implementing, assessing, and managing cybersecurity capabilities and services for federal systems.[83]

b. Current methods for managing risk

The three types of risk called out at the symposium are not comprehensive, and different types of risk are constant concerns for various stakeholders in the T&E and acquisition communities. Managing schedule, cost, and performance risk is an integral part of the job for program managers. Following are some examples:

- The Navy and Marine Corps regularly assess the risk, after DT is completed, that the system under test will satisfactorily complete OT.
- Using Real Time Casualty Assessment systems in OT allows testers to achieve realistic force-on-force operational behaviors without putting soldiers at risk by using real bullets.
- DoD has thus far not been willing to conduct a cyber-attack on an aircraft in flight, deeming the risk of a mishap larger than the risk of not having done the test; in only a few cases have cyber-attacks on a moving ground combat vehicle been conducted.
- Open-air and HWIL simulators allow us to test aircraft survivability equipment without shooting live missiles at manned aircraft, and diligent adherence to policies and procedures allows testers to shoot weapons at targets on land and sea ranges while minimizing risk of damage to people or property.
- M&S allows programs to explore the flight envelope of missile systems and plan live test events that minimize the risk of failures.

This list, which is not exhaustive, highlights a few of the many ways programs manage risk.

c. Types of statistical risk

Unlike many of the risks discussed at the symposium and in Section 2.H.2.a above, statistical risks are well defined and quantifiable, and these are often the types of risk that test teams focus on. Type I and Type II risks are commonly discussed in TEMPs and Test Plans. Type I is the risk of declaring that a factor (e.g., system configuration, altitude, airspeed) is influential when it is not, and Type II is the risk of failing to identify important factors that drive system performance.[84] Test teams also care about Type III risks, which

[83] NIST (2010), NIST (2018)

[84] Simpson, et al. (2013), Montgomery, et al. (2010)

are the risks associated with failing to properly define the battlespace completely and correctly, resulting in omission of important factors from the test design.[85] Mitigating Type III risks by selecting the more extreme test point must be balanced against potential increases in risk to the safety of the operators. Examples where these types of risk must be balanced include F-15E Suite 7 and C-130 Dragon Spear.[86]

d. Risks for reliability

Building test programs around risk remains challenging. However, current methodologies and typically available data make it possible to examine a program's risk of failing to achieve a prescribed reliability requirement. Tools such as reliability growth curves and assessments of reliability growth potential are designed to help analysts determine what is achievable in a given timeline and with a given set of resources. Risk characterization can help determine whether a reliability test program can continue as designed or whether it should be interrupted and reset after a pause to determine and implement major redesigns. Previous risk and reliability growth projection methodology was simply forecasting using the parameter estimates in the model curve; the methodology in current reliability growth projection methodologies, however, is more detailed.[87]

Although this more detailed guidance has the potential to help programs, DoD has yet to see an increase in the proportion of programs achieving their reliability requirements.[88]

3. Summary

T&E decision makers should always consider risk, and there are many types of risks to consider. Unfortunately, quantifying many of these risks has proven to be challenging. Better-defined types of risk are commonly considered and used to help determine the scope and volume of testing. Other types of risk are considered on a qualitative basis by programmatic decision makers. Tools for quantifying risk for reliability are available, although programs do not always make full use of them. The principal challenge remains in identifying the types of risk relevant to T&E and finding ways to quantify those risks for decision makers.

[85] Kimball (1957)

[86] Hutto (2013)

[87] DoD MIL-HDBK-189C, 2011

[88] Avery, et al. (Forthcoming)

I. DA 9 - Greater statutory and regulatory flexibility to address "the unexpected"

1. Notes from the MORS/ITEA symposium

Some of the discussion centered on increasing the number of options available to programs and organizations to address the unexpected occurrences that programs are frequently faced with. One symposium participant suggested that existing regulations should be reconsidered in light of the changing acquisition environment. Existing policies and regulations were reported to be inconsistent and often misinterpreted; increasing the flexibility on dollars spent on testing, rather than procurement, presumably would allow programs to better adjust to the unexpected.

2. Progress since 1994

Allowing for greater regulatory flexibility can clearly have both good and bad effects. More flexibility can allow programs to better adjust based on their circumstances, but removing regulatory structure can result in programs repeating the errors of the past.

a. Middle Tier Acquisition

Acquisition policy continues to change rapidly, and the current environment is in some ways much more challenging than it was 25 years ago. Since the 1994 symposium, DoD has tried many approaches for rapidly developing and deploying effective systems. (See DA 1 for more discussion.) In the past few years, multiple new pathways have been created to smooth the way for fielding new technologies and providing acquisition executives with greater flexibility.[89] Middle Tier Acquisition pathways provide programs with options on rapid prototyping and rapid fielding. Although these are not new ideas to the DoD acquisition community, formalized pathways defined explicitly in the statute are a step in the right direction. Supporting agencies such as DOT&E have published policies defining how MTA programs will work within existing regulatory and oversight structures, helping program offices understand their T&E obligations. Reducing ambiguity concerning these new acquisition pathways helps to mitigate the confusion that can accompany new regulatory flexibility.[90] The effects that these new policies will have on the acquisition system are yet to be seen.

3. Summary

Program managers always want increased flexibility. Recent changes such as the introduction of Middle Tier Acquisition pathways are intended to provided new avenues

[89] FY16 NDAA

[90] DOT&E (2019, 1)

to quickly take systems from development to the field. However, it is not clear whether these pathways are new options within the existing structure or a fundamental change to DoD acquisition.

