

# Modeling and Simulation for Test and Evaluation Guidebook



May 2025

Office of the Director,  
Developmental Test, Evaluation, and Assessments  
Office of the Under Secretary of Defense  
for Research and Engineering

Office of the Director, Operational Test and Evaluation

Washington, D.C.

**CLEARED**  
**For Open Publication**

May 14, 2025

Department of Defense  
OFFICE OF PREPUBLICATION AND SECURITY REVIEW

Distribution Statement A. Approved for public release. Distribution is unlimited.

## **Modeling and Simulation for Test and Evaluation Guidebook**

Office of the Director, Developmental Test, Evaluation, and Assessments  
Office of the Under Secretary of Defense for Research and Engineering  
3030 Defense Pentagon  
Washington, DC 20301-3030  
[osd.r-e.comm@mail.mil](mailto:osd.r-e.comm@mail.mil)  
<https://www.cto.mil/dtea>

Office of the Director, Operational Test and Evaluation  
1700 Defense Pentagon  
Washington, DC 20301-1700  
<https://www.dote.osd.mil/>

Distribution Statement A. Approved for public release. Distribution is unlimited.  
DOPSR Case # 25-T-1804.

## Executive Summary

The Office of the Director, Developmental Test, Evaluation, and Assessments in collaboration with the Office of the Director, Operational Test and Evaluation developed this Modeling and Simulation (M&S) for Test and Evaluation (T&E) Guidebook to provide information on the use; development; and verification, validation, and accreditation (VV&A) of M&S activities that support the developmental test (DT) objectives of Department of Defense systems. M&S VV&A during DT is a critical element of continuous validation of systems from conception through deployment and operation. The guidebook presents definitions, processes, and techniques to encourage higher-quality implementation and utilization of M&S in developing defense capabilities.

This guidebook is intended to serve as a desk reference for test engineers and analysts, as well as project managers who provide oversight of engineering and analysis activities. It is not intended to prescribe any methods or techniques as standard or preferred; rather, this guidebook will describe recommended best practices for use, development, and VV&A of M&S activities; the situations to which they are best suited; and how to collect and utilize their outputs. In addition to laying out the technical details of these methods, the guidebook summarizes the key points, including benefits, of using each method. These key points give project managers and other decision makers a fundamental understanding of the techniques being used and what they should expect from them.

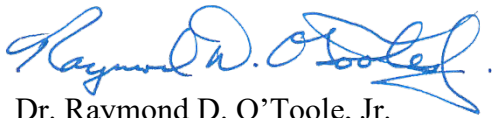
The guidebook emphasizes scientific test and analysis techniques, including design of experiments (DOE), for VV&A of M&S and analysis of its output. Topics covered include the design of physical experiments and deterministic computer experiments and how to combine data from physical experiments with M&S. The chosen DOE method dictates the information that can be gathered. Careful consideration must be applied when choosing a test approach as different DOE methods serve different purposes. This guidebook describes some of these designs and when they are appropriate.

Finally, the guidebook looks at the analysis of M&S results, with a focus on VV&A. Decisions based on M&S depend greatly on what information was gathered and how it was analyzed. It is imperative that experimenters, testers, systems engineers, and project managers mutually understand what constitutes quality test design and analysis so that informed decisions can be made to support good systems engineering. The guidebook includes the following topics:

- Exploratory data analysis.
- Hypothesis testing.
- Statistical uncertainty.

- Unbalanced data.
- Sequential analysis.
- Metamodel techniques.
- Predicted probability validation.
- Time series analysis and functional data analysis.
- Methods for uncertainty quantification.
- Model validation levels.

07 MAY 2025



Dr. Raymond D. O'Toole, Jr.  
Director, Operational Test and Evaluation  
(Acting)

COLLINS.CHRISTO  
PHER.CLAY.10509  
67561

Digitally signed by  
COLLINS.CHRISTOPHER.CLAY.  
1050967561  
Date: 2025.04.25 12:45:50 -04'00'

Mr. Christopher C. Collins  
Director, Developmental Test, Evaluation,  
and Assessments

## Contents

1	Introduction.....	1
1.1	Purpose of the Guidebook .....	1
1.2	Scope of the Guidebook .....	2
1.3	Role of M&S in T&E .....	2
2	M&S Processes.....	4
2.1	VV&A Processes.....	4
2.1.1	Accreditation for Specific Use.....	7
2.2	The Role of Statistical Analyses in M&S VV&A.....	8
2.2.1	Uncertainty Quantification .....	10
2.2.2	Evaluating Model Adequacy .....	12
2.2.3	Combining Statistical and Subject Matter Expertise .....	14
2.3	Participation in M&S Development and Execution .....	15
2.3.1	Modeling and Simulation Working Group.....	15
2.3.2	Connection to Capability Evaluation and Decision Support .....	16
3	Design of Experiments to Support VV&A.....	17
3.1	Design of Physical Experiments.....	17
3.1.1	Plan Phase.....	19
3.1.2	Design Phase.....	19
3.1.3	Test Phase.....	22
3.1.4	Analyze Phase.....	22
3.1.5	DOE Process Example.....	22
3.1.6	Useful Software and Resources.....	25
3.2	Design of Computer Experiments .....	25
3.2.1	Purpose of Space-Filling Designs in VV&A.....	27
3.2.2	Space-Filling Designs.....	27
3.2.3	Determining Sample Size .....	31
3.2.4	Useful Software and Resources.....	31
3.3	Methods for Comparing Data from Physical Experiments and M&S.....	32
4	Analysis .....	34
4.1	Exploratory Data Analysis .....	34
4.1.1	Run-Sequence Plot .....	36
4.1.2	Lag Plot.....	37
4.1.3	Histogram .....	38
4.1.4	Normal Probability Plot.....	38

4.2 Statistical Methods for Comparing Physical and M&S Data .....	40
4.2.1 Basics of Hypothesis Testing.....	40
4.2.2 Two-Sample T-Test.....	41
4.2.3 Analysis of Variance .....	42
4.2.4 Goodness-of-Fit Tests.....	44
4.2.5 Fisher’s Combined Probability Test .....	48
4.2.6 Summary of Statistical Analysis Methods for Comparing Physical and M&S Data.....	48
4.3 Methods for Uncertainty Quantification.....	50
4.3.1 Useful Software and Resources.....	55
4.3.2 Example .....	55
4.4 Advanced Methods and Additional Techniques.....	61
4.4.1 Unbalanced Data.....	61
4.4.2 Sequential Analysis .....	62
4.4.3 Predicted Probability Validation.....	65
4.4.4 Time Series Analysis and Functional Data Analysis.....	73
4.4.5 Metamodel Techniques for Analysis of M&S Data .....	78
4.4.6 Model Validation Levels .....	81
5 Conclusion .....	84
Appendix A: Catalog of Popular Experiment Designs .....	85
Appendix B: Policy and Guidance for M&S in T&E .....	89
Appendix C: Training .....	94
Acronyms.....	95
References.....	98

## Figures

Figure 2-1. V&V Activities and Outcomes .....	12
Figure 3-1. Input-Process-Output Diagram of a System or Process Under Testing .....	17
Figure 3-2. DOE Process for Building Test Designs .....	18
Figure 3-3. Test Design Trade-Off Diagram .....	20
Figure 3-4. Survivability Testing Example of an Infrared Countermeasure System .....	23
Figure 3-5. Power Calculations for Model Terms (Main Effects and All First-Order Interactions) Versus the Number of Runs for the Test Design .....	24
Figure 3-6. Notional Example Analysis from the Results of Using a Designed Experiment .....	25
Figure 3-7. Projection of Space-Filling Designs.....	26
Figure 3-8. Maximin Design.....	27

