INSTITUTE FOR DEFENSE ANALYSES

Determining How Much Testing is Enough: An Exploration of Progress in the Department of Defense Test and Evaluation Community

Rebecca Medlin, Project Leader

Matthew R. Avery
James R. Simpson
Heather M. Wojton

February 2021

Approved for Public Release.

Distribution Unlimited

IDA Document NS D-21561

Log: H 2021-000046

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

The Institute for Defense Analyses is a nonprofit corporation that operates three Federally Funded Research and Development Centers. Its mission is to answer the most challenging U.S. security and science policy questions with objective analysis, leveraging extraordinary scientific, technical, and analytic expertise.

About This Publication
This work was conducted by the Institute for Defense Analyses (IDA) under a Separate Contract Task C9082 , CRP Cross-Divisional Statistics and Data Science Working Group," 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 Dean Thomas from the Operational Evaluation Division.

For more information:
Rebecca Medlin, Project Leader
rmedlin@ida.org • (703) 845-6731

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

Rigorous Analysis | Trusted Expertise | Service to the Nation

INSTITUTE FOR DEFENSE ANALYSES

IDA Document NS D-21561

Determining How Much Testing is Enough: An Exploration of Progress in the Department of Defense Test and Evaluation Community

Rebecca Medlin, Project Leader

Matthew R. Avery
James R. Simpson
Heather M. Wojton

Executive Summary

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. 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 years, there are still no simple answers to the question, “How much testing is enough?”

Determining How Much Testing is Enough: An Examination of Progress in the Department of Defense Test and Evaluation Community

The test and evaluation (T&E) community within the Department of Defense (DoD) has been grappling with the question of "How much testing is enough?" for years. Testers, constrained by the limits of funding and technology, must answer this question while meeting the real-world needs of warfighters and the budget- and schedule-related needs of program managers. More testing helps us answer questions about system performance and can identify previously unknown system flaws or limitations. Knowing when enough testing has been done, however, is also important.

A certain number of test runs or data points alone is not sufficient to know how much testing is enough. The particulars of the test are vital, including how operationally realistic* each test event is, how complete and accurate the data collected during testing will be, and the breadth of scenarios we want to evaluate using the test data. In general, for systems that are to be employed in only a limited set of conditions, we learn most quickly from realistic tests, with complete and accurate data collection. If parts of the test are less realistic, or if some of the data cannot be collected, more testing will be required. For systems that are to be employed in a wide variety of conditions, however, it may not be possible to use the data collected under one set of conditions to evaluate system performance in other conditions. More testing might be required in this case as well. Answering the question of "How much testing is enough?" is a complex challenge requiring nuanced, case-dependent answers.

In 1994, the Military Operations Research Society (MORS) and the International Test and Evaluation Association (ITEA) held a joint three-day symposium entitled, "How Much Testing Is Enough?"[1] to discuss this question.† The symposium consisted of keynote

---

* While distinctions are drawn between operational testing (OT) and developmental testing (DT), with perhaps integrated testing (IT) being considered something of a middle ground, the testing landscape is better thought of as a multi-dimensional plane than a one-dimensional continuum. Relevant dimensions include system configuration, operating conditions, and personnel. In general, the closer a test event is to the real-world scenario in which a system will be employed, the more information we can potentially glean.

† Throughout this paper, mentions of "the 1994 symposium," "the symposium," "symposium participants," etc. refer to information discussed in this report about the 1994 symposium.

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 and made recommendations, although according to the symposium leaders, it may have produced more questions than it answered.

It has been 27 years since this symposium took place, and the question that inspired it remains crucially relevant in the T&E 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 T&E community has made progress, some in which we have not, and others where the changing world forces us to reexamine old solutions and address new challenges.

Rather than offering a detailed retelling of 27 years of changes in policy and priority within DoD, this paper describes holistic progress in answering the question of "How much testing is enough?" It covers areas in which the T&E community has made progress, areas in which progress remains elusive, and issues that have emerged since 1994 that provide additional challenges. The selected case studies used to highlight progress are especially interesting examples, rather than a comprehensive look at all programs since 1994. A more complete look at progress and future directions in determining "How Much Testing is Enough?" was published by IDA in 2020.[2]

A. Recommendations from 1994 Symposium

The symposium report's Executive Summary highlights major areas of discussion from the symposium, including the application of statistics within the T&E discipline, integration of modeling and simulation (M&S) with testing, how to pool or share data across DT, OT, and M&S, and challenges evaluating the costs and benefits of testing. There is overlap in these areas, and they do not cover each topic discussed at the symposium, but they provide a starting point for summarizing the concerns of and recommendations from symposium participants.

