MITRE

MTR210454

MITRE TECHNICAL REPORT

Advance M&S in Acquisition T&E

Sponsor: OUSD(RE) DTE&A
Division: OSD Division
Department: N222- Defense Systems Engineering
Contract No.: W56KGU-18-D-0004

McLean, VA

Authors: Luis A. Cortes
Melissa Kim Wong
Angel Cortes-Morales
David Wells
Jason Daly

19 November 2021

This page intentionally left blank.

Executive Summary

This document provides a characterization of the current state of modeling and simulation (M&S) in acquisition test and evaluation (T&E) as viewed from the lens of a small number of acquisition programs as well as recommendations to advance M&S in acquisition T&E. *Advance M&S in acquisition T&E* is an initiative (Initiative # 13) established in FY21 by Director, Developmental Test, Evaluation, and Assessment [D(DTE&A)] to help acquisition programs think about how to employ M&S earlier and more often in the acquisition lifecycle to improve the delivery of capability. Initiative #13 is rooted in three of the ten D(DTE&A)'s *Test Vision Key Takeaways*:

- M&S will continue to grow as a critical component of an overall test program to explicitly establish close alignment with M&S validation, verification, and accreditation (VV&A) activities via collection and analysis of relevant objective quality evidence (OQE).
- Ongoing digital engineering activities and model-based systems engineering (MBSE) methods present significant opportunities to leverage early capability insight via M&S analysis and testing but presents challenges on how to fully leverage automated testing within traditional testing frameworks.
- Early investment and validations of threat characterizations (M&S and live fire) will be critical to establishing confidence in system effectiveness via testing.

The envisioned goals of Initiative #13 are to:

- Improve the understanding of how acquisition programs use M&S across the system's acquisition lifecycle.
- Help shape policy and guidance for M&S in T&E, particularly for developmental test and evaluation (DT&E).
- Improve the use of statistical engineering techniques for M&S VV&A, including uncertainty quantification methodologies.
- Improve the relevance of M&S in decision support frameworks.
- Help shape policy and guidance for assessing model maturity.
- Establish an M&S framework for DT&E.
- Identify digital engineering strategies to improve M&S in T&E.

M&S assists developers and decision makers in a wide range of technical processes such as analysis of alternatives, developing the system concept, requirements evaluation, production and manufacturing, test and evaluation, systems integration, training, logistics, risk management, experimentation, and assessing the entire capability space. Because of the focus in the Department of Defense (DoD) for delivering integrated, network-centric systems-of-systems (SoS) that provide the material solution of the needed capability, the effective use of M&S in systems engineering has become essential in meeting those challenges. For instance, M&S is a catalyst for the Service's portions of Joint All-Domain Command and Control (JADC2)—the Navy's Project Overmatch, the Army's Project Convergence, and the Air Force's Advanced Battle Management System (ABMS).

The use of M&S and T&E is characterized by an interdependency. On one side, data from test events is used to inform the development and sufficiency of the M&S. On the other side, data collected during M&S events is used to inform T&E planning, assessments, and critical decisions in the system's lifecycle. This document focuses on that interdependency.

M&S will continue to be a critical component of an overall test program strategy. To minimize the risks associated with M&S-informed decisions, the interdependency between T&E and M&S must be strengthened—continue to focus on improving the effort of using data from test events for the development and VV&A of the M&S as well as using data collected during M&S events to inform T&E assessments and critical decisions in the lifecycle. This report provides detailed recommendations to strengthening those relationships. In summary, MITRE recommends DTE&A collaborate with Director, Operational Test and Evaluation (DOT&E), Office of the Undersecretary of Defense for Research and Engineering (OUSD(R&E)), or Services, as appropriable, to:

- Encourage acquisition programs to improve the content of the M&S Catalog on models, testbeds, and their uses to allow for a better re-use of models across programs.
- Update the DOT&E Test and Evaluation Master Plan (TEMP) Guidebook with consistent test design guidance for live-fire test and evaluation (LFT&E), DT&E, and operational test and evaluation (OT&E), particularly for M&S-driven tests.
- Adopt the recommendations provided to DTE&A Chief Engineer (provided via separate correspondence) to update the engineering guidebooks.
- Identify and sponsor opportunities to improve the T&E workforce knowledge, skills, and abilities in the overall use of M&S in acquisition T&E and to select the appropriate statistical method for M&S V&V.
- Leverage the ASME VVUQ standard to further guide policy and develop best practices tailored to the needs of the DoD T&E community.
- Work with the IDSK-EF Working Group in FY22 on integrating M&S strategies into the IDSK-EF framework. Potential M&S objectives include (1) formulation of M&S needs (including intended use); (2) validation of Conceptual Model; (3) definition of M&S requirements; (4) design; (5) implementation; (6) implementation verification; (7) design verification; (8) validation; and (9) accreditation.
- Address the gap of M&S maturity assessment over the system's acquisition lifecycle.
- Plan for and allocate resources to train managers, developers, analysts, and users about M&S concepts such as the types of uncertainty (e.g., epistemic vs aleatory), uncertainty characterization and sensitivity analysis.
- Define a general cradle-to-grave framework that can provide adequate guidance for M&S users or reviewers of M&S documentation and that can be tailored to each AAF pathway.
- Lead and develop a generic M&S VV&A architecture in systems modeling language to assist the program's digital transformation.
- Continue collaborating with OSD, Agency, Service acquisition stakeholders, other Federally Funded Research Centers (FFRDCs), and University Affiliated Research Centers (UARCs) in developing a roadmap for growing M&S capabilities and resources that enhance current DT&E M&S procedures for acquisition pathways.

Acknowledgement

Many thanks to our MITRE colleagues Dr. Lillianne Troeger, Mr. Ronal Kacsmar, Ms. Katie Lilevjen, Mr. Gregory Chesterton, and Mr. David Cleary for their contributions to this report. Thanks to Dr. Shamik Das, Dr. Ernie Page Jr., and Dr. Saurabh Mittal for their reviews and feedback. We also wish to thank Dr. Katherine Morse (John Hopkins University/Applied Physics Laboratory), Mr. David Sparrow (Institute for Defense Analysis), and Mr. Nicholas Jones (Scientific Test and Analysis Techniques Center of Excellence) for their feedback. Thanks to Naval Surface Warfare Center Port Hueneme Division for contributing to the Model Maturity Assessment Workshops, from which content was extracted for this report, and the sections addressing uncertainty quantification and digital engineering strategies.

This page intentionally left blank.

Contents

List of Figures

List of Tables

Advance M&S in Acquisition T&E

1. Introduction

This work was part of the Director, Developmental Test, Evaluation, and Assessment [D(DTE&A)] strategic initiatives for fiscal year (FY) 2021. These initiatives were intended to support core objectives and requirements and provide opportunities to influence senior leader perceptions of DTE&A within the Department of Defense (DoD). This paper is the result of MITRE's work on Initiative #13, which focused on the use of models and simulations (M&S) for developmental test and evaluation (DT&E). Initiative #13 is the follow-on work to a DTE&A FY19 Naval Warfare study [1] that produced recommendations to advance policy and guidance for where M&S can be used in DT&E and to provide support to DT&E staff specialists examining M&S usage in Test and Evaluation Master Plans (TEMPs). The study provided valuable insight into M&S usage based on a small sample of Navy programs evaluated with a limited scope and with a relatively small level of effort. The study identified factors that contribute to the success and efficiency of M&S validation, verification, and accreditation (VV&A), proposed additions and changes to the DTE&A Action Officer Handbook, and recommended to further explore M&S usage in test and evaluation (T&E) to identify and document best practices for M&S in T&E.

Initiative #13 is rooted in three D(DTE&A)'s *Test Vision Key Takeaways* [2] addressing M&S:

- M&S will continue to grow as a critical component of an overall test program to explicitly establish close alignment with M&S VV&A activities via collection and analysis of relevant objective quality evidence (OQE).
- Ongoing digital engineering activities and model-based systems engineering (MBSE) methods present significant opportunities to leverage early capability insight via M&S analysis and testing but presents challenges on how to fully leverage automated testing within traditional testing frameworks.
- Early investment and validations of threat characterizations (M&S and live fire) will be critical to establishing confidence in system effectiveness via testing.

This document provides recommendations aimed at advancing M&S in T&E with the envisioned outcome of helping acquisition programs successfully improve capability delivery using M&S. The specific areas and goals addressed by the report are:

- M&S use across acquisition programs – Build perspectives on the role of M&S in the DT&E strategy, sufficiency of M&S process, VV&A policy and guidance, and model maturity assessment.
- M&S Policy and Guidance for M&S in DT&E – Identify gaps in M&S and M&S VV&A guidance and policy and help shape DoD Instruction (DoDI) 5000.61 and Military Standard (MIL-STD) 3022 updates. Continue focusing on the broad issue of M&S alignment with T&E by providing M&S-focused updates to systems engineering and T&E guidebooks.
- Use of statistical methods for M&S VV&A – Show how statistical methods are applied to ensure M&S is adequately informing DT&E and operational test and evaluation (OT&E) objectives. Propose best practices on the application of statistical methods for M&S VV&A during DT&E phases. Further the idea of building strategies to accommodate uncertainty quantification (UQ) estimates for M&S.

- Integration of M&S into evaluation frameworks – Develop a coordinated approach for including M&S strategies into evaluation frameworks as means to enhance decision-making in the lifecycle.
- Model maturity assessment methodologies – Survey and evaluate methodologies for assessing model maturity, identify best practices and gaps, and make recommendations that address DoD needs for assessing models.
- M&S framework – Identify an M&S framework that could be used to augment the acquisition program's M&S and T&E strategies.
- Digital engineering strategies for M&S in T&E – Identify digital engineering strategies that could help improve M&S VV&A over the lifecycle.

This report contains four more sections and three appendices. Section 2 provides definitions of M&S-related terms used in the report, outlines M&S and VV&A activities, and describes the role M&S plays in DoD T&E. Section 3 summarizes findings related to the way a small number of major acquisition programs use M&S for T&E, top policy and guidance that governs the application of T&E and M&S for DoD, and the application of statistical techniques for M&S VV&A. Section 4 leans forward and captures transformational enablers for M&S in T&E such as the integration of M&S into evaluation frameworks, methods used to assess the level of maturity of models, description of a methodology that could be used to improve the pedigree of M&S, and digital engineering strategies that could help improve M&S VV&A.. Section 5 provides recommendations. Appendix A provides a summary of the model maturity assessment methods. Appendix B provides information to assist users tasked with generating M&S documentation or the authority charged with reviewing. Appendix C provides a list of acronyms used throughout the report.

Indirectly, Sections 3 and 4 address aspects of M&S-centric trends observed in DT&E assessments conducted from 2016 through 2021 for 60 programs under DT&E oversight [3]. Those trends include:(1) maturity of M&S; (2) availability of M&S tools; (3) VV&A of M&S tools; (4) focus on data required for DT&E; (5) modern threat representative targets; (6) stability and maturity of systems entering T&E; (7) realization of performance risks and late discoveries in T&E; and (8) requirements change during development.

2. Modeling and Simulation in T&E

Institute of Electrical and Electronics Engineers (IEEE) *Standard Glossary of Modeling and Simulation Terminology* [4] defines model as "an approximation, representation, or idealization of selected aspects of the structure, behavior, operation, or other characteristics of a real-world process, concept, or system". Similarly, a simulation is "a model that behaves or operates like a given system when provided a set of controlled inputs." *DoD Modeling and Simulation Glossary* (DoD 5000.59-M) [5] defines modeling as "the application of standard, rigorous, structure methodology to create and validate a model". Similarly, DoD 5000.59-M defines simulation as "a method for implementing a model over time". Thus, M&S refers to "the use of models, including emulators, prototypes, simulators, and stimulators, either statically or over time, to develop data as a basis for making managerial or technical decisions". Simulations, hardware-in-the-loop (HWIL) components, and real systems can be integrated into a distributed synthetic environment in which to test. A distributed synthetic environment can enable a cost-effective development and sustainment of systems.

Models can be categorized in many ways, some of which include [4]:

- Conceptual model – a description of what the model or simulation will represent, the assumptions limiting those representations, and other capabilities needed to satisfy the user's requirements. A collection of assumptions, algorithms, relationships, and data that describe a developer's concept about the simulation.
- Physical model – a model whose physical characteristics resemble the physical characteristics of the system being modeled.
- Mathematical model – a symbolic model whose properties are expressed in mathematical symbols and relationships.
- Numerical model – a mathematical model in which a set of mathematical operations are reduced to a form suitable for solution by simpler methods such as numerical analysis or automation.
- Predictive model – a model in which the values of future states can be predicted or are hypothesized.
- Physics-based model – a model that uses part of physics to represent the system and its environment.
- Process model – a model that defines the functional decomposition and the flow of inputs and outputs for a system.
- Deterministic model – a model in which the results are determined through known relationships among the states and events, and in which a given input will always produce the same output.
- Stochastic model – a model in which the results are determined by using one or more random variables to represent uncertainty about a process or in which a given input will produce an output according to some statistical distribution.
- Metamodel – a model of a model or simulation; an abstraction that uses functional decomposition to show relationships, paths of data and algorithms, ordering, and interactions between model components and subcomponents.

The use of M&S in DoD systems engineering is not new. M&S, as an enabler of warfighting capabilities, is used to: analyze and inform acquisition decisions; adoption of new tactics, techniques, and procedures (TTPs); processing of intelligence data; and testing of systems before their deployment [6]. For many years, M&S have assisted developers and decision makers in a wide range of technical processes such as analysis of alternatives, developing the system concept, requirements evaluation, production and manufacturing, test and evaluation, systems integration, training, logistics, risk management, experimentation, and assessing the entire capability space. Because of DoD's renewed interest in delivering integrated, network-centric systems-of-systems (SoS) that provide the material solution of the needed capability, the effective use of M&S in systems engineering has become essential in meeting those challenges, particularly for situations that require an understanding of the system derived from many costly events. This use of M&S adds simulation systems to the toolbox of traditional decision-support systems. There are many reasons to use models and simulations, but fundamentally M&S helps save lives, save taxpayers' dollars, and improve the operational readiness of warfighting capabilities [6].

There is an interdependency between M&S and T&E. On one side, data from T&E events is used to guide the development and VV&A of the M&S. Once the M&S is accredited, data collected during M&S events is used to inform T&E assessments and critical decisions in the lifecycle. This interdependency is a central theme in this report.

To minimize the risk associated with M&S-informed decisions in T&E, M&S needs to provide a valid and correct forecasting of system performance under operationally representative conditions that could be impractical to achieve in live testing due to the large number of resources required for the evaluation. Because M&S can only approximate the actual system regardless of the investment, confidence in M&S must be justified before accepting its results to inform decisions. To ensure that M&S results are appropriate for a specific purpose, DoD follows rigorous VV&A processes. *DoD Modeling and Simulation (M&S Verification, Validation, and Accreditation (VV&A)* (DoDI 5000.61) [7] provides the official definition for these processes. Rouche [8] relates them to disciplines:

- Verification – the process of determining that a model implementation and its associated data accurately represent the developer's conceptual description and specifications. The process of verification can be thought as to answer the question "*Did I build the thing right?*" Verification is related to mathematics and computer science, in that its purpose is to prove that the equations that model the phenomena are solved correctly in the code. This definition implicitly includes the simulator as part of the model implementation.
- Validation – the process of determining the degree to which a model and its associated data provide an accurate representation of the real world from the perspective of the intended uses of the model. This process can be thought as to answer the question "*Did I build the right thing?*" Validation is related to science and engineering in that its purpose is to prove the M&S produces results that represent physical measurements and satisfy the model intended use.
- Accreditation – the official certification that a model, simulation, or federation of models and simulations and its associated data is acceptable for use for a specific purpose. Accreditation can be thought as to answer the question "*Is the M&S believable enough to be used?*" Accreditation is related to engineering and engineering management, in that its purpose is to determine whether the code produces results that are acceptable for the intended use.

A lot of emphasis has been placed on validating the prediction capability of computational models that are deterministic in nature. However, because most of the models used in T&E are stochastic in nature, the goal of the validation process is to identify regions of the operational space where live data can be matched to simulation outputs for each specific use of the models. Along with the quantification of uncertainty (statistical and knowledge uncertainty) that contributes to the differences between model output and live data, the definition of intended use of the model and the bounds (i.e., acceptability criteria) for which the results are considered valid are the key aspects of the validation process for T&E. The growing dependency in M&S is leading to a greater reliance on predictive modeling for decision-making, which increases the visibility for the need to quantify the sources of uncertainty that could influence those decisions.

An important attribute of M&S is the credibility of its results to inform decision-making. However, the credibility of M&S-based results cannot be measured directly. There are many factors that potentially contribute to the credibility of M&S. According to the *VV&A Recommended Practices Guide (RPG)* [9], which provides guidance applicable to the full spectrum of M&S employed for military and defense applications, the credibility of M&S is established by assessing the M&S's capability, accuracy, correctness, and usability. National Aeronautics and Space Administration (NASA) *Standard for Models and Simulations* (NASA-STD-7009) [10], which is a generic standard for all M&S, identifies eight factors that potentially contribute to the credibility of M&S

results: data pedigree, verification, validation, input pedigree, uncertainty quantification, results robustness, M&S history, and M&S process/product management. Other M&S maturity assessment methods, which are designed to either assess specific types of M&S or for special use, identify other important attributes to be considered when assessing the credibility of M&S-based results. For example, *Predictive Capability Maturity Model* [1210], which is intended for large-scale computational models, identifies six technical attributes that contribute to the credibility of M&S results computational simulations: representation and geometry fidelity, physics and material model fidelity, code verification, solution verification, model validation, and uncertainty quantification. The differences between the attributes and methods highlights the need for a framework that can help guide the T&E community in how to engineer credible M&S.

Whether the M&S capability is integrated or stand-alone, the M&S *VV&A RPG* lists applicable VV&A activities. Those activities include:

- Verify M&S requirements – confirm the requirements for the simulation match those needed for the current problem, and are correct, consistent, clear, and complete.
- Develop Accreditation Plan – identify all the information needed to perform the accreditation assessment and their priorities, tasks, schedules, participants, etc., in coordination with simulation development and verification and validation (V&V) plans.
- Develop Verification and Validation Plan – identify the objectives, priorities, tasks, and products of the V&V effort; establish schedules, allocate resources; etc. in coordination with simulation development and accreditation plans.
- Validate conceptual model – confirm the capabilities indicated in the conceptual model embody all the capabilities necessary to meet the requirements.
- Verify design – determine the design is consistent with the conceptual model and contains all the elements necessary to provide just the needed capabilities.
- Verify implementation – determine the code is correct and is implemented correctly on the hardware.
- Validate results – determine the extent to which the simulation addresses the requirements of the intended use.
- Collect and Evaluate Accreditation Information – information needed for the assessment is collected from the V&V effort and other sources and evaluated to determine its completeness.
- Perform Accreditation Assessment – fitness of the simulation is assessed using all the evidence collected from the V&V effort and other sources, and an accreditation report and recommendations are prepared for the User.