Many of the discussants from 1994 sought increased flexibility in the way funds were spent across years and programs (e.g., more flexibility to move money between "testing" and "procurement"). Policies since 1994 designed to increase flexibility, however, tend to focus more on reducing the number of traditional acquisition events (Milestones, Touchpoints, etc.) required prior to fielding systems. As a result, these efforts may not be addressing the concerns of the symposium participants.

3. Modern Challenges

1. Notes from the MORS/ITEA symposium

In hindsight, the technology challenges of 25 years ago appear less daunting than the modern challenges, but symposium participants showed foresight in some critical areas. M&S was discussed thoroughly, with a particular focus on how best to use the newer capabilities (such as distributed simulation) and how best to represent simulated environments visually. A more general comment regarding technology challenges was apropos to current discussions within DoD: "T&E must quickly adapt to the changing environment in DoD. New demands and challenges facing the T&E community must be addressed. Changes in the acquisition process will probably require quicker response and greater flexibility by the T&E community. Also, the T&E community should take advantage of the new acquisition policies and new technologies to do its job more effectively." This remains as true today as it was at the time.

2. New challenges since 1994

Over the past 25 years, the T&E community and the DoD acquisition system community as a whole have seen systems become significantly more complex while still being expected to perform in highly contested environments. The result has been a greater reliance on M&S to supplement live testing. Modern systems are almost universally software intensive and software centric, which increases the need to develop cyber tactics and ensure adequate cybersecurity. Most recently, system designs are taking advantage of artificial intelligence and machine learning to operate in a more automated fashion and even autonomously. This has created new challenges, some of which the T&E community is still working to overcome.

a. Software-intensive systems and automated software T&E

Software testing in the commercial sector has developed apace since 1994. Combinatorial tests help maximize coverage of deterministic systems while minimizing the required number of test points. [91] In the unique circumstances where combinatorial tests can be applied to DoD programs, they have the potential to provide a great cost and time savings over traditional approaches.[92]

[91] Kuhn et al. (2010); Higdon (2017); Dahmann (1997)

[92] Freeman, et al. (2017)

The importance of agile/development operations software development in industry as a standard procedure has permeated DoD contractors' practices, as have software-intensive systems (e.g., C4ISR programs) that are largely internal government development programs. With sprints and scrums, T&E practices have had to adapt to more frequent testing, often using automated software test tools that allow for more extensive coverage of the software while running overnight and on weekends.[93]

The use of software test automation and an integrated, systems-of-systems testing approach is also becoming more prominent, although tremendous challenges still remain in motivating, enabling, and resourcing multiple programs and systems to perform these family-of-systems tests. Although early efforts have met with mixed success, the long lists of lessons learned generated by experience should facilitate better processes in future system-of-systems tests. Agile development of increasingly software-intensive systems places an even greater premium on early testing with user involvement. This is reinforced by the current desire for fielding systems faster; the more that we can do in early testing, the sooner the development is likely to succeed.[94]

b. Cybersecurity testing

Cybersecurity test and evaluation is now required for many software-centric systems,[95] but unfortunately the methods and what constitutes adequate testing remain works in progress. The current cybersecurity T&E process consists of sequenced phases of test based on guidance from DoD and AT&L. This approach was built up over time and is designed to help programs manage cybersecurity risk intelligently. Dedicated cybersecurity testing organizations (e.g., 346 Test Squadron, 96th Cyberspace Test Group, 16th Air Force (Air Forces Cyber), and U.S. Cyber Command) now exist across DoD and the Services, testifying to the significant progress made in this area.[96]

With regard to the question of "How much testing is enough?", answers remain unclear. Current testing calls for Cooperative Vulnerability and Penetration Assessments and Adversarial Assessments, but does not describe quantitative approaches for determining test duration. Personnel qualified to perform cybersecurity testing are in constant demand, and increasing or fast-tracking testing for one system often requires the delaying or reducing of testing for other systems. Some efforts to establish a framework

[93] Simpson et al. (2018)

[94] Ibid.

[95] DOT&E (2014, 2), DOT&E(2016, 1), DOT&E(2018, 2)

[96] USD(AT&L) (2015), DoD (2018)

for answering the question of “How much testing is enough?” for cybersecurity are ongoing,[97] though these efforts have yet to see large-scale adoption within DoD.

One possible consideration for cybersecurity test duration could be the target the adversary is attacking, along with the effect they are attempting. By defining these “cyber fires,” the test team can scope the test by considering the likelihood and severity of each attack, providing a more rigorous basis for determining the amount of testing required. At present, programs appear to test for “the usual time” rather than scoping the test for the particular system. Using information about mission capabilities, vulnerabilities, and cyber effects can help test teams design efficient tests tailored for the higher-risk scenarios.[98]

c. Artificial intelligence and autonomous systems

There is no reference to either autonomy or artificial intelligence in the 1994 report. The testing of autonomous systems, or AI-enabled elements of these systems, poses a number of unsolved problems.

AI is increasingly common and increasingly relied upon in DoD systems, meaning decisions will be made in increasing frequency by so-called “black box” systems. Those decisions might take different forms, such as planning engagements, identifying or categorizing objects, or even just deciding whether to go left or right. Performance evaluations for these systems must provide assurance that systems will make these decisions appropriately.