Figure 3-9. Latin Hypercube Designs: (a) Maximin Optimized and (b) Highly Correlated Design .....	28
Figure 3-10. Sliced Latin Hypercube Design.....	29
Figure 3-11. Fast Flexible Filling Design Generated from Clustering .....	29
Figure 4-1. Anscombe’s Quartet: A Set of Four Data Sets with Identical Means, Variance, R-Squared, Correlations, and Linear Regression Lines .....	35
Figure 4-2. Run-Sequence Plot Showing an Overall Trend in Observations Along With Two Outliers ...	36
Figure 4-3. Lag Plot of Notional Data Showing a Linear Relationship, Indicating Autocorrelation.....	37
Figure 4-4: Histogram Showing Distribution of Notional Data.....	38
Figure 4-5. Normal Probability Plot .....	40
Figure 4-6. ANOVA Illustration for a Single Factor .....	44
Figure 4-7. Cumulative Probability Functions for Measured and Simulated Data .....	45
Figure 4-8. Propagation of Input Uncertainties to Obtain Output Uncertainties .....	53
Figure 4-9. Example P-Box .....	54
Figure 4-10. Increase in Predictive Uncertainty Due to the Addition of Model Form Uncertainty .....	55
Figure 4-11. Missile Warning System .....	56
Figure 4-12. 3-D Graphical Representation of the Regression Model Developed Using DOE for MWS Example .....	58
Figure 4-13. SPRT Applied to Analyze a System’s Failure Rate .....	64
Figure 4-14. Example of Implementation of Sequential DOE.....	65
Figure 4-15. Example of a Group Average Plot .....	67
Figure 4-16. Example of PPV Histogram .....	68
Figure 4-17. Example of PPV Calibration Curves.....	69
Figure 4-18. Example of Metrics with Confidence Intervals Indicating Good Calibration, Good Discrimination, and Good Overall Performance.....	73
Figure 4-19. Example Output from the CORA Validation Methodology.....	76
Figure 4-20. Visualizing the Accuracy and Variability Components of the Fidelity Metric.....	82
Figure 4-21. Volume Coverage Versus Density Coverage for Continuous Factors .....	83

## Tables

Table 2-1. Outlines of Four Core VV&A Documents .....	6
Table 2-2. Information Supporting VV&A and Associated Statistical Concepts .....	9
Table 3-1. Recommended Experiment Designs for Common Test Objectives .....	20
Table 3-2. Comparison of Five Space-Filling Design Types.....	30
Table 3-3. Simulation Design Recommendations.....	33
Table 4-1. Data Set for Lag Plot Example .....	37

## Contents

Table 4-2. Data Table for Normal Probability Plot Example .....	39
Table 4-3. Values of $c(\alpha)$ for Different Significance Levels.....	46
Table 4-4. Example of a Contingency Table .....	47
Table 4-5. Statistical Analysis Methods According to Distribution of Responses and Amount of Live Test Data .....	50
Table 4-6. $2^4$ Design for MWS Test .....	57
Table 4-7. Calculated Upper Confidence Intervals for MWS Example .....	59
Table 4-8. Calculated Upper Prediction Intervals for MWS Example .....	60
Table 4-9. Calculated Upper Tolerance Interval for MWS Example .....	61
Table 4-10. PPV Methods for Comparing M&S Probabilities with Binary Test Results.....	66
Table 4-11. Notional Pairs of Observed Binary Data and Predicted Probabilities .....	71
Table 4-12. Validation Metrics and the Types of Data to Which They Are Targeted.....	77
Table 4-13. Referent Authority Level Scale .....	82
Table B-1. Top-Level DoD T&E Policy and Guidance.....	89
Table B-2. Service and OSD Policy and Guidance on VV&A of M&S for T&E .....	92



This page is intentionally blank.

# 1 Introduction

## 1.1 Purpose of the Guidebook

This guidebook provides guidance on the development; verification, validation, and accreditation (VV&A); and use of modeling and simulation (M&S) in support of developmental test and evaluation (DT&E), live fire test and evaluation (LFT&E), and operational test and evaluation (OT&E) of Department of Defense (DoD) systems. M&S can complement physical testing and improve the outcomes of the campaign of learning across the acquisition life cycle. The guidebook presents definitions, processes, and techniques to promote high-value utilization of M&S in developing and evaluating defense capabilities. The guidebook serves as a desk reference for test engineers and analysts responsible for VV&A of M&S. It also provides program leadership insight into the value of M&S in support of test and evaluation (T&E) and how to integrate M&S into an effective and efficient T&E program.

For DT&E, this guidebook is part of a broader set of guidance developed to support the implementation of a DT&E paradigm shift. Key enablers include model-based systems engineering (MBSE) and digital engineering (DE) (authoritative source of truth); incorporation of technological innovations; a supportive infrastructure; and a transformed culture. This paradigm shift leverages the principles of Agile and scales them to move DT&E holistically from a serial set of activities to an integrative framework focused on capability and outcome-focused testing, Agile scalable evaluation, and enhanced test design facilitating an ongoing campaign of learning. Fully leveraging M&S will provide the agility required to support an iterative DT&E campaign, consistent with the “constructive” contribution to a full spectrum of live, virtual, and constructive testing.

For OT&E and LFT&E, this guidebook complements DoD Manual (DoDM) 5000.102, “Modeling and Simulation Verification, Validation, and Accreditation for Operational Test and Evaluation and Live Fire Test and Evaluation.” M&S is commonly used in OT&E and LFT&E evaluations when testing an end-to-end capability in a live environment is unsafe, infeasible, or prohibitively expensive. This guidebook supports the policy captured in DoD Instruction (DoDI) 5000.98, “Operational Test and Evaluation and Live Fire Test and Evaluation,” which promotes an expanded use of M&S in OT&E and LFT&E including using the latest advances in science and technology and digital technologies (e.g., DE and MBSE) in OT&E and LFT&E. The guidebook provides details on processes, test design methods, and statistical analysis methods that support the updated policy and guidance.

### 1.2 Scope of the Guidebook

The guidebook begins by discussing the use of M&S for T&E as part of the capability development and management processes needed to ensure that the models are used for their intended purpose across the acquisition life cycle. Section 1 of the guidebook provides a high-level description of the role that M&S plays in T&E and, more broadly, how M&S and T&E combine with systems engineering practices to support system development. Section 2 describes the processes required to verify, validate, and accredit models for use in T&E and the best practices associated with following those processes.

The guidebook then describes the use of scientific test and analysis techniques (STAT) for the development of M&S for T&E, including design of experiments (DOE) for VV&A of M&S and analysis of its output in support of T&E. The guidebook describes experimental designs, when they are appropriate, and their intended results. Section 3 details the development of a test design for M&S in T&E.

The guidebook then provides techniques for the analysis of M&S results, with a focus on VV&A. It is imperative that testers, systems engineers, analysts, and program managers (PMs) mutually understand what constitutes quality test design and analysis to ensure that the program collects the data needed to validate models and quantifies the risks associated with using M&S for T&E. Section 4 details these best statistical practices for analyzing the results of the model.

Appendix A provides a catalog of popular experiment designs; Appendix B outlines policy and guidance for M&S in T&E; and Appendix C provides information about relevant training.

The guidebook is not intended to prescribe any methods or techniques as standard or preferred; rather, the guidebook describes best practices for use, development, and VV&A of M&S activities; the situations to which they are best suited; and how to collect and utilize their outputs. In addition, the guidebook summarizes the key points, including benefits and risks, of each method.

Although the guidebook is not a textbook that describes methods in detail, technical personnel should find sufficient explanation to apply the methods, along with recommended references for further details.

### 1.3 Role of M&S in T&E

Incorporating M&S into DoD acquisition can help PMs balance the cost, schedule, and performance of acquisition in accordance with DoD Directive (DoDD) 5000.01, “The Defense Acquisition System,” and DoDI 5000.98. The Government Accountability Office (GAO) recommends using knowledge-based acquisition practices that include M&S and, as stated in the

GAO Special Report, “Weapon Systems Annual Assessment: Knowledge Gaps Pose Risks to Sustaining Recent Positive Trends,” GAO found that “attaining high levels of knowledge before significant commitments are made during product development drives positive acquisition outcomes.”