In this brief summary, we are regularly forced to generalize, since the symposium participants didn't reach consensus on most topics. We attempt to indicate the degree of support each recommendation or proposition had, and in cases where there was clear disagreement among the symposium participants, we describe each view in turn. More details can be found in the original symposium report from 1994.

Throughout the symposium report, there is broad agreement that statistical methods, particularly Design of Experiments (DOE), could be leveraged to improve T&E and address the question of "How much testing is enough?" More than half of the pages in the proceedings used some variant of the word "statistical," and the symposium report specifically recommends that "experimental design techniques and other established statistical approaches should be better exploited." The consensus view was that statistical tools could be used to design and scope tests (e.g., DOE and power analysis), analyze early test data in order to inform programmatic decisions (e.g., through analysis of costs and benefits of additional testing), and combine data from multiple sources (e.g., Bayesian methods).

Symposium participants were generally optimistic about the use of M&S for T&E, and most participants agreed that M&S was an essential part of the T&E domain. Among the strongest proponents of using M&S, 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. Other symposium participants noted that some systems could not be tested in live environments, making M&S critical to T&E. Some preached caution, noting that M&S is not a replacement for live testing, and that M&S does not solve the problem of the high cost of T&E. One participant said, "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."

Questions about combining or "pooling" data were closely related to the use of M&S and the application of statistical methods. Despite substantial interest and discussion, there was no emergent consensus about what conditions made it appropriate to use DT data to inform OT, or what the best approaches were for incorporating data from very early test events or technology demonstrations. Participants supported calls for better coordination between DT and OT communities, but did not propose specific policies. While participants agreed that it was important to share data, they did not explain the mechanisms that would facilitate data sharing or how sharing would create specific benefits in terms of cost, schedule, or analytic accuracy. Using data from prior phases of test or from older systems could, in theory, "shorten the acquisition process, reduce test time, and get technology into the field faster," but it wasn't clear which specific actions programs and analysts should be taking to achieve these benefits.

For most programs, the question of "How much testing is enough?" is a question of time and money. To help justify test costs and use quantitative methods for scoping tests, some symposium participants advocated quantifying the value of testing as well as the costs. By understanding the marginal benefit of an additional run and comparing that to the marginal cost, program managers could make better, analytically rigorous decisions about their test programs. By quantifying the risks associated with *not* testing, decision-

makers could make choices based on balancing costs and acceptable levels of risk. DOT&E was asked to put together a database of T&E costs, but this proved to be more challenging than symposium attendees anticipated.

B. Progress Since 1994

There has been substantial progress since 1994 in DoD's approach to scoping tests. Below are areas where the symposium participants anticipated the direction in which the T&E community would move, as well as some cases in which the hopes of participants were not realized.

1. Statistical Methods

The subject area in which the T&E community of today looks the most like what the 1994 symposium participants rightly envisioned is the adoption of statistical tools and methods. The efficacy of statistical tools for application to T&E is well established. Adoption of these methods has improved our understanding of system performance and made testing more efficient. Experimental test design methods are routinely applied in testing to identify and systematically vary the relevant factors to adequately cover the test space. The 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 Operational Test Agencies (OTAs) has made sure that gains from using statistical methods will continue to accrue across DoD in the coming years.

The OT community in particular, spurred by leadership from DOT&E, has embraced statistical methods. DOT&E commissioned a report in 2010 on the use of DOE for planning operational test events,[3] finding that few programs used it at the time. Leadership did not require that reports include uncertainty estimates, and analytical tools such as operating characteristic curves were not used consistently to size tests. Programs and testers substituted intuition or rules of thumb for criteria based on producer and consumer risk, or uncertainty and risk. 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.[4,5,6]

DOE is now widely used across the T&E domain. Service test plans now include experimental design as a standard practice; the DT community has invested in DOE expertise;[7] and the T&E community has devoted resources to training its workforce in statistical methods such as DOE.[8] Programs that use DOE (such as the F-35 Joint Strike Fighter (JSF)) have seen benefits on the order of millions of dollars in cost savings.[9]