Taken alone, the traditional ways testers identify how much testing is enough will be insufficient for AI-enabled systems. Techniques such as DOE assume that system performance or behavior observed under a finite set of conditions will generalize well to similar but unobserved sets of conditions. With an AI-enabled system, especially when the decision-making algorithm is a black box, those inferences may not be valid. The dimensions of interest for these systems are those that change what the appropriate decision is in a situation. There might be many such dimensions for a decision, which alone makes testing hard,[99] but even more pernicious is the issue of correlated but irrelevant information.[100] Before we can make inferences along our dimension of interest, we have

[97] For example, Avery and Gilmore (2019)

[98] Whetstone et al. (forthcoming)

[99] Haugh, Sparrow, & Tate (2018)

[100] For example, the terrain along the edge of most roads is higher than the road itself. (This might be a sidewalk or a berm.) An AI system that is not given human-provided context might learn to define the edge of the road as having a rise in elevation. The system might perform fine in test scenarios that have this terrain feature but could fail in live operations if it encounters a road that does not have this common-but-not-universal feature.

to be confident that the system actually bases its decision on the feature(s) of interest and not the irrelevant correlation.

If we want to design a test of the system's performance, we must understand how the system makes decisions. Without that understanding, we will be unable to make inferences, and the decision spaces are functionally impossible to test exhaustively.[101] The first stage of testing will likely need to be a series of sequential experiments designed to produce understanding of how the system makes decisions. These will be exploratory,[102] and there are currently no techniques to identify how much of this is enough. Once this is complete to an acceptable level of risk, techniques such as DOE should be employed to efficiently test system decision performance. Deciding how much testing is enough without understanding how a system makes its decisions would be like trying to design a test of an aircraft's flight envelope without anyone understanding aerodynamics. Doing so would invite blind acceptance of risk.

At the time of this writing, several efforts are underway specifically to develop test methods for AI-enabled and/or autonomous systems within the defense community. For example, STAT COE has been running a workshop on the topic for several years;[103] the Services' test and research organizations have been hosting knowledge-sharing meetings on the topic; DT&E has prepared a course for Defense Acquisition University;[104] and IDA has a forthcoming framework for designing tests of AI-enabled systems.

3. Summary

As the battlefield evolves and threats become more capable, requirements for new weapon systems become more and more demanding. Many existing systems require more sophisticated hardware and especially software to succeed in this modern environment. Some of the areas requiring better test and evaluation processes and solutions are for highly software-intensive systems. Testing these systems as stand-alone systems is insufficient, but testing integrated systems of systems adds complexity to the operational environment. Testing must occur early on in software development at the pace of agile/development operations, and must make use of automation. Routine testing of new systems in system-of-system configurations will be necessary to prevent integration-related failures during post-fielding operations. Cyber testing processes that will result in efficient and effective testing across a broad range of possible penetration conditions are still being developed

[101] Zacharias (2019)

[102] Haugh et al., 2018

[103] E.g., Ahner & Parson, 2016

[104] CLE 002, "Introduction to the Test & Evaluation (T&E) of Autonomous Systems"

and refined. Autonomous systems leveraging artificial intelligence/machine learning will soon become commonplace and present new challenges to our testers.

4. Conclusions

The similarities between the challenges today and those outlined in the 1994 symposium report are striking. Program offices desire more flexibility and fewer requirements for testing. Test programs are too often stovepiped, and the DT and OT communities struggle to coordinate because of the differing organizational goals. Everyone agrees that risk management should inform testing, but stakeholders disagree about the types of risk that are the most important and how to evaluate them. In the symposium summary, it is not hard to find passages that could be said just as easily today as 25 years ago.

There are clear areas in which progress has been made, the most notable of which is the adoption of statistical techniques throughout the OT community. The Service OTAs and DOT&E both advocate the use of statistical tools to help in the designing of tests. As these initiatives continue, an impressive body of case studies has been built up attesting to the benefits of using tools such as experimental design. But there is still room for improvement, particularly on the analysis side. In cases where they can be applied, Bayesian techniques and sequential test designs offer further opportunities to gain efficiencies in testing. These tools work hand in hand with improved cooperation between DT and OT. As DoD continues to emphasize integrated testing approaches, this coordination becomes even more important for ensuring that both communities are able to achieve their goals from shared test events.

In other areas highlighted by the 1994 symposium, progress has been slower. Efforts to estimate the cost of testing (operational testing in particular) have failed to produce useful methods for analyzing the cost of T&E or conducting cost/benefit analyses. No database of test costs currently exists. Widely accepted and applied approaches for characterizing risk remain elusive beyond traditional statistical measures. Recently, DoD undertook broad efforts to provide program managers with more flexibility to bring new technology to the battlefield. The Middle Tier Acquisition pathways created by the FY16 NDAA may prove successful in reducing acquisition timelines for new technologies, but we do not yet have enough examples to confirm this.

New challenges have emerged since 1994 that make determining the right amount of testing even more challenging. The amount of software in modern systems is orders of magnitude higher by some metrics than the systems of the mid-1990s, and software updates come ever faster. Finding ways to automate software testing is critical if T&E is to keep up with the pace of innovation. Designing cybersecurity tests presents novel challenges, both in terms of the level of acceptable risk and in terms of defining how much data can be gathered from a fixed amount of testing. The T&E community is also starting to tackle the problem of designing tests for autonomous systems and AI via courses and workshops.

One of the main takeaways from the 1994 symposium was that there was no single answer to the question, "How much testing is enough?" This remains true today and will be for the foreseeable future. It is critical that we continue to improve the way we address this question to ensure that resources are spent efficiently and that testing is effective in providing critical information to warfighters and decision makers.