3. Characterizing the Current State

DTE&A identified the use of M&S in T&E as an area where acquisition programs inconsistently follow guidance and best practices. This section provides a characterization of the use of M&S as viewed from the lens of a small number of acquisition programs, focuses on the sufficiency of M&S and VV&A policy and guidance, and the use of statistical methods for M&S V&V to ensure M&S is adequately informing DT&E and OT&E objectives.

3.1 Use of M&S in T&E Across Acquisition Programs

MITRE surveyed the TEMPs of five acquisition programs to identify and characterize (1) the current state and trends in which M&S is used in T&E; (2) the specific aspects of testing model support; (3) interrelationships or interdependencies among M&S elements; and (4) the degree of VV&A rigor being applied to ensure models are a "good representation" of warfighting systems and operational conditions. Key findings include:

- The primary references to M&S in the reviewed TEMPs are consistent with the concept of interdependency between M&S and T&E that calls for using live test events to capture data to inform M&S development and VV&A and for using data obtained from M&S events to inform T&E assessments. However, to a large extent, M&S referenced for DT&E is often only in terms of collecting data from live events to support VV&A of models that would later inform OT&E instead of informing DT&E objectives.
- The levels of M&S documentation organization and completeness are inconsistent among TEMPs.
- Some TEMPs make explicit reference to TEMPs and T&E activities of other programs that will be leveraged for T&E support, data collection, and model VV&A, while others make no such references.
- Some acquisition programs are using the same federation of models. There is a need to improve the global view of the role M&S plays in the test strategy of individual acquisitions programs when re-using models because TEMPs: (1) do not provide a clear, consistent way to trace relationships to previous tests that contribute to model VV&A; and (2) do not provide descriptions of the level of effort required for VV&A after the intended use of a model changed.
- The acquisition programs surveyed are not assessing the maturity level of M&S with existing model maturity methodologies, but rather following the guidance from their operational test agency (OTA) such as Commander, Operational Test and Evaluation Force (COMOPTEVFOR) [13].
- Given the inconsistencies of how programs discuss M&S in their TEMPs and how they implement the M&S processes, there is a need to update M&S policy and guidance, particularly with respect to TEMP content.

3.2 Policy and Guidance for M&S in DT&E

While programs use data from live test events to inform M&S VV&A activities and data from M&S events to inform T&E assessments, the inconsistencies across TEMPs outlined in the proceeding suggests that there is a need to review M&S policy and guidance. DTE&A also identified the need to continue focusing on the broad issue of M&S alignment with T&E by providing M&S focused updates to systems engineering and T&E guidebooks.

3.2.1 M&S V&V Policy, Guidance, and Best Practices

MITRE reviewed the instructions listed in Table 1, which outline DoD policies and guidance that govern the management of capability acquisitions and T&E, including M&S use and M&S VV&A. The core documents within DoD for M&S VV&A policy and guidance are DoDI 5000.61, MIL-STD-3022, and M&S VV&A RPG.

Table 1. Top-level DoD T&E policy and guidance

Policy	Tittle	Effective Date	Description
DoDD 5000.01	*The Defense Acquisition System* (Change 2)	31 Aug 2018	Provides guiding principles for the management of capability acquisitions.
DoDI 5000.02	*Operation of the Acquisition System*	23 Jan 2020	Provides the policies to support the Defense Acquisition System.
DoDI 5000.89	*Test and Evaluation Instruction*	19 Nov 2020	Establishes policy and procedures across the Adaptive Acquisition Framework (AAF) for DT&E, OT&E, live fire test and evaluation (LFT&E), and integrated test and evaluation (IT&E).
DoDD 5000.59	*DoD Modeling and Simulation (M&S) Management*	15 Oct 2018	Updates policy and responsibilities for DoD M&S management and establishes the DoD M&S Steering Committee.
DoDI 5000.59-M	*DoD Modeling and Simulation (M&S) Glossary*	19 Mar 2014	Provides uniform M&S terminology for use by DoD.
DoDI 5000.61	*DoD Modeling and Simulation (M&S Verification, Validation, and Accreditation (VV&A)*	15 Oct 2018	Establishes VV&A policies, procedures, and guidelines for M&S applications, standards, and databases managed by DoD components as well as policy, responsibility, and procedures for M&S VV&A used to support decision-making.
DoDI 5000.70	*Management of DoD Modeling and Simulation (M&S) Activities*	15 Oct 2018	Implements DoDD 5000.59, assigns responsibilities for the M&S SC, establishes Director, DoD Modeling and Simulation Office (MSCO), and extends discovery metadata policy to key DoD M&S tools, data, services, data assets, models, and simulations.
MIL-STD-3022	*Department of Defense Standard Practice; Documentation of Verification, Validation, and Accreditation (VV&A) for Models and Simulations (Change 1)*	1 Apr 2012	Provides a common framework for documenting information produced during the VV&A processes by establishing templates for documenting VV&A planning, implementation, and reporting. This standard practice may be cited as a contractual requirement in contracts.
MSE Core Document	*VV&A Recommended Practices Guide (RPG)*	27 Oct 2020	Facilitates the application of DoD directives, policies, and guidelines promote effective and efficient VV&A processes for the full spectrum of M&S products employed in DoD.
DTM 19-007	*Directive-Type Memorandum (DTM) 19-007, "Developmental Test and Evaluation Sufficiency Assessments"*	19 Jul 2019	Establishes policy, assigns responsibilities, and provides guidance on sufficiency assessment of DT&E for MS B and MS C, including the adequacy of M&S.
DOT&E TEMP Guidebook	*Director, Operational Test and Evaluation (DOT&E) Test and Evaluation Master Plan (TEMP) Guidebook Version 3.1*	19 Jan 2017	Provides guidance for the content of the TEMP.

DoDI 5000.61 establishes policy, assigns responsibility, and prescribes procedures for the VV&A of models, simulations, distributed simulations, and their associated data. It mandates the minimum set of items to document as part of the VV&A process: (1) identification of the date performed and the person(s) or organization performing VV&A; (2) identification of the version and/or release of the model, simulation, or associated data being verified, validated, or accredited; (3) identification of the intended use of the model, simulation, or associated data being VV&A; (4) list of, or reference to, the M&S requirements and associated accreditation criteria for the model, simulation, or associated data being VV&A; (5) list of, and description of, the VV&A accreditation assessment activities; (6) summary of results, including the M&S limitations risks, potential impacts, and assumptions of the models, simulations and/or the associated data undergoing V&V; (7) summary of the results of the accreditation assessment; and (8) identification of the user and/or accreditation authority and record of the accreditation decision.

MIL-STD-3022 establishes templates for the core set of M&S VV&A documents as well as a framework for sharing information throughout the VV&A processes. The core set of VV&A documents consists of the following documents, which are outlined in Table 2. The Accreditation Plan focuses on (1) defining the criteria to be used during the accreditation assessment; (2) defining the methodology to conduct the accreditation assessment; (3) defining the resources needed to perform the accreditation assessment; and (4) identifying issues associated with performing the accreditation assessment. The V&V Plan focuses on (1) defining the methodology for scoping the V&V effort to the application and the acceptability criteria; (2) defining the V&V tasks to that will produce information to support the accreditation assessment; (3) defining the resources needed to perform the V&V; and (4) identifying issues associated with performing the V&V. The V&V Report focuses on (1) documenting the results of the V&V tasks; (2) documenting M&S assumptions, capabilities, limitations, risks, and impacts; (3) identifying unresolved issues associated with V&V implementation; and (4) documenting lessons learned. The Accreditation Report - focuses on (1) documenting the results of the accreditation assessment; (2) documenting the recommendations in support of the accreditation decision; and (3) documenting lessons learned during accreditation. The RPG provides guidance to facilitate the application of DoD M&S directives, policy, and guidelines to promote effective and efficient VV&A processes for the full spectrum of M&S products employed in DoD. The RPG also describes the interrelated processes that make up VV&A, roles and responsibilities of participants, special topics associated with VV&A, tools and techniques, and reference material on related areas. RPG contains an excellent overview of the VV&A process with links to the four core M&S VV&A products (Accreditation Plan, Verification and Validation Plan, Verification and Validation Report, and Accreditation Report). RPG addresses the process, artifacts, and players' roles and responsibilities. However, the diagrams are complex and hard to follow (overall problem solving process and flow diagram for the VV&A of a legacy simulation), and the common sections contain a lot of redundant material as shown in Table 2. The potential exists to reduce the amount of redundant information once DoD starts transitioning to a digital engineering environment. The analogy provided in the RPG comparing building a new house to developing a new simulation (versus buying an existing home, analogous to reusing a legacy simulation), is illustrative and helpful.

Table 2. Outline of Core M&S VV&A Documents

Section	Accreditation Plan	V&V Plan	V&V Report	Accreditation Report
	Executive Summary	Executive Summary	Executive Summary	Executive Summary
1	Problem Statement	Problem Statement	Problem Statement	Problem Statement
2	M&S Requirements and Acceptability Criteria	M&S Requirements and Acceptability Criteria	M&S Requirements and Acceptability Criteria	M&S Requirements and Acceptability Criteria
3	M&S Assumptions, Capabilities, Limitations & Risks/Impacts	M&S Assumptions, Capabilities, Limitations & Risks/Impacts	M&S Assumptions, Capabilities, Limitations & Risks/Impacts	M&S Assumptions, Capabilities, Limitations & Risks/Impacts
4	Accreditation Methodology	V&V Methodology	V&V Task Analysis	Accreditation Assessment
5	Accreditation Issues	V&V Issues	V&V Recommendations	Accreditation Recommendations
6	Key Participants	Key Participants	Key Participants	Key Participants
7	Planned Accreditation Resources	Planned V&V Resources	Actual V&V Resources Expended	Actual Accreditation Resources Expended
8			V&V Lessons Learned	Accreditation Lessons Learned
Suggested Appendices	A - M&S Description B - M&S Requirements Traceability Matrix C - Basis of Comparison D - References E - Acronyms F - Glossary G - Accreditation Programmatics H - Distribution List	A - M&S Description B - M&S Requirements Traceability Matrix C - Basis of Comparison D - References E - Acronyms F - Glossary G - V&V Programmatics H- Distribution List I - Accreditation Plan	A - M&S Description B - M&S Requirements Traceability Matrix C - Basis of Comparison D - References E - Acronyms F - Glossary G - V&V Programmatics H- Distribution List I - V&V Plan J – Test Information	A - M&S Description B - M&S Requirements Traceability Matrix C - Basis of Comparison D - References E - Acronyms F - Glossary G – Accreditation Programmatics H- Distribution List I - Accreditation Plan J - V&V Report

Note: Common sections between the four templates in blue

The review of those documents revealed some possible gaps related to the use of M&S in T&E and to M&S VV&A that need to be addressed for a better application of the VV&A process:

- While the RPG and MIL-STD-3022 templates provide useful information to generate the four essential VV&A reports, the information is generic. Guidance can be specifically

tailored to assist T&E personnel, which often are non-M&S specialists, and to improve the quality and timeliness of M&S VV&A products and templates.

- Guidance for verification methods is too generic. The efficiency and effectiveness of verification testing can be improved by adding the use of software testing methods and statistical methods to the repertoire of verification techniques.
- There is a lack of guidance on how to apply statistical methods for the validation of M&S, including uncertainty quantification.
- Improve policy to enable a smooth transition to a digital engineering environment.

Coinciding with MITRE's review of those instructions, Office of Under Secretary of Defense for Research and Engineering [OUSD(R&E)] Engineering, Policy, and Systems (EPS) Modeling and Simulation Enterprise (MSE) started the process of updating DoDI 5000.61 and MIL-STD-3022 to include concepts and terminology consistent with current practices, better templates for documentation that support those with VV&A responsibilities, and references to additional guidance (e.g., recommended practices) needed for effective, efficient M&S VV&A implementation. To capture as many perspectives as possible on how the M&S VV&A policy should be updated, OUSD(R&E) EPS MSE convened a 3-hr technical-level workshop on April 14, 2021, to capture concerns of VV&A practitioners and shape plans to address them. Participants reviewed the current policy and associated guidance, presented challenges in their Services and communities with VV&A, and began to identify common areas that should be addressed with OSD leadership, including changes to the DoDI.

As supplement to the M&S VV&A templates, MITRE provided OUSD(R&E) EPS MSE a first set of inputs that outlines information that could help in formulating supplemental guidance for the development, management, and VV&A of M&S. While the information complements the templates and RPG, it takes it a step further for the benefit of a user tasked with generating the documentation, or the authority charged with reviewing it to determine completeness and acceptability. Key subject areas addressed include formulating the problem statement, M&S requirements traceability and acceptability criteria, M&S development and structure, M&S capabilities and limitations, concept model validation, M&S design and implementation verification, basis for comparisons, accreditation assessment, configuration management, data needs, resources, and V&V tests and analysis. This paradigm was subsequently modified and incorporated into the process described in Section 4.3. MITRE will continue to engage with OUSD(R&E) EPS as they get further along in their review and our process matures.

MITRE also reviewed *Director, Operational Test and Evaluation (DOT&E) TEMP Guidebook* and the memos from DOT&E that are outlined in Table 3. Two of those memos, *Guidance on the validation of models and simulation used in Operational Test and Live Fire Assessments* and *Clarification on guidance on the validation of models and simulation used in Operational Test and Live Fire Assessments*, set expectations on elements of test design that must be addressed in the TEMP. Those expectations include to (a) describe the M&S capability, its intended use, and its validation and accreditation approach to include elements of design of experiments techniques for M&S VV&A; and (b) compare live and M&S outcomes using statistical methods and a comprehensive strategy to assess M&S output across the operational domain for which the M&S will be accredited. While appropriate for DT&E, the guidance provided in the DOT&E Memos is not completely detailed in the DOT&E TEMP guidebook.

Table 3. DOT&E Memos on guidance for the use of design of experiments in T&E

Title	Effective Date	Content
Guidance on the use of design of experiments (DOE) in Operational Test and Evaluation	19 Oct 2010	Sets expectations on elements of experiment design that must be addressed in the TEMP.
Case studies for the use of design of experiments (DOE) in Developmental Test and Evaluation (DT&E)	21 Jan 2011	Examines the applicability of DOE to DT&E activities, from early engineering analysis to final verification of requirements.
Flawed application of design of experiments (DOE) to Operational Test and Evaluation (OT&E)	26 Jun 2013	Reinforces the expectations set forth in the October 2010 memo and addresses areas for improving the use of DOE.
Best practices for assessing the statistical adequacy of experimental designs used in Operational Test and Evaluation	23 Jul 2013	Identifies statistical figures of merit that should be used to determine the adequacy of a test design.
Inadequacy of recently proposed test design for the P-8A Increment 2 and Multi-static Active Coherent (MAC) Programs	24 Sep 2013	Identifies deficiencies in the test design concept for evaluating the P-8A Increment 2 and MAC systems.
Guidance on the validation of models and simulation used in Operational Test and Live Fire Assessments	14 Mar 2016	Sets expectations for the TEMP and Test Plan to describe the M&S capability, its intended use, and its VV&A approach.
Clarification on guidance on the validation of models and simulation used in Operational Test and Live Fire Assessments	17 Jan 2017	Re-emphasizes the expectations for the M&S validation strategy, including the quantitative comparison of live and M&S outcomes using statistical methods across the operational domain for which the M&S is accredited.

The premises of DOT&E TEMP Guidebook for test design are to capture the complexity of the system, all aspects of mission execution, and span the operational space. Similarly, the premises for the content of the TEMP are to tailor it to the different phases of test and acquisition milestones. The TEMP Guidebook also provides guidance for developing the appropriate TEMP content for the application of design of experiments methodologies throughout the various phases of test while focusing on the delivery of content for milestones A, B, and C. While the section for LFT&E in the TEMP Guidebook provides a comprehensive list of test design requirements, only about one-half of those requirements were leveraged in the sections for DT&E and OT&E. Both DT&E and OT&E test design activities will significantly benefit if content from the LFT&E section was also available in the DT&E and OT&E sections. Additionally, those requirements are also applicable to tests where an M&S capability has been used to generate the data for assessment or for its VV&A. Thus, it seems useful to combine content from the LFT&E section of the TEMP Guidebook with content from DOT&E Memos to provide a single source of test design requirements applicable to all phases of test, including M&S-driven tests.

3.2.2 Alignment Between M&S and T&E

To address the goal of continuing to focus on the broad issue of M&S alignment with T&E, MITRE provided M&S-focused findings and recommendations to DTE&A's Chief Engineer for updates to the *Engineering for Defense Systems Guidebook*, *Systems Engineering Guide*, and *T&E Enterprise Guidance* drafts. The M&S-focused recommendations were aimed at ensuring the processes involving the use of M&S tools and their VV&A are consistent and mutually reinforced across program management, systems engineering, and T&E. A high-level description of the critical findings is provided below. Those findings are representative of the hurdles the community is trying to overcome. In many cases, the documents fail to describe a deliberate use of M&S in T&E and codependences between program management, systems engineering, T&E, and M&S.

- Both the T&E and system engineering guidebooks failed to ensure M&S requirements are captured early in the lifecycle.
- Properly validated M&S is a requirements verification method.
- Lack of reference to the validation of conceptual models.
- There are misalignments and little leverage between M&S, T&E, and the systems engineering process.
- The benefits of M&S need to be articulated better.
- The role of M&S in decision-support systems needs to be articulated better.
- The use of statistical methods for M&S validation is not specifically called out.

3.3 Statistical Methods for M&S VV&A

Evaluations are increasingly relying on M&S to supplement live testing. For these evaluations to be useful, models must represent the systems they simulate and the environment and conditions in which the systems operate. Statistical methods are key on validating how well the models represent the systems they simulate. This section illustrates how statistical methods are applied to ensure M&S is adequately informing DT&E and OT&E objectives, proposes best practices on the application of statistical methods for M&S VV&A during DT&E phases, and furthers the idea of building strategies to accommodate uncertainty quantification (UQ) estimates in M&S V&V.

3.3.1 Statistical Methods for V&V

To illustrate how programs are currently approaching the use of M&S in T&E, MITRE leveraged a broad-scope assessment of the M&S validation strategy of a program to inform a decision to proceed to another phase of testing. Assessment results (at the Controlled Unclassified Information level) were presented to Director, DTE&A on 21 April 2021.

The assessment centered around a key measure that required data from M&S events to inform the evaluation. This measure requires such a significant amount of ship and combat system resources that traditional live-fire only tests are unaffordable. Therefore, the program was relying on M&S data to quantify the measure. Because stakeholders that rely on M&S for decision-making need valid results to ensure that capabilities meet mission requirements, the M&S must be verified, validated, and accredited.

The report provides a detailed assessment of, and identified risk for, each of the three phases of the M&S validation strategy. Phase 1 consisted of using historical validation data and Monte Carlo simulations to generate pre-test predictions to inform early decision-making. Phase 2 consisted of

using statistical hypothesis tests to compare data from missile firing tests to data from M&S pre-test predictions to inform test readiness decision. Phase 3 consisted of using statistical techniques to compare data generated from M&S runs-for-record and data from missile firing tests.

