SYM-AM-24-069

Excerpt from the Proceedings of the Twenty-First Annual Acquisition Research Symposium

Acquisition Research: Creating Synergy for Informed Change

May 8–9, 2024

Published: April 30, 2024

Prepared for the Naval Postgraduate School, Monterey, CA 93943.

The research presented in this report was supported by the Acquisition Research Program at the Naval Postgraduate School.

To request defense acquisition research, to become a research sponsor, or to print additional copies of reports, please contact any of the staff listed on the Acquisition Research Program website (www.acquisitionresearch.net).

ACQUISITION RESEARCH PROGRAM
DEPARTMENT OF DEFENSE MANAGEMENT
NAVAL POSTGRADUATE SCHOOL

Advantages of Using Complex Decision Support Tools in Planning Multi-Modal Test Programs

Milo Taylor—earned a Master of Science degree with the thesis option at Iowa State University where he studied applied mathematics. His thesis research involved implementing numerical methods for solving relativistic fluid problems. He is a modeling and simulation analyst with nearly a decade of experience in radar, autonomy, and electronic warfare. Taylor is currently a research scientist at the Georgia Tech Research Institute (GTRI) in the Electro-Optical Sensor Laboratory (EOSL). His recent work interests include Space EW and cislunar operations. [Milo.taylor@gtri.gatech.edu]

Craig Arndt—currently serves as a principal research engineer on the research faculty of the George Tech Research Institute (GTRI) in the System Engineering Research division of the Electronic Systems Lab. Arndt is a licensed Professional Engineer (PE), and has more than 40 years of professional engineering and leadership experience. Arndt holds engineering degrees in electrical engineering, systems engineering, and human factors engineering and a Master of Arts in strategic studies form the US Naval War college. He served as Professor and Chair of the engineering department at the Defense Acquisition University, and as technical director of the Homeland security FFRDC at the MITRE Corporation. In industry he has been an engineering manager, director, vice president, and CTO of several major defense companies. He is also a retired naval officer [Craig.arndt@gtri.gatech.edu]

David Zurn— received his bachelor's and Master of Science degrees in Electrical Engineering from the Georgia Institute of Technology in 1985 and 1990 respectively. Since joining GTRI's Electronic Systems Laboratory (ELSYS) in 2003, he has worked on a variety of EW-related research efforts including Radar Warning Receiver hardware and software development and test, Missile Warning System hardware and software test, and development of Hardware in the Loop (HITL) test solutions tailored to EW applications. Zurn is currently serving as the Division Chief of the Test Engineering Division within ELSYS. Zurn is a lecturer for the RWR Design short course offered through GT's Professional Education program. His recent research interest areas are Cognitive EW T&E and Space EW T&E. [David.zurn@gtri.gatech.edu]

Abstract

Emerging systems being tested in complex environments require the development of alternate test modalities, including hardware in the loop (HITL) and modeling and simulation (M&S) environments. The investment in these modalities are often significant. For example, testing the survivability of space system uplinks requires difficult over the air (OTA) testing or the development of threat models, orbital models, and propagation models tied together in a HITL or M&S testbed accurately simulating the problem. If properly designed, these testbeds could meet developmental or operational test requirements and potentially be used across a range of space acquisition programs. This research highlights challenges and approaches for developing alternate test modalities and proposes a multi-modal decision support tool for understanding the usage of the testbeds and evaluating tradeoffs. A specific example is explored for the space EW test use case.

Key words: Hardware in the Loop (HITL), Modeling and Simulation (M&S), Decision Support Tools

Executive Summary

This research investigates the challenges associated with evaluating the suitability of alternate test modalities for testing complex systems. We address the difficulty of testing complex systems and recognize that most complex systems are tested in operational test environments, which are referred to as over the air (OTA). Alternate test modalities (HITL and M&S) should also be considered for broader usage. However, it is not always clear where alternate modalities can be used and how much advantage can be achieved with alternate modalities.

A process and decision support tool would be effective for test planners in addressing these uncertainties. A framework is presented for the multi-mode test tool, along with an examination of a space uplink survivability example that is used with a first order implementation of the tool. The example allows us to identify several key uses for the tool:

- Evaluating trades between test objectives (quality, coverage, difficulty) and test modalities
- Understanding test use case to test modality mapping
- Test resource planning—understanding benefits and usages of alternate modalities.

The tool is discussed in the broader context of the system engineering process and common decision support tools. Finally, challenges and a potential way-forward with a tool of this type are discussed.

Background

There are a wide range of methods for testing complex systems throughout their life cycle. The selection of different methods to use at different times has been developed over time and is now incorporated into policies and procedures at different acquisition and test organizations. Most of these practices were established well before the advent of digital engineering processes and do not take into consideration the capabilities of alternate test modalities, including hardware in the loop (HITL) and modeling and simulation (M&S), incorporating digital twins and other digital representations of the system under test.

The introduction of digital engineering has changed both the methods we used to develop defense systems and the timelines associated with the development of those systems. The reduction in the time it takes the Department of Defense (DoD) to develop, test, and field a system is a high priority for the DoD acquisition leadership.

Traditionally, test and verification of a defense system is a complex and time-consuming enterprise. As defense systems continue to grow in complexity, the resources needed to test and verify these systems also grow. There is however a need to reduce, not grow the timelines for testing.

There are a number of different ways that systems can be tested. These different test methods or modes offer the opportunity to verify system performance in different ways. However, the test modes are significantly different. The principle modes are live range testing (referred to as over the air (OTA) testing for this report), HITL, and M&S. Each of these modes have advantages and disadvantages. Moreover, these different test modes can be conducted at different times in the life cycle of the system: development, manufacturing, and operation.

The need to reduce testing time and to accelerate system development and fielding favors early testing using HITL and M&S tools where possible.

The test planner faces challenges in determining the suitability of different modes of testing at different times in the life cycle of the system. Tools and processes are required to help the test planner understand the tradeoffs involved in using alternate modes and the test cases for which each mode is most suitable. If used correctly, these tools can inform test plans that meet program cost and schedule while maintaining high confidence in the performance of these systems.

Test Modalities

Over the Air

Over the air (OTA) testing is a traditional testing mode in which the system under test is placed in a real-world environment. For example, at the Redstone Test Center, the Open-Air Range provides field testing for sensor and seeker systems. Such test ranges allow for the characterization of system targets and environments the system may operate in. System components may also be tested in a controlled environment. An environmental chamber at such a range further provides a real-world simulation of conditions the system under test may face (U.S. Army, 2024). In the context of space vehicles, the live test range may be the actual orbit of the satellite. Uses of OTA for in-orbit testing include future efforts in upgrading sensors on existing satellites whose primary function is space domain awareness (Albon, 2024).