Appendix A
Acronyms

ACTD	Advanced Capability Technology Demonstration
AFIT	Air Force Institute of Technology
AI	Artificial Intelligence
AMRAAM	Advanced Medium Range Air-to-Air Missile
AMSAA	Army Materiel Systems Analysis Activity
AT&L	Acquisition, Technology, and Logistics
ATD	Advanced Technology Demonstration
C4ISR	Command, Control, Communications, Computers, Intelligence, Surveillance, and Reconnaissance
COE	Center of Excellence
CONOPS	Concept of Operations
COTF	Commander, Operational Test and Evaluation Force
DA	Discussion Area
DASD(SE)	Deputy Assistant Secretary of Defense for Systems Engineering
DATAWorks	Defense and Aerospace Test and Analysis Workshop
DAU	Defense Acquisition University
DMSCO	Defense Modeling and Simulation Coordination Office
DoD	Department of Defense
DoDi	Department of Defense Instruction
DOE	Design of Experiments
DON	Department of the Navy
DOT&E	Director, Operational Test and Evaluation
DT	Developmental Test
DT&E	Developmental Test and Evaluation
DTIC	Defense Technical Information Center
DTSE&E	Director for Test Systems Engineering and Evaluation
DUSD(A&T)	Deputy Under Secretary of Defense for Acquisition and Technology
EA	Electronic Attack
EW	Electronic Warfare
EWO	Electronic Warfare Officer
FDSC	Failure Definition Scoring Criteria
FOT&E	Follow-on Test and Evaluation
HLA	High Level Architecture
HOTT	H-1 Operational Test Team
HWIL	Hardware-in-the-Loop
IDA	Institute for Defense Analyses
IOT&E	Initial Operational Test and Evaluation
IR	Infrared

IT	Integrated Test
ITEA	International Test and Evaluation Association
JAGM	Joint Air-to-Ground Missile
JCTD	Joint Capability Technology Demonstration
JSF	Joint Strike Fighter
JUON	Joint Urgent Operational Need
JWA	Joint Warfighting Assessment
KPP	Key Performance Parameter
LAIRCM	Large Aircraft Infrared Countermeasures
LFT&E	Live Fire Test and Evaluation
LUT	Limited User Test
LVC	Live, Virtual, and Constructive
M&S	Modeling & Simulation
MALD	Miniature Air-Launched Decoy
ML	Machine Learning
MORS	Military Operations Research Society
MRAP	Mine-Resistant, Ambush Protected
MTA	Middle Tier Acquisition
NDAA	National Defense Authorization Act
NIE	Network Integration Evaluation
OPNAV	Office of the Chief of Naval Operations
OR	Operations Research
OSD	Office of the Secretary of Defense
OT	Operational Test
OT&E	Operational Test and Evaluation
OTA	Operational Test Agency
PMA	Program Management Administration
R&M	Reliability and Maintainability
RAM	Reliability, Availability, and Maintainability
RAM-C	Reliability, Availability, and Maintainability-Cost
SDB II	Small Diameter Bomb II
SME	Subject Matter Expert
STAT	Scientific Test and Analysis Techniques
STEM	Science, Technology, Engineering, and Mathematics
T&E	Test and Evaluation
TEMP	Test and Evaluation Master Plan
TSS	Targeting Sight Sensor
UON	Urgent Operational Need
USD(A&T)	Under Secretary of Defense for Acquisition and Technology
VV&A	Verification, Validation, and Accreditation

Appendix B
Acknowledgments

The writing process of this report included soliciting contributions from many individuals with first-hand knowledge of the programs and events discussed herein. We would like to acknowledge the contributions of those individuals, without whom this report would be less interesting and less complete. In alphabetical order:

- Mark Couch
- Brent Crabtree
- Theresa Daily
- Karl Glaeser
- Greg Hutto
- Ken Mathiasmeier
- Dave Sparrow
- Brad Thayer
- Steve Thorsen
- Shawn Whetstone
- Marion Williams

We are also indebted to our technical review committee, who provided valuable feedback as well as additional references and expertise. The input we received from Art Fries, Dean Thomas, Michael Gilmore, and Dan Porter improved the clarity of the report and provided valuable context. We appreciate their thoughtfulness, thoroughness, and candor.

References

53rd Wing, United States Air Force (2009), “AIM-9X Block II Pk Performance Set: Test Point Considerations.” Presentation to the AIM-9X Block II Missile Simulation Working Group.

Aegis Ballistic Missile Defense Program Office (2015), “Aegis BL 9.B1 and Aegis BL 9.C1 Element Verification and Validation Plan.”

Ahner, d. and Parson, C. (2016), “Workshop Report: Test and Evaluation of Autonomous Systems,” SATA COE Report 01-2016

Air Force Operational Test and Evaluation Center (2018), “Small Diameter Bomb Increment II Multiservice Initial Operational Test and Evaluation Modeling and Simulation Accreditation Plan.”

Albon, C. (2019), “Lockheed awaiting F-35 IP protest decision that delayed key IOT&E testing phase.“ *Inside Defense Now*.

Aldridge, J., Crabtree, B., Shaw, S., and Warner, C. (2005), “Operational Assessment of the USMC H-1 Upgrades Program,” IDA Technical Report P-3959.

Anderson-Cook, C., and Lu, L. (editors) (2012), “Statistical Engineering – Forming the Foundations” *Quality Engineering*, 24, 2, 110-132

Assistant Secretary of the Army, AL&T (2018), “Army Report on the Director, Operational Test & Evaluation Annual Report, Fiscal Year 2017,” DoD Technical Report.

Avery, K. and Gilmore, J. (2019). “Applying DOE to Cyber Testing.” Informal IDA briefing 2019-24903.

Avery, K., and Freeman, L. (2017), Statistical Techniques for Modeling and Simulation Validation, IDA Non-Standard Document 8694

Avery, K., Freeman, L., Parry, S., Whittier, G., Johnson, T., Flack, A., and Wojton, H. (2019), “Handbook on Statistical Design & Analysis Techniques for Modeling & Simulation Validation.” IDA Technical Report NS D-10455.