M&S can reduce the cost of T&E by allowing the testing of systems and processes in a virtual environment before physical prototypes are built (Krasner 2015; Boehm 2021). M&S can also be used to replicate real-world scenarios that would be too expensive, dangerous, or time-consuming to recreate physically (McDermott and Van Aken 2020). It can be used to train personnel on new systems or processes, reducing the cost of training and the potential for errors during the actual operation (McGinnity 2016). M&S can serve as a powerful tool for conducting comprehensive and rigorous tests and evaluations to reduce program risk and improve system performance. It can simulate extreme environmental conditions, high-stress situations, rare events to test the system’s robustness and reliability, or unobserved user behaviors to evaluate the system’s performance and effectiveness (Beling et al. 2021).

Despite these potential benefits, programs must be careful when incorporating M&S into acquisition. M&S aims to provide valid forecasting of system performance under operationally representative conditions to minimize the risk associated with M&S-informed decisions in T&E. However, M&S can only approximate the actual system regardless of the investment. Therefore, programs must establish confidence in M&S before accepting its results to inform decisions (for more information, see the North Atlantic Treaty Organization Guidelines for M&S Use Risk Identification, Analysis, and Mitigation), and PMs must decide the level of risk they are willing to accept. Best practices include using statistical techniques to quantify the uncertainty of how well those models represent reality.

Statistical methods play an important role in verification and validation (V&V) of models. It can be unclear, however, what methods are most appropriate and what information these methods provide to the PM. A 2020 MITRE review on the use of M&S for T&E found that guidance for verification methods is too generic and that the efficiency and effectiveness of verification testing can be improved by adding the use of statistical methods (Cortes and Ortiz 2020). Therefore, Sections 2.2.1 and 4 of this guidebook describe methods for V&V of models and how these methods reduce the risk of using M&S for programmatic decisions.

## 2 M&S Processes

Successfully applying M&S to T&E requires a series of detailed steps and a deep understanding of the system or process being modeled. These steps are vital for ensuring the accuracy of the results and providing the most information to decision makers. Section 2 discusses the processes needed to successfully include M&S in T&E.

The VV&A process covers the crucial steps needed to ensure that the model is designed correctly, accurately mirrors the real-world system it aims to simulate, and is fit for its intended purpose. These steps are required by DoD and are the key to establishing confidence in the credibility and reliability of the M&S.

STAT methods are central to understanding the simulation results, spotting patterns, and making informed decisions. Using the correct statistical methods allows for a comprehensive understanding and characterization of model responses, including their statistical uncertainty, across all input variations. In conditions where live data cannot be collected, statistical models can be useful in extrapolating expected performance with associated uncertainty. Ultimately, it is crucial to have subject matter expertise to interpret whether the statistical analysis results and the discrepancies they expose between live test data and M&S have any practical significance. Subject matter expertise is also useful in evaluating models in regions where live data cannot be collected and determining the value of any statistical extrapolations into those regions.

Programs that use M&S for T&E must include all relevant stakeholders in M&S development. This involvement ensures that the simulation results are meaningful and helpful to all stakeholders.

### 2.1 VV&A Processes

Programs that use M&S to inform decisions should certify the models and simulations for their intended use through the VV&A process. The purpose of VV&A of M&S in T&E is to ensure that all stakeholders have a clear understanding of the M&S capabilities and limitations and the uncertainty around the simulation results. DoD adheres to stringent VV&A processes to ensure the appropriateness of M&S results for a specific purpose, as defined in DoDI 5000.61, “DoD Modeling and Simulation Verification, Validation, and Accreditation.” Verification involves checking that the model has been correctly implemented and operates as intended, and validation ensures that the model accurately represents the real-world system or scenario it is designed to simulate. This process helps to identify and correct any errors or inconsistencies in the model, ensuring that the results it produces are as accurate and reliable as possible. V&V activities are not done at a single point in time but rather are life cycle processes in which any pertinent new

evidence (e.g., new live test data) is to be used to refine and validate the model on an ongoing basis.

Accreditation, on the other hand, is the formal recognition that a model or simulation is suitable for a specific purpose. It involves an assessment of the model's credibility, considering factors such as its accuracy, reliability, and relevance to the scenario being simulated. This process provides assurance that the model or simulation can be trusted to provide accurate and reliable results in the context of its intended use.

DoDI 5000.61 establishes policy, assigns responsibilities, and prescribes procedures for the VV&A of models, simulations, distributed simulations, and associated data and establishes the basis for credible M&S across DoD. Each Service published additional guidance to clarify roles and responsibilities within its organization. In addition to referencing DoDI 5000.61, the Service guidance also references Military Standard (MIL-STD) 3022, "Documentation of Verification, Validation, and Accreditation (VV&A) for Models and Simulations." MIL-STD-3022 recommends standardizing core VV&A documents and provides templates for each document. Table 2-1 provides outlines of the four core VV&A documents (Accreditation Plan, V&V Plan, V&V Report, and Accreditation Report); information that is common and shared across the documents appears in italicized font in the table. Typically, different groups prepare and use these documents at varying times, and a significant portion of the information contained in each is common and should be shared. The templates offer a unified structure and the ability to interface between the four documents promoting uniformity and efficiency. Furthermore, the plans and reports generated using these templates act as a communication tool among the participants in the VV&A processes. The PM, T&E Working-Level Integrated Product Team, and M&S Working Group (MSWG) should consider generating information equivalent to that identified in MIL-STD-3022 using digital and model-based engineering techniques rather than creating documents. The PM must ensure that both the creators and the receivers of the information have the appropriate tools and skills to receive the information in the new formats. The program may find that digital delivery methods are faster, are less error-prone, and enable reuse of M&S internal and external to the program, thus improving capability delivery across DoD.

For OT&E and LFT&E, DoDM 5000.102 provides details on M&S VV&A. The manual highlights that M&S and the related VV&A strategy, including the required resources, will be outlined in the Test and Evaluation Master Plan (TEMP)/T&E strategy. DoDM 5000.100, "Test and Evaluation Master Plans and Test and Evaluation Strategies," outlines the information required in TEMPs/T&E strategies, as well as in V&V plans.

Table 2-1. Outlines of Four Core VV&amp;A Documents

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

Source: MIL-STD-3022

The Institute for Defense Analyses (IDA) Handbook on Statistical Design and Analysis Techniques for M&S Validation recommends the following nine-step approach to developing the V&V process (Wojton et al. 2019). See the IDA Handbook for more details on each step, as well as its linkage to statistical elements of M&S VV&A.

1. Develop the intended use statement.
2. Identify the response variables or measures.
3. Determine the factors that are expected to affect the response variable(s) or that are required for operational evaluation.
4. Determine the acceptability criteria.

5. Estimate the quantity of data that will be required to assess the uncertainty within the acceptability criteria.
6. Iterate the Model-Test-Model loop until desired model fidelity is achieved.
7. Verify that the final instance of the simulation accurately represents the intended conceptual model (verification process).
8. Determine differences between the model and real-world data for acceptability criteria of each response variable using appropriate statistical methods (validation process).
9. Identify the acceptability of the model or simulation for the intended use.

### 2.1.1 Accreditation for Specific Use

Programs that use M&S for T&E will need to accredit their model(s) for a specific use; this is not a broad acceptance of the model or modeling tool for any use. DoDI 5000.61 defines accreditation as “the official certification that a model, simulation, or distributed simulation is acceptable for use for a specific purpose.” The Services offer carveouts for general-use models that may be accredited for a class of applications, but these models must also be accredited for their specific use when using the model for T&E. For example, Air Force Instruction 16-1001 states that the Deputy Chiefs of Staff may accredit common-use models (models provided to multiple DoD Components) and general-use models (models or representations that are common to many models or simulations). Similarly, Army Regulation 5-11 states that M&S can be accredited for a generic set of applications by the Army official with general oversight responsibility for that class of applications and that as long as the M&S application falls within the guidelines of the class accreditation, the entire M&S need not undergo V&V for a new application. Finally, Secretary of the Navy Instruction (SECNAVINST) 5200.46 states that M&S services, tools, and data should be reused to the extent possible. The Director, Operational Test and Evaluation (DOT&E) TEMP Guidebook mandates that, for OT&E, existing M&S capabilities previously accredited for other applications must complete another VV&A process and be accredited for each new intended use. DoDM 5000.102 highlights that operational test agencies accredit models for OT&E objectives and LFT&E organizations accredit models for LFT&E intended uses. However, as stated in the DOT&E TEMP Guidebook, previous VV&A may simplify the process because the previous efforts have been documented and the new VV&A effort typically can focus on the changes.