Due to the nature of OTA, the clearest advantage is the level of fidelity afforded by implementing real systems and effectors.

However, considerations for OTA testing stem mostly from its level of fidelity and the fact that the system under test and other test resources exist in the real world. For example, one must consider emissions control (EMCON) when dealing with multiple electromagnetic signals that may interfere or be exposed to unauthorized monitoring.

Hardware in the Loop

Hardware in the loop (HITL) testing is a T&E solution that provides a blend of real-world components and simulation facilities. For example, for nearly 30 years, the U.S. Army Redstone Test Center (RTC) has served as a U.S. Army Test and Evaluation Command to provide T&E support and facilities for various customers. In terms of HITL, the RTC provides T&E for missile seekers by combining traditional T&E with virtualized HITL and M&S environments (U.S. Army, 2024).

An example of HITL in satellite testing is known as a FlatSat. A FlatSat is a “high fidelity electrical and functional representation” of the satellite bus (Amason, 2008). For NASA’s Solar Dynamics Observatory (SDO), it is a test bed for integration and test, flight software, and flight operations.

Some benefits of HITL testing include: ability to perform repetitive tests, non-destructive tests where applicable, and closed-loop testing to minimize external factors. Some benefits of a FlatSat in particular include pre-launch flight software development and verification due to the fact that a physical representation is being tested on the ground in a lab.

Modeling and Simulation

Digital modeling and simulation (M&S) testing is the means of using digital models of the system, its processes, and the environment to test system performance. Digital M&S testing is a mode in which the tests are fully implemented digitally. However, the models of the systems involved and environmental factors are driven by data and intelligence.

Each component of a digital model—whether the system itself or the environment it will operate in—may vary in its scope or fidelity. For example, in the context of DoD applications, M&S software suites may be primarily suited for different fidelities such as mission-level modeling in the Advanced Framework for Simulation, Integration and Modeling (AFSIM) model (*AFSIM, AFRL*, 2024) or at the campaign level as in the Synthetic Theater Operations Research Model (STORM; *STORM, AFRL*, 2023). However, with growing computational capabilities and years of software development, some of these tools may allow the user to operate with others at varied levels of fidelity, or span these levels themselves.

Some advantages of digital M&S testing include the relatively low cost to represent the system and the ability to test under various levels of complexity. Some sources of difficulty can be alleviated if there is a standardization of M&S principles and practices. The recent increased DoD adoption of digital twins for M&S has addressed some of these difficulties and is accelerating usage of M&S for test.

In 2023, the Office of the Director, Operational Test and Evaluation (DOT&E) Strategy Implementation Plan laid out five strategic pillars. In particular, the plan raises Pillar 4, "Pioneer T&E of weapon systems built to change over time" which focuses on standardizing and promoting the use of digital tools. This plan specifically calls for increased usage of digital twins as well as other tools in digital engineering (Director, Operational Test and Evaluation, 2023c).

According to the Digital Twin Consortium, a digital twin is "a virtual representation of real-world entities and processes, synchronized at a specified frequency and fidelity" (Digital Twin Consortium, 2020). Digital twins may come in a myriad of different forms, but an important feature is the synchronization with its real-world counterpart over the system's life cycle.

General Framework for Planning Multi-Modal Tests

We've defined the need for efficient testing as systems grow more complex and acknowledged that modern systems require a mix of test modalities, which we've defined as OTA, HITL, and M&S. Each modality brings with it advantages and disadvantages typically in the form of fidelity, coverage, and cost. The challenge for the test planner is to determine which individual test use cases are best-suited to specific modalities. The current process for developing test plans looks at available test resources and through interaction with experts, determines the best approach for planning individual test use cases and determining the best modality for each use case. This becomes both difficult and inefficient however, as the size and complexity of systems grow. In the following section we propose a general approach for aiding the tester in selecting the "best" test mode for specific test use cases.

The design of any test is a tradeoff between competing objectives such as:

- Quality
    - Fidelity—the level of detail which the test replicates in the operational environment that the system will be operating in.
    - Repeatability—the ability to produce the same results if the test is conducted multiple times with no change in parameters
    - Reliability/Confidence—the measure of how well results represent the real world and the sensitivity to external factors—a function of the number of test data points collected (this is determined as experimental design)
- Coverage—the part or percentage of the system performance envelope that the test verifies.
- Difficulty
    - Cost—a measure of the affordability of a test
    - Schedule—a measure of the timeliness of a test
    - Risk—an assessment of whether the test will function as intended and provide usable data

An ideal test program will experience high quality, extensive coverage, with low difficulty. This is not achievable because some factors improve at the expense of others. The fundamental tradeoffs most designers encounter due to test resource limitations are:

- As quality increases, coverage generally decreases
- As quality increases, difficulty increases

- As coverage increases, difficulty increases

Trade-offs are always present in a given test design, particularly for the specific test mode chosen. Moreover, we can generally say that OTA, HITL, M&S test modes typically come with the following objective tradeoffs:

1. OTA—quality high, coverage low, difficulty high
2. HITL—moderate quality, moderate coverage, moderate level of difficulty
3. M&S—lowest quality, highest coverage, lower difficulty.

Test resource planners should understand these objective trade-offs to determine the "best" mode for each use case. The following framework is proposed to aid in this determination.

Define a Use Case set, composed of a set of Test Categories:

USE_CASE (i,j,k) = {FUNC(i), ENV(j), ENG(k)}

Where Test Categories are defined as:

CAT = {FUNC, ENV, ENG} with

FUNC = {$func_1$, $func_2$, … $func_x$}, a set of Functions or Functional modes,
ENV = {env_1, env_2, … env_y}, a set of Environmental variants,
ENG = {eng_1, eng_2, … eng_z}, a set of Engagement variants.

Next define a set of Test Modes over which to evaluate the Use cases:

MODE = {OTA, HITL, M&S}.

Finally, define a Test Objective set that supports the evaluation of use cases. The Test Objective set is defined as:

OBJ = {Quality, Coverage, Difficulty}.

Note that we've simplified Test Objectives for this general framework. In reality, the Quality and Difficulty objectives should be decomposed into the components described above, scored, and combined to provide overall Quality and Difficulty scores.