Avery, M., Fealing, C., Wojton, H., Carter, K. (Forthcoming). “DOT&E 2019 Major Activities and Historic Trends.” IDA Technical Report D-10874.

Avery, M., Thomas, D., Goodman, A., Thayer, B., Crabtree, B., Anderson, C., Pechkis, D., DeWolfe, D., Heuring, E., Wojton, H., Gonzales, J., Bell, J., Erikson, W., Clutter, J., Avery, K., Freeman, L., Luhman, M., Shaw, M., Dickinson, R., Shaw, S., Satyapal, S., Movit, S., Johnson, T., Lillard, V. (2017), “The Value of Statistical Thinking in Test and Evaluation.” IDA D-8600.

Box, G. (1993), “Sequential Experimentation and Sequential Assembly of Designs,” *Quality Engineering*, 5, 2: 321-330.

Cohen, M., Rolph, J., and Steffey, D., eds. (1998), *Statistics, Testing, and Defense Acquisition: New Approaches and Methodological Improvements.* National Academy Press. (Sponsor: DOT&E)

Comfort, G., Aldridge, J., Crabtree, B., and Shaw, S. (2003), "Evaluation of the H-1 Upgrade Early Operational Assessment." IDA Technical Report P-3813.

Commander, Operational Test and Evaluation Force (2003), "AIM-9X Multi-service Operational Test and Evaluation Modeling, Simulation and Accreditation Plan."

Crabtree, B., Aldridge, J., Eusanio, L. Shaw, S., and Volpe, V. (2007). "Interim Assessment of the H-1 Upgrades (AH-1Z and UH-1Y) Aircraft." IDA Technical Report P-4249.

Dalal, S., Poore, J., and Cohen, M., eds. (2003) *Innovations in Software Engineering for Defense Systems*, National Academies Press. (Sponsor: DOT&E and AT&L)

Dahmann, J. (1997). "The Department of Defense High Level Architecture". Proceedings of the 1997 Winter Simulation Conference.

DATAWorks Archive (https://testscience.org/archive/)

Defense Acquisition University (2019). CLE 002 "Introduction to the Test and Evaluation (T&E) of Autonomous Systems."

Defense Modeling and Simulation Coordination Office (2011), "DMSCO M&S VV&A Recommended Practices Guide Core Document."

Defense Science Board (2008). *Developmental Test and Evaluation*. DTIC Number ADA482504

Department of Defense (2018), "Cybersecurity Test and Evaluation Guidebook, Version 2.0." DoD Guidebook.

Department of Defense Operational Test Agencies, 2019 memorandum. "Operational Test Agencies Six Core Test Principles."

Deputy Assistant Secretary of Defense for Systems Engineering (2018), Digital Engineering Strategy.

Dickinson, R., Freeman, L., and Simpson, B. (2015). "Statistical Methods for Combining Information: Stryker Family of Vehicles Reliability Case Study." *Journal of Quality Technology*. 46(4):400-415.

Dickinson, R., Freeman, L., Wilson, A., and Simpson, B. (2012). "Statistical Methods for Combining Information: Stryker Reliability Case Study." IDA Technical Report NS D-4721.

Director, Operational Test and Evaluation (1996), "Guidelines for Conducting Operational Test and Evaluation for Software-Intensive System Increments." DoD memorandum.

Director, Operational Test and Evaluation (1999), "Policy for Operational Test and Evaluation of Information Assurance." DoD memorandum.

Director, Operational Test and Evaluation (2009), “Extended Range Multi Purpose Unmanned Aircraft System's Quick Reaction Capability Early Fielding Report.” DoD memorandum.

Director, Operational Test and Evaluation (2010), “Assessment of the Mine Resistant Ambush Protected (MRAP) Family of Vehicles.“ DoD memorandum.

Director, Operational Test and Evaluation (2010), “Extended Range Multi-Purpose Unmanned Aircraft System Operational Assessment.” DoD memorandum.

Director, Operational Test and Evaluation (2010), “Live Fire and Operational Test and Evaluation Report on the Mine Resistant Ambush Protected (MRAP) - All-Terrain Vehicle (M-ATV).” DoD memorandum.

Director, Operational Test and Evaluation (2010), “Guidance on the use of Design of Experiments (DOE) in Operational Test and Evaluation.” DoD memorandum.

Director, Operational Test and Evaluation (2011), “Testing Doesn't Cost - It Pays.” Presentation. https://www.dote.osd.mil/pub/presentations/TestingDoesntCost-ItPays_final26APR11.pdf

Director, Operational Test and Evaluation (2011), “The Marginal Cost to Programs of Operational Test and Evaluation.” Presentation. https://www.dote.osd.mil/pub/presentations/201104MarginalCost_ofOTE.pdf

Director, Operational Test and Evaluation (2013). “Aegis Weapon System Modeling and Simulation and Aegis DDG 51 Flight III Probability of Raid Annihilation (PRA) Study.” DoD memorandum.

Director, Operational Test and Evaluation (2013). “Analysis of Small Arms Failures That Occur During Combat Helmet Testing.” DoD memorandum.

Director, Operational Test and Evaluation (2013), “(U) Requirement for Use of a Self-Defense Test Ship for Operational Testing of the DDG-51 Flight III Equipped with the Air and Missile Defense Radar.” DoD memorandum.

Director, Operational Test and Evaluation (2013), “Best Practices for Assessing the Statistical Adequacy of Experimental Designs Used in Operational Test and Evaluation.” DoD memorandum.

Director, Operational Test and Evaluation (2013), “Flawed Application of Design of Experiments (DOE) to Operational Test and Evaluation (OT&E).” DoD memorandum.