Statistical methods for analysis have also become more widely used. Several reports by the National Research Council (NRC) (and sponsored by DOT&E and AT&L) addressed various statistical challenges in DoD T&E.[10,11,12,13,14,15,16] These reports contributed to awareness of the importance of statistical methods in the T&E domain, and encouraged the adoption of existing methods and development of new ones. Focused efforts in developing statistical methods for T&E have borne fruit; DoD engineers and operations analysts have adopted existing statistical methods and adapted or enhanced them as needed. Where no statistical methods exist, they create new methods tailored to the specific problems (referred to as statistical engineering).[17,18,19] Analysts share these new tools across similar programs, or programs with comparable methodological challenges.[20]

DoD has invested in trainings and workshops to upgrade and maintain the level of statistical skills in the T&E community. 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 more than 7,300 students participating. As a result of effective training and hiring practices, the part of the T&E workforce with statistical skills has increased over the past 10 years.[21] 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.[22]

2. Modeling and Simulation

Since 1994, the maturity and capabilities of M&S have grown tremendously. M&S is now widely used for requirements verification, as well as for demonstrating and validating system performance data. It is increasingly used to provide evidence for capability in regions of the operational space where live testing is not conducted. Validated M&S systems are crucial for understanding initial subsystem, system, and system-of-systems capabilities of DoD programs. They are also relied upon for assessment of the more mature systems, as well as follow-on testing, including operational training and tactics development.[23,24]

Model validation remains a challenge, but DoD's approaches continue to mature. Guidance from DOT&E has clarified the requirements and limitations involved for M&S

data to be used for OT&E. A thorough validation process makes it easier to have a model accredited for a specific purpose, including OT. Although traditional methods for validation are useful, they are not sufficient to validate models for use in OT. For programs that will rely heavily on M&S, the best approach remains a test strategy that integrates M&S validation with live test events. A comprehensive strategy should include assessment of M&S output across the entire operational domain for which the M&S will be accredited, making use of statistical methods, sensitivity analyses, and SME review to ensure that the simulated results are consistent with reality.[25,26] To provide the T&E workforce with tools and knowledge to effectively use M&S in OT&E, DOT&E commissioned IDA to produce a handbook,[27] technical report,[28] and short course[29] 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.

Although M&S is critical for testing many 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 contributed to delays in the completion of IOT&E.[30] In other cases, such as the Aegis Combat System, issues with the existing M&S system limited its utility for OT&E.[31,32] These examples highlight the risk of relying on undeveloped or unvalidated M&S for T&E.

The most optimistic symposium participants believed that models could be built, combined, and reused, allowing testers to gather new data at little to no additional cost. That vision has not been achieved, but similar ideas have been attempted. 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.[33] 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 and is still used as a standard today across a variety of domains beyond M&S. Today, 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.[34] The success of this initiative has not yet been determined.

3. Pooling Data

Symposium participants hoped that by combining data across DT, OT, and M&S, we could improve test efficiency and help reduce the total amount of resources and time required for testing. Although this is generally true, we still do not have systematic, generalizable approaches, and data pooling remains a case-by-case proposition.

The potential benefits of pooling data remain too large to ignore. Data from legacy systems can help provide a baseline understanding of a new (but similar) system, or provide a basis for comparison. By using data collected on the legacy system, we can better

describe the new system without increasing our test budget. When a system receives an update, data from an older version can be used to inform assessments of the new version. Formal statistical approaches, including frequentist methods such as the Gray-Lewis Test[35] as well as Bayesian approaches,[‡] can be used to quantitatively assess whether data sets should be combined. For example, data collected in DT on a less mature version of the system can often provide substantial value for OT assessments. [36] The literature includes numerous case studies highlighting the utility of pooling or sharing data.[37,38,39]

Despite this, pooling data remains controversial. In 2004, a report from the NRC, 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."[40]

Many organizations have created guidance or policies focused specifically on sharing data to improve system reliability evaluation. The Army Materiel Systems Analysis Activity's guidance on reliability growth planning and tracking curves[41] and the National Research Council's 2015 report are two examples,[42] but there are many others.

One success story is the B61 program, in which 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, as well as data from early developmental testing.[43] 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. In such cases, the contractor may not be willing to provide the data to the government testers. In the 1990s, a common cost-saving tactic used by the government was to not purchase the intellectual property (IP) rights of system models and software. Although this may have saved money in the short run, it meant that when systems were updated, or when testers wanted to save costs by using data from M&S, the government had to pay every time the model was used.