The assessment revealed the need to build upon the existing M&S platform and reprogram it using an MBSE approach to create a digital twin and strengthen the interaction between the physical product and the digital twin. MBSE provides for the creation, verification, and validation of models. A digital twin not only could be useful to improve the rigor, effectiveness, and efficiency of future tests but it also will improve overall the systems engineering process and allow for testing of scenarios that might otherwise be cost prohibitive or unsafe for testers. There is a need to continue developing digital twins and their associated engineering processes. Strengthening the interaction between the physical product and the digital twin improves the flow of technical and operational data between the physical product and the digital twin, which would allow for much richer testing and replication of errors or operational situations to cost effectively increase confidence in and inform decisions.

The assessment also revealed or corroborated some observations that could be applicable to many programs. This information can be used to develop or improve guidance for acquisition programs. The observations include:

- The development of M&S is not necessarily driven by T&E requirements, which includes the intended use of the model. This often results in limitations, later in the lifecycle of the system, due to the inability of simulating the system's operational environment. Thus, there is a need to incorporate T&E requirements, including the identification of intended use, early in the development of M&S.
- While there is a basic understanding of which statistical methods should be used to validate M&S, the knowledge is not widespread and the selection of methods for specific applications are often left to program analysts. Thus, there is a need to improve the workforce knowledge on how to select appropriate methods for the validation of M&S.
- While the final M&S test design was adequate as a stand-alone design, the opportunities to match M&S output and live fire data were limited, which could present a risk for the validation process and ultimately the program. Thus, there is a need to start collecting data for M&S VV&A as early as possible during system development.

3.3.2 Statistical Methods for M&S V&V in DT&E

To understand the state-of-the-field related to use of statistical methods available for M&S V&V, MITRE reviewed Institute for Defense Analysis (IDA)'s *Handbook on Statistical Design & Analysis Techniques for Modeling & Simulation Validation* [1515] to determine the applicability of those techniques to DT&E. The goal of the handbook was to "aid the T&E community in developing test strategies that support data-driven model validation and uncertainty quantification". The handbook is a collection of best practices that describe the VV&A process as it relates to using M&S for operational testing, methods for analyzing the simulation, methods for comparing M&S output and live data, some methods for quantifying associated uncertainties, and the application of design of experiment techniques to both live and simulation environments. To the largest extent, the T&E community considers the handbook the authoritative source for developing strategies to validate M&S with statistical techniques. IDA has been delivering workshops and tutorial in those techniques to the DoD T&E community since 2018—before the

handbook was published. Those techniques have been adopted by academia and taught in short courses by contactors. The techniques outlined in the handbook are adequate for DT&E.

3.3.3 Uncertainty Quantification in M&S V&V

Uncertainty is present in almost every acquisition decision. For example, there are situations in which the characteristics of the system or the variability of the operational environment are not exactly known. Likewise, there are situations in which the operating conditions are not fully understood. Yet, there is a need to quantify those sources of uncertainty to determine their effect on likely outcomes. UQ is the field of science that deals with the quantitative characterization and reduction of uncertainties. To understand the influence of UQ in M&S V&V, MITRE conducted a literature review on UQ methods and participated in the online seminar *Simulation Credibility for Decision Making – The Importance of Verification, Validation, and Uncertainty Quantification (VVUQ)* sponsored by *The International Association for the Engineering Modeling, Analysis and Simulation Community (NAFEMS)*. The concept of VVUQ, which has been gaining popularity within the community, emphasizes the role of UQ within V&V processes.

The NAFEMS presentations showcased practical aspects of applying VVUQ across different sectors, including industry, regulatory agencies, and research organizations. Many of the presenters referred to ASME standard VV10, which serves as a guide for V&V activities within the scope of computational solid mechanics. Overall, the presentations provided an overview of the role of UQ within M&S VV&A and the challenges faced when attempting its implementation within VV&A programs. In particular, the breadth of presentations showcased (1) the history, status, and future work around the development of VVUQ standards, some of which are tailored to different fields; (2) different aspects of conducting VVUQ activities within their respective fields; (3) technical topics such as UQ methods and robust optimization; (4) non-technical aspects of VVUQ for building credibility; (5) frameworks for assessing simulation maturity levels and selecting levels of rigor for validation; (6) the connection between VVUQ activities and risk in decision-making; (7) technical, cultural, and policy-related challenges in establishing credibility of computational models; and (8) differences between key factors within industry and within regulatory agencies. Participation in the online seminar prompted MITRE to conduct a literature review to further guide recommendations around UQ within the scope of acquisition M&S V&V.

The UQ literature review shows that definitions for the term “uncertainty quantification” vary considerably, most likely because UQ is not considered to be a standalone field of study. Often the various definitions of UQ emphasize technical aspects yet tends to omit its role within the scope of informed decision-making as well as its position in relation to other key analysis methods, such as risk analysis. Generally, UQ can be defined as the study of all sources of error and uncertainty [16]. However, from a practical perspective UQ can be defined as the process of characterizing uncertainties in model outputs based on uncertainties in certain model inputs.

Uncertainty tends to be categorized as either aleatory, epistemic, mixed, or ontological uncertainty. Aleatory uncertainty refers to uncertainty due to the inherent variability of a system, which can be collected and expressed in statistical terms. Epistemic uncertainty refers to uncertainty due to resolvable lack of knowledge, and it is often related to having little or no empirical data. Mixed uncertainty refers to the combination of both aleatory and epistemic uncertainty. Ontological uncertainty refers to uncertainty attributed to an unresolvable lack of knowledge—transparent and unquestionable assumptions outside of the experience base or normal rules.

Within the context of scientific computing, aleatory, epistemic, and mixed uncertainties can be quantified. Sources of uncertainty generally include [17] model inputs (uncertainty due to model parameters, geometry, initial conditions, and any other feature that must be provided to the model), numerical approximations (uncertainty due to computational considerations, such as discretization error, iterative convergence error, roundoff error, and bugs within computer code), and model form (uncertainty ascribed to the assumptions, approximations, and mathematical formulations built into a model). In practice, it is not a trivial task to segregate aleatory from epistemic uncertainty. Both types of uncertainties should be treated independently given their distinct statistical nature.

Sources of uncertainty associated with model inputs, numerical approximations, and model forms should all be quantified using, for example, some of the frameworks available [16][17][18]. While the details of such frameworks may slightly differ from one another depending on which aspects of UQ are in focus, most center around three key steps (1) assess sources of uncertainty – identify and characterize all sources of uncertainty in the model; (2) propagate uncertainties – quantify the effect of input uncertainties on output variability; and (3) assess output variability – evaluate the estimated variabilities and their potential impact.

The lack of maturity as a field has led to several misconceptions. One misconception is that UQ is strictly synonymous with uncertainty propagation. Another common misconception is that UQ can be used to tell if a model's predictions are "right" or "true". UQ as such cannot tell us that, instead it tells us that if we decide to accept the validity of a model to a certain extent, we must then accept the validity of conclusions drawn from it up to the degree suggested by the UQ analysis [16]. UQ can, however, play a key role in a V&V plan to build confidence in the predictive capabilities of complex models and simulations for a particular intended use [19]. Finally, the term UQ is often used interchangeably with other associated yet distinct methods, particularly with sensitivity and risk analyses. UQ aims to quantitatively assess all sources of uncertainty, while sensitivity analysis (SA) aims at quantifying the relative importance of each input parameter to the output of a model [20]. SA is often complementary to UQ, particularly in situations where there are many uncertain parameters which lead to high computational expense. Both types of analysis can be used as part of a greater assessment, such as in risk or performance analysis [21][22].

The literature describes a wealth of methods available for uncertainty propagation. The proper choice is application specific, meaning that factors such as prior knowledge, time, computing resources, and data availability all play an important role in deciding which method to choose. Propagation methods can be broadly categorized by the type of uncertainty being propagated—namely aleatory, epistemic, or mixed. The list of propagation methods is extensive, but it includes (1) common sampling methods for aleatory uncertainty such as Monte Carlo and design of experiments; (2) surrogate methods for aleatory uncertainty such as polynomial chaos expansion [23], stochastic collocation [24], low-rank tensor approximations [25], principal component analysis [26], Karhunen-Loeve expansion [27], and other types associated with modern machine learning techniques such as Kriging [28], deep learning neural networks [29], and support vector machines [30]; and (3) methods for epistemic uncertainty such as interval analysis [31] and Dempster-Shafer evidence theory [32].

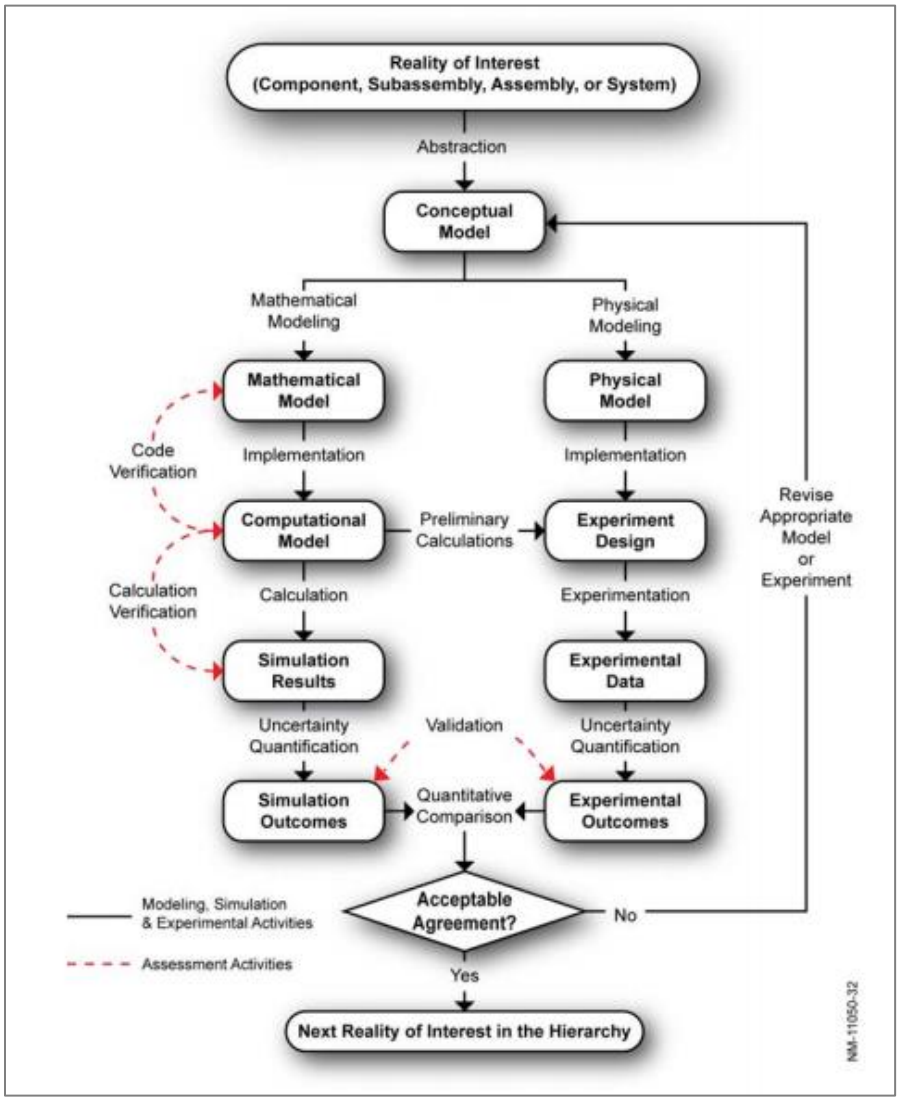


Figure 1. V&V activities and outcomes, ASME Std. VV10

The growing role of M&S within T&E programs will lead to greater reliance on predictive modeling for decision-making in acquisition. It is through V&V that stakeholders and decision-makers can thoroughly assess the correctness and credibility of results obtained from M&S. This credibility-building exercise aims to assess the limitations of predictive modeling tools within the scope of their intended purpose. From a verification perspective, UQ provides a quantitative measure of the extent to which a computational model represents its underlying mathematical model [19]. From a validation perspective, UQ allows for uncertainties in both computational and experimental results to be accounted for when the two are compared to establish the extent to which M&S results truly resemble real-world data. While UQ can take considerable effort to carry out properly, incorporating thorough UQ analyses into the V&V process provides a frame of reference for confidently making decisions based on M&S results. ASME Standard VV10 [19] highlights the activities and outcomes of V&V including UQ, as shown in Figure 1. UQ is centerpiece to establishing agreement between the real world and M&S.

In summary, the key findings related to the application of UQ for DoD M&S VV&A are:

- UQ is gaining positive reception despite limited adoption within the DoD M&S V&V process. UQ requires new policy and standardization with DoD M&S VV&A processes.
- UQ is not a mature field of study, but rather a complex set of tools and practices applied across scientific computing disciplines. There is a need to improve the workforce knowledge, skills, and abilities in the field. Similarly, there is a need to generate case studies that can be used as models for different applications.

- There is no current guidance within the DoD T&E community on how to apply the concept of UQ to M&S V&V.
- One key message from NAFEMS seminar is that VVUQ simply does not matter if it is not communicated in a way that can be understood by key stakeholders and decision-makers, including communication around resources and VVUQ outcomes. Thus, it is important for DTE&A to effectively communicate the desired VVUQ outcomes.
- Another key lesson learned from the NAFEMS seminar is that the elements that underpin credibility in M&S are personnel training, quality control of M&S processes, and maturity assessments for M&S results.
- ASME is actively developing VVUQ standards across different fields and domains. DTE&A has an opportunity to leverage this knowledge base to further guide policy and develop best practices tailored to the needs of the DoD.

4. Defining the Future State

This section leans forward towards the future state of M&S in DT&E. It focuses on the integration of M&S into the Integrated Decision Support Key – Evaluation Framework (IDSK-EF) to enhance decision-making with M&S, identification of methodologies to assess the maturity of models, identification of state-of-the-art statistical techniques for M&S VV&A, integration with digital engineering, and development of a DT&E M&S framework.

4.1 Integration with Evaluation Frameworks

Part of the strategy to define the future state of M&S was to develop a coordinated approach for including M&S strategies into the unified evaluation framework (UEF) decision support system to enhance the use of M&S in DT&E. Specific objectives were to (1) work on M&S content for the UEF to assist DTE&A improve the DT&E M&S planning process; and (2) identify M&S measures and ensure alignment between contractor developmental test and evaluation (CDT&E), DT&E, and OT&E measures and objectives.

The UEF was the evolution of the developmental evaluation framework (DEF), which was developed in 2014 to assist acquisition programs focus their DT&E strategy on decisions, capabilities, and evaluations. The DEF has been instantiated into DoDI 5000.02, the *Defense Acquisition Guidebook*, and the *DTE&A Guidebook for Staff Specialists*. To date, DTE&A has engaged with over 120 DoD-wide acquisition programs to build a DEF for their TEMPs.

Throughout FY21, the UEF evolved into the IDSK-EF [33] concept. The IDSK-EF notional concept is intended to integrate into the acquisition system to help DoD acquire systems that support the warfighter in accomplishing their mission. The construct articulates the need to inform decisions and knowledge points throughout the lifecycle with operational and technical evaluations that are fed by data obtained from live, virtual, or constructive events, including M&S.

The IDSK-EF combines the best-of-breed of the DEF, integrated evaluation framework (IEF), mission-based test design (MBTD), and MBSE concepts into an All Services, three-stage planning approach, single-scope, mission-based, evaluation-focused framework that addresses all—operational, technical, and programmatic—T&E information decisions needs for application across the AAF T&E continuum. The IDSK identifies the decisions, decision support questions (DSQs), and critical operational issues (COIs) that will be informed with the evaluation of

operational and technical capabilities linked to the mission-oriented measures and attributes that are identified in the evaluation framework (EF). The final stage, captured in the TDF, links the measures to test and M&S events descriptions, factors and levels, statistical test designs, vignettes, resources, etc. The IDSK-EF is a step forward for test planning and analysis in that principles of statistical test design, a critical enabler for M&S VV&A, are incorporated for the first time into a DT&E decision support system. The IDSK-EF accommodates decisions and objectives involving both (1) T&E needs to inform VV&A of M&S and threat models; and (2) M&S needs to inform T&E assessments.

The IDSK-EF is still under development. The IDSK-EF team and Services are working to identify pilot programs that will help define the content and structure of the IDSK-EF. We will work with the IDSK-EF in FY22 on identifying a good opportunity to implement M&S strategies into the framework. Potential M&S objectives to be addressed in the IDSK-EF include (1) formulation of M&S needs (including intended use); (2) validation of Conceptual Model; (3) definition of M&S requirements; (4) design; (5) implementation; (6) implementation verification; (7) design verification; (8) validation; and (9) accreditation. These nine objectives are consistent with the process outlined in Section 4.3.

4.2 Model Maturity Assessment Methods

The strengths and limitations of models to represent a system become clear thru an assessment of model maturity. The extent to which acquisition programs are assessing the maturity of models to support T&E decisions is not completely understood. MITRE conducted a literature search and evaluated the ten model readiness assessment methods below to identify best practices that should be considered by DTE&A when adopting or developing a model readiness approach and to make recommendations to address DoD needs for assessing models. Key features for each methodology can be found in **Error! Reference source not found.**

- NASA-STD-7009A Credibility Assessment Scale (CAS) (NASA, Baseline 2008, Change 1 2016) [10]
- Simulation Software Technology Readiness Levels (SSTRL) (The MITRE Corporation 2020) [34]
- Predictive Capability Maturity Model for Computational Modeling and Simulation (PCMM) (Sandia National Laboratories, Baseline 2007, Rev 4 2013) [10, 35]
- Model Readiness Levels: A Mathematical Construct for Validation and Thrust (MRL) (STAT COE, 2021) [36, 37]
- International Council on Systems Engineering (INCOSE) Model-Based Capabilities Matrix and User's Guide v1.0 (MBCM) (INCOSE 2020) [38]
- NATO General Methodology for Verification and Validation to Support Acceptance of Models, Simulations and Data (GM-VV) [39]
- Simulation Interoperability Readiness Levels (SIRL) SISO-REF-076-2020 [40]
- Model Assurance Levels (MAL) (The Aerospace Corporation, 2020) [41]
- Risk Based Methodology for Verification, Validation and Accreditation M&S Use Risk Methodology (MURM) (The Johns Hopkins University Applied Physics Laboratory, 2011) [42]
- Google Model Cards (GMC) (Google, 2020) [43,44]

On 4 August 2021 and 9 September 2021, The MITRE Corporation hosted two cross-organizational workshops on model readiness assessment methodologies. The objectives of the workshop were to collect information to inform the development of DoD guidance and policy on model readiness assessment methods and to address challenges posed by the T&E community like the need for an agreed upon model readiness assessment framework and the transition towards a digital engineering environment. Participants included representatives from DTE&A, DOT&E, OUSD(R&E) MSE, NASA, MITRE, the Aerospace Corporation, Institute for Defense Analyses (IDA), The Johns Hopkins University Applied Physics Laboratory (JHU/APL), and Scientific Test and Analysis Techniques Center of Excellence (STAT COE).