Ultimately, it is up to the PM to identify how to accredit the model for use in T&E in conjunction with the accreditation authority as specified in the applicable Service policy and guidance (see Table B-2 in Appendix B). DoDM 5000.102 requires that the M&S strategy, including the data needed for V&V activities, is included in the TEMP and/or T&E strategy to ensure that adequate



data are collected to support VV&A activities. This requirement highlights the need for PMs to understand how models are verified and validated using statistical techniques.

### **2.2 The Role of Statistical Analyses in M&S VV&A**

Statistical methods enable the characterization of models' responses across the full range of their inputs, including their associated uncertainties. Subject matter expertise is also required to determine whether the results of statistical analyses and the differences between the live test data and M&S that the analyses reveal have a practical impact on system performance or mission execution. Portions of this section have been summarized from Chapter 2, Section E and Chapter 3, Sections B through E of the IDA Handbook (Wojton et al. 2019); see that reference for additional details.

Statistical concepts and analyses play important roles throughout the VV&A process described in Section 2.1 of this guidebook. Table 2-2 links information needed during the VV&A process to the appropriate statistical concepts needed in the process. For example, M&S responses should be compared with live test data in an iterative model-test-model approach, ideally beginning before live data are available. The M&S responses used both for characterizing the behavior of the responses and for comparison purposes should be selected using statistical design approaches discussed subsequently in this guidebook: classical DOE if the M&S is stochastic, and design for computer experiments (including space-filling designs) if the M&S is deterministic (or nearly deterministic). See Section 3 of this guidebook for additional details on DOE. Before making comparisons with live data, the M&S responses should be characterized and understood using variation analysis and metamodeling (statistical emulation); see Section 4.4.5.

**Table 2-2. Information Supporting VV&A and Associated Statistical Concepts**

<b>VV&amp;A Information</b>	<b>Statistical Concepts</b>
M&S Requirements and Intended Use	Response variables Factors Stochastic versus deterministic M&S
Data Requirements	Classical DOE Design for computer experiments
Iterating Model-Test-Model	Classical DOE Design for computer experiments Variation analysis Metamodels Comparing live test data to M&S responses
Verification Analysis	Parametric emulators
Validation Analysis	Parametric emulators Comparing live test data to M&S responses
Uncertainty	Hypothesis testing and interval estimation

Source: IDA Handbook (Wojton et al. 2019)

Variation analysis. Variation analysis includes sensitivity analysis and Monte Carlo analysis. Monte Carlo analysis can be applied to M&S responses that are stochastic. This type of analysis involves running the model many times using the same inputs to obtain the variation in responses that occur. Sensitivity analysis can be used for both stochastic and deterministic M&S. It can involve using large changes in model inputs to generate the data needed to develop a parametric emulator, or it can involve using small changes in inputs to determine whether the resulting changes in responses are reasonable (i.e., correspondingly small and predictable). By characterizing the changes in inputs that drive significant changes in responses, variation analysis helps quantify uncertainty and guides where data from live testing and/or M&S improvements should be sought or made.

Metamodels (also called statistical emulators). Metamodels (see Section 4.4.5) use M&S responses generated using statistical design approaches to build a statistical model of the M&S. The metamodel can be used to predict responses and their variation across a range of input conditions, including those not explicitly simulated, without repeatedly rerunning the model. Both the choice of design approach and the type of statistical model used depend on whether the M&S is stochastic or deterministic. As explained in the IDA Paper, “Metamodeling Techniques for Verification and Validation of Modeling and Simulation Data” (Haman and Miller 2022), metamodels of deterministic, discrete responses should be built using nearest neighbor or decision tree interpolators. Metamodels of deterministic (or near-deterministic), continuous responses should adopt Gaussian process (GP) interpolation. Continuous or discrete stochastic

responses should use a generalized additive model (GAM). Analogous to variation analysis, metamodels help characterize uncertainty and aid in the selection of live test points. Inputs that strongly affect responses suggest the need for additional live test data, and those that do not suggest the opposite.

Comparing live test data to M&S responses. After characterizing the M&S responses, live test data should be compared with those responses to identify and understand the differences between the two. Differences between means, variance, or other aspects of the distributions of the two sets of data should be analyzed. The appropriate statistical methods to use depend on whether the responses are binary or continuous, and if the latter, whether their distributions are skewed or symmetric. The nature of the factors most affecting the responses also matter, including whether there are no varying factors across the data (univariate); there are factors that affect means, but there is no designed experiment associated with them (distributed); or there is a designed experiment with factors that determine the differences between the M&S responses and live test data (see Section 4.2.6).

The purpose of the analysis is to determine whether the differences between the M&S responses and live test data are statistically significant. To do so, the statistical techniques used generally test hypotheses: the null hypothesis ( $H_0$ ), that there is no difference between a response and its corresponding live test data; and the alternative hypothesis ( $H_1$ ), that there is a difference. The data are analyzed to either reject the null hypothesis—meaning there is a statistically significant difference between the M&S and live test data—or to fail to reject the null. Section 4 provides additional discussion on some of the hypothesis testing techniques that are most relevant to validating models. However, it is important to note that hypothesis tests are a tool to guide decision making and not a decision. The overall accreditation decision should leverage not only hypothesis test results but also the uncertainty quantification (UQ) across the domain of the intended use. Subject matter expertise is critical for assessing the practical impact of any identified differences and/or the ability of the model to extrapolate into areas that cannot be executed in live testing.

### 2.2.1 Uncertainty Quantification

As models and simulations are increasingly used to support T&E and acquisition program decisions, it is important to remember that they are imperfect representations of the real-world system. In other words, “all models are wrong, but some are useful” (Box 1976). UQ focuses on quantifying the uncertainty present in model predictions and live data independently and the contributions of specific sources to overall uncertainty, whereas validation is focused on quantifying the uncertainty in the difference between models’ predictions and live data. Both are necessary for decision makers to fully understand the risk accepted when making decisions based on simulation output; this is especially important when the M&S intended use includes

extrapolation and prediction outside the envelope in which the real-world system can be tested. By understanding the uncertainties, the risk can be quantified for using M&S as surrogates, and informed decisions can be made about using M&S to improve outcomes. To understand the uncertainties, T&E is performed to verify, validate, and accredit the M&S.

Figure 2-1 illustrates the role UQ plays in the V&V process. It is through this V&V process that stakeholders can thoroughly assess the correctness and credibility of M&S results. From a verification perspective, UQ provides a quantitative measure of the extent to which a computational model represents its underlying mathematical model. From a validation perspective, UQ accounts for uncertainties in both computational and experimental results when the two are compared to establish the extent to which the M&S results resemble the real-world data.

There are generally two broad categories of uncertainty: statistical and knowledge (Wojton et al. 2019).<sup>1</sup> Statistical uncertainty is due to inherently random effects; can be better characterized by accumulating more samples (i.e., more replicates); and is generally characterized by a probability distribution. An example of statistical uncertainty is measurement error. In contrast, knowledge uncertainty is due to a lack of information and thus can be reduced by collecting data under new conditions, which provides new information, but generally cannot be characterized by a probability distribution.<sup>2</sup> An example of knowledge uncertainty is the unknown performance capabilities of a foreign weapon system due to limited intelligence data.