The T&E community has yet to identify consistent, effective ways to integrate information from technology demonstrations with other DT and OT data. Participants in the 1994 symposium hoped that these demonstrations might be able to provide useful data

---

‡ 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 T&E, but data from technology demonstrations are generally not useful. Because these events are unstructured and involve early prototypes of systems, it is rare for any of the information collected at such events to be relevant for evaluations of the final system. Today, programs using Middle Tier of Acquisition pathways use "touchpoint" events during program development. These events are both more structured and focused than technology demonstrations and, as a result, provide more useful information for system evaluation. Successful rapid development programs (such as the Mine-Resistant, Ambush Protected (MRAP) vehicle[44] and the MQ-1C Gray Eagle[45,46]) relied on focused coordination between OT and DT communities to ensure that everything that could be tested early was tested early and efficiently. Careful planning, not unstructured demonstrations, is critical for data pooling and efficient testing.

4. Costs and Benefits

One area where little progress has been made is quantitative evaluation of risk. Many participants in the 1994 symposium believed that using rigorous methods to quantify the marginal costs and benefits of additional testing would make it simple to answer the question of "How much testing is enough?" The intervening years have demonstrated that quantifying costs and benefits from testing is difficult.

A 2013 IDA study[47] 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
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."

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.[48,49,50,51] While this study identified many qualitative benefits from testing (e.g., critical system deficiencies were identified in a test environment rather than when the system was being used in combat), these do not translate easily to monetary terms.

C. New Challenges

Systems and the environments in which we expect them to perform have grown more complex since the 1994 symposium, and the challenges of testing them have grown as well. Modern systems are almost universally software intensive and software centric. Today's systems are networked, which increases the need to develop cyber tactics and ensure adequate cybersecurity. Most recently, system designs are taking advantage of artificial intelligence (AI) 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.

1. Software-Intensive Systems

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. Many programs have adopted an agile/development operations approach, which is already widely used in industry. T&E practices have had to adapt to more frequent testing, often using automated software test tools to collect data more frequently and cheaply than humans could. Modern systems rarely operate in a vacuum, making integrated, systems-of-systems testing vital. Early efforts have met with mixed success, but the long lists of lessons learned generated by experience should facilitate better processes in future system-of-systems tests. Combined with the persistent desire to field systems faster, successful program development is increasingly reliant on early testing.[52]

In some cases, specialized test designs can make software testing more efficient. For systems with large sets of configurations and deterministic outcomes, combinatorial tests identify the smallest set of test points that will cover relevant system configurations.[53,54,55] In the unique circumstances where combinatorial tests can be applied to DoD programs, they have the potential to provide great cost and time savings over traditional approaches.[56]

2. Cybersecurity Testing

Cybersecurity test and evaluation was not discussed in the 1994 symposium but is now required for most DoD systems.[57,58] Dedicated cybersecurity testing organizations now exist across DoD and the Services.[59] The current cybersecurity T&E process consists of sequenced phases of test based on guidance from DoD and AT&L. This approach was evolved iteratively and is designed to help programs manage cybersecurity risk intelligently.

Unfortunately, methods for determining what constitutes enough testing remain works in progress. Current protocols do not describe quantitative approaches for determining test duration. At present, programs appear to test "for the usual time" rather than scoping the test for the particular system. Some efforts to establish a framework for quantitatively scoping cybersecurity tests are ongoing,[60,61] though these efforts have yet to see large-scale adoption within DoD.

3. Artificial Intelligence and Autonomous Systems

Unsurprisingly, there is no reference to either autonomy or AI in the 1994 report. However, AI is increasingly common and increasingly relied upon in DoD systems, and DoD is already making decisions about how much testing these "black box" systems require.

Traditional approaches are 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 using a black box decision-making algorithm, 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,[62] but even more challenging is the issue of correlated but irrelevant information.[§] Before we can make inferences along our dimension of interest, we have to be confident that the system actually bases its decision on the dimensions of interest and not the irrelevant correlation. Without understanding how a system makes its decisions, we will be unable to make inferences, and the decision spaces cannot be tested exhaustively. As a result, initial system testing should first establish how the system makes decisions. Subsequent testing to evaluate system performance will make use of tools such

---

[§] 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.

as DOE, but there are currently no techniques to identify how much of this initial exploratory testing is enough.