A cadre of nationally recognized subject matter experts provided talks on the methodologies listed above plus on the application of statistical methods for M&S VV&A. The presentations and discussions provided valuable insight into the need to assess model maturity throughout the lifecycle, capabilities and limitations of each method, context for their interpretation, best use and appropriateness of each method, safeguards for the adoption of new methods, and the role assessment plays in simulation-informed decision making. Key findings and a summary from the workshop and literature review are provided later in the report[1].

4.2.1 Purpose and Scope

The purpose and scope of the methods vary widely. CAS, PCMM, MAL, MRL, MURM and SIRL all aim to help decision makers better understand the risks associated with using M&S results. In addition, the draft whitepaper for the MRL also states a second purpose which is to "provide developers with clearer standards by which to develop their models"; however, given its draft state, the mechanism by which the MRL meets this second purpose is currently unclear. The SIRL is slightly different in that it focuses on interoperability between groups of simulations rather than readiness of a single simulation. SSTRL is primarily intended to serve as a common language for communicating with contractors developing simulations and stimulators. GM-VV provides a generic framework to justify why identified models, simulations, data, outcomes, and capabilities are acceptable for deployment in the intended operational context of use. In contrast, MBCM is intended to help organizations plan for and develop processes to implement digital engineering or model-based capabilities. Thus, MBCM is the least relevant to the task of assessing readiness of a given model; for this reason, it will not be discussed in detail in subsequent sections.

Many of the selected methodologies are intended for specific types of models. PCMM is intended for large-scale computational models for integro-differential or partial differential equations. SSTRL is intended for testbed simulations or stimulations. MRL is intended for models that will be used to inform decisions with the analysis of data from operational tests or activities. MAL was originally developed for design and architecture software models that are written in UML; it is currently being extended to other models such as system engineering models. SIRL was initially intended for training simulations; however, the developers believe the methodology should be applicable to interoperability between all types of simulations.

[1] All presentations are available by contacting MITRE (jdaly@mitre.org).

CAS, MURM and GM-VV are the most broadly applicable methodologies. CAS (along with the associated processes described in NASA-STD-7009A) was originally developed for use with all M&S and applications throughout NASA. Similarly, MURM has no limitation on intended application but rather states that it "should be capable of application to any M&S no matter the category, type, domain, or application". GM-VV is intended for any NATO M&S systems design, development, and employment process, and it is applicable to any M&S scope, technology, and application domain.

GMC were developed for human-centered machine learning models for computer vision and natural language processing applications. The developers believe the method can be adapted for other applications. Unlike the previous methods, GMC are not an assessment method. Instead, they provide a framework to increase transparency in communicating the intended use, limitations, and trade-offs of a machine learning model.

4.2.2 Methodology Maturity

CAS, PCMM, and MURM were originally developed before wide adoption of MBSE and modern-day digital engineering practices. Nevertheless, valuable lessons can be learned from these methods. CAS and PCMM are by far the most mature methodologies. They have been applied to numerous programs for over a decade by NASA and Sandia National Laboratories, respectively; both methodologies have been refined during that time based on lessons learned from their application. NASA-STD-7009A is currently undergoing another update with Revision B expected to be published in 2022. The companion handbook, NASA-HDBK-2009A [11], provides detailed guidance to assist in complying with NASA-STD-7009A. The PCMM has undergone three updates since it was originally developed with the most recent in 2013. In 2011, a spreadsheet tool to assist in completing a PCMM assessment was developed but it is not publicly available.

MURM was developed in 2009-2011 as a Deputy Assistant of Defense for Systems Engineering (DASD(SE)) sponsored task. However, the MURM has not been widely applied to DoD programs. The MURM developers believe one of the reasons hindering broader use is that the method is complex and time consuming—likely to be two of the underpinning reasons of why MRL methods are not broadly used by DoD programs. Development of interactive tools to help users apply the method could help in that regard. However, it is important to note it is not solely a technical challenge to broader adoption; programs still need to prioritize and plan for the time and resources to apply MURM (or any other assessment method).

GM-VV was developed under the auspices of the Simulation Interoperability Standards Organization (SISO), and guides were published in 2012-2013. The methodology has been applied to numerous NATO programs with many case studies described in related documents.

SIRL, SSTRL, MAL, and MRL methods are under development or have been developed more recently, some with MBSE and digital engineering practices in mind. However, that also means there is very limited information about use of these methods in practice. The draft SIRL standard is under development (draft open for comment as of the writing of this report) and it is unknown at this time if it is being used in any programs. TRMC used the SSTRL on a Navy torpedo program and reported that having a common language with the contractor for defining simulation readiness levels was helpful in maturing the simulation capability from SSTRL 3-4 to SSTRL 5-6. The MAL has been used on one customer program as of the writing of this report, but the details are not publicly available. The Aerospace Corporation developed tools to support Software MALs

including an Evaluator Tool which computes the raw detailed scores directly from a native Rhapsody® model format and a Scorer Tool which computes the MAL score for overall software based on manual inputs of MAL detailed criteria. Initial development of Enterprise MAL tools has begun. However, these tools are not yet publicly available. MRL is the least mature methodology as it is still under development. As of the writing of this report, STAT COE has released a document for limited distribution that is pending approval for unlimited distribution.

Google Model Cards are based on a Google research paper published in 2019, and Google made a pre-release version of the Model Card Toolkit (MCT) in 2020. The MCT version 1.0.0 was released 2 August 2021. The MCT streamlines and automates the creation of Model Cards; however, in its current state it may not be applicable to the broad set of M&S.

4.2.3 Factors Assessed

Given the diversity in scope and purpose, it is not surprising there is also significant variation in the factors considered by the methods. Nevertheless, there are many factors which are common to multiple methods, and they are as follows:

- Verification (CAS, PCMM, MAL, GM-VV)
- Validation (CAS, PCMM, MRL, MURM, MAL, GM-VV)
- Results Uncertainty (CAS, PCMM)
- Results Sensitivity/Robustness (CAS, PCMM)
- Fidelity (PCMM, SSTRL, MRL)
- Referent Data Pedigree/Authority (CAS, MRL)

There are some factors unique to a given methodology which are also relevant to DoD. Notably, only MAL explicitly considers Security and Vulnerability as a distinct factor. Also, the CAS originally included People Qualifications as a factor. However, when the NASA standard was revised in 2016 this was removed as a factor in the CAS and moved to a reporting requirement in the overall NASA M&S process.

GMC, which was primarily focused on machine learning (ML)-based models, considers different factors from the other methods. The factors are tailorable based on the project. Some examples include cultural, demographic, phenotypic and intersectional groups.

4.2.4 Scoring Approaches

The scoring approaches used by these methods can be grouped into one of two broad categories: those that do not aggregate the factors assessments into one overall score and those that do. CAS, PCMM and MURM specifically recommend *against* score aggregation for a variety of reasons. NASA-STD-7009A Change 1 eliminates attempts to aggregate into a single score for the CAS. This change "simplifies and clarifies reporting, makes it less abstract, and eliminates the problems/limitations associated with non-numerical aggregation." The PCMM strongly recommends against score aggregation because the factors are conceptually independent and "using the average value would be analogous to claiming the breaking strength of a chain by averaging the strength of each chain in the link." Nevertheless, PCMM acknowledges the pressures to condense information for decision makers, and if aggregation must be done then the minimum, average and maximum score should be reported (referred to as an "aggregation triple"). The recommendation against score aggregation, particularly the aggregation of ordinal ranking scales,

has been endorsed by many of the experts that participated in the August and September workshops as well as other experts [46].

Another unique feature of the CAS and PCMM methods that distinguishes them from the other methods is the explicit separation of assessed maturity of the M&S from the required maturity of the result. Determining the required maturity necessarily includes a risk assessment which is highly variable depending on the program, decision-maker, etc. Trying to incorporate requirements into the assessment scale itself would be difficult and hard to interpret. Instead, these methods base their scales on intrinsic and contextual information quality.

MURM calculates an overall Use Risk score for each M&S capability (e.g., software elements, hardware elements, data) assessed. However, each assessed capability is not equal and aggregation across capabilities is not recommended (although it does state further research is needed into the mathematical logic behind aggregation). Furthermore, the recommended scorecard for reporting MURM results includes all the factors in addition to the overall Use Risk score for each assessed capability.

SIRL aggregates scores *within* the factors, but the scores are not aggregated *across* factors. Thus, separate scores are reported for each of the five factors considered. The draft standard notes the SIRL scores are independent and intended to enable comparisons *between* simulations; a single roll-up SIRL score for an individual simulation is neither informative nor useful.

GM-VV also falls into the group that does not advocate for score aggregation; in fact, it does not recommend any scoring at all. Instead, GM-VV calls for defining acceptability criteria for each factor based on the intended use of the M&S. Then, an acceptance recommendation is made based on those acceptability criteria. Evidence supporting the acceptance recommendation is documented in an Argumentation Structure which provides traceability and transparency.

MAL and SSTRL belong to the second group where overall scores are calculated or assessed. However, there is a distinction even within this group. SSTRL consists of an ordinal scale (2-9) with a qualitative description for each level in the same vein as the original Technology Readiness Levels; the factors considered are not scored independently and then aggregated. MAL, on the other hand, scores factors and sub-factors separately and then aggregates into an overall score. The aggregation formulas for the MAL have been published, but the transformation of the score to the final MAL level is not yet publicly available.

Again, GMC takes a fundamentally different approach from the other methods as there is no scoring or scale recommended. Instead, a Model Card consists of descriptive explanations which should communicate the intended use, limitations, and trade-offs of the machine learning model. The accuracy of a Model Card depends on the integrity of the creator. Therefore, the recommendation is a Model Card should be just one of many tools/methods for assessing the machine learning model.

There was significant discussion during both workshops about the pros and cons of aggregation into one overall score. Previous studies identified shortcomings with applying to TRLs to software and M&S activities [39]. These findings influenced the scoring approaches in the PCMM and CAS. The consensus among the workshop participants was these same shortcomings apply to model readiness assessment and aggregation into one overall score is not recommended.

4.2.5 Summary

The key findings after the literature review and two workshops are:

- None of the assessment methodologies themselves describe how to develop and mature a model. The GM-VV is a generic framework for M&S V&V more than a readiness assessment methodology. Only NASA-STD-7009A describes an overarching M&S development process, and the CAS is one step within that overall process.
- The scope and purpose of the methods vary widely, and not all the methods are relevant to the problem of assessing model readiness.
    - The MBCM is intended for organizational processes and transformation and thus is not relevant to assessing model readiness.
    - The CAS, MURM and GM-VV were each developed for use with all types and applications of M&S.
    - The remaining six methods have narrower and varying scopes and thus are relevant in varying degrees.
- Maturity of the methods themselves varies widely.
    - The CAS and PCMM are by far the most mature having been applied to numerous programs for over a decade by NASA and Sandia National Laboratories, respectively; both methodologies have undergone revisions during that time based on lessons learned from their application.
    - The MRL and SIRL are the least mature as they are still under development.
- Despite the diversity in scope and maturity, there are many factors which are common across multiple methods.
    - Verification (CAS, PCMM, MAL, GM-VV)
    - Validation (CAS, PCMM, MAL, GM-VV)
    - Results Uncertainty (CAS, PCMM)
    - Results Sensitivity/Robustness (CAS, PCMM)
    - Fidelity (PCMM, SSTRL, MRL)
    - Referent Data Pedigree/Authority (CAS, MRL)
- The definitions of Verification and Validation are not necessarily the same.
    - The CAS, PCMM and GM-VV use DODI 5000.61 and RPG definitions.
    - It is unclear from the literature whether the MAL uses the DODI 5000.61 or IEEE 1012 definitions.
- One notable gap in the previous list of key factors is Security and Vulnerability. Only the MAL explicitly considers this factor.
- The GMC is the only method focused on human-centered machine learning models. Like cybersecurity, machine learning and artificial intelligence are relatively new dimensions to be considered in the M&S context. It is important these are not overlooked when considering adoption of a model readiness assessment methodology.
- The scoring approaches used by these methods can be grouped into one of two broad categories: those that do not aggregate the factors assessments into a single overall score (CAS, PCMM, MURM, GM-VV, SIRL) and those that do (SSTRL, MAL). Previous studies concluded that a single TRL number is not appropriate for M&S activities, and the workshop participants generally agreed those concerns also apply to model assessment methodologies which aggregate into a single overall score.

- The CAS and PCMM explicitly separate assessed maturity of the M&S from the required maturity of the result (unlike the other methods). This was a conscious decision when the methods were developed because determining the required maturity necessarily includes a risk assessment which is highly variable depending on the program, decision-maker, etc. Trying to incorporate requirements into the assessment scale itself would make it hard to interpret.
- Training for managers, developers and analysts is crucial to the successful application and adoption of most of the methodologies reviewed.

4.3 Systems Engineering Framework for M&S

DTE&A identified the need for a framework that enhances the way programs use M&S, regardless of the pathway selected. To address this need, MITRE performed a literature search on different frameworks that focus on the lifecycle of M&S:

- *Best practices for the development of models and simulations* [51] synthesized the results of an assessment of seven major systems engineering frameworks into a new framework that consisted of five phases: requirements development, conceptual analysis, product design, product development, and product testing.
- *Standard for Models and Simulations (NASA-STD-7009A)* [10] establishes requirements and provides recommendations for the design, development, and use of M&S, recommendations for the analysis and presentation of M&S-based results, and guidance for the use of a CAS for assessing the credibility of M&S-based results. CAS involves eight factors that were grouped into three broad categories. The three categories and subordinate factors are M&S development (data pedigree, verification, and validation), M&S use (input pedigree, uncertainty quantification, and results robustness), and supporting evidence (use history, and M&S process/product management).
- *AcqNotes for Modeling and Simulation* [52] outlines eleven steps involved in developing a model, designing a simulation experiment, and performing simulation analysis based on the paper entitled *Introduction to Modeling and Simulation* [53]: identify the problem, formulate the problem, collect and process real system data, formulate and develop a model, validate the model, document model for future use, select appropriate experimental design, establish experimental conditions, perform simulation runs, interpret and present results, and recommend further courses of action.
- *Handbook on Statistical Design & Analysis Techniques for Modeling & Simulation Validation* [15] defines the VV&A process in nine steps: develop the intended use statement, identify the response variables or measures, determine the factors that are expected to affect the response variable(s) or that are required for operational evaluation, determine the acceptability criteria, estimate the quantity of data that will be required to assess the uncertainty within the acceptability criteria, iterate the model-test-model loop until desired model fidelity is achieved, verify that the final instance of the simulation accurately represents the intended conceptual model (verification), determine differences between the model and real-world data for acceptability criteria of each response variable using appropriate statistical methods (validation), and identify the acceptability of the model or simulation for the intended use (accreditation).
- *Model-Based Engineering Diamond* [54] depicts diamond-shape, integrated engineering pathway for the development of a physical system described separately with the classical

systems engineering "Vee" on the lower V and the mirror reflection of its digital twin on the top V. The mirror image of the digital twin progresses through a sequence of steps that correlate exactly to the progression of the physical system. This does not imply that each M&S step occurs in parallel with or at the same time as its mirror in the physical system. Because models can inform decisions in any step in a system's lifecycle, an entire M&S lifecycle can exist within one or more steps of the system engineering process.

While all the frameworks reviewed contribute to different aspects of the M&S lifecycle, and although all the pieces required to build a comprehensive best practice exists somewhere in those ecosystems, there is no single framework or best practice that integrates all aspects of the M&S lifecycle--development, evolution, integration, maturity, VV&A, and use. Thus, the idea evolved to outline a framework that leverages acquisition guidelines, M&S standards such as NASA-STD-7009A Change 1 and SIRL (in draft), best practices and standards in systems engineering, the application of technology readiness levels (TRLs), guidance from RPG for M&S VV&A, the use of statistical methods for M&S VV&A, and the application of digital engineering principles. Figure 2 illustrates such a framework using as example the Major Capability Acquisition (MCA) pathway, which is centerpiece for the AAF illustrated in Figure 3. Figure 3 outlines the various acquisition pathways that afford acquisition authorities and PMs opportunities to develop strategies and employ processes that match the characteristics of the capability being acquired. Each acquisition pathway is tailored to the unique characteristics and risk profile of the capability being acquired. *DoD Test and Evaluation Instruction* (DoDI 5000.89) establishes the T&E policy and procedures across the AAF, which requires the TEMP to include a strategy starting at MS-B that identifies live T&E and M&S events to generate the data for the evaluation. For M&S-generated data to inform programmatic and technical risks as well as major decisions, M&S must be fully accredited.