To begin the evaluation, each combination of Test Category (CAT) and Mode (MODE) are scored for each Objective (OBJ). As an example, a score would be assigned to func1, using the OTA Mode, for the Quality test objective. This scoring would be performed for all members of the FUNC, MODE, OBJ sets, ENV, MODE, OBJ sets, and ENG, MODE, OBJ sets, creating a three-dimensional scoring array SCORE with CAT, MODE, and OBJ vectors. A representation of the array is shown below with Test Category and Mode shown for each Objective.

Quality Scores		Mode		
		OTA	HITL	M&S
CAT	func1	Q_{1O}	Q_{1H}	Q_{1M}
	.	Q_{2O}	Q_{2H}	Q_{2M}
	env1	.	.	.
	.	.	.	.
	eng1	.	.	.
	.	Q_{nO}	Q_{nH}	Q_{nM}

Coverage Scores		Mode		
		OTA	HITL	M&S
CAT	func1	C_{1O}	C_{1H}	C_{1M}
	.	C_{2O}	C_{2H}	C_{2M}
	env1	.	.	.
	.	.	.	.
	eng1	.	.	.
	.	C_{nO}	C_{nH}	C_{nM}

Difficulty Scores		Mode		
		OTA	HITL	M&S
CAT	func1	D_{1O}	D_{1H}	D_{1M}
	.	D_{2O}	D_{2H}	D_{2M}
	env1	.	.	.
	.	.	.	.
	eng1	.	.	.
	.	D_{nO}	D_{nH}	D_{nM}

Figure 1. Scoring Array

Next a set of Use cases, combining the FUNC, ENV, and ENG Test categories is created. The total number of Use cases depends on the number of combinations of the variants in each Test category. A simple example might be a system with four discrete Functional modes

(FUNC), three Environmental variants (ENV), and two Engagement (ENG) variants. In this case the system would have a maximum total of 512 (2^9) total potential use cases. Some combination of variants might not make sense so this is the maximum number of Use cases. For each use case, a score is calculated for each objective. The score combines the Scoring array entries for the Test Categories included in that Use case. Scores are calculated as follows:

Quality_Score (Use Case, Mode) = SCORE (FUNC (Use case), Mode, Quality) + SCORE (ENV (Use case), Mode, Quality) + SCORE (ENG (Use case), Mode, Quality)

Coverage Score (Use Case, Mode) = SCORE (FUNC (Use case), Mode, Coverage) + SCORE (ENV(Use case), Mode, Coverage) + SCORE (ENG(Use case), Mode, Coverage)

Difficulty_Score (Use Case, Mode) = SCORE (FUNC (Use case), Mode, Difficulty) + SCORE (ENV (Use case), Mode, Difficulty) + SCORE (ENG (Use case), Mode, Difficulty)

Finally, a Utility score is calculated for each Mode in each Use case. The Utility score is a weighted sum of the objective scores defined above.

Utility (Use Case, Mode) = Quality_Score (Use Case, Mode)*w1 + Coverage_Score (Use Case, Mode)*w2 + Difficulty_Score (Use Case, Mode)*w3

The w1, w2, w3 weight values are determined based on the test type and system complexity. For example, training, developmental testing (DT), and operational testing (OT) each have increasing weight placed on quality with OT designated as highest required quality. Increasing system complexity might place higher weight on coverage.

For each use case, the Utility scores for each of the three Test modes are compared. The highest Utility score indicates the "optimal" test mode for that use case.

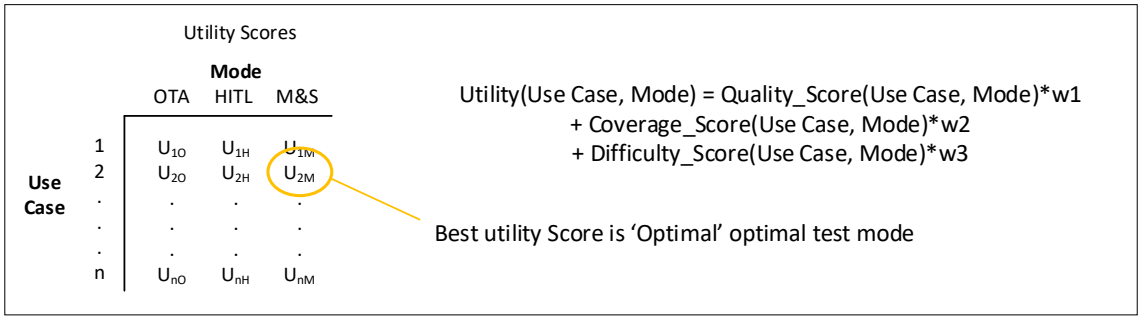

Figure 2. Total Utility Function

A simplified schematic of the process is shown below.