Director, Operational Test and Evaluation (2014). “Follow-on Operational Test and Evaluation (FOT &E) Report for the Lot 4 AH-64E Apache Attack Helicopter.“ DoD memorandum.

Director, Operational Test and Evaluation (2014), “Procedures for Operational Test and Evaluation of Cybersecurity in Acquisition Programs,” DoD Memorandum.

Director, Operational Test and Evaluation (2014), “Reasons Behind Program Delays.” Presentation.

https://www.dote.osd.mil/pub/presentations/ProgramDelaysBriefing2014_8Aug_Final-77u.pdf

Director, Operational Test and Evaluation (2014), FY 2013 Annual Report

Director, Operational Test and Evaluation (2014), “RQ-7BV2 Shadow Tactical Unmanned Aircraft System Follow-On Operational Test and Evaluation and One System Remove Video Terminal Operational Assessment.” DoD memorandum.

Director, Operational Test and Evaluation (2016), “Cybersecurity Operational Test and Evaluation Priorities and Improvements.” DoD Memorandum.

Director, Operational Test and Evaluation (2016), “Guidance on the Validation of Models and Simulation used in Operational Test and Live Fire Assessments.” DoD memorandum.

Director, Operational Test and Evaluation (2016), “Realistic Operational Test and System Requirements.” Presentation. https://www.dote.osd.mil/pub/presentations/2016/Value%20of%20OT_V8.pdf

Director, Operational Test and Evaluation (2017), “Clarifications on Guidance on the Validation of Models and Simulation used in Operational Test and Live Fire Assessments.” DoD memorandum.

Director, Operational Test and Evaluation (2017), FY 2016 Annual Report

Director, Operational Test and Evaluation (2018), “Department of Navy Large Aircraft Infrared Countermeasures (DON LAIRCM) Advanced Threat Warner (ATW) MV-22 Installation.” DoD memorandum.

Director, Operational Test and Evaluation (2018), “Procedures for Operational Test and Evaluation of Cybersecurity in Acquisition Programs.” DoD Memorandum.

Director, Operational Test and Evaluation (2018), FY 2017 Annual Report

Director, Operational Test and Evaluation (2019), “Operational and Live-Fire Test and Evaluation Planning Guidelines for Middle Tier of Acquisition Programs.” DoD memorandum.

Director, Operational Test and Evaluation (2019), CVN 78 Gerald R. Ford-Class Nuclear Aircraft Carrier.” FY2018 DOT&E Annual Report.

Director, Test, Systems Engineering and Evaluation (1996), “A Description of the DoD Test and Evaluation Process for Electronic Warfare Systems.” DTIC.

DoD Handbook, (2011), *Department of Defense Handbook on Reliability Growth Management*, MIL-HDBK-189C.

Dominy, J., Forrest, J., Barnett, T., Rogers, B., Williams, M. (2013), “Cost of Operational Test and Evaluation: Phase I Report.” IDA Technical Report P-4970.

DOT&E and USD(AT&L), 2007 memorandum. “Policy for Assessing Technical Risk of Entry into Initial Operational Test and Evaluation.”

DOT&E and USD(AT&L), 2008 memorandum. “Definition of Integrated Testing.”

DOT&E, USD(AT&L), ASD(C3I) and Joint Staff (2000), "Promulgation of DoD Policy for Assessment, Test, and Evaluation of Information Technology System Interoperability." DoD memorandum.

Fisher, R. (1952), "Sequential Experimentation," *Biometrics*, 8: 183-187.

Freeman, L. (2018), "Revolutionizing T&E with New Methods." ITEA Conference Paper.

Freeman, L., Goodman, A., Peek, D., Bell, J., Avery, M., Roberts, M., Hueckstaedt, R. (2016). "2015 Reliability Assessment." IDA Technical Report D-8152.

Freeman, L., Ryan, A., Kensler, J., Dickinson, R., and Vining, G. (2013), "A Tutorial on the Planning of Experiments." Quality Engineering, 25, 4, 315-332.

Freeman, Laura; Wilson, Alyson; Browning, Caleb; Fronczyk, Kassandra; and Medlin, Rebecca (2015), "Bayesian Hierarchical Models for Common Components Across Multiple System Configurations." IDA document: D-5514.

Fronczyk, K., Dickinson, R., Wilson, A., Browning, C., Freeman, L. (2015). "Bayesian Hierarchical Models for Common Components Across Multiple System Configurations." IDA Technical Report NS D-5514.

FY16 NDAA, Section 804. "Middle Tier of Acquisition for Rapid Prototyping and Rapid Fielding."

Gehrig, J., Brown, C., and Finfera, J. (1994), MORS/ITEA Mini-Symposium, "How Much Testing Is Enough?"

Gilmore, J. (2013), "A Statistically Rigorous Approach to Test and Evaluation." *ITEA Journal*, 34, 225-229.

Hare, L. (2019), "The Foundation of Statistical Engineering." *Quality Progress*, August 2019, 48-51.

Higdon, J. (2017), "Designing Experiments with Software-Intensive Systems: A 2-day Short Course." 96th Cyberspace Test Group.

Hoecherl, J. (2018), "AH-64 Decision." Email Correspondence to IDA.

Hoerl, R., and Snee, R. (2010), "Closing the Gap: Statistical Engineering Links Statistical Thinking, Methods, Tools." *Quality Progress*, May 2010, 52-53.

Holt, P. (2016), "Test Capability Accreditation," AFOTEC Course 301.

Huffman, M., Harmon, D., Selling, N., Hutto, G., Dailey, T., Zessin, B., and Ortiz, F. (2019), "B61: Building a Statistically Defensible Integrated Flight Test." 96 TW Presentation.