The UQ process consists of three key steps:

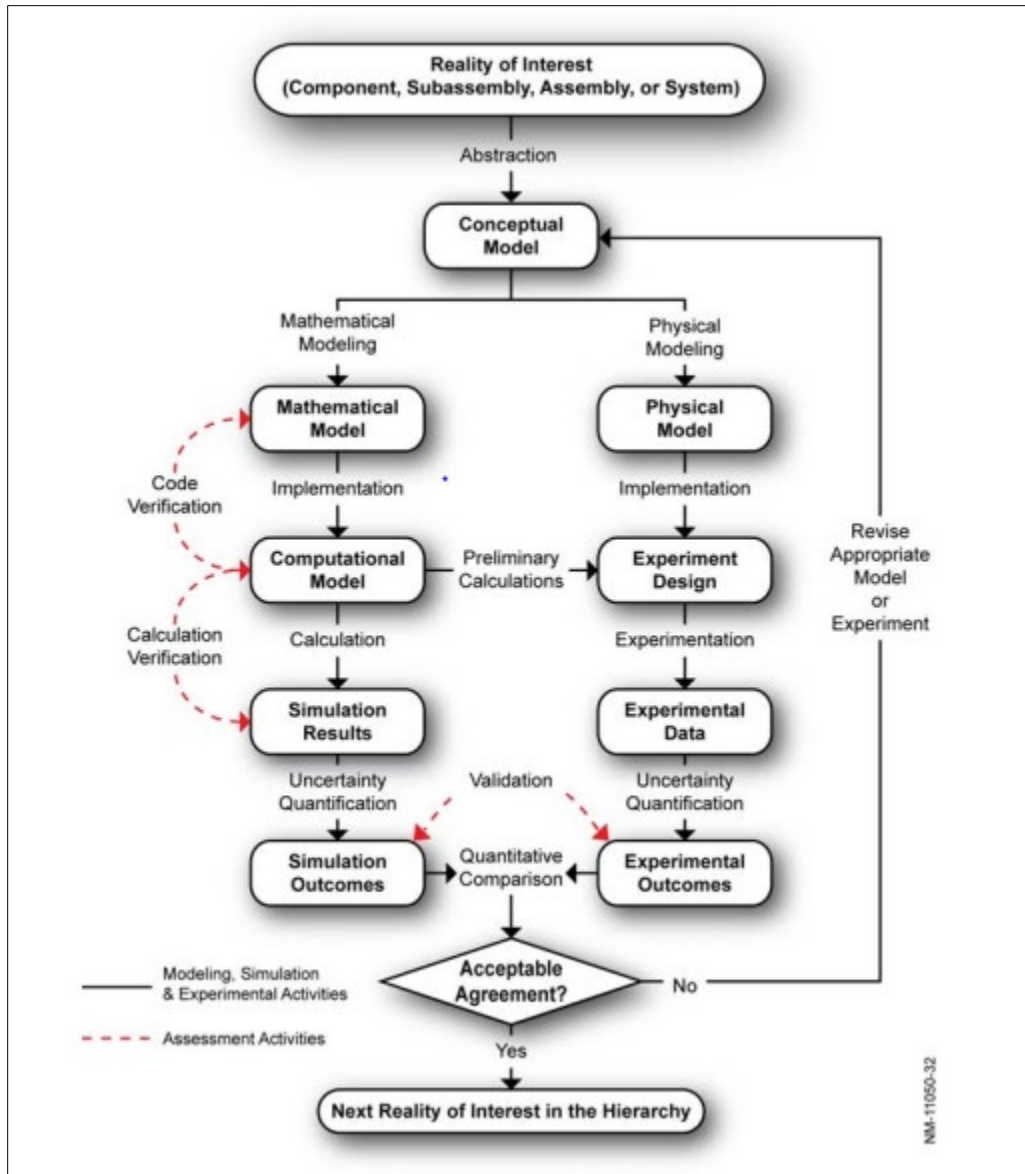
1. Assess sources of uncertainty: Identify and quantify all sources of uncertainty in the model.
2. Propagate uncertainties: Quantify the effect of input uncertainties on output variability.
3. Assess output variability: Evaluate the estimated variabilities and their potential impact.

Information on UQ methods, including recommended best practices for each step in the UQ process, can be found in Section 4.3.

---

<sup>1</sup> Aleatory and epistemic are also commonly used terms for statistical and knowledge uncertainty, respectively.

<sup>2</sup> Knowledge uncertainty is typically represented as an interval with no associated probability density function (PDF). However, it may be represented as a PDF that reflects the subject matter expert's degree of belief (Cortes et al. 2021). Methods for eliciting subjective probability functions are outside the scope of this guidebook; an introduction to the topic can be found in the book *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis* (Morgan and Henrion 1990).



**Figure 2-1. V&V Activities and Outcomes**

### 2.2.2 Evaluating Model Adequacy

Whether a model is “good enough” to be validated and accredited depends on its intended use, the referent authority, its fidelity, the responses it is expected to generate, and the scope of its inputs that can be used to generate those responses. This section of the guidebook is a synopsis of the discussion in Chapter 1, Section D of the IDA Handbook (Wojton et al. 2019); see that reference for additional details.

Intended use. A model’s intended use constitutes the specifics of how it will be used as part of a test program. Intended use comprises the model’s hierarchy (its fidelity and type consistent with

its purpose); its response variables (the outputs it will generate for the purposes of evaluation); the range of its input conditions and their intersection with the conditions over which live testing data will be available; and the differences between live data and model responses that are permissible (acceptability criteria). Identifying in model responses the kinds of uncertainty and their magnitudes that are relevant for the purposes of evaluation involves the details of the model's implementation, the system being tested, and the model's intended use.

Referent authority. The referent is that against which the model is being compared, and the referent authority is the strength of credibility of a referent's claim to be a high-fidelity representation of reality. Live data of the system would have the highest referent authority, but it may not always be available, especially in the early development phases of a program. In those situations, other models or subject matter expert (SME) judgment, which are less authoritative, may serve as the referent. Section 4.4 describes an approach for assessing referent authority.

Fidelity. Fidelity is the degree to which the representation within a simulation is similar to a real-world object, feature, or condition in a measurable or perceived manner. The DoD M&S Glossary defines fidelity as the accuracy of the representation when compared to the real world. Model hierarchy and fidelity span physics and engineering models (e.g., the Probability of Raid Annihilation Testbed (Thomas and Dickinson 2015), which comprises multiple physics and engineering sub-models); engagement models (e.g., the Enhanced Surface-to-Air Missile Simulation (ESAMS) (Baty et al. 1988)); mission-level models (e.g., SUPPRESSOR (Whitehurst et al. 1997)); and campaign-level models (e.g., the Logistics Composite Model (Boyle 1990)). Validation should be consistent with the model's hierarchy and fidelity. Mission-level models can comprise aggregations of many engagement models. That aggregation includes all those models' uncertainties, implying that mission-level models can have greater uncertainty in their responses compared to a single engineering or engagement model. A model's uncertainty should be considered rigorously when using its responses for conducting evaluations. Approaches for doing so include comparing inputs and responses between the model and live testing, including their means, variances, and other parameters. Techniques for accomplishing these comparisons are discussed in Section 4 of this guidebook.

If a series of models is used to evaluate performance of a complex system comprising multiple subsystems, live testing is unlikely to generate enough data to enable the use of simple statistical methods to quantify overall uncertainty in the models' responses. Other techniques such as Monte Carlo error propagation in conjunction with SME judgment can be used to quantitatively and qualitatively assess the validity of the models' responses and uncertainty (Ogilvie 1984).

Responses. Selecting model responses appropriate to the test program and what the test is measuring is necessary for successful validation and accreditation. Responses such as overall mission success or force ratios following engagement are often accompanied by substantial

uncertainty and are unlikely to be capable of straightforward estimation using live test data. These attributes make such responses inappropriate choices for conducting rigorous validation and accreditation. A better approach for validating and accrediting a model as “good enough” is to use specific performance measures that strongly affect the overall mission success such as ranges at which an aircraft is detected when engaging threats or number of successful engagements versus total engagements.

Scope. The scope of a model’s validity is defined by the range of inputs that can be used to generate responses for comparison with live test data and for face validation (subjective judgment by SMEs of whether model responses are appropriate). The initial determination of scope is accomplished by SMEs identifying the input conditions that should be important determinants of system performance. Scope can then be adjusted based on the responses obtained and their comparison with live test data.