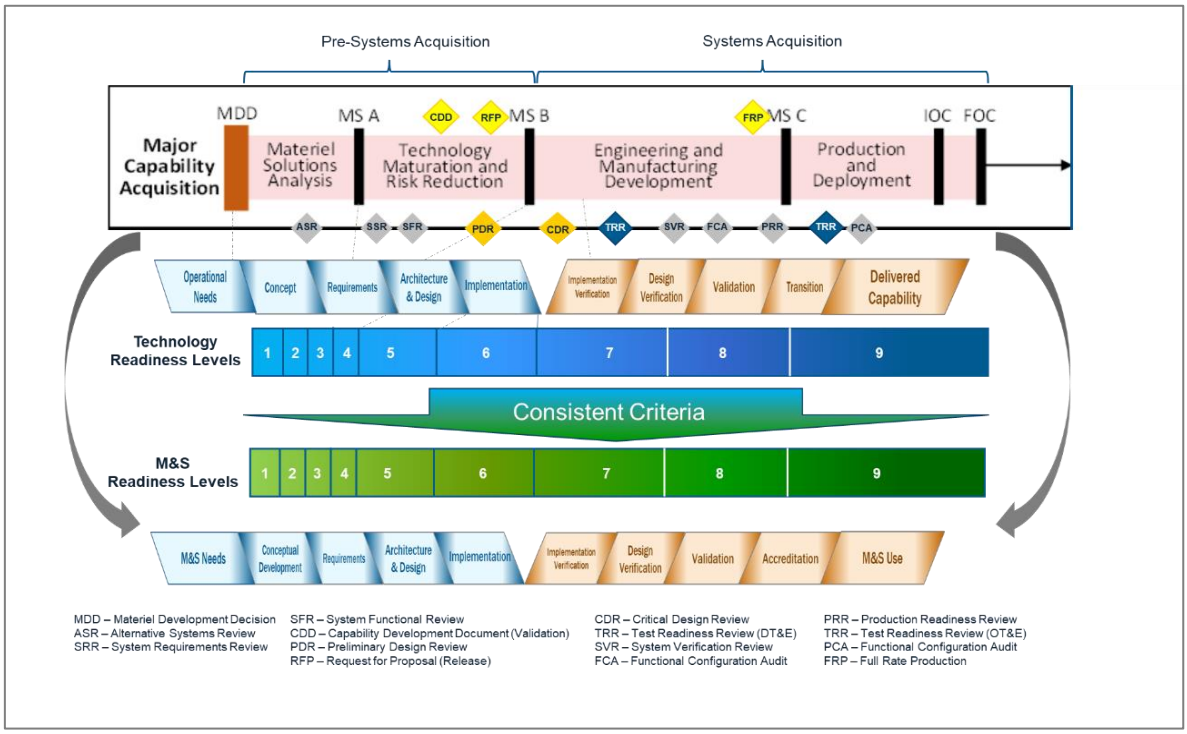


Figure 2. Systems engineering, M&S engineering, and readiness levels

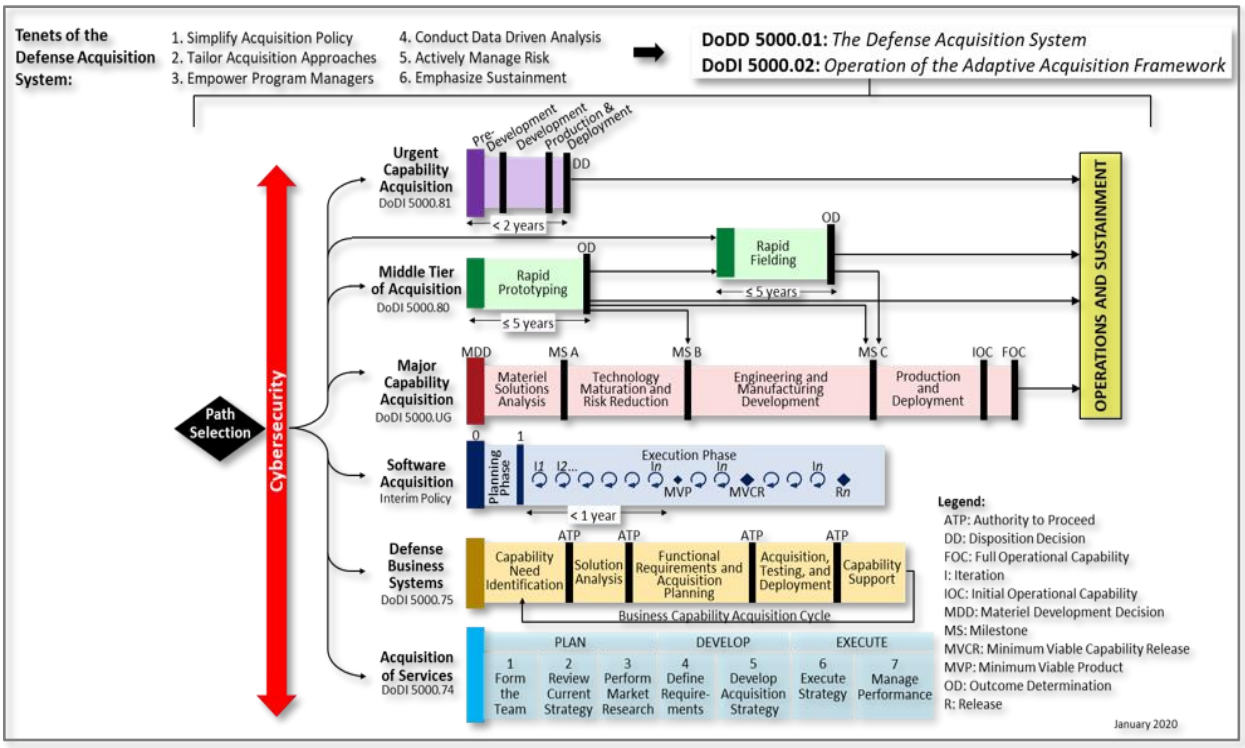


Figure 3. The Adaptive Acquisition Framework (AAF)

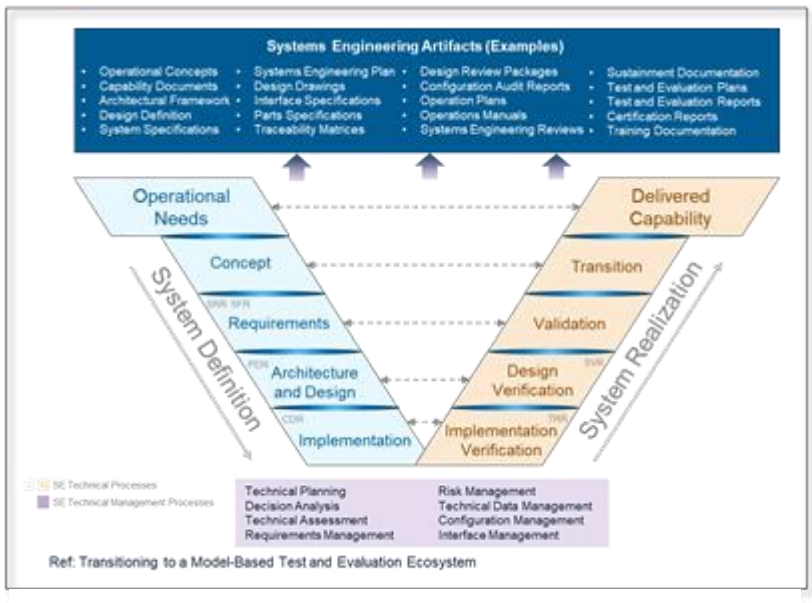


Figure 5. Adaptation of System Engineering "Vee" to M&S

Forsberg and Mooz [45] envisioned "the technical aspect of the project lifecycle" as a V-shaped diagram that starts with user needs on the upper left side and ends with a user-validated system in the upper right side. Because the "Vee" process model clearly illustrates the relationships between activities of system design [47], it has been widely adopted by almost every systems engineering community. While advances in systems engineering over time, different interpretations of the model, and the tailoring of the systems engineering process has led to variations of the systems engineering "Vee", the fundamental concepts and principles that were established over 30 years ago remain essentially unchanged. Figure 5 illustrates a typical systems engineering "Vee".

Standard for the Application of Systems Engineering on Defense Programs (IEEE 15288.1) [49], adopted by DoD in 2015, provides a common framework that describes both the technical processes and the technical management processes that are typical for the full system life cycle.

Standard for Technical Reviews and Audits on Defense Program (IEEE 15288.2) [50] is a companion to IEEE15288.1 that provides definition, requirements, and evaluation criteria for the technical reviews and audits within DoD systems engineering.

The "V-model" is a graphical representation of the main engineering activities in the systems development lifecycle and a valuable tool for visualizing the management of the systems engineering process. It is a variant of the traditional waterfall model for system development in which the bottom half of the waterfall model is bent upwards so the activities on the right verify and validate the development activities on the left hand. The layered technical processes on both sides of the "Vee" produce artifacts that are familiar to the systems engineering community. On the left side of the "Vee" definition and decomposition activities produce details about the design while on the right side of the "Vee" verification and validation activities produce details about the use. The technical management processes are listed in the bottom of the "Vee".

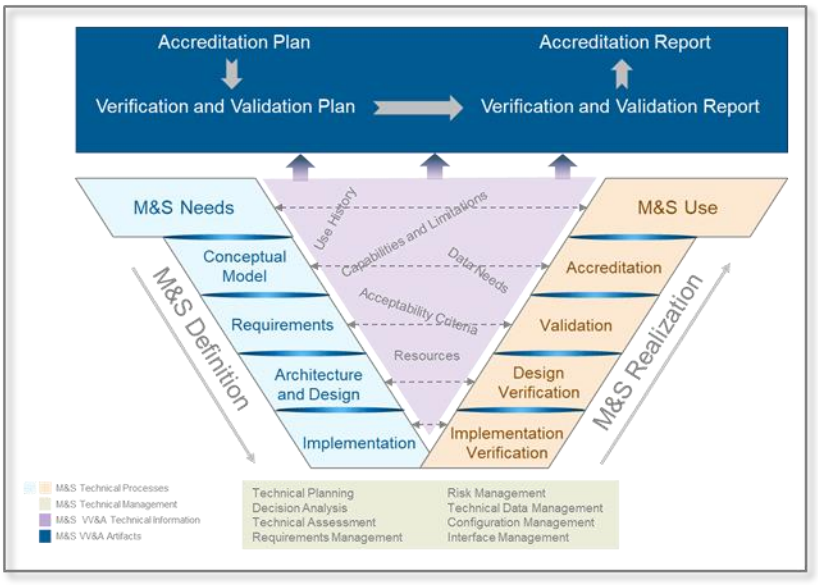


Figure 4. Adaptation of system engineering "Vee" to M&S

Because the success of an M&S program or strategy hinges on the soundness of requirements, affordability, and executability of the development or acquisition strategy, the adaptation of the systems engineering process to M&S is an appealing alternative to achieve that goal. Figure 4 illustrates an adaptation of the systems engineering "Vee" to the development lifecycle of M&S. In theory, there is no difference between engineering a system and engineering a model (in practice there is!). Like for the systems engineering "Vee", the left side produces details about the design while the right side produces details about the VV&A process. Those details are captured in the MIL-STD-3022 accreditation plan, V&V plan, V&V report, and accreditation report.

Table 4, developed by MITRE, provides a description of the desired state for each model development level plus an equivalent description in terms of TRLs. Appendix B outlines some top-level considerations, in the form of questions, that can help users tasked with generating M&S documentation, or authorities reviewing it, determine the level of development of the M&S. This information is not intended to be unique, but rather to complement guidance provided by RPG or other standards.

The concept illustrated in Figure 2 could be applied to the development of one or more models needed to inform decisions in the lifecycle. For instance, it could be used to develop a single simulation for a single intended use. It could also be used to acquire and incorporate into the lifecycle a single simulation developed commercially. Likewise, it could be used to develop or acquire multiple simulations for different intended uses. In this case, the proposed SIRL standard is essential to ensure the interoperability between the various models.

Table 4. M&S Engineering Levels

Level	Equivalent TRL Description	M&S Desired State
M&S Needs	Scientific research begins to be translated into applied research and development.	M&S is needed for generating data to quantify key measures that will be used to evaluate system capability and the intended use of the model is clearly defined and documented. The role M&S will play in the program, objectives that will be fulfilled, decisions that will be made based on the M&S results, and consequences resulting from no or erroneous M&S outputs are understood. Components of the system to be modeled, configurations of interest, and standards to be used are identified.
Conceptual Model	Invention begins. Once basic principles are observed, practical applications can be invented. Applications are speculative and there may be no proof or detailed analysis to support the assumptions.	A conceptual model that represents the intended use is defined and built. System specifications, performance data, and referent data inform the definition of the conceptual model. Acceptability criteria is defined, and potential sources of uncertainty identified. Specific segments of the conceptual model, M&S requirements, acceptability criteria, quantitative and qualitative measures of performance, and authoritative sources of truth are correlated to validate the conceptual model.
Requirements Definition	Active research and development are initiated. This includes analytical studies and laboratory studies to physically validate analytical predictions of separate elements of the technology.	Requirements are defined, preferably in a system modeling language. A complete description of the model exists, including functionality, basic parameters, information flow, sources of uncertainty, configuration (stand-alone or federated), and data required to populate and execute the model (i.e., input and output variables, constants, operational conditions, descriptive metadata, runtime requirements, and authoritative sources for each item) are defined. The scope of the VV&A effort is established based on intended model use, acceptability criteria, and uncertainty quantification.
Design and Development	Basic technological components are integrated to establish that they will work together.	The conceptual model is translated into schematics and diagrams, including model-based systems engineering simulations that are consistent with the intended use of the model. Hardware and software specifics are fully defined. Information about how the model and simulations are organized, constructed, and executed are available. The M&S development plan is published, which outlines basic assumptions, capabilities and limitations, and risks associated with development, testing, and M&S use.
Implementation	Fidelity of breadboard technology increases significantly. The basic technological components are integrated with reasonably realistic supporting elements so it can be tested in a simulated environment.	The model is translated into a programming language, preferably systems modeling language like SysML. The simulation model is documented, and a user manual is being developed. Standards are being applied throughout the different domains.
Implementation Verification	Representative model or prototype system is tested in a relevant environment. Represents a major step up in a technology's demonstrated readiness.	The model is correct, complete, consistent with functional requirements, and simulation execution are described. Qualitative assessment of uncertainty for model outputs are documented. Specific segments of the design are correlated to the conceptual model and adherence to standards and best practices are evaluated. Both the *Accreditation Plan* and *Verification and Validation Plan* are published, which serve as reference against which to measure M&S representations.
Design Verification	Demonstration of a prototype near, or at, planned operational system and in an operational environment.	All requirements are verified and documented in the *Verification and Validation Report.* Simulation tests verify the model code is free of bugs, the model assumptions are valid, and the simulation execution behaves as intended (i.e., results obtained from statistical analysis match, relative to acceptability criteria, outputs for use cases that are familiar to designers and users).
Validation	Technology has been proven to work in its final form and under expected conditions.	The M&S represents the real world for its intended use. Simulation tests and statistical techniques are used to characterize the similarity between model and system outputs with respect to the model intended use, within the acceptability criteria, and under quantified realistic operational conditions. Validation results are documented in the *Verification and Validation Report.*
Accreditation and Use	Actual application of the technology in its final form and under mission conditions.	The model is approved for its intended use. Simulation tests are conducted in representative real-world operational conditions and inferential statistical techniques and uncertainty characterization methods quantify model outputs to inform decisions.

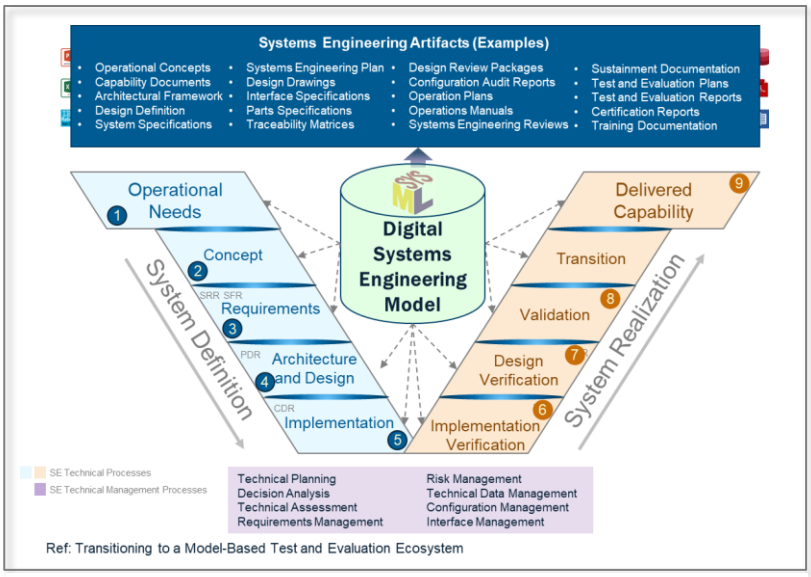


Figure 5. Systems engineering "Vee" in digital environment

The digital engineering strategy, outlined in the following section, is envisioned to improve the speed of and quality of execution of the systems engineering process. Figure 6 illustrates a systems engineering "Vee" in a digital environment. The desired state envisions the integration of systems engineering models across multiple disciplines or domains and throughout the lifecycle. Artifacts are instantiations of integrated digital systems engineering models, and their content is synchronized, consistent, and concurrent with system maturity.

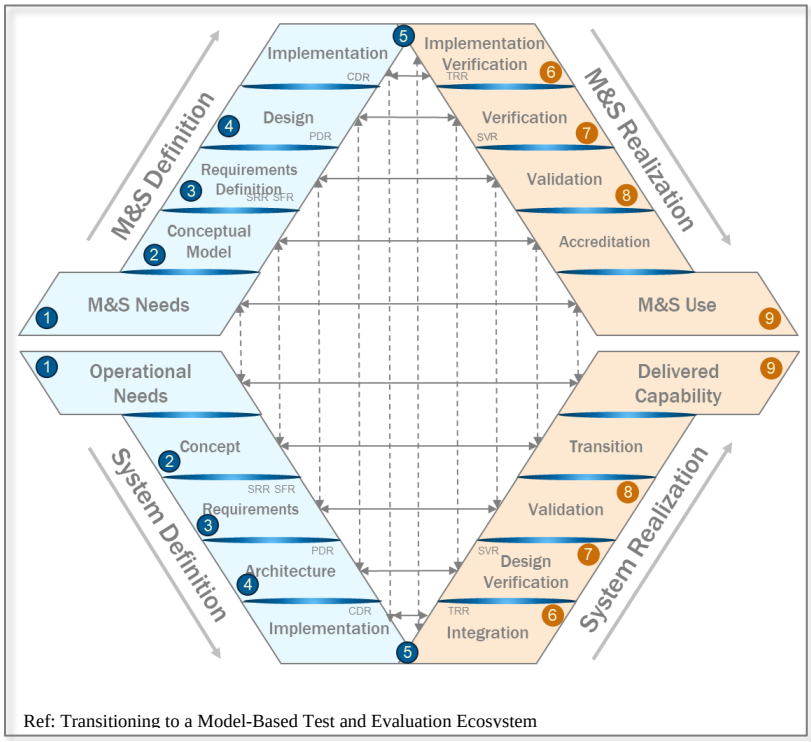


Figure 6. Correspondence between M&S and system development

Figure 7 illustrates the situation where a model and a system are concomitantly developed, although not necessarily in a synchronous fashion. The lower "Vee" represents the classical systems engineering "Vee" for the system while the top "Vee" represents the mirror image of the system engineering "Vee" but for the model development. The horizontal lines represent the typical relationships between the left and right sides of the corresponding "Vees" while the vertical lines represent connections between the phases of the system and model "Vees". As in the case of the Model-Based Engineering Diamond, the M&S steps do not need to occur in parallel with or concurrently as its mirror in the physical system. An entire M&S lifecycle can exist within one or several steps of the system engineering process since M&S can inform decisions in any step in a system's lifecycle. For example, an M&S can be developed to support

architectural trade-offs during system definition; thus, the lifecycle of M&S will be dedicated to only to that intended use.

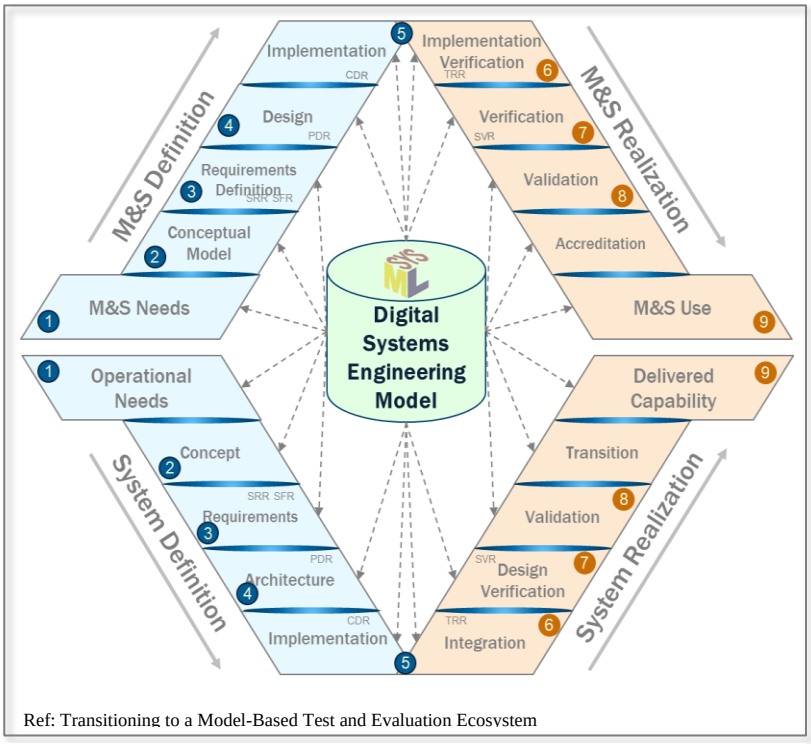


Figure 7. Correspondence between M&S and system development in a digital environment

Figure 8 illustrates the correspondence between model development and system development in a digital environment. The interpretation of Figure 8 is like the interpretation of Figure 7 except that in this case the relationships between left and right sides of "Vees" as well as between model and system "Vees" are achieved via the digital systems engineering model.