Hutto, G., Simpson, J. and Schroeder, K. (2018), "Case: F-35 Vibration AMRAAM Qualification Testing – Block 2B and 3F." 96TW Eglin AFB.

Johnson, R., Hutto, G., Simpson, J., and Montgomery, D. (2012), "Designed Experiments for the Defense Community." *Quality Engineering*, 24, 1, 60-79.

Johnson, T., Freeman, L., Hester, J., and Bell, J. (2014), "A Comparison of Ballistic Resistance Testing Techniques in the Department of Defense." *Institute of Electrical and Electronics Engineers (IEEE).*

Joint Attack Munition System Project Office (2017), "JAGM T&E WIPT – 14 June 2017." PEO Missiles and Space.

Kass, R. (2015), "Twenty-one Parameters for Rigorous, Robust, and Realistic Operational Testing." *ITEA Journal*, 36, 121-138.

Kuhn, D. R., Kacker, R. N., and Lei, Y. (2010), "Practical Combinatorial Testing." NIST Special Publication 800-142.

Kyle, S. (2017), "Common M&S Threat Environment Model for Long-Range Strike (LRS) Family of Systems (FoS) OT&E." OSD DOT&E Air Warfare.

Landman, D., Simpson, J., Vicroy, D. and Parker, P. (2007), "Response Surface Methods for Efficient Complex Aircraft Configuration Aerodynamic Characterization." *Journal of Aircraft*, 44, 4, 1189-1195.

Lawrence, C. (2019), "T&E Renaissance Forum Shines Light on Challenges and Advances." Defense Visual Information Distribution Service.

Leach, L. (2014), *Critical Chain Project Management, 3d ed.* Artech House, Boston MA.

Meely, M. (2017), "Apache V6 US Army Redstone Test Center." US Army Apache Program Office October WIPT Presentation.

Montgomery, D., Runger, G., and Hubele, N. (2010), *Engineering Statistics, 5th ed.*, Wiley, NY.

Nair, V., and Cohen, M., eds. (2006), *Testing of Defense Systems in an Evolutionary Acquisition Environmen*t, National Academy Press. (Sponsors: DOT&E and AT&L)

National Institute of Standards and Technology (2010). *Guide for Applying the Risk Management Framework to Federal Information Systems*. NIST Special Publication 800-37 Rev 1.

National Institute of Standards and Technology (2018). *Guide for Applying the Risk Management Framework to Federal Information Systems*. NIST Special Publication 800-37 Rev 2.

National Research Council (2002), *Modeling and Simulation in Manufacturing and Defense Acquisition: Pathways to Success*, National Academies Press, ISBN-10 0309084822.

National Research Council (2002), *Reliability Issues for DoD Systems: Report of a Workshop*, National Academies Press. (Sponsors: DOT&E and AT&L)

National Research Council (2004), *Improved Operational Testing and Evaluation and Methods of Combining Test Information for the Stryker Family of Vehicles and Related Army Systems: Phase II Report*, National Academies Press. (Sponsor: ATEC)

National Research Council (2012), *Industrial Methods for the Effective Development and Testing of Defense Systems*, National Academy Press. (Sponsors: DOT&E and AT&L)

National Research Council (2012), *Testing of Body Armor Materials: Phase III*, National Academy Press. (Sponsor: DOT&E)

National Research Council (2014), *Review of Department of Defense Test Protocols for Combat Helmets*. (Sponsor: DOT&E)

National Research Council (2015), *Reliability Growth: Enhancing Defense System Reliability*, National Academy Press. (Sponsors: DOT&E and AT&L)

O'Bryon, J. (2006), "DoD's Modeling and Simulation Reform in Support of Acquisition: Stop Kicking the M&S Can Down the Road," Defense Acquisition University Press.

Ortiz, F. (2019), "Design of Experiments Integration with Simulation Development for the USAF B61 Tail Kit Mod 12 Life Extension Program," STAT COE Report-11-2019.

PEO Integrated Warfare Systems (2015), "Standard Missile-6 Block I Follow-on Operational Test and Evaluation Modeling and Simulation Verification and Validation Plan."

Peterson, M. (2006), "Advanced Concept Technology Demonstration (ACTD) and the transition to the Joint Capability Technology Demonstration (JCTD) business model." Deputy Under Secretary of Defense (Advanced Systems & Concepts) Presentation.

Program Manager PMA-276 (2005), "Test and Evaluation Master Plan No. 1435 (Revision B) for the USMC H-1 Upgrades Program, DoD Technical Report.

Reese, C., Wilson, A., Hamada, M., Martz, H., Ryan, K. (2004), "Integrated Analysis of Computer and Physical Experiments, *Technometrics*, pp. 153-164.

Rolph, J. and Steffey, D., eds. (1994), Statistical Issues in Defense Analysis and Testing: Summary of a Workshop, Washington, DC: National Academy Press. (Sponsors: DOT&E AND PA&E)

Roth, D. Kitto, W. and Taylor, M. (2010), "Improving Air Force Flight Test Center Developmental Test and Evaluation Through Increased Use of Statistical Methods," AIAA 2010-1766, USAF T&E Days Conference Paper.

Secretary of the Navy (2019), Draft i3960.1

Simpson, J (2019), "Sequential Testing and Simulation validation for Autonomous Systems." Workshop presentation.

Simpson, J. Listak, C., Hutto, G. (2013), "Guidelines for Planning and Evidence for Assessing a Well-Designed Experiment," *Quality Engineering*, 25, 4, pp. 333-355.

Simpson, J., and Wisnowski, J. (2001), "Streamlining Flight Test with the Design and Analysis of Experiments," *Journal of Aircraft*, 38, 6, 1110-1116.