Several efforts within the defense community are underway specifically to develop methods for testing AI-enabled and/or autonomous systems. For example, 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;[63] and IDA has published a framework for designing tests of AI-enabled systems.[64]

D. Conclusions

It is striking how similar many of the challenges discussed in the 1994 symposium report are to those that the T&E community faces today. 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. It is not hard to find passages from the 1994 report that could have just as easily been said today. Given their persistence and focus, these issues may be structural, and we should not expect a new technique or minor policy shift to resolve them.

In the areas where the T&E community has made progress since 1994, we should remain optimistic and continue to push the state of the art forward. Investing in training the T&E workforce in quantitative literacy and statistical techniques, as well as developing tools and techniques to effectively use M&S for testing, should remain priorities. As new challenges continue to emerge, these competencies will continue to play a crucial role in T&E.

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.

---

[1] J. Gehrig, C. Brown and J. Finfera, "MORS/ITEA Mini-Symposium, *How Much Testing Is Enough?*" Military Operations Research Society, 1994.

[2] M. Avery, J. Simpson, and H. Wojton, ""How Much Testing is Enough?" 25 Years Later," IDA Technical report P-10994, December 2019.

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

[4] Director, Operational Test and Evaluation, "Guidance on the use of Design of Experiments (DOE) in Operational Test and Evaluation." DoD memorandum, October 2010.

[5] Director, Operational Test and Evaluation, "Flawed Application of Design of Experiments (DOE) to Operational Test and Evaluation (OT&E)." DoD memorandum, June 2013.

[6] Director, Operational Test and Evaluation, "Best Practices for Assessing the Statistical Adequacy of Experimental Designs Used in Operational Test and Evaluation." DoD memorandum, July 2013.

[7] STAT Center of Excellence, "Scientific Test and Analysis Techniques Center of Excellence," Presentation to AFMC/CC, 2019.

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

[9] G. Hutto, J. Simpson, and K. Schroeder, "Case: F-35 Vibration AMRAAM Qualification Testing – Block 2B and 3F." Presentation to 96TW Eglin AFB, January 2018.

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

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

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

[13] National Research Council, "Review of Department of Defense Test Protocols for Combat Helmets", National Academy Press, 2014. (Sponsor: DOT&E)

[14] V. Nair, and M. Cohen, eds. "Testing of Defense Systems in an Evolutionary Acquisition Environment", National Academy Press. (Sponsors: DOT&E and AT&L)

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

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

[17] C. Anderson-Cook, and L. Lu, eds., "Statistical Engineering – Forming the Foundations," Quality Engineering, 2012, 24, 2, 110-132

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

[19] L. Hare, "The Foundation of Statistical Engineering." Quality Progress, August 2019, 48-51.

[20] L. Freeman, "Revolutionizing T&E with New Methods." ITEA Conference Paper, March 2018.

[21] D. Thomas, "Analysis of OTA Workforce," IDA OED Memorandum for DOT&E Science Advisor, May 2017.

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

[23] Aegis Ballistic Missile Defense Program Office, "Aegis BL 9.B1 and Aegis BL 9.C1 Element Verification and Validation Plan," 2015.

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

[25] Director, Operational Test and Evaluation, "Guidance on the Validation of Models and Simulation used in Operational Test and Live Fire Assessments." DoD memorandum, March 2016.

[26] Director, Operational Test and Evaluation, "Clarifications on Guidance on the Validation of Models and Simulation used in Operational Test and Live Fire Assessments." DoD memorandum, January 2017.

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

[28] K. Avery and L. Freeman, "Statistical Techniques for Modeling and Simulation Validation", IDA Technical Report NS D-8694, September 2017.

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

[30] C. Albon, , "Lockheed awaiting F-35 IP protest decision that delayed key IOT&E testing phase." Inside Defense Now, November 2019.

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

[32] Director, Operational Test and Evaluation, "(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, 2013.

[33] J. Dahmann. "The Department of Defense High Level Architecture," Proceedings of the 1997 Winter Simulation Conference, December 1997.

[34] Deputy Assistant Secretary of Defense for Systems Engineering, "Digital Engineering Strategy," June, 2018.