### **2.2.3 Combining Statistical and Subject Matter Expertise**

Methods for designing the combined M&S and live testing experiments and analyzing their results, enabling quantitative evaluation of the differences between modeling results and live testing data, are described in this guidebook (see Sections 3 and 4) and in detail in the references found in those sections. These methods combine classical experimental designs (Johnson et al. 2012) with computer experiments (Santner et al. 2003; Fang et al. 2005; Kleijnen 2008), and, based on the variance in the differences, can inform estimates of uncertainty. This section of the guidebook is a synopsis of the discussion in Chapter 1, Section E of the IDA Handbook (Wojton et al. 2019); see that reference for additional details.

Statistical methods cannot account for all sources of uncertainty. For example, knowledge uncertainty is due to a lack of information and thus cannot be reduced by collecting more replicates from either live testing or M&S, as previously described in Section 2.2.1 of this guidebook. Knowledge uncertainty can be reduced by collecting data that provide new information, for example, incorporating better theoretical descriptions into models, which involves the use of subject matter expertise. Moreover, the M&S may be used to explore system performance against future threats and in complex combat scenarios that cannot be replicated in live testing and that include substantial supposition on the part of SMEs. In these cases, collecting live test data is impossible and therefore evaluation of the results and their uncertainty can of necessity be substantially qualitative. Nonetheless, it is necessary to document the qualitative judgments reached and their bases, and judgments by SMEs should be obtained using rigorous approaches.

Metamodels (or statistical emulators) (see Section 4.4.5) are a means of summarizing M&S responses across a wide span of inputs without using the simulation itself to generate each of the

responses (Haman and Miller 2022). Constructing a metamodel requires inputs from both subject matter and statistical experts regarding the factors that should most affect simulation responses and therefore be incorporated in the metamodel. Metamodels enable SMEs to judge whether simulations are quickly and efficiently generating reasonable responses across their full range of inputs, including those not covered by live testing. Metamodels also provide data on the variation in simulation responses, thus informing judgments regarding uncertainty.

In summary, both statistical and subject matter expertise are needed to comprehensively assess M&S results and use those results in T&E and to support decision making.

### 2.3 Participation in M&S Development and Execution

To make the best use of M&S across the continuum of testing, programs need to develop the M&S VV&A strategy early in the development process when the program first identifies the need for M&S. This may mean developing a preliminary strategy as early as Milestone A to capture data from contractor testing. This includes identifying the intended use(s) of the model and output data, the required fidelity, and the methods and required data for V&V of the model. Representatives from the operational user, technology development, testing, and accreditation communities should provide input into the M&S VV&A strategy. The strategy is documented in the Decision Support Evaluation Framework and Integrated Decision Support Key (IDSK) (see Section 2.3.2), and it should address the VV&A activities, including required data, and the use of M&S output to support decisions across the program life cycle.

#### 2.3.1 Modeling and Simulation Working Group

Programs should establish an MSWG under the T&E Working-Level Integrated Product Team. The size and capabilities of the MSWG should reflect the size and complexity of the simulation effort. The MSWG members should include representation from the following:

- **Operational user**, who defines the problem and the intended use of the capability under development.
- **Developer**, who designs and implements the M&S.
- **Testers**, both those who conduct the V&V test of the M&S and those who use M&S output to support T&E events.
- **Accreditation agent**, who conducts the accreditation assessment of the M&S.

The MSWG should meet at least quarterly or when an event occurs that is likely to affect the M&S VV&A strategy. Significant changes to the intended use, required fidelity, system or test



scope, schedule, and resource availability are examples of events that necessitate MSWG review and revision of the M&S VV&A strategy.

### **2.3.2 Connection to Capability Evaluation and Decision Support**

Testing and M&S provide data for capability evaluation to inform decisions throughout the capability life cycle. IDSK development uses a decision-evaluation-data thought process. Articulating the thought process into the IDSK provides a T&E strategy able to inform technical, programmatic, and operational decision makers. By making capability-evaluation-informed decisions, the risk that the capability does not meet technical requirements and mission needs is minimized.

The IDSK aligns the generation of data from contractor, developmental, integrated, and operational testing and M&S events to the technical and operational capability evaluation and to the technical, programmatic, and operational decisions. The data may be collected from physical testing or from appropriately verified and validated M&S. For OT&E and LFT&E policy requirements, see DoDI 5000.98. The forthcoming DoDI 5000.DT, “Developmental Test and Evaluation,” will provide DT&E policy requirements.

M&S can be both an IDSK data provider (e.g., supplying data or results for capability evaluation) and an IDSK data consumer (e.g., using test data to inform an M&S accreditation decision). For example, M&S can provide data to evaluate a capability that cannot be tested and must be evaluated using M&S, such as probability of raid annihilation. The M&S used to evaluate probability of raid annihilation may also have been a consumer earlier in the life cycle; the VV&A plan for the M&S requires evaluation of physically collected data to make the accreditation decision.

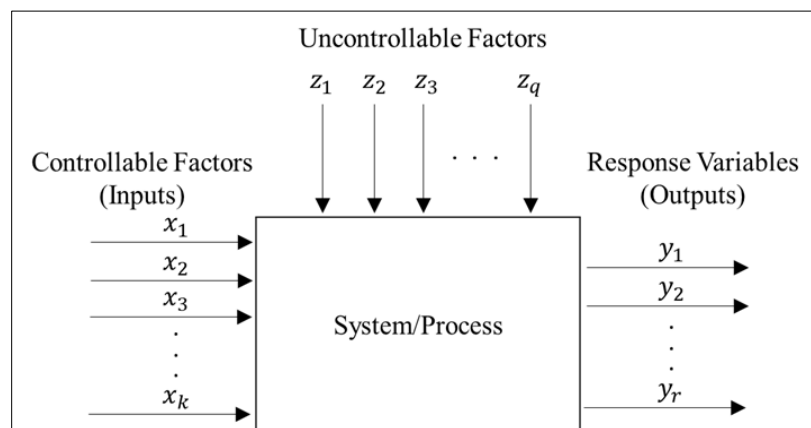
### 3 Design of Experiments to Support VV&A

As previously described in Section 2.2, statistical analyses should underpin the M&S VV&A process. Section 3 describes methods for designing experiments to produce the required data for those statistical analyses. The M&S VV&A should compare M&S output to live data whenever possible, and this requires a combination of both physical experiments (Section 3.1) and computer experiments (Section 3.2).

#### 3.1 Design of Physical Experiments

DOE methodologies integrate well-defined and structured scientific strategies with statistical techniques to gather knowledge about a system and then transform it into information that guides decision quality. The DOE process drives testers to determine exactly what questions need to be asked, how many resources are needed to answer those questions, and which analysis techniques are required to interpret the data and answer the test questions.

Figure 3-1 depicts a generic input-process-output (IPO) diagram of a system or process.



**Figure 3-1. Input-Process-Output Diagram of a System or Process Under Testing**

The core principle of DOE is to make deliberate and systematic changes to the controllable factors (i.e., inputs) of a system and observe and estimate their effect on the response variables (i.e., outputs). The uncontrollable factors are variables that cannot be controlled in the test environment. In operational tests, these tend to be environmental factors, such as weather and temperature. In some cases, they are operator or user based (e.g., user experience, training, demographics). Uncontrollable factors are typically dealt with via replication, randomization, and blocking.