4.4 Digital Engineering Strategies to Improve M&S

DoD Digital Engineering Strategy [55] sets five goals:

- Formalize the development, integration, and use of models to inform enterprise and program decision making.
- Provide an enduring, authoritative source of truth.
- Incorporate technological innovation to improve the engineering practice.
- Establish a supporting infrastructure and environments to perform activities, collaborate, and communicate across stakeholders.
- Transform the culture and workforce to adopt and support digital engineering across the lifecycle.

Consistent with *DoD Digital Engineering Strategy*, DoDI 5000.89 requires programs to digitally represent the system in a mission context and to use (to the largest extent possible) a digital ecosystem that integrates authoritative sources of models, data, and test artifacts (e.g., test cases, plans, deficiencies, and results) to improve efficiencies across the integrated test and evaluation (IT&E) continuum. T&E needs to transform and adopt digital engineering strategies to improve both the M&S V&V process as well as T&E assessments over the lifecycle. Figure 9 and Figure 9 illustrate ideas related to strategies that could help with the digital transformation. Because the digital transformation is on its infancy, the ideas illustrated in those figures are just conceptual.

4.4.1 M&S V&V Conceptual Diagram

Figure 9 illustrates a conceptual diagram in a systems modeling language (SysML) platform, Cameo System Modeler ™ in this case, for the M&S V&V process. SysML is a general-purpose language that can be used for the analysis, design, and verification of complex systems using graphical notations as well as to gain efficiencies and effectiveness in other activities of the T&E ecosystem.

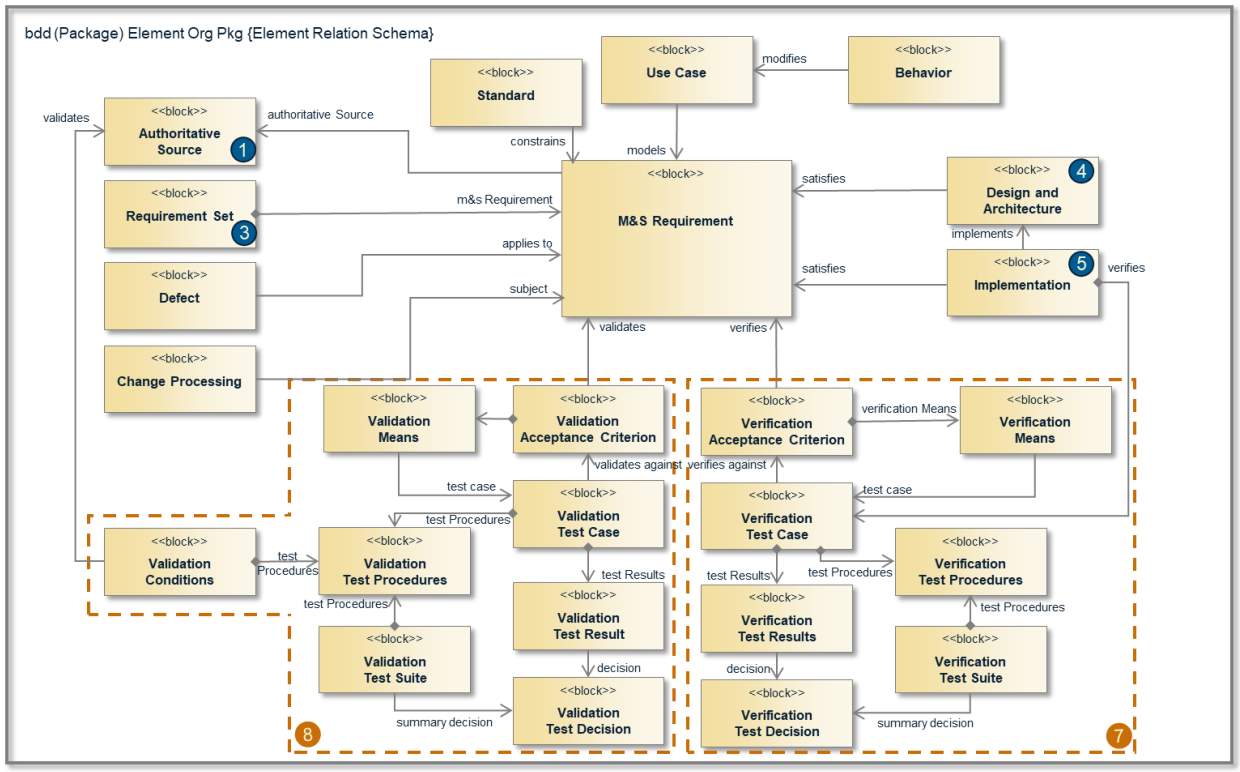


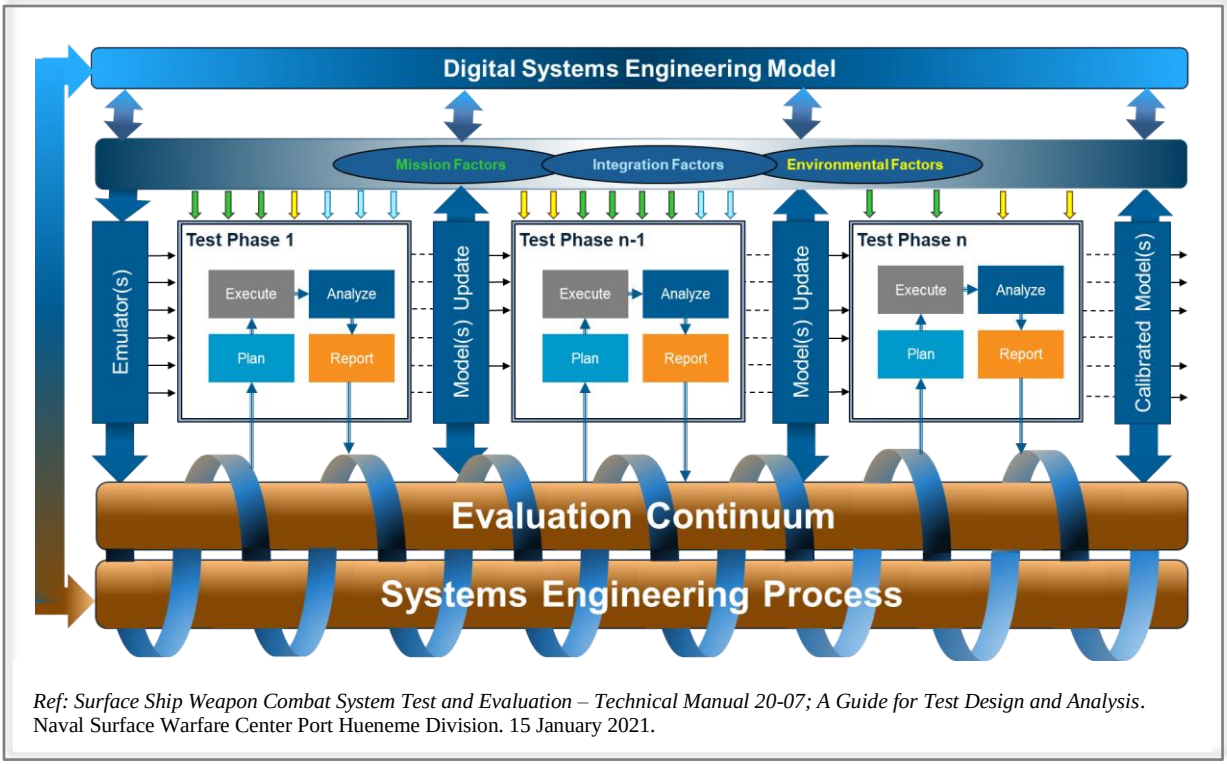
Figure 8. Concept SysML diagram strictly for M&S V&V

4.4.2 Sequential Test Strategy

Figure 9 [56] depicts a framework for sequential T&E to efficiently integrate successive phases of test (any of which could be a live or simulated event) through the system lifecycle to immediately learn crucial information about the capability before proceeding to the next phase of test or to the next phase of system development. Figure 10 illustrates how a capability model (a.k.a. authoritative source of truth) matures over time through a series of updates that occur at the end of each phase of a sequence of tests (e.g., CT&E, DT&E, IT&E, TECHEVAL, and IOT&E). The initial capability model could be a mathematical model, a crude physics model, or a DoD Architectural Framework (DODAF) view. A consistent set of response variables is evaluated through the different phases of test while the factor space (combination of environmental, mission, or integration factors is adapted to both the stage of development of the capability and test resources (it may not be necessary or practical to follow this approach all the time).

The response variables may include key performance parameters (KPPs), critical technical parameters (CTPs), or mission-oriented response variables (MORVs). Those responses are identified early in system development and are quantified from test to test to evaluate how the

system matures over time. While sometimes the definition of those measures evolves, their calculation is often identical from test to test. Specific measures can also be used for certain tests.



Ref: Surface Ship Weapon Combat System Test and Evaluation – Technical Manual 20-07; A Guide for Test Design and Analysis. Naval Surface Warfare Center Port Hueneme Division. 15 January 2021.

Figure 9. Framework for sequential test and evaluation

The complexity of the factor space typically increases over time. The factors and conditions are screened between events to update the model with only those factors that influence system performance. The factor screening process maximizes the knowledge gained from a phase of testing and makes T&E more efficient and effective. The model updates are also consistent with the Model-Test-Model (MTM) paradigm, in which the combination of live and model test results can be used to inform the next step in model evolution until the model acceptability criteria can be achieved. The Evaluation Continuum could represent planning steps in the IDSK-EF as well as independent developmental and operational evaluations and reporting activities.

5. Summary and Recommendations

M&S will continue to be a critical component of an overall test program strategy. To minimize the risks associated with M&S-informed decisions, the interdependencies between T&E and M&S must be strengthened. That is, focus should continue to be placed on improving the effort of using data from test events for the development and VV&A of the M&S as well as using data collected during M&S events to inform T&E assessments and critical decisions in the lifecycle. Below are detailed recommendations to strengthening those relationships.

The survey of the TEMPs of five major acquisition programs described in Section 3.1 revealed the need to improve the global view of the role M&S plays in the test strategy of acquisitions programs when re-using models. MITRE recommends DTE&A and DOT&E encourage acquisition programs to use the Defense M&S Catalog [57]—which provides a "card catalog" level of detail

about M&S tools, data, and services—to allow for (1) a better understanding of interdependencies and interrelationships of models across programs; and (2) improving the re-use of models.

The review of policy and guidance in Section 3.2 revealed the need to update M&S policy and guidance with respect to M&S content in the TEMP. MITRE recommends DTE&A and DOT&E jointly update the DOT&E Guidebook with consistent guidance for LFT&E, DT&E, and OT&E. As a starting point, guidance should require programs:

- Describe both the M&S and how it will be used to generate data to inform T&E decisions.
- Identify mission-focused measures that will be quantified with M&S data to evaluate system performance and how will they inform test objectives throughout the lifecycle.
- Provide an initial list (for Milestone A) or refined list (Milestone B and Milestone C) of factors and levels that could influence each of the mission-focused measures and how they will be varied or controlled during each stage of testing.
- Describe the use of statistical techniques to estimate resources throughout the lifecycle, including Milestone A Request for Proposal (RFP).
- Identify data needed by the models (anticipated inputs) and the types of output expected.
- Identify and describe M&S resource requirements in Part IV of the TEMP.
- Provide updated or completed test designs to support resourcing throughout the lifecycle and to account for any new information.
- Identify organizational responsibilities for M&S VV&A, including responsibilities of T&E WIPT for test design purposes.
- Describe the overall M&S "flow" (e.g., where the output of one model will be required as input for another) if multiple models will be used.
- Describe the techniques that will be used for M&S V&V.

MITRE also reviewed drafts of *Engineering for Defense Systems Guidebook*, *Systems Engineering Guide*, and *T&E Enterprise Guidance* and recommended a consolidation and restructuring of the guidance to describe the program management, system engineering, T&E, and M&S activities that enable successful execution of each AAF model pathway. M&S-focused recommendations were aimed at ensuring processes related to the use of M&S tools and their VV&A are consistent and mutually reinforced across program management, systems engineering, and T&E. They are also intended to remove some of the hurdles the community is facing. MITRE recommends DTE&A adopt the recommendations provided to DTE&A Chief Engineer for updates to the engineering guidebooks.

As discussed in Section 3.3, numerous instructions set expectations for documenting M&S (e.g., intended use, capabilities, VV&A methodologies, etc.). Similarly, documents like the *Handbook on Statistical Design & Analysis Techniques for Modeling & Simulation Validation* describe numerous statistical techniques available for M&S V&V. However, these regulatory and guidance documents do not prescribe the V&V approach appropriate for each application. Since test designers and analysts have numerous V&V techniques at their disposal, their desired degree of rigor and formality often guides them towards preferred V&V methods. Each V&V method has unique requirements, so their use will often be guided by the underlying model types, and may be constrained by available data, skills, and resources. Test designers and analysts must always choose the test design or statistical method that fits the problem. The incorrect selection of a test design or statistical method for M&S V&V can have significant impact on a program's cost, schedule, and performance. MITRE recommends DTE&A work with DOT&E and Services to

improve the workforce's knowledge, skills, and abilities for selecting the appropriate statistical method for M&S V&V.

The growing dependency programs have in M&S is leading to a greater reliance on predictive modeling for decision-making. UQ provides a quantitative measure of the extent to which model results can vary based on the real-world variability of inputs and provides a frame of reference for making decisions based on M&S results. UQ is not a mature field and there is no guidance within the DoD T&E community on how to apply the concept of UQ to M&S VV&A. ASME is actively developing VVUQ standards across different fields and domains. MITRE recommends DTE&A (1) leverage the ASME VVUQ standard to further guide policy and develop best practices tailored to the needs of the DoD T&E and M&S communities, including the characterization of uncertainty within the context of risk; and (2) identify and sponsors training opportunities to improve the T&E workforce knowledge, skills, and abilities in the field and generates case studies to help the community improve the understanding of this topic.

The IDSK-EF, discussed in Section 4.1, accommodates decisions and objectives involving both (1) T&E needs to inform VV&A of M&S and threat models; and (2) M&S needs to inform T&E assessments. The IDSK-EF concept is a step forward for test planning and analysis in that statistical test design, a critical enabler for M&S VV&A, is incorporated for the first time into a DT&E decision support system. The IDSK-EF concept is still under development (white paper disclosure occurred on 16 September 2021). The IDSK-EF team and Services are working on identifying pilot programs that will help define the content and structure of the IDSK-EF. MITRE recommends DTE&A continue working with the IDSK-EF in FY22 on identifying a good opportunity to implement M&S strategies into the framework. Potential M&S objectives to be addressed in the IDSK-EF include (1) formulation of M&S needs (including intended use); (2) validation of Conceptual Model; (3) definition of M&S requirements; (4) design; (5) implementation; (6) implementation verification; (7) design verification; (8) validation; and (9) accreditation.

The literature search and workshops on model readiness assessment methods discussed in Section 4.2 validated the need to understand the level of model maturity throughout the lifecycle (requirements, capabilities and limitations, V&V methods, uncertainty quantification, etc.), yet none of the assessment methodologies describe how to develop and mature a model though the lifecycle. Only *Standard for Models and Simulations*, NASA-STD-7009A Change 1, establishes practices to ensure requirements are applied to design, development, and use of models. Even though many assessment factors are common across multiple methods, the scope and purpose of the methods vary widely, and even the maturity of the methods themselves varies. MITRE recommends DTE&A address the gap of M&S maturity assessment over the complete lifecycle. Additionally, if a model readiness assessment methodology is selected for adoption and implementation, MITRE recommends DTE&A includes requirements to:

- Define the purpose, scope, criteria and use cases against which any model readiness assessment methodology will be evaluated, including cybersecurity and machine learning/artificial intelligence considerations.
- Plan for and allocate resources to train managers, developers, analysts, and users about M&S concepts such as the types of uncertainty (e.g., epistemic vs aleatory), uncertainty characterization and sensitivity analysis; NASA found that having informed M&S

users/customers significantly helps with successful adoption and implementation of the assessment methodology.

As discussed in Section 4.3, there are several frameworks that contribute to different phases of the M&S lifecycle. While all the pieces required to build a comprehensive best practice exists somewhere in these ecosystems, particularly in NASA-STD 7009A Change 1, there is no other single framework or best practice that integrates all aspects of the M&S lifecycle—design, development, evolution, integration, VV&A, and use. The conceptual systems engineering framework outlined in the section for the MCA pathway, which is the brainstorm product of a small number of engineers, suggest there is potential to develop such a comprehensive best practice or standard. MITRE recommends DTE&A partner with DOT&E and OUSD(R&E) EPS to define a general cradle-to-grave framework that can provide adequate guidance for M&S users or reviewers of M&S documentation and that can be tailored to each AAF pathway. In particular, the framework should focus on using T&E requirements to drive the development of M&S (including the identification of the intended use) as well as starting collection of data for V&V as early as possible during system development.

The combination of *DoD Digital Engineering Strategy* and DoDI 5000.89 requires a transformation of the DoD acquisition paradigm—digitally represent the system in a mission context and use a digital ecosystem to integrates authoritative sources of models, data, and test artifacts to improve efficiencies across the IT&E continuum. The digital engineering transformation present both significant opportunities and challenges on how to fully leverage automated testing within traditional testing frameworks. Because of its infancy, it also represents a challenge for defining the new T&E paradigm in a digital environment, which requires buying-in from OSD and Services to be effective and to fulfill the promise of efficiently and effectively developing systems. Since the work for Initiative #13 was limited in scope and level of effort, Section 4.4 provided one strategy related to the use of T&E to inform the M&S VV&A process and one strategy to inform T&E assessments with M&S data. Only the strategy related to using M&S data to inform T&E assessments has been partially discussed with the Navy T&E community. The SysML model for M&S VV&A has not been yet presented to the community. MITRE recommends DTE&A continue supporting the development of these two strategies in FY22.

DOT&E articulated their science and technology strategic plan in January 2021, which established a vision in key focus areas. While Initiative # 13 permeates over all DOT&E's focus areas, it overlaps three specific initiatives: Scalable M&S Capabilities, Guidebook for M&S, and M&S VV&A. MITRE recommends DTE&A work with DOT&E in the development of the M&S Guidebook and continue collaborating with OSD, Agency, Service acquisition stakeholders, other Federally Funded Research Centers (FFRDCs), and University Affiliated Research Centers (UARCs) in developing a roadmap for growing M&S capabilities and resources that enhance current DT&E M&S procedures for acquisition pathways.