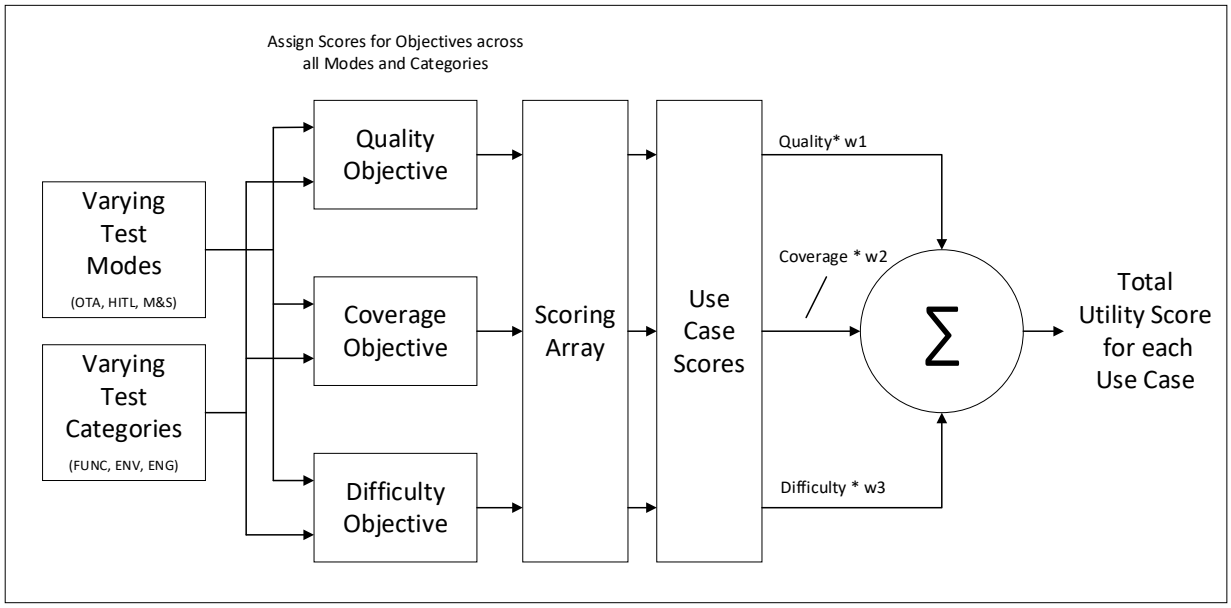

Figure 3. Simplified Utility Scoring Schematic

Multi-Modal Test Tool—Uplink Survivability Test Example

Counterspace threats come in various forms ranging from direct-ascent anti-satellite weapons to cyber-attacks. Electronic warfare (EW) poses a particularly unique set of challenges to successful operations in space. Indeed, the most modern militaries consider EW to be an essential facet of warfare, and have incorporated jamming and anti-jamming counterspace capabilities (Defense Intelligence Agency, 2022). Likewise, military powers have incorporated EW to secure navigational and informational superiority. In offensive electronic warfare, the objective is to disrupt, deny, degrade, destroy, or deceive communications or target acquisition.

At a basic level, a satellite communications (SATCOM) set up is composed of three segments (NASA, 2024):

1. Space segment: a collection of space vehicles
2. Link segment: the functional segment consisting of signals from the ground (uplink), transmitting data down to the ground (downlink), and transmitting and receiving data to and from other satellites (crosslink)
3. Ground segment: assets located on the ground (or in sometimes air, land, and sea) such as ground control or user terminals

The ground segment may be decomposed further into the control terminals, user terminals, and infrastructure.

An adversarial actor may interfere with SATCOM by introducing jamming or spoofing. Uplink survivability (ULS) testing determines how well the satellite under test (SUT) performs in the presence of such jamming of the uplink signals. In uplink jamming, a threat system specifically interferes with a signal originating from the ground segment meant for the space segment. The purpose is to deny or degrade the reception at the satellite receiver in order to prevent communication, increase error rates, or decrease throughput.

ULS Test Description

The SUT is a blue satellite in orbit (any orbit type) capable of transmitting and receiving either a data link or telemetry, tracking, and command (TT&C). The emphasis is on testing and evaluating the system's operational capability in the presence of a corrupted uplink signal.

On the blue side, the ULS test case is comprised of the SUT (a representation of the space vehicle, or the vehicle itself), a ground control terminal, relevant infrastructure depending on the test modality, test control instrumentation, and possibly a user downlink terminal.

On the threat side, the ULS test case is comprised of at minimum a simulator for the threat jammer. The threat jammer may employ either basic noise jamming, or more advanced techniques. The uplink survivability test case may also be generalized to consider the survivability of a constellation of satellites against multiple sources of interference for an M v N engagement.

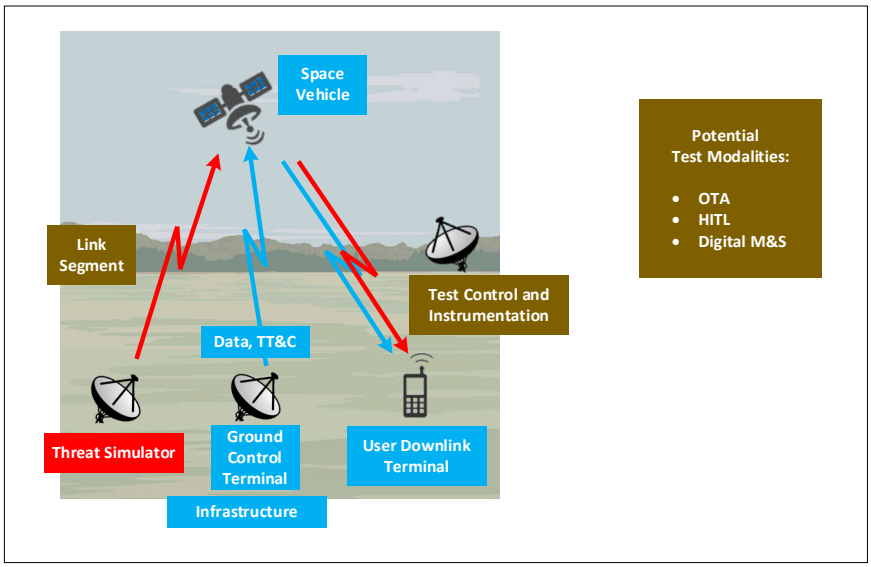

Figure 4. Basic ULS Test Components

Framework Applied to ULS test

Next, the general framework for planning multi-modal tests defined above will be applied to the ULS test problem. Recall, the framework is intended to identify optimal test modes for each test use case.

The first step is to establish use cases for the ULS test example. A use case is a basic test case composed of categories that reflect a specific operating condition of the test article. An example for ULS might be testing a data link in the presence of basic interference, in a clear environment, with a single interference source. We can designate the following categories for ULS testing:

- Function—Data link, TT&C link
- Interference Type—Basic, Advanced
- Environment—Clear, Obscured
- Engagement—1v1, M v N

Note that we've added Interference Type to the three test categories defined above (Function, Environment, Engagement). Interference could have been added as an additional sub-category of environment. We chose to break it out into its own category because it's a key component in

ULS testing. Note also that this is a simplified example—a real test design would incorporate a wider variety of types for each category.

Next, test objectives are established. For this example, we'll use Quality, Coverage, and Difficulty as first order test objectives. As noted above, formal detailed analysis would decompose these into sub-objectives (for example, Quality would be decomposed into Fidelity, etc.) for a more accurate assessment.

The method described above is used for scoring each objective based on the Function, Interference Type, Environment, and Engagement categories. The scoring is done for the three test modes, for each category. The actual scoring should be done by test designers with knowledge in the test domain and knowledge in the three test modes. For this exercise, we scored by assigning numeric values from 1–9. To simplify this initial analysis, the three modes were ranked with 7 assigned to the highest-ranking mode, 5 assigned to the next highest, and 3 assigned to the lowest mode. Scores of 1 or 9 were assigned to "edge cases" where an extreme score is justified.