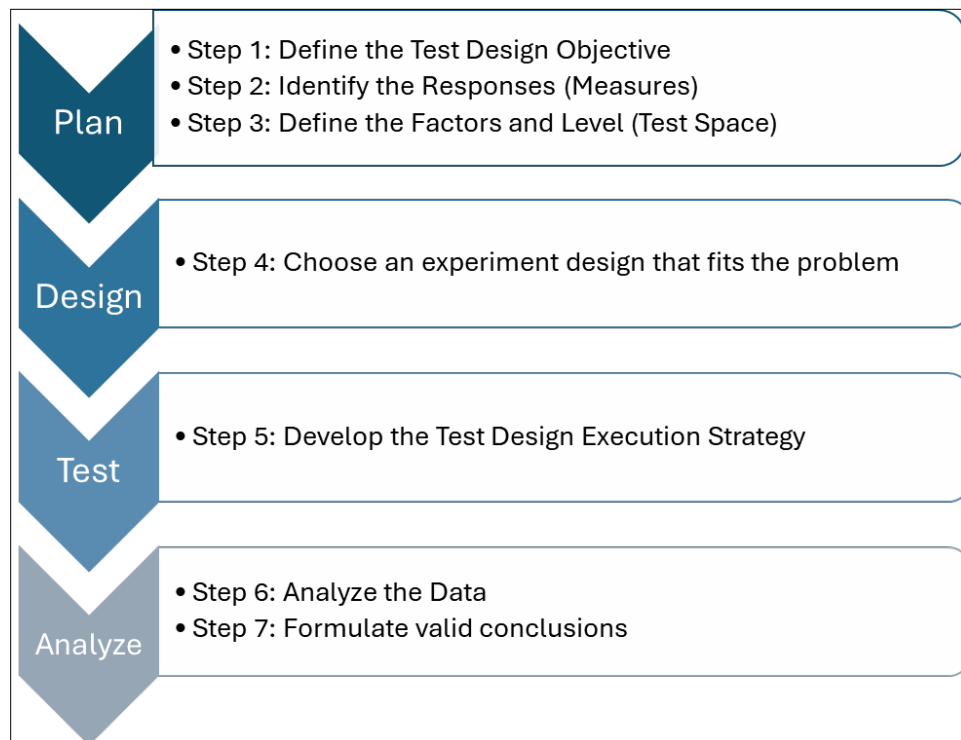
- Replication: Performing test runs multiple times to allow for an estimation of uncertainty.
- Randomization: Deliberately executing test runs randomly to eliminate potential biases.

- Blocking: A technique for removing the effects of unwanted nuisance factors from the analysis.

For a more detailed summary of how uncontrollable factors are handled, see the book *Design and Analysis of Experiments* (Montgomery 2013) and the MITRE Technical Report, “Modern Test Design and Analysis Playbook” (Cortes and Ortiz 2020).

DOE methodologies help generate test efficiencies by simultaneously maximizing information gains and minimizing resources. Because the high cost of performing certain physical experiments is one of the key reasons why M&S is developed, DOE helps determine the optimal real-world test scenarios needed to ground the M&S. DOE balances test risk and resources by building designs tailored for specific test questions and objectives. Developing a good test design is about not only how many test points are collected but also which test points are collected and their relationship to each other. As a result, DOE provides cleaner and richer analyses that build predictive models to incorporate variability and uncertainties.

Figure 3-2 shows the phases and the steps recommended to develop efficient and effective test designs (Coleman and Montgomery 1993).



**Figure 3-2. DOE Process for Building Test Designs**

#### 3.1.1 Plan Phase

During the Plan phase, a working group consisting of systems SMEs and DOE/statistics experts works through steps 1–3:

**Step 1: Define the objective of the test and the various questions to be answered.**

In the case of M&S, the objective is to validate that the simulation provides realistic results and that the effect of factors on performance is indeed accurate.

**Step 2: Identify the responses.**

Determine which response variables should be used to help validate M&S results. These response variables must match some output of the M&S to allow some comparison. See “Responses” in Section 2.2.2 for some insight on selecting good response variables.

**Step 3: Define the factors and their level to be examined.**

Identify all potential controllable and uncontrollable factors that could influence the response, and then determine their data type and how best to control them for the sake of testing, establishing whether blocking or other randomization restrictions are involved. Select practical factor levels that can produce the maximum amount of information with the minimum number of runs, and identify unfeasible factor combinations (i.e., constraints and disallowed combinations).

#### 3.1.2 Design Phase

During the Design phase, the DOE expert takes the information provided up to this point and develops a test design that balances test risk (i.e., power and confidence, see below) with test resources.

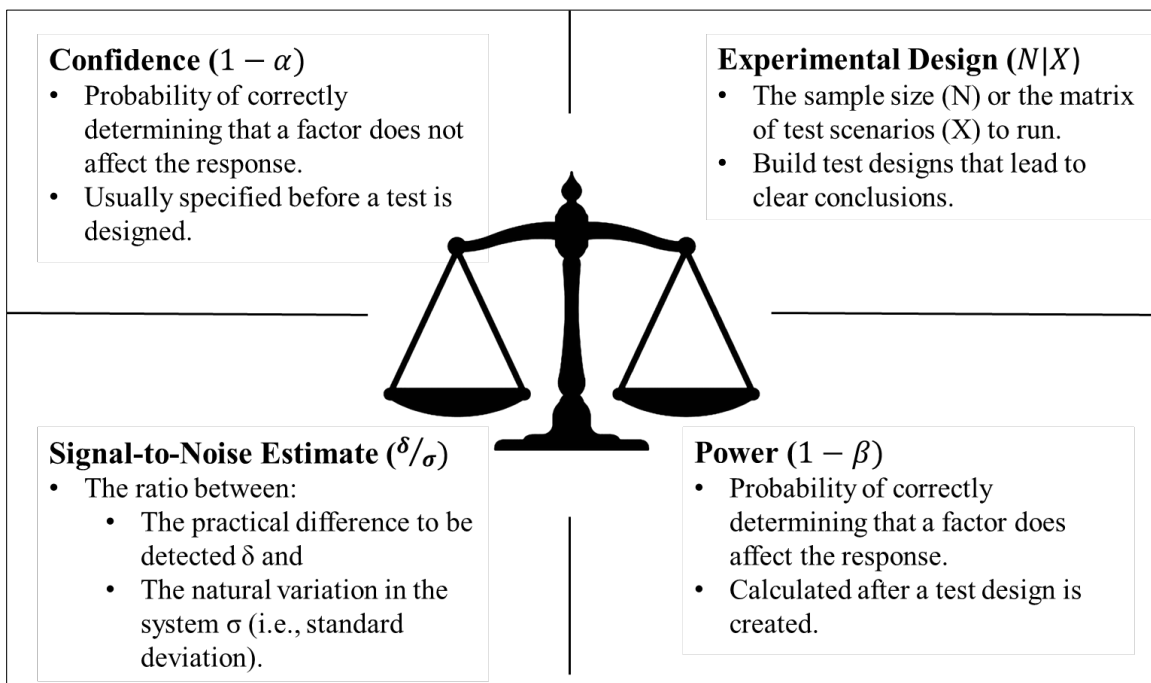
**Step 4: Choose an experiment design that fits the problem.**

Choosing the right test design depends on the objective of the experiment. Table 3-1 shows common objectives for physical experiments that will be used to verify or validate M&S. The table also lists recommended test designs based on the objective. For more information on these test designs, see the book *Design and Analysis of Experiments* (Montgomery 2013) and the MITRE Technical Report, “Modern Test Design and Analysis Playbook” (Cortes and Ortiz 2020).

**Table 3-1. Recommended Experiment Designs for Common Test Objectives**

Objective	Description	Recommended Experimental Designs
Screening	Used to quickly identify the most influential factors on the response(s).	<ul style="list-style-type: none"> <li>• Plackett-Burman</li> <li>• Fractional Factorial</li> <li>• Definitive Screening</li> </ul>
Characterization	Used to help get accurate estimates of the effect that factors and their interactions have on the response(s). Can also be used to build a statistical model for prediction.	<ul style="list-style-type: none"> <li>• Factorial</li> <li>• Response Surface</li> <li>• Optimal</li> </ul>
Optimization	Used to build a statistical model for prediction. This statistical model is then used to find an optimal solution.	<ul style="list-style-type: none"> <li>• Response Surface</li> <li>• Optimal</li> </ul>

Once the appropriate experimental design approach has been chosen, the test design must be properly sized to achieve an acceptable level of rigor. Properly sizing the test design depends on balancing multiple aspects—confidence, power, signal-to-noise estimate, and the number of replications of the experimental design; see Figure 3-3.