6. References

1. Prochnow, D., et.al. (2019) Naval Warfare Modeling and Simulation (M&S) Recommendations for Test and Evaluation (T&E). MITRE Technical Report 190573
2. Director, Developmental Test, Evaluation, and Assessment [D(DTE&A)] *Test Vision Key Takeaways*. 11 February 2021.
3. *Summary of Key Risk Area Trends Identified in DTE&A Program Assessments (2016-2021)* Office of the Secretary of Defense, Developmental Test, Evaluation, and Assessments. 26 March 2021
4. IEEE Standard Glossary of Modeling and Simulation Terminology (IEEE Std 6103-1989). *(1989) Institute of Electrical and Electronics Engineers. New York, NY.*
5. DoD Modeling and Simulation Glossary (DoD 5000.59-M)
6. Defense Modeling and Simulation Enterprise (MSE) Website. https://www.msco.mil Retrieved on 25 August 2021.
7. DoD Modeling and Simulation (M&S) Verification, Validation, and Accreditation (VV&A) (DoDI 5000.61)
8. *Rouche, P. J. (2009).* Fundamentals of Verification and Validation. *Hermosa Publishing.*
9. VV&A Recommended Practices Guide (RPG)
10. NASA-STD-7009 (2008) and NASA-STD-7009A (2016).
11. NASA Handbook for Models and Simulations: An Implementation Guide for NASA-STD-7009A (2019).
12. *Oberkampf, W., Trucano, T., & Pilch, M. (2007).* Predictive Capability Maturity Model for Computational Modeling and Simulation. Sandia National Laboratories. SAND2007-5948.
13. Use of Modeling and Simulation in Operational Test (COMOPTERVFORINST 5000.1C), Commander, Operational Test and Evaluation Force
14. Cortes, L. A. (2021) Assessment of the Modeling and Simulation Validation Strategy for Over-the-Horizon Weapon System Probability of Hit. The MITRE Corporation. MP210028.
15. Wojton, H. M. et. al. (2019) Handbook on Statistical Design & Analysis Techniques for Modeling & Simulation Validation. Institute for Defense Analyses, NS D-10455, Feb. 2019.
16. T. J. Sullivan, Introduction to Uncertainty Quantification. Springer International Publishing, 2015. doi: 10.1007/978-3-319-23395-6.
17. C. J. Roy and W. L. Oberkampf. A comprehensive framework for verification, validation, and uncertainty quantification in scientific computing, Comput. Methods Appl. Mech. Eng., vol. 200, no. 25, pp. 2131–2144, Jun. 2011, doi: 10.1016/j.cma.2011.03.016.
18. *B. Sudret, S. Marelli, and J. Wiart.* Surrogate models for uncertainty quantification: An overview, in 2017 11th European Conference on Antennas and Propagation (EUCAP). *Mar. 2017, pp. 793–797. doi: 10.23919/EuCAP.2017.7928679.*

19. "Standard for Verification and Validation in Computational Solid Mechanics," American Society of Mechanical Engineers, VV10-2019, 2020. [Online]. Available: https://www.asme.org/codes-standards/find-codes-standards/v-v-10-guide-verification-validation-computational-solid-mechanics

20. B. Sudret, "Global sensitivity analysis using polynomial chaos expansions," Reliab. Eng. Syst. Saf., vol. 93, no. 7, pp. 964–979, Jul. 2008, doi: 10.1016/j.ress.2007.04.002.

21. A. Saltelli, M. Ratto, S. Tarantola, and F. Campolongo, "Sensitivity analysis practices: Strategies for model-based inference," Reliab. Eng. Syst. Saf., vol. 91, no. 10, pp. 1109–1125, Oct. 2006, doi: 10.1016/j.ress.2005.11.014.

22. J. C. Helton, "Uncertainty and sensitivity analysis techniques for use in performance assessment for radioactive waste disposal," Reliab. Eng. Syst. Saf., vol. 42, no. 2, pp. 327–367, Jan. 1993, doi: 10.1016/0951-8320(93)90097-I.

23. D. Xiu, Numerical Methods for Stochastic Computations: A Spectral Method Approach. Princeton University Press, 2010. doi: 10.2307/j.ctv7h0skv.

24. D. Xiu, "Stochastic Collocation Methods: A Survey," in *Handbook of Uncertainty Quantification*, R. Ghanem, D. Higdon, and H. Owhadi, Eds. Cham: Springer International Publishing, 2016, pp. 1–18. doi: 10.1007/978-3-319-11259-6_26-1.

25. K. Konakli and B. Sudret, "Global sensitivity analysis using low-rank tensor approximations," *Reliab. Eng. Syst. Saf.*, vol. 156, pp. 64–83, Dec. 2016, doi: 10.1016/j.ress.2016.07.012.

26. G. Roma *et al.*, "A Bayesian framework of inverse uncertainty quantification with principal component analysis and Kriging for the reliability analysis of passive safety systems," *Nucl. Eng. Des.*, vol. 379, p. 111230, Aug. 2021, doi: 10.1016/j.nucengdes.2021.111230.

27. R. Tipireddy, D. A. Barajas-Solano, and A. M. Tartakovsky, "Conditional Karhunen-Loève expansion for uncertainty quantification and active learning in partial differential equation models," *J. Comput. Phys.*, vol. 418, p. 109604, Oct. 2020, doi: 10.1016/j.jcp.2020.109604.

28. T. J. Santner, B. J. Williams, and W. I. Notz, *The Design and Analysis of Computer Experiments*, 2nd ed. New York: Springer-Verlag, 2018. doi: 10.1007/978-1-4939-8847-1.

29. M. Abdar *et al.*, "A review of uncertainty quantification in deep learning: Techniques, applications and challenges," *Inf. Fusion*, May 2021, doi: 10.1016/j.inffus.2021.05.008.

30. R. Trinchero, M. Larbi, H. M. Torun, F. G. Canavero, and M. Swaminathan, "Machine Learning and Uncertainty Quantification for Surrogate Models of Integrated Devices with a Large Number of Parameters," *IEEE Access*, vol. 7, pp. 4056–4066, 2019, doi: 10.1109/ACCESS.2018.2888903.

31. D. R. Jones, M. Schonlau, and W. J. Welch, "Efficient Global Optimization of Expensive Black-Box Functions," *J. Glob. Optim.*, vol. 13, no. 4, pp. 455–492, Dec. 1998, doi: 10.1023/A:1008306431147.

32. J. Li, S. Luo, and J. S. Jin, "Sensor Data Fusion for Accurate Cloud Presence Prediction Using Dempster-Shafer Evidence Theory," *Sensors*, vol. 10, no. 10, pp. 9384–9396, Oct. 2010, doi: 10.3390/s101009384.

33. Beers, S. and Medlin, R. Decision Support, Evaluation, and Test Planning with Integrated Decision Support Key (IDSK) – Evaluation Framework (EF) – Test Design Framework (TDF) (Draft White Paper; 2 September 2021)

34. Torres, G., Sheehan, M., and Pickett, K. (2020) The Need for Simulation Software 'Technology Readiness Levels' (TRLs) in the Development of Simulation Software for Training and Test and Evaluation Systems, In Proceedings of the 37th International Test and Evaluation (ITEA) Symposium 15-18 September 2020. Virtual Symposium.

35. Hills, R., Witkowski, W., Urbina, A., Rider, W., & Trucano, T. (2013). Development of a *F*ourth Generation Predictive Capability Maturity Model. Sandia National Laboratories. SAND2013-8051.

36. Ahner, D., et. al. (2021) *A Conceptual Framework for the Establishment of Model Readiness Levels.* STAT COE Report 11-2021; 6 October 2021.

37. Ahner, D., (2021), Test & Evaluation in a Digital Engineering Enabled World, August 4, 2021, presentation at Model Readiness Workshop.

38. INCOSE Model-Based Capabilities Matrix and User's Guide, v1.0 (2020).

39. NATO Generic Methodology for Verification and Validation to Support Acceptance of Models, Simulations and Data (2015).

40. Simulation Interoperability Readiness Levels (*SISO-REF-076-2020 v1.0) (2020). Simulation Interoperability Standards Organization.* St. Petersburg, FL.

41. *J. S. Fant, R. G. Pettit and D. Gayek, A Quantitative Approach for Calculating Model Assurance Levels, 2019* IEEE 22nd International Symposium on Real-Time Distributed Computing (ISORC), *2019, pp. 69-76, doi: 10.1109/ISORC.2019.00020.*

42. *Youngblood S., Stutzman M., Pace D., and Pandolfini P. (2011).* Risk Based Methodology for Verification, Validation and Accreditation M&S Use Risk Methodology, *NSAD-R-2011-011, Technical Report, The Johns Hopkins University Applied Physics Laboratory.*

43. *Google Model Cards, Retrieved from https://modelcards.withgoogle.com/about on 24 August 2021.*

44. *Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT'19). ACM, New York, NY, 220–229. DOI: 10.1145/3287560.3287596*

45. Forsberg, K., and Mooz, H. The Relationship of Systems Engineering to the Project Cycle. Joint Conference of National Council on Systems Engineering (NCOSE) and American Society for Engineering Management (ASEM), 21-23 October 1991, Chattanooga, TN.

46. Kujawski, E. (2010) The trouble with the System Readiness Level (SRL) index for managing the acquisition of defense systems. Presentation for the 13th Annual Systems Engineering Conference, National Defense Industrial Association. Sand Diego, CA. October 2010.

47. Blanchard, B. S., and Fabrycky, W. J. (1998). *Systems Engineering and Analysis*. (3rd ed.) Prentice Hall, Inc., Upper Saddle River, NJ.

48. Cortes, L.A., White, K. Transitioning to a Model-Based Test and Evaluation Ecosystem.

49. IEEE Standard for Application of Systems Engineering on Defense Programs (IEEE 15288.1). (2014). *Institute of Electrical and Electronics Engineers. New York, NY.*

50. IEEE Standard for Technical Reviews and Audits on Defense Program (IEEE 15288.2) (2015). *Institute of Electrical and Electronics Engineers. New York, NY.*

51. Best Practices for the Development of Models and Simulations. The Johns Hopkins University Applied Physics Laboratory. NSAD-R-2010-037. 2010

52. AcqNotes Simulation Modeling Steps. Retrieved on 20 July 2021 from https://acqnotes.com/acqnote/tasks/simulation-modeling-steps

53. Maria, A. (1997). Introduction to Modeling and Simulation. In *Proceedings of the 1997 Winter Simulation Conference* (Andradóttir, S., Healy, K. J., Withers, D. H., and Nelson, B. I. Editors).

54. Hatakeyama, S. J., et. al. *An Alternate View of the Systems Engineering "V" in a Model-Based Engineering Environment*. 2018 AIAA SPACE and Astronautics Forum and Exposition. 17-19 September 2018. DOI: 10.2514/6.2018-532.

55. *Department of Defense Digital Engineering Strategy*. (2018). Office of the Deputy Assistance Secretary of Defense for Systems Engineering. Washington, DC.

56. Surface Ship Weapon/Combat Systems Test and Evaluation Guide - Technical Manual 20-07; A Guide for Test Design and Analysis. Naval Surface Warfare Center Port Hueneme Division. 15 January 2021.

57. https://www.msco.mil/DoDTools/DoDEnterpriseManagementTools/MSCatalog.aspx

58. Clay, R. L., S. J. Marburger, M. S. Shneider and T. G. Trucano (2007), "Modeling and Simulation Technology Readiness Levels." Sandia National Laboratories, SAND2007-0570, Albuquerque, New Mexico.

59. Smith, J. (2004), "An Alternative to Technology Readiness Levels for Non-Developmental Item (NDI) Software." Carnegie Mellon, Software Engineering Institute, CMU/SEI-2004-TR-013, ESC-TR2004-013, Pittsburgh, PA.

60. Chesnava C., et al., Towards and Argument Interchange Format. The Knowledge Engineering Review, 2007, Cambridge University Press.

61. Willmott S., et al., AIF: Argumentation Interchange Format Strawman Model Version 0.8, 16 December 2005.

62. Rahwan I., et al., On Building Argumentation Schemes Using the Argument Interchange Format

This page intentionally left blank

Appendix A Model Maturity Assessment Methods

Table A-1. Summary of Selected Model Assessment Methods

Purpose & Scope	Factors	Scoring	Scale
NASA-STD-7009A Credibility Assessment Scale (CAS)			
Intended to ensure that NASA decision makers are informed about the credibility of M&S results in terms of a common process and a common language. No limitations on the type of M&S or application.	1. M&S Development 1.1. Data Pedigree 1.2. Verification 1.3. Validation 2. M&S Operations 2.1. Input Pedigree 2.2. Results Uncertainty 2.3. Results Robustness 3. Supporting Evidence 3.1. Use History 3.2. M&S Process/Product Management	Five ordinal levels (0-4) are defined for each factor. The definitions for many of the levels consist of multiple conditions; all conditions must be satisfied to achieve that level.	No thresholds are defined in the standard, but instead they are determined by individual program needs/requirements. The assessed level for each of the 8 factors are reported (i.e., there is no aggregation). See **Error! Reference source not found.** for an e xample of how to report CAS results.
Predictive Capability Maturity Model (PCCM)			
M&S efforts that rely heavily on large-scale computer codes to solve complex, nonlinear partial differential equations (PDEs) or integro-differential equations	1. Representation and Geometric Fidelity 2. Physics and Material Model Fidelity 3. Code Verification 4. Solution Verification 5. Model Validation 6. Uncertainty Quantification and Sensitivity Analysis	Four levels (0-3) are defined for each factor and sub-factors. Each level may contain multiple descriptors/conditions. A partial score (e.g., 1.5) may be assigned when some, but not all, of the conditions for a level are achieved.	No thresholds are defined in the standard, but instead they are determined by individual program needs/requirements. The assessed level for each of the 6 factors are reported (i.e., there is no aggregation). The results are reported in a similar fashion as the CAS.
Simulation Software Technology Readiness Levels (SSRL)			
M&S environments used for training and T&E (i.e., testbeds)	1. Fidelity of testbed to system under test (SUT) 2. Fidelity of testbed to represent operational environment around SUT 3. Fidelity of testbed to represent real-time data rate to SUT 4. Testbed architecture 5. Controllability and robustness of testbed	No explicit scoring for factors. Instead, a qualitative definition for each level is given.	Ordinal levels 2-9 reflecting increasing maturity.

Table A-1 Summary of Selected Model Assessment Methods (Cont)

Purpose & Scope	Factors	Scoring	Scale
Model Assurance Level (MAL)			
Used to concisely express the assurance the model is providing to the program and the risks associated with the model. Initially developed for software models written in UML. Currently being extended to systems engineering models.	1. General Model-based Engineering 2. Structural Content 3. Behavioral Content 4. Data Content 5. Non-Functional 6. Security and Vulnerability 7. V&V Performed 8. Model-based Testing 9. Implementation Synchronization Each of the above MAL Attributes is broken down into MAL Characteristics (35 total). Each of the 35 characteristics is further decomposed into MAL Detailed Criteria. The full list of Detailed Criteria is not provided in the paper.	1. Each of the Detailed Criteria are assigned a score. Details on how to score are not provided in the paper. 2. Each Detailed Criteria is also assigned a weight based on engineering judgment and experience. 3. The MAL Characteristic Weighted Score is calculated: $C_{weighted} = \sum_{1}^{n} DC_{weighted}$ 4. The maximum possible score for each MAL characteristic is calculated: $C_{max} = \sum_{1}^{n} DC_{max}$ 5. The MAL characteristic percentage is calculated: $C_{percent} = \frac{C_{weighted}}{C_{max}} * 100$ 6. The overall MAL quality attribute percentage is calculated: $QA_{percent} = \frac{\sum C_{weighted}}{\sum C_{max}} * 100$ 7. The MAL level is determined by comparing the $QA_{percent}$ scores for each attribute against the MAL baseline scores for each level. Baseline scores are based on MBE methods and judgment. The full list of baseline scores is not provided in the paper.	Three MAL levels, each with sublevels: 1. Sparse Models 1.1. Sparse Analysis Model 1.2. Sparsely V&V Space Analysis Model 1.3. Sparse Detailed Design Model 1.4. Sparsely V&V Detailed Design Model 2. Basic Models 2.1. Basic Analysis Model 2.2. Sparsely V&V Basic Analysis Model 2.3. Completely V&V Basic Analysis Model 2.4. Basic Detailed Design Model 2.5. Sparsely V&V Basic Detailed Design Model 2.6. Completely V&V Basic Detailed Design Model 3. Advanced Models 3.1. Advanced Analysis Model 3.2. Sparsely V&V Advanced Analysis Model 3.3. Completely V&V Basic Analysis Model 3.4. Advanced Detailed Design Model 3.5. Sparsely V&V Advanced Detailed Design Model 3.6. Completely V&V Advanced Detailed Design Model 3.7. Advanced Detailed Design Model used to drive implementation testing 3.8. Advanced Detailed Design Model incorporated into operational system

Table A-1. Summary of Selected Model Assessment Methods (Cont)

Purpose & Scope	Factors	Scoring	Scale
Model-Based Capability Matrix (MBCM)			
A tool for organizational transformation and development to help organizations that have decided to implement digital engineering. The scope may be entire enterprise, program/product line, project/product, or some other organizational level. Models may be descriptive or analytical.	Forty-two model-based capabilities which can be mapped to either roles or OSD DE Strategy Goals [2].	Five stages (0-4) are defined for each factor.	No standard for weighting or aggregating capabilities is specified. Tailoring of the matrix is encouraged, and organizations are free to implement weighting /aggregation methods if they so choose. As a result, comparing MBCM assessments between organizations is not meaningful.
Model Readiness Levels (MRL)			
Purpose is to 1) provide developers with clearer standards by which to develop their models and 2) provide decision makers with construct to understand risk when using M&S information to make decisions. Scope is operational analysis models.	1. Fidelity 2. Referent Authority 3. Scope a. Input b. Output	$MRL = f(Fidelity, Referent\ Authority$ Note: Formula is still under development.	Scale is still under development.
NATO General Methodology for Verification and Validation (GM-VV)			
Purpose is to provide a generic framework to justify why identified models, simulations, data, outcomes, and capabilities are acceptable for deployment in the intended operational context of use. Scope is NATO M&S systems design, development, and employment processes; applicable to any M&S scope, technology, and application domain.	1. Utility 2. Validity 3. Correctness 4. Verification & Validation Quality	No scoring method is defined. Acceptability criteria are defined for each factor based on the intended use for a given M&S.	No scale is defined. This method describes a tailorable framework based on systems engineering practices. An acceptance recommendation is based on the unique acceptability criteria for a given M&S use. Evidence supporting the acceptance recommendation is provided in an Argumentation Structure. The Argumentation Structure ontology is defined by the Argumentation Interchange Format [60,61,62] which can be instantiated in many different formats. The provided examples use traceability matrices, Goal Structuring Notation and Claim Argument Evidence.