	Use Case Category																							
	Link						Interference						Environment						Engagement					
	Data			Control			Basic			Advanced			Clear			Obscured			1v1			MvN		
Objective	**O**	**H**	**M**	**O**	**H**	**M**	**O**	**H**	**M**	**O**	**H**	**M**	**O**	**H**	**M**	**O**	**H**	**M**	**O**	**H**	**M**	**O**	**H**	**M**
Quality	9	5	3	9	5	3	9	5	3	9	7	3	7	5	3	3	7	3	7	5	3	1	5	5
Coverage	3	5	7	3	5	7	3	5	7	3	5	7	3	5	7	3	5	7	3	5	7	1	1	9
Difficulty	1	5	7	1	5	7	3	7	5	1	7	5	1	5	7	1	5	7	3	5	7	1	1	9
Scoring:	Rate 7,5,3 with 7 = best; allow extremes for edge cases (9,1)																							

Figure 5. Scoring Array for ULS Example

Next, intermediate scores are calculated for each use case, for Objective and Test mode, as shown in Figure 6 below. The intermediate score sums the Scoring array entries for the Test Categories included in that use case.

In the final step, we calculate Utility values for each use case by combining scores determined for each Category and Mode in the Scoring Array. The Total Utility is a simple weighted sum of Category scores for each Objective for each Mode as described in the framework description above. See Figure 6 for results using even weighting values (w1=w2=w3=5) for each objective.

Use Case					Intermediate Scores									Total Utility Scores		
					Quality			Coverage			Difficulty					
Use Case #	Function	Interference	Environment	Engagement	OTA	HITL	M&S	OTA	HITL	M&S	OTA	HITL	M&S	OTA	HITL	M&S
1	Data Link	Basic	Clear	1v1	32	20	12	12	20	28	8	22	26	260	310	330
2	Data Link	Advanced	Clear	1v1	32	22	12	12	20	28	6	22	26	250	320	330
3	Control Link	Basic	Clear	1v1	32	20	12	12	20	28	8	22	26	260	310	330
4	Control Link	Advanced	Clear	1v1	32	22	12	12	20	28	6	22	26	250	320	330
5	Data Link	Basic	Obscured	1v1	28	22	12	12	20	28	8	22	26	240	320	330
6	Data Link	Advanced	Obscured	1v1	28	24	12	12	20	28	6	22	26	230	330	330
7	Control Link	Basic	Obscured	1v1	28	22	12	12	20	28	8	22	26	240	320	330
8	Control Link	Advanced	Obscured	1v1	28	24	12	12	20	28	6	22	26	230	330	330
9	Data Link	Basic	Clear	MvN	26	20	14	10	16	30	6	18	28	210	270	360
10	Data Link	Advanced	Clear	MvN	26	22	14	10	16	30	4	18	28	200	280	360
11	Control Link	Basic	Clear	MvN	26	20	14	10	16	30	6	18	28	210	270	360
12	Control Link	Advanced	Clear	MvN	26	22	14	10	16	30	4	18	28	200	280	360
13	Data Link	Basic	Obscured	MvN	22	22	14	10	16	30	6	18	28	190	280	360
14	Data Link	Advanced	Obscured	MvN	22	24	14	10	16	30	4	18	28	180	290	360
15	Control Link	Basic	Obscured	MvN	22	22	14	10	16	30	6	18	28	190	280	360
16	Control Link	Advanced	Obscured	MvN	22	24	14	10	16	30	4	18	28	180	290	360

Test Objective	Weight
Quality	5
Coverage	5
Difficulty	5

Figure 6. Use Cases and Utility Values for ULS Example—Equal Objective Weighting

ULS Test Case Observations

Green highlighted cells indicate optimal Test Modes for each use case. Red highlighted cells indicate lowest total score for each use case. Note that in this example, all of the use cases selected M&S as the optimal mode. Closer examination of the intermediate scores show that the M&S modes scored consistently higher for Coverage and Difficulty. This, combined with the fact that all Objectives were weighted equally (at 5), drives consistently higher total scores for M&S.

Based on this first example we can see that total utility scores heavily depend on Objective weightings. To explore this further, a first order sensitivity analysis by weighting was conducted. This sensitivity analysis yields the following results for test modes with highest utility score:

1. **Evenly weighted**—M&S for nearly all use cases, with several HITL
2. **Heavy weighting towards Quality**—mix of OTA, HITL use cases
3. **Heavy weighting towards Coverage**—all M&S
4. **Heavy weighting towards Difficulty**—all M&S
5. **OT weighting** (high Quality, low Coverage, medium Difficulty)—mix of OTA, HITL, M&S
6. **DT weighting** (high Quality, high Coverage, medium Difficulty)—mix of OTA, HITL, M&S

Case		Weight			# Modes w/Highest Utility Score		
		Quality	Coverage	Difficulty	OTA	HITL	M&S
1	Evenly weighted	5	5	5	0	2	16
2	Quality weighted	8	2	2	8	8	0
3	Coverage weighted	2	8	2	0	0	16
4	Difficulty weighted	2	2	8	0	0	16
5	Operational Test	8	2	3	4	10	2
6	Developmental Test	7	4	3	2	6	8

Figure 7. Sensitivity Analysis Showing Variance in Optimal Mode Selection as a Function of Weighting

The sensitivity analysis underscores the basic Fidelity versus Coverage tradeoff inherent in most tests. We expect that if the Fidelity (Quality) is heavily weighted and coverage is rated low (Case 2 in table above), then OTA would be a preferred choice. If Coverage is heavily weighted (Case 3) then M&S is the preferred choice.

The results are shown in Figure 8 for the set of Operational Test weights, which reflect a more realistic weighting scheme.