[35] H. Gray and T. Lewis, "On a Test for Equality of Means of Two Independent Poisson Distributions," IEEE Transactions on Reliability, R-17 (3), September 1968.

[36] L. Freeman, A. Wilson, C. Browning, K. Fronczyk, and R. Medlin, "Bayesian Hierarchical Models for Common Components Across Multiple System Configurations." IDA document: D-5514, June 2015.

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

[38] S. Steiner, R. Dickinson, L. Freeman, B. Simpson, and A. Wilson, "Statistical Methods for Combining Information: Stryker Family of Vehicles Reliability Case Study," Journal of Quality Technology, 46(4):400-415, November 2017.

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

[40] National Research Council "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, 2004. (Sponsor: ATEC)

[41] L. Crow, "Reliability Growth Planning, Analysis and Management," 2011 Reliability and Maintainability Symposium, January, 2011.

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

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

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

[45] Director, Operational Test and Evaluation, "Extended Range Multi Purpose Unmanned Aircraft System's Quick Reaction Capability Early Fielding Report," DoD memorandum, September 2009.

[46] Director, Operational Test and Evaluation, "Extended Range Multi-Purpose Unmanned Aircraft System Operational Assessment." DoD memorandum, August 2010.

[47] J. Dominy, J. Forrest, T. Barnett, B. Rogers, and M. Williams, "Cost of Operational Test and Evaluation: Phase I Report." IDA Technical Report P-4970, May 2013.

[48] Director, Operational Test and Evaluation, "Testing Doesn't Cost - It Pays." DOT&E Presentation, April 2011.

[49] Director, Operational Test and Evaluation, "The Marginal Cost to Programs of Operational Test and Evaluation," DOT&E Presentation, April 2011.

[50] Director, Operational Test and Evaluation, "Reasons Behind Program Delays," DOT&E Presentation, August 2014

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

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

[53] D. Kuhn, R. Kacker, and Y. Lei, "Practical Combinatorial Testing." NIST Special Publication 800-142, 2010.

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

[55] J. Dahmann. "The Department of Defense High Level Architecture," Proceedings of the 1997 Winter Simulation Conference, December 1997.

[56] L. Freeman, "Revolutionizing T&E with New Methods." ITEA Conference Paper, 2018.

[57] Director, Operational Test and Evaluation, "Cybersecurity Operational Test and Evaluation Priorities and Improvements." DoD Memorandum, July 2016.

[58] Director, Operational Test and Evaluation, "Procedures for Operational Test and Evaluation of Cybersecurity in Acquisition Programs." DoD Memorandum, April 2018.

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

[60] K. Avery and J. Gilmore, "Applying DOE to Cyber Testing," Informal IDA briefing 2019-24903, April 2019.

[61] S. Whetstone, T. Botting, and R. White, "Cybersecurity and Operational Test and Evaluation: Fundamental Concepts for IDA Research Staff Members," IDA Technical Report NS P-10784, June 2020.

[62] B. Haugh, D. Sparrow, and D. Tate. "The Status of Test, Evaluation, Verification, and Validation (TEV&V) of Autonomous Systems," IDA Technical Report P-9292, November 2018.

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

[64] D. Porter, M. McAnally, C. Bieber, H. Wojton, and R. Medlin, "Trustworthy Autonomy: A Roadmap to Assurance." IDA Technical Report NS P-10768, May 2020.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
02-2021	Final	

4. TITLE AND SUBTITLE

Determining How Much Testing is Enough: An Exploration of Progress in the Department of Defense Test and Evaluation Community

5a. CONTRACT NUMBER

Separate Contract

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Matthew R. Avery (OED); James R. Simpson (OED); Heather M. Wojton (OED)

5d. PROJECT NUMBER

5e. TASK NUMBER

C9082

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

NS D-21561

H 2021-000046

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

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

10. SPONSOR/MONITOR'S ACRONYM(S)

IDA

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for Public Release. Distribution Unlimited.

13. SUPPLEMENTARY NOTES

Project Leader: Rebecca Medlin

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. 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 years, there are still no simple answers to the question, “How much testing is enough?”

15. SUBJECT TERMS

Design of Experiments (DOE); ITEA; Test and Evaluation

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			Rebecca Medlin
Unclassified	Unclassified	Unclassified	Unlimited	20	**19b. TELEPHONE NUMBER** *(Include area code)* 703-845-6731