Table A-1. Summary of Selected Model Assessment Methods (Cont)

Purpose & Scope	Factors	Scoring	Scale
M&S Use Risk Methodology (MURM)			
Identify the risk involved in using M&S information and optimize VV&A resource use while minimizing the risk of using M&S. No limitations on the type of M&S or application.	Causes of inappropriate M&S application: 1. Lack of clarity in intended use (Clarity) 2. Adverse impact on decision if capability not achieved (Importance) 3. Incorrect recommendation to use/not use M&S Results relative to that capability (Confidence) Examination Technique: V&V Evidence Confidence:	States for Causes & Effects are subjective and tailorable, although recommendations are given. Weights and probabilities for each state are based on the maximum information entropy principle. Probabilities are then used in a formula to calculate the Use Risk score for the M&S capability: $UR = P[(Causes \wedge Effects) \wedge (Causes \Rightarrow Effects)]$ Where: $\wedge - logical\ conjunction\ (and)$ $\Rightarrow\ - implication\ (if\ then)$	No scale or thresholds are defined or recommended. Instead, a scorecard is reported for each M&S capability assessed (see example in **Error! Reference source not found.**). The scorecard is intended to be updated regularly as the program progresses and the M&S capabilities mature.

Table A-1. Summary of Selected Model Assessment Methods (Cont)

Purpose & Scope	Factors	Scoring	Scale
SISO Simulation Interoperability Readiness Levels (SIRL)			
Provides a framework for decision makers and developers to make evidence-based assessments of the feasibility and risk of attempting to integrate simulations. Initially, the scope was intended for training applications. SIRL team members believe the framework is broadly applicable to all simulation domains.	Defines five interoperability levels: 1. Conceptual 2. Modeling 3. Simulation Control 4. Data 5. Technical/Syntactic	Specific types of engineering evidence are defined for levels 2-5 (Conceptual requires further research). Defines utility curves for each type of engineering evidence. Weights for each item of engineering evidence are developed by the project/program based on the intended use of the simulation. Weights within each level must sum to 1. Weighted scores are then summed for each level.	Scores for each of the five levels are reported. There is no aggregation across the levels. Scores may be reported in a spider/radar chart (like the CAS) to visualize alignment between sims. SIRL assessment scores are compared by level between simulations. Unlike other methods, SIRL does not assess the readiness of a single simulation, but instead assesses the feasibility and risk of a simulation integrating with other simulations before committing resources to do so. Note: high scores for a given level do not denote the likelihood of simulation interoperability, but that there is enough evidence to make that decision.
Google Model Cards			
Provides a framework for clarifying intended use cases for machine learning models and to minimize their usage in contexts for which they are not well-suited. Scope is human-centered machine learning models for computer vision and natural language processing, but the method can be adapted for any trained machine learning model.	Factors are tailorable by project, but some examples include: 1. Cultural groups 2. Demographic groups 3. Phenotypic groups 4. Intersectional groups	There is no scoring with this method. Model cards consist of descriptive explanations for each of the following sections: 1. Model Details 2. Intended Use 3. Factors 4. Metrics 5. Evaluation Data 6. Training Data 7. Quantitative Analysis 8. Ethical Considerations 9. Caveats and Recommendations	There is no scale with this method. There is no standardization, and the usefulness of this method depends on the integrity of the model card creator. Therefore, it should be used in conjunctions with other methods and tools to understand the risks and limitations with using a machine learning model.

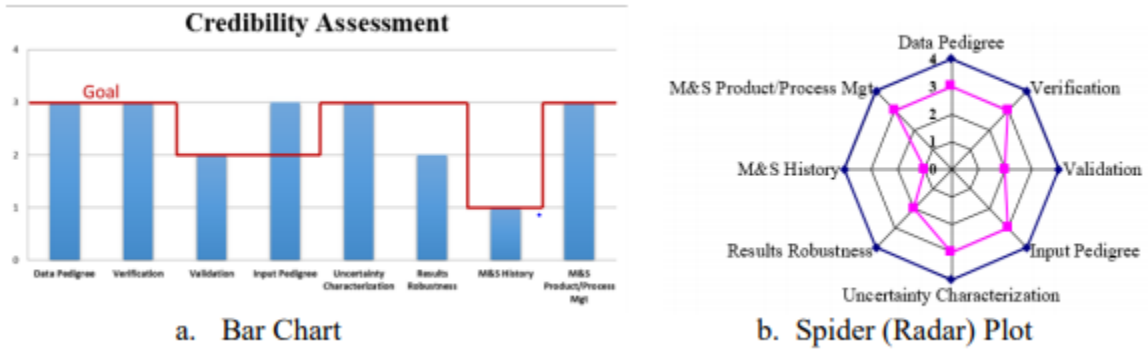


Figure Error! No text of specified style in document.**-1. NASA-HDBK-7009A Credibility Assessment Reporting Examples**

Table Error! No text of specified style in document.**-1. Example MURM Scorecard**

Capability	Clarity	Importance	Activities & Examination Technique	Recommen-dation	V&V Evidence Confidence	M&S Use Risk
CAP_1	Lucid; A	Medium; D	SET L: 3,3,0,1	TBD	Very Low; E	P(Causes)=0.736 P(Effects)=TBD UR=TBD
CAP_{2A}	Partial; B	High; F	SET L: 3,3,0,3	TBD	Medium; C	P(Causes)=0.942 P(Effects)=TBD UR=TBD
CAP_{2B}	Lucid; A	Medium; D	SET F: 3,3,5,0	TBD	Low; D	P(Causes)=0.688 P(Effects)=TBD UR=TBD
CAP_3	Partial; B	Low; C	SET L: 3,3,0,5	TBD	High; B	P(Causes)=0.673 P(Effects)=TBD UR=TBD
…						
CAP_{N-1}						
CAP_N						

This page intentionally left blank

Appendix B Considerations for M&S in T&E

Table Error! No text of specified style in document.**-1. Considerations for M&S use in T&E**

	M&S Maturity Considerations for T&E
M&S Needs	• Is the intended model use clearly defined (i.e., aspects of the real world—threat, environment, combat system, etc.—the M&S represents)? • What knowledge points, decision(s), or program objectives will be informed or impacted by the M&S output(s) and how are they linked to linked to the program's acquisition strategy? • Can the information provided by M&S be acquired through other means like live testing? Why or why not? • Is there intent to use M&S output generated with conditions for which there is no live test data available? If so, why? • Is there an M&S available for the intended use? If so, has it been accredited, by whom, when, for what environment, and what are its limitations? (Reference the V&V Plan, V&V Report, Accreditation Plan, Accreditation Report, TEMPs, and SEPs for programs that have applied it.) • Has the existing M&S been updated based on feedback and empirical data? If so, provide details of the updates and the evolution. • Which specific M&S outputs will be used to support program objectives and to quantify key measures of capability? • What is the role of the model in supporting the mission or program objectives? (i.e., is it intended to support a particular lifecycle activity?) • Who will be using the M&S (i.e., system developers, systems engineers, testers, trainers, etc.) and its outputs? • Is the M&S a component of a larger integrated or federated environment? • What does "success" look like in terms of quality of M&S outputs and impact on decision-making?
Conceptual Model Development and Validation	• Is the conceptual model accurately described, including all its elements? • Does the model information exist in an appropriate and accessible form? • Is the conceptual model validation methodology adequately described? • Is the validation approach consistent with the M&S intended use? • How was the conceptual model validated (aim at answering "Does the conceptual model represent the real world?)? • Are the operational conditions that bound the validity of the M&S for its intended use listed? • Is the level of fidelity (acceptability criteria) defined for each element? • What underlying assumptions about acceptability criteria were made? • Which sources establish the acceptability criteria? • Does the V&V Plan and the V&V Report describe the analysis performed to validate the conceptual model, including statistical analysis? • Does the V&V Report describe how the results support the conclusion that the conceptual model accurately represents the real world in terms of the model's intended use? • Are statistical methods used to describe system performance (i.e., fitting performance data to theoretical distributions)?

	M&S Maturity Considerations for T&E
Requirements Definition	• What requirements must the M&S meet, and under what conditions? • What does M&S "goodness" look like in terms of inputs, outputs, and execution, in meeting those requirements and the acceptability criteria? • Does the high level design identify the data the simulation must accept? • Is there a clear correlation between operational measures, technical requirements, objectives, and decisions to be informed with the M&S? • Are the acceptability criteria clearly stated, in quantitative terms? • Who determined and approved the acceptability criteria? How? Has it been socialized with, and accepted by, all stakeholders? • Are all input, output, constants, and operational data clearly defined, including units of measure and the range of values for each data item? • Are the operational measures and conditions specified in, and consistent with, the requirements? • Are the M&S integration environment requirements (e.g., Lab, Facility, SIL, connectivity) adequately addressed? • Why was this type of model selected? • Describe the computing environment (SysML, C, Java, etc.). • Describe the source, content, and format of the M&S input and output data. • Are there data assumptions that should be verified or periodically revisited? • Is the focus of the M&S VV&A effort clearly articulated?
Design and Development	• What is the plan to complete the development of the M&S? • If this application is an existing M&S, when was the model last updated, how, and why? • Have Modular and Opens System Approach (MOSA) design principles been applied to the M&S (i.e., modular design, use of interface standards, etc.)? • What are the limitations and implications regarding inputs, outputs, execution, and context (i.e., what can it do, and what can it not do)? • For federated models, are capabilities and limitations that influence the overall kill chain or the interoperability of the federation of models considered? • For a federated M&S, will all the components be ready when needed? • Are there risks associated with M&S development (i.e., resourcing, administration, coordination, scheduling, execution, data, etc.)? Are there contingent plans to mitigate those risks? • Are there any constraints that may result in inadequate information, inadequate technical knowledge, unavailable data, inadequate methodologies, or inadequate test environments to support the M&S VV&A assessments?
Implement ation	• Is the implementation acceptable as defined by pre-established acceptability criteria? • Does the user manual describe how to operate the simulation model, how to set up input data values, and how to analyze model results?

	M&S Maturity Considerations for T&E
Implementation Verification	• Is there coordination between the Requirements, M&S, T&E, and Statistical Engineering IPTs to ensure adequate data from both M&S and live test is collected and properly analyzed? • Have the Requirements, M&S, T&E, and Statistical Engineering IPTs considered data needs in their planning? • How are errors or out-of-range input and output values handled? • How is the collection of M&S output data discussed? • Does the *Implementation Verification Plan* include methods to check for errors in the processed or reduced data? • Does the plan describe the means to demonstrate the M&S software is free of design errors? • Is the M&S design a correct implementation of the conceptual model? • Is the M&S software free of implementation errors? • Does the M&S software comply with standards? • What results do the implementation verification tests show? • Is the implementation acceptable as defined by pre-established acceptability criteria? • What authoritative resources are used in the verification process (i.e., SMEs; mathematical or statistical techniques; etc.)?
Design Verification	• Is the M&S "built right" (i.e., does the M&S implementation accurately reflect the design and conceptual model? • Are the verification methods, including software verification, clearly described in the plan? • Are there plans to check for the quality of inputs (whether from another federated element or manual)? • Were adequate design reviews conducted prior to testing? If so, which and by whom? Who participated in the reviews? • Are the verification test cases and analyses plan adequately described (test case identification, schedule, resources required for execution, test procedures, entrance criteria, go/no go criteria, exit criteria, justification for sample size, test architecture, objectives, traceability to requirements, acceptability criteria, prerequisites, inputs, statistical analysis methods, key participants, assumptions, and constraints)? • Besides data, which documents will be gathered and how will it be analyzed? • Is there traceability between the test cases, requirements, and the decisions to be made? • Is there sufficient description about the data to be collected (and how much) and how the data will be used for verification? • Do the tests procedures (or checklists) document how the tests will be executed and the critical data necessary to evaluate the execution of the test? • Are the M&S, T&E, and Statistical Engineering IPTs collaborating in the generation, review, and acceptability of test plans, data collection plans, etc.? • Are the results of any benchmark tests, acceptance tests, model-model comparisons, etc. included or referenced in the plans? • For the final report, are deviations from procedures, sources of data, sample size, data analysis techniques, unresolved anomalies or discrepancies encountered during the execution of the test, causes of discrepancies, and corrective procedures adequately described? • Who confirms the output data as valid and acceptable? • Describe the consequences of any noted deviations from the original plans, factors and conditions, relevant measures, statistical figures of merit (such as collinearity, power, and confidence), and acceptability criteria. • For federated models, outline how the results influence the entire kill chain. • Who reviews the reports? Are all stakeholders included in the distribution lists?

	M&S Maturity Considerations for T&E
Validation	• Is the M&S the "right thing" (i.e., does the M&S accurately represent the real world? What gives confidence in its validity? • Which validation methods are used? Do they include model-to-live-data comparisons? Model-to-model comparisons? • Are statistical techniques used for the comparison? If so, which one(s)? Are they adequately described? Does the design span the operational space (i.e., factor levels) to the extent of what can be executed in live testing? If not, why? Does the statistical design provide figures of merit? • What are the sources of the live data for validation? • Are the input variables and operational conditions (i.e., factors) that will exercise the M&S described? How were the ranges of those conditions (i.e., factor levels) set? • Are the validation test cases and analyses plan adequately described (test case identification, schedule, resources required for execution, test procedures, entrance criteria, go/no go criteria, exit criteria, justification for sample size, test architecture, objectives, traceability to requirements, acceptability criteria, prerequisites, inputs, statistical analysis methods, key participants, assumptions, and constraints)? • Is there traceability between the test cases, output variables, requirements, acceptability criteria, uncertainty quantification, and the decisions to be made? • What methods were used to establish compliance with acceptability criteria? • Are the validation results aligned with known, trusted references for the operational context of interest? • Is there sufficient description about the data to be collected (and how much) and how the data will be used for validation? • Are uncertainty quantification methods described? • Do the tests procedures (or checklists) document how the tests will be executed and the critical data necessary to evaluate the execution of the test? • Are the M&S, T&E, and Statistical Engineering IPTs collaborating in the generation, review, and acceptability of test plans, data collection plans, etc.? • For the final report, are deviations from procedures, sources of data, sample size, data analysis techniques, unresolved anomalies or discrepancies encountered during the execution of the test, causes of discrepancies, and corrective procedures adequately described? • Describe the consequences of any noted deviations from the original plans, factors and conditions, relevant measures, statistical figures of merit, and acceptability criteria. • For situations involving federated models, outline how the results influence the entire kill chain. • Who reviews the reports? Are all stakeholders included in the distribution lists?

	M&S Maturity Considerations for T&E
Accreditation Assessment and Use	• Is the process for accepting M&S for a specific use completely laid out? • What is the outcome of the accreditation assessment? • What is the operational envelope for which the M&S is accredited? • What are the risks associated with the accreditation recommendations? • What forms the basis for the accreditation assessment (i.e., events, techniques, comparisons, participants involved, milestones achieved, statistical analysis, information sources (how, when, form and from whom), etc.)? • How was objectivity preserved in performing the assessment? • How does the data and information maps to the needs of the accreditation process? • Was any information or data needed for accreditation not used or unobtainable? Why? What is the impact on the accreditation assessment? • Is there additional information used that was not originally planned for? What is the impact on the accreditation assessment? • What is the intended analytic approach for post-test analysis of data? • Which measures will be quantified? Which input variables (factors) are likely to influence it, and at which levels? • Is the experiment design to gather data adequate (sample size, confidence, statistical signal-to-noise ratio, and statistical power)? • Is the analysis methodology adequate? • Are the conclusions drawn from the V&V processes adequate to resolve issues relevant to the accreditation effort and to provide recommendations relevant to M&S use? • What adjustments needed to be made for future events? • What lessons were learned during the VV&A process

This page intentionally left blank

Appendix C Acronyms

Table Error! No text of specified style in document.**-1. Acronyms**

Term	Description
IDSK-EF	Integrated Decision Support Key – Evaluation Framework
D(DTE&A)	Director, Developmental Test, Evaluation, and Assessment)
AAF	Adaptive Acquisition Framework
CAS	Credibility Assessment Scale
CDT&E	Contractor Development Test & Evaluation
COI	Critical Operational Issues
COMOPTEVFOR	Commander, Operational Test and Evaluation Force
CTP	Critical Technical Parameters
DASD(SE)	Deputy Assistant of Defense for Systems Engineering
DEF	Developmental Evaluation Framework
DoD	Department of Defense
DODAF	DoD Architectural Framework
DoDI	DoD Instruction
DOE	Design of Experiments
DSQ	Decision Support Questions
DT&E	developmental test and evaluation
DTE&A	Developmental Test, Evaluation, and Assessment
DTM	Directive-Type Memorandum
EF	Evaluation Framework
EPS	Engineering, Policy, and Systems
FFRDC	Federally Funded Research and Development Center
FY	Fiscal Year
GMC	Google Model Cards
GM-VV	General Methodology for Verification and Validation
HDBK	Handbook
IDA	Institute for Defense Analysis
IDA	Institute for Defense Analyses
IEEE	Institute of Electrical and Electronics Engineers
IEF	Integrated Evaluation Framework
INCOSE	International Council on Systems Engineering
IOT&E	Initial Operational Test and Evaluation

Term	Description
IT&E	Integrated Test and Evaluation
JHU/APL	Johns Hopkins University Applied Physics Laboratory
KPP	Key Performance Parameters
LFT&E	Live Fire Test and Evaluation
M&S	Models and Simulations
MAL	Model Assurance Levels
MBCM	Model-Based Capabilities Matrix
MBSE	Model-Based Systems Engineering
MBTD	Mission-Based Test Design
MCA	Major Capability Acquisition
MCT	Model Card Toolkit
MIL-STD	Military Standard
MORV	Mission-Oriented Response Variables
MRL	Model Readiness Level
MSCO	Modeling and Simulation Office
MSE	Modeling and Simulation Enterprise
MTM	Model-Test-Model
MURM	M&S Use Risk Methodology
NAFEMS	International Association for the Engineering Modeling, Analysis and Simulation Community
NASA	National Aeronautics and Space Administration
NSM	Naval Strike Missile
NSWC PHD	Naval Surface Warfare Center, Port Hueneme Division
OQE	Objective Quality Evidence
OT&E	Operational Test and Evaluation
OTA	Operational Test Agency
OTH WS	Over-the-Horizon Weapon System
PCMM	Predictive Capability Maturity Model for Computational Modeling and Simulation
PEO IWS	Program Executive Office Integrated Warfare Systems
RFP	Request for Proposal
RPG	Recommended Practices Guide
SIRL	Simulation Interoperability Readiness Levels
SISO	Simulation Interoperability Standards Organization
SoS	Systems-of-Systems
SSTRL	Simulation Software Technology Readiness Level

Term	Description
STAT COE	Scientific Test and Analysis Techniques Center of Excellence
SysML	Systems Modeling Language
T&E	Test and Evaluation
TDF	Test Design Framework
TEMP	Test and Evaluation Master Plans
TRL	technology readiness levels
TRMC	Test Resource Management Center
TTP	Tactics, Techniques, and Procedures
UARC	University Affiliated Research Center
UEF	Unified Evaluation Framework
UQ	Uncertainty Quantification
V&V	Verification and Validation
VV&A	Validation, Verification, and Accreditation
VVUQ	Verification, Validation, and Uncertainty Quantification