Use Case					Intermediate Scores									Total Utility Scores		
					Quality			Coverage			Difficulty					
Use Case #	Function	Interference	Environment	Engagement	OTA	HITL	M&S	OTA	HITL	M&S	OTA	HITL	M&S	OTA	HITL	M&S
1	Data Link	Basic	Clear	1v1	32	20	12	12	20	28	8	22	26	304	266	230
2	Data Link	Advanced	Clear	1v1	32	22	12	12	20	28	6	22	26	298	282	230
3	Control Link	Basic	Clear	1v1	32	20	12	12	20	28	8	22	26	304	266	230
4	Control Link	Advanced	Clear	1v1	32	22	12	12	20	28	6	22	26	298	282	230
5	Data Link	Basic	Obscured	1v1	28	22	12	12	20	28	8	22	26	272	282	230
6	Data Link	Advanced	Obscured	1v1	28	24	12	12	20	28	6	22	26	266	298	230
7	Control Link	Basic	Obscured	1v1	28	22	12	12	20	28	8	22	26	272	282	230
8	Control Link	Advanced	Obscured	1v1	28	24	12	12	20	28	6	22	26	266	298	230
9	Data Link	Basic	Clear	MvN	26	20	14	10	16	30	6	18	28	246	246	256
10	Data Link	Advanced	Clear	MvN	26	22	14	10	16	30	4	18	28	240	262	256
11	Control Link	Basic	Clear	MvN	26	20	14	10	16	30	6	18	28	246	246	256
12	Control Link	Advanced	Clear	MvN	26	22	14	10	16	30	4	18	28	240	262	256
13	Data Link	Basic	Obscured	MvN	22	22	14	10	16	30	6	18	28	214	262	256
14	Data Link	Advanced	Obscured	MvN	22	24	14	10	16	30	4	18	28	208	278	256
15	Control Link	Basic	Obscured	MvN	22	22	14	10	16	30	6	18	28	214	262	256
16	Control Link	Advanced	Obscured	MvN	22	24	14	10	16	30	4	18	28	208	278	256

Test Objective	Weight
Quality	8
Coverage	2
Difficulty	3

Figure 8. Use Cases and Utility Values for ULS Example—OT Objective Weighting

Basic trends in the intermediate scores correspond with our understanding of tradeoffs associated with the basic test modes:

- For the Quality objective, OTA scored highest—this is expected given that OTA most closely resembles real-world conditions. Note that scores for OTA Quality decrease for obscured use cases and MvN use cases because these are more difficult to replicate in the OTA mode.
- For the Coverage objective, M&S scored highest for all uses cases. It is expected that properly constructed M&S environments should provide the best coverage.
- For the Difficulty objective, M&S scored highest for all use cases. It is assumed that once the M&S environment is set-up, the difficulty associated with running these tests is lowest for all test modes. Note that these scores assume that the M&S environment has been constructed, Blue and Red models developed, and the combined environment has been verified and validated. This is likely a faulty assumption for ULS M&S tools given the current state of M&S tool development in the EW domain.

The optimal test modes selected for test use cases make intuitive sense when considering each use case in detail:

- Use cases 1–4 selected OTA because Quality was scored highly because these use cases called for a clear environment.
- Use cases 5–8 selected HITL because these cases called for an obscured environment, which is difficult to achieve consistently in an OTA mode, but can be more readily simulated in a HITL setup. This drove Quality scores for HITL higher and OTA lower.

- Use cases 9–16 selected HITL and M&S because these use cases require MvN engagement which disadvantages OTA scoring. The MvN use cases require multiple orbital SUTs and interference sources, which are difficult to achieve with an OTA test.

Multi-Modal Test Tool—General Observations

The multi-modal test tool we've presented is suitable for evaluating basic trades between Quality, Coverage, and Difficulty for different test modes and test use cases. It is not intended for detailed test planning such as assigning specific use cases to specific modes. More granularity is required for the test use case definition. This should be done by carefully examining the basic test scenario and defining a wider range of test categories, in much greater detail. Additionally, the Objective functions need to be refined. As presented, "Quality" encompasses a variety of components (Fidelity, Repeatability, Confidence/Reliability) as does Difficulty. These need to be split into distinct attributes that can be scored separately, then combined into the appropriate objective. Note however that the number of objectives used to calculate the total utility function needs to remain small. Adding additional, non-critical objectives to the utility function will dilute the effect of critical objectives in expressing utility.

A more fully developed multi-modal test tool could be used by test planners to:

- Understand the key modalities required for a test campaign, which would drive test planning and near-term test resource development
- Assign specific use cases to specific test modes/resources as part of test planning
- Understand the impact of emphasizing one test objective over another (i.e., Quality versus Coverage, etc.)

It is essential to recognize, however, that the tool is entirely dependent on subject matter experts (SMEs) providing accurate scoring. Fundamentally the tool relies on the SMEs to score individual categories versus objective functions. It does not capture the actual relationship between a category and objective function. Indeed, subtle tradeoffs such as fidelity versus coverage are captured through SME scoring, not through tool design.

Given the critical nature of scoring in achieving reliable results, users of the tool need to pay close attention to scoring methods. SMEs should be carefully chosen, and should independently assign scores which are then compared for variance. If the variance between SME scoring is significant then the scoring should be re-evaluated. Other methods for validating scoring should be developed as well.

The tool relies on the user assigning appropriate weighting to the Quality, Coverage, and Difficulty objectives. The utility function is heavily dependent on weighting. In fact, a biased user of the tool can achieve whatever result is desired by arbitrarily adjusting weights. It is recommended that prior to scoring, careful consideration be given to establish weighting values appropriate to the test application. Application of this tool for an operational test may prioritize Quality over Coverage and Difficulty, but a test designed for science and technology (S&T) application may favor lower Difficulty at the expense of Quality.

The multi-modal test tool could also be used for long term test resource development. Consider an acquisition organization developing a new capability and trying to determine test resources required. The tool could be used to understand the Quality, Coverage, and Difficulty trades associated with the resource test modes. This could drive resource planners to develop specific solutions in key modes as indicated by the tool. It has been observed that test resource developers generally support investment in OTA resources but are reluctant to invest in HITL and M&S resources because it is difficult to see the value provided by these alternate modes. This tool illuminates the trade-space for the three test modes, exposing benefit for alternate

modes. Additionally, the tool may help more directly with resource acquisition—understanding use cases and test mode mapping can aid in developing capabilities, needs, and requirements for the test environments associated with these modes.

To facilitate its use for resource development, it may make sense to develop a "Development Difficulty" objective associated with test resource development. The Difficulty objective described above relates to difficulty in actually performing the test and not in developing a resource. Introducing the Development Difficulty objective ensures that the utility of each mode also reflects development difficulty. This may not be an issue for existing test resources, and if that's the case, the user can weight Development Difficulty at zero. If, however, a resource does not exist or significant effort is required to develop components of a resource, the overall utility of the resource is negatively impacted. For example, using M&S may provide great coverage with little relative difficulty in running the test. Yet if significant effort and risk is involved in developing and validating Blue and Red models, the M&S mode may not in reality provide much utility. Indeed, the user of the tool can be misled about the utility of various modes if they are assumed to provide benefit but end up requiring substantial resources to develop.