Simpson, J., Listak, Hutto, G. (2013), "Guidelines for Planning and Evidence for Assessing a Well-Designed Experiment," *Quality Engineering*, 25- 4.

Simpson, J., Wisnowski, J. and Pollner, A. (2018), "Automated Software Test Implementation Guide for Managers and Practitioners," STAT COE Report 05-2018.

Simpson, J., Wisnowski, J., and Doane, S. (2019), "Statistical Methods for Modeling and Simulation Verification and Validation," 3-day Short Course.

STAT Center of Excellence (2019), "Scientific Test and Analysis Techniques Center of Excellence." Presentation to AFMC/CC.

Steyn, H. (2001), "An Investigation into the Fundamentals of Critical Chain Project Scheduling," International Journal of Project Management, 19, 6, pp. 363-369.

Tate, D. (2019), "What Counts as Progress in the T&E of Autonomy?" Workshop presentation.

Thomas, D. (2017), "Analysis of OTA Workforce," IDA OED Memorandum for DOT&E Science Advisor.

Thomas, D., Bieber, C., Wojton, H., Snavely, J, and Freeman, L. (2017). "Analysis of OTA Workforce." IDA memo to DOT&E.

Thomas, D., Christofek, L., Keese, H., Lillard, V., Mathiasmeier, K., Overholser, E., Rabinowitz, S, Shaw, M., Simpson, B., Warner, C. (2010), "DOE in TEMPS, T&E Concepts, Test Plans and BLRIPS," IDA Technical Report D-4142.

Tucker, A., Hutto, G., and Dagli, C. (2010), "Application of Design of Experiments to Flight Test: A Case Study," *Journal of Aircraft*, 47, 2, 458-463.

Undersecretary of Defense (Advanced Technology and Logistics) (2015), "DoD Program Manager's Guidebook for Integrating the Cybersecurity Risk Management Framework (RMF) into the System Acquisition Lifecycle." DoD Guidebook.

Warner, C. (2013). Report on the Test Science Roadmap for the Director, Operational Test and Evaluation (DOT&E). https://www.dote.osd.mil/Portals/97/pub/reports/20130711TestScienceRoadmapReport.pdf?ver=2019-09-03-161501-357

Wauer, George, G., Wallace, C. G., and Wright, S. J. (2001), "DOT&E Initiatives to Improve System-of-Systems Interoperability," *ITEA Journal*, pp. 11-18.

Whetstone, Shawn, Botting, Tye, and White, Richard (Forthcoming) "Cybersecurity and Operational Test and Evaluation: Fundamental Concepts for IDA Research Staff Members," IDA NS-P10784

Williams, M. and Sikora, J. (2002), "Military Operations Research Society Workshop Series on Simulation Validation," Workshop Summary Presentation.

Wu, C., and Tian, Y. (2014) "Three-Phase Optimal Design of Sensitivity Experiments." Journal of Statistical Planning and Inference 149: 1-15.

Zessin, C., Oelrich, M., Hutto, G., Romeo, A., Rios, A., and Simpson, J. (2017), "Alternative Modeling Strategies for Characterizing Radio Performance," Quality and Reliability Engineering International, 33, 215-224.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*
12-2019

2. REPORT TYPE
IDA Publication

3. DATES COVERED *(From - To)*

4. TITLE ANDSUBTITLE

"How Much Testing is Enough?" 25 Years Later

5a. CONTRACT NUMBER
HQ0034-19-D-0001

5b. GRANT NUMBER
____ ____ ____

5c. PROGRAM ELEMENT NUMBER
____ ____ ____

6. AUTHOR(S)

Matthew R. Avery (OED); James R. Simpson (OED);

5d. PROJECT NUMBER
BD-09-2299

5e. TASK NUMBER
229990

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Institute for Defense Analyses
4850 Mark Center Drive
Alexandria, Virginia 22311-1882

8. PERFORMING ORGANIZATION REPORT NUMBER

P-10994
H 2019-000627

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Director, Operational Test and Evaluation
1700 Defense Pentagon
Washington, DC 20301-1882

10. SPONSOR/MONITOR'S ACRONYM(S)
DOT&E

11. SPONSOR/MONITOR'S REPORT NUMBER

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release. Distribution Unlimited

13. SUPPLEMENTARY NOTES
Project Leader, Heather Wojton

14. ABSTRACT

In 1994, MORS (Military Operations Research Society) and ITEA (International Test and Evaluation Association) co-sponsored a mini-symposium to tackle the question, "How Much Testing is Enough?" in test and evaluation (T&E). Participants from the symposium produced a report detailing the discussions and recommendations from the three-day event. Twenty five years later, the question that inspired the symposium is still hotly debated within the T&E community. The intervening years have seen substantial progress made in areas like the use of experimental design for sizing tests, combining data from developmental and operational test to improve efficiency, and the wide-spread adoption of modeling and simulation in T&E. Less progress has been made in cost transparency and integration of technology demonstrations with the T&E process. Since 1994, new challenges like cybersecurity and autonomy have emerged, presenting new challenges to determining the right amount of testing. Despite the many improvements made over the past twenty-five years, there are still no simple answers to the question, “How much testing is enough?”

15. SUBJECT TERMS
Autonomy;Design of Experiments (DOE);DT&E(Developmental Test and Evaluation); Integrated Testing and Evaluation; Operational Test and Evaluation

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** Heather Wojton (OED)
a. REPORT Unclassified	**b. ABSTRACT** Unclassified	**c. THIS PAGE** Unclassified	Unlimited	75	**19b. TELEPHONE NUMBER** *(include area code)* (703) 845-6811