Multi-Modal Test Tool in the System Engineering Context

The systems engineering process is critical to all aspects of the development and testing of DoD systems. Within this process are a number of key steps which include requirements development, design, and test. Testing and verification are critical to the system engineering process because the testing and evaluation are needed in the design and development process and also needed to verify the performance of the end product and the manufacturing of the system before deployment. As a result, there are a wide range of tests that are conducted throughout the system life cycle. Testing and evaluation can be done in many different ways and times. To better understand the scope of different means of testing, we can look at testing along several different dimensions. First look at the purposes for testing. Testing can be performed to a) determine subsystem performance in design, b) determine overall system performance in design, and c) verify performance of the system for operational suitability, survivability, and effectiveness. Second, different actors conduct the testing, including designers, manufacturers, and different government organizations (including users). Third, testing is conducted at different times, including before the program starts (for legacy system parts), during development, during formal system verification, and after the system is in operation. The multi-modal test tool should be considered for test planning for all of these test types.

The key element of the defenses systems engineering process are captured in the systems engineering "V," as shown in Figure 9.

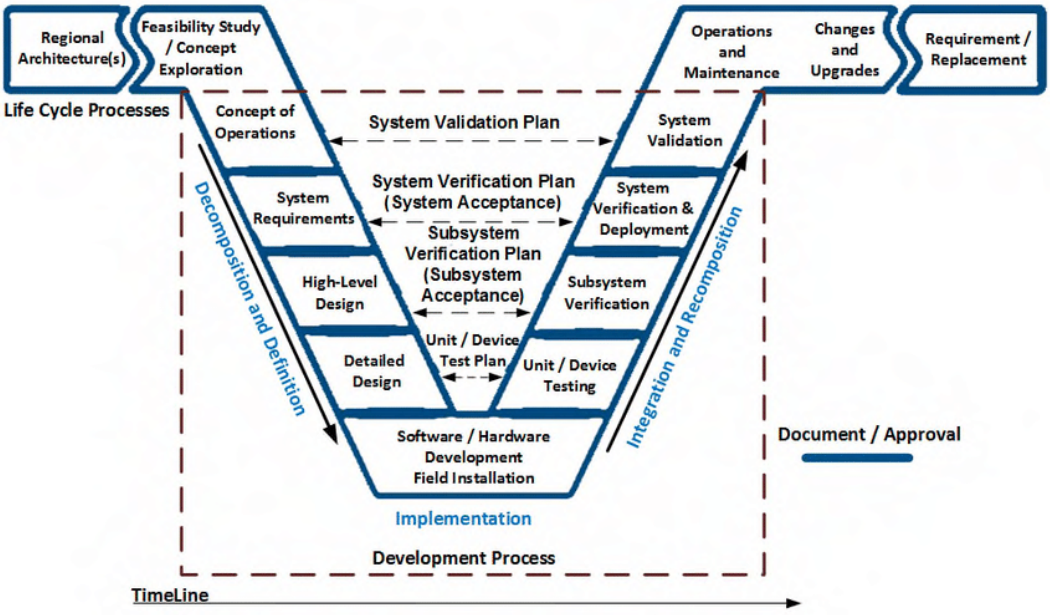

Figure 9. System Engineering "V"

As we can see from the systems engineering "V," the test process needs to be engaged throughout the life cycle. Although most tests and other verification activities are conducted after the system has been designed and built, the planning for test and verification should be an integral part of the development of system requirements and all aspects of the design and realization process. This leads to opportunities to do different tests very early in the development process and different modeling and simulation opportunities.

Decision Support Tools

The multi-modal test tool is a simple form of a more complex decision support tool. It may be beneficial to apply lessons learned from existing decision support tools as the multi-modal tool is more fully developed. Some basic background relating to decision support tools is presented below.

Decision support tools are used across a wide range of different domains to help analyze different courses of action. Over the past few years decision support tools have advanced significantly. The most common technology that has emerged is multi-dimensional decision frontiers. This mathematical analysis allows the user to evaluate complex multi-dimensional trade spaces. Trade space analyses are needed to support key decision makers, and some questions critical to informing these decisions are not well-addressed via traditional, more globally focused analyses. Systems engineering questions unique to a given system or problem will often require similarly unique analytical workflows supported by contextually relevant data. Multiple specific systems engineering insights can be gleaned from exploration of specific analysis pathways rather than over-simplified global analysis. To address this issue, the Georgia Tech Research Institute (GTRI) has sought to tie analytical components (building blocks such as sensitivity analyses, regression models, etc.) to data pipelines relevant to the question we are trying to ask. The question in the case of optimizing different combinations of testing methods and testing times is what combination of test parameters will maximize the quality of the test while minimizing cost and schedule.

As it is used here, context can refer to how the additive value of a system varies between stakeholders or temporal differences in a system's application over its life cycle that impact its perceived usefulness. A major understanding from GTRI's efforts is that generating a trade space from various models is not a trivial task if the goals are to achieve flexibility, scalability (often via properly orchestrated modularity), and efficiency of the process. Also, a use case has a specific path through a networked workflow. In addition, these goal characteristics are often defined according to life cycle stage or blur across several stages—care must be taken to operationalize appropriately. Specifying the precise way in which any analytical construct applies to trade space analysis and also its specific life cycle context is critical to future synthesis with other methods. Composability and traceability of constructs is key to future maturation using other methods in tandem. GTRI discovered through this work that the degree of modularity and the extent of the abstract description necessary to define the problem in a way that is directly executable is strongly linked and tremendously important to usability by a person and reusability in a computational environment. An example decision support tool is included in Figure 10 (Ender, 2014).

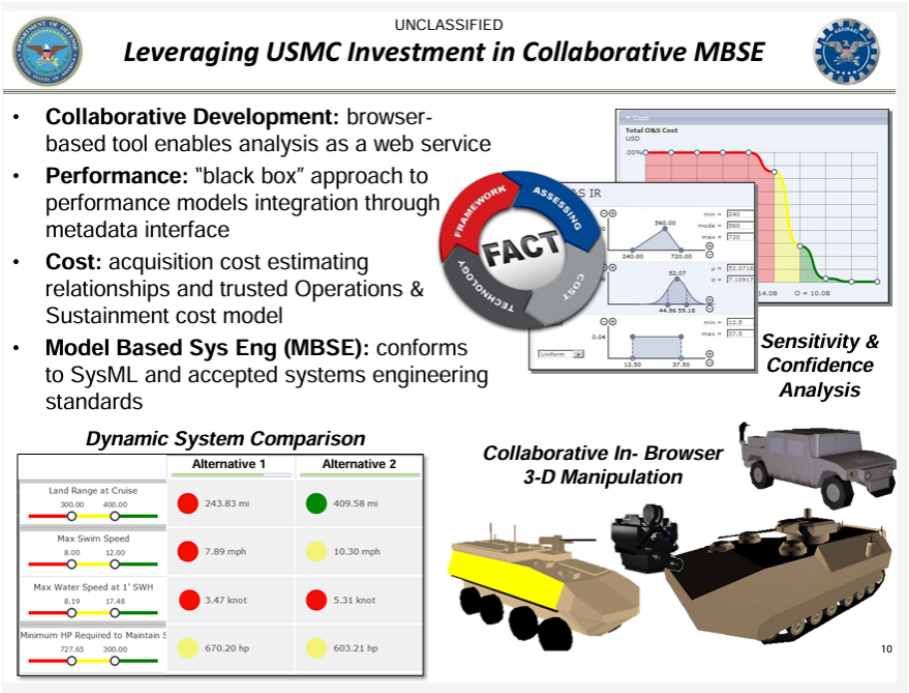

Figure 10. Example Decision Support Tool

Challenges

Several challenges must be addressed to more fully develop the multi-mode test tool.

The example tool developed for this research and presented above is specific to the ULS test application. How readily can the tool be generalized for other test applications? Ideally a fully developed tool could be re-used for different test applications without major re-work. It appears that the framework, process, and objective functions are generizable. However, the specific test use cases and their component categories are unique to a given test application and any future version of this tool must allow the user to specify use cases and categories in an efficient way. Interestingly, the basic OTA, HITL, and M&S test modes as defined may vary depending on system complexity and type. For this research we've implicitly defined the test modes according to their degree of virtualization:

- OTA—comprises no virtualization (as close to real world as possible)
- HITL—comprises no virtualization of the test article, with other elements (environment, interference) virtualized
- M&S—virtualizes all components

In reality, as a system gets more complex, the OTA-HITL-M&S distinctions may become a continuum of modalities where various components of the system or parts of the test scenario (aspects of the environment for example) are virtualized in a way that makes sense for efficient testing. The tool may need to be designed to account for this.

Another critical challenge is performing verification and validation (V&V) for the decision support tool. Verification (establishing that the tool is performing the way it's been designed) should be straightforward, using synthetic data. Validation on the other hand needs to establish that the tool is meeting the needs of the test planner. This involves determining whether the methods and processes underlying the tool are providing useful predictions. Ideally, validation would compare the tool test mode recommendations to historical data, but specific use case/test mode mapping data may be limited for specific test applications. Alternatively, independent review of tool output across a variety of test applications may be required for validation.

Conclusion

A multi-modal test decision support tool could be effective for aiding test planning and test resource development planning for complex systems. A practical framework has been created for a multi-mode test tool, which if developed into a formal decision support tool, could be used to:

- Evaluate trades between test objectives (Quality, Coverage, Difficulty) and OTA, HITL, and M&S test modalities.
- Understand the test use case to test modality mapping.
- Aid in test resource planning by highlighting benefits and usages of alternate modalities.

The authors recommend continuing this research by fully developing the ULS test tool presented here. There is an opportunity to collaborate with test planners and test resource acquisition professionals who are currently engaged in determining ULS test methods and doing specific ULS test planning. These SMEs could provide inputs for tool development and scoring. Ideally the tool outputs would help to inform their decision making and lead to more efficient testing. The team should also leverage existing decision support tool research, identifying well-developed frameworks and interfaces, enabling the efficient development of a multi-mode test tool.

References

AFSIM, AFRL. (2024). https://www.wpafb.af.mil/News/Art/igphoto/2001709929/

Albon, C. (2024, March 27). *Space Force to upgrade sensors for in-orbit testing, training*. C4ISRNET. https://www.c4isrnet.com/battlefield-tech/space/2024/03/27/space-force-to-upgrade-sensors-for-in-orbit-testing-training/

Amason, D. (2008). *SDO FlatSat facility*. Goddard Space Flight Center.

Bingen, K. A., Johnson, K., & Young, M. (2023). *Space threat assessment 2023.* Center for Strategic & International Studies.

Defense Intelligence Agency. (2022). *2022 challenges to security in space: Space reliance in an era of competition and expansion.*

Defense Science Board. (2021). *GEMS: Gaming, exercising, modeling, & simulation.* Department of Defense.

Department of Defense. (2007). *Directive 5000.59.*

Digital Twin Consortium. (2020, December 3). *Digital Twin Consortium defines digital twin*. https://www.digitaltwinconsortium.org/initiatives/the-definition-of-a-digital-twin/

Director, Operational Test & Evaluation. (2023a). *FY 2023 annual report.*

Director, Operational Test & Evaluation. (2023b). *Test and evaluation resources.*

Director, Operational Test & Evaluation. (2023c). *DOT&E strategy implementation plan—2023.*

Ender, T. (2014). Engineered resilient systems tradespace enabled decision making. *Engineered Resilient Systems Tradespace Enabled Decision Making*.

NASA. (2024, February 12). *State-of-the-art of small spacecraft technology*. https://www.nasa.gov/smallsat-institute/sst-soa/soa-communications/

National Institute of Standards and Technology. (2021). *Considerations for digital twin technology and emerging standards.* https://nvlpubs.nist.gov/nistpubs/ir/2021/NIST.IR.8356-draft.pdf

Office of the Undersecretary of Defense for Research and Engineering. (2023). *DoD instruction 5000.97 digital engineering.*

Rohde & Schwarz. (2023). *An overview of space electronic warfare.*

STORM, AFRL. (2023). https://intelligencecommunitynews.com/air-force-to-host-storm-industry-day-2023

Systems Engineering Research Center. (2019). *Technical report SERC-2019-TR-012.*

U.S. Army. (2024, April 2). *Test capabilities*. https://www.atec.army.mil/rtc/resources.html

Wright, M. (2008). Lunar reconnaisance orbiter FlatSat. *IEEE 13th European Test Symposium IEEE Computer Society and Test Technology Technical Council .* Verbania, Italy: Institute of Electrical and Electronics Engineers.

Acquisition Research Program
Department of Defense Management
Naval Postgraduate School
555 Dyer Road, Ingersoll Hall
Monterey, CA 93943

www.acquisitionresearch.net