**Figure 3-3. Test Design Trade-Off Diagram**

**Confidence ( $1 - \alpha$ )** is the probability of correctly determining that a factor does not influence the response. The level of confidence is chosen based on the level of acceptable risk in conjunction with the other test design trade-offs. Higher levels of confidence may be a better choice in some

situations, for example in a characterization test for a critical system function that will be used for model validation. A PM will need to work with testers to identify the confidence level that balances risk with cost and schedule.

**Experimental design** refers to the chosen data points or the matrix of experimental runs that will be executed and includes the number of replications for each run. The experimental design considers the desired model terms (e.g., main effects, two-factor interactions (2FIs)) to be estimated and any potential constraints in factor space that cannot be examined.

**Signal-to-noise estimate** is the ratio of the change in each response variable ( $\delta$ ) that it is important to detect divided by the estimate of the experimental error ( $\sigma$ ) for each response variable.  $\delta$  represents the smallest operationally meaningful difference in outcomes that the study should be able to detect. Choosing  $\delta$  requires considering the practical effects that need to be observed to change the judgment or assessment of the system under study. For example, a missile might be considered an improvement over a previous variant only if its miss distance is at least 5 meters less on average than the previous variant's miss distance; a decrease of 1 meter, in comparison, is not considered important, so  $\delta = 5$ .  $\sigma$  is the natural variation in the response, a property of the system under study and not under control of the testers. (It may even be determined automatically by the distribution of the response variable.) For example, a missile's miss distance may have a standard deviation of 10 meters. Some of the quantities of interest, such as the natural variation in performance or baseline performance, may need to be estimated from historical data, similar test events' dry runs, repeatability studies, experience, or arguing from first principles. The signal-to-noise ratio (SNR),  $\delta/\sigma$ , may be an input to statistical software for computing power; however, this ratio should be derived by considering  $\delta$  and  $\sigma$  separately, considering what the behavior of the system could be and what operationally meaningful changes in behavior the test should detect (in the example above, the SNR is 0.5). Furthermore, planning analysis and documents should discuss  $\delta$  and  $\sigma$  separately to ground test planning in terms of operationally meaningful performance metrics.

It is important to distinguish between statistical significance and practical significance in model validation. For example, a hypothesis test may find that differences between the model and referent are statistically significant, but those differences are so small that they have no practical implications for system performance. In this case, the model may still be considered valid for its intended use. The bigger the sample sizes of the model and referent, the easier it is to determine whether any differences are statistically significant. Therefore, in addition to considering the signal-to-noise estimate, a program should also consider what difference is of practical consequence, such as with acceptability/accreditation criteria, and tests should be sized to have high power to find differences of that size.

**Power** ( $1 - \beta$ ) is the probability of not committing a Type II error or the probability of correctly determining that a factor influences the response. Power is calculated after the other aspects have been determined. Higher-power designs mean lower risk, but time and resource constraints may mean that lesser-power designs are used, with commensurate increase in the risk of reaching incorrect conclusions regarding system performance.

Once these considerations have been determined, numerous statistical programming packages (e.g., R, JMP, Python) are available to help determine the appropriate test design matrix.

#### 3.1.3 Test Phase

During the Test phase, the experimental plan is implemented, and data is collected. Standard test execution protocols are employed to ensure the integrity of the data.

##### **Step 5: Develop the test design execution strategy.**

Develop a plan to address the uncontrollable and undesirable factors that can influence the results and lead to an invalid conclusion. Replication, randomization, and blocking approaches are employed.

#### 3.1.4 Analyze Phase

During the Analyze phase, prediction models are created to estimate the system's performance capabilities and associated uncertainties and to explain what factors are important (e.g., statistically impact performance) and which are not.

##### **Step 6: Analyze the results.**

Use statistical methods (e.g., analysis of variance (ANOVA), regression; see Section 4) to identify and estimate the effects of significant factors on performance and to answer an array of questions about the system under test and the M&S performance.

##### **Step 7: Formulate valid conclusions.**

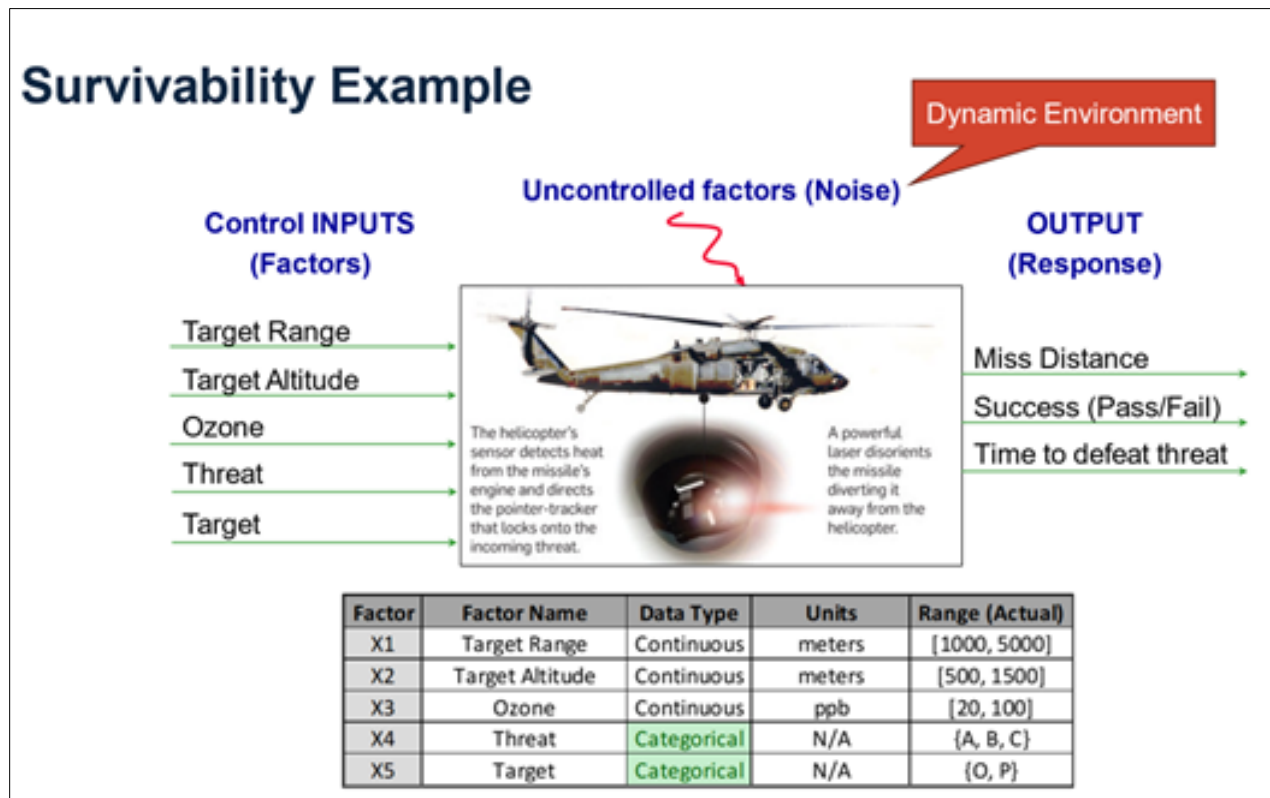
In terms of the test objectives and questions defined in the Plan phase, report on the results and draw insights and implications for further research from the analysis.

#### 3.1.5 DOE Process Example

To further illustrate the DOE process, consider the following survivability testing example of an infrared (IR) countermeasure (CM) system. Rotary-wing aircraft are vulnerable targets for IR missiles. To defeat incoming IR missiles, an aircraft will deploy flares that drop away from itself, providing a more attractive target for the incoming missile and diverting the missile away from

the aircraft. Although flares are effective, they can be deployed only once during a mission. To address this limitation, a nonexpendable solution and IR CM system have been developed. The CM receives a cue from the missile warning system (MWS) and then tracks and acquires the incoming missile by directing laser energy onto the missile, thus causing it to miss the aircraft.

Figure 3-4 shows a notional IPO diagram potentially based on participants from a DOE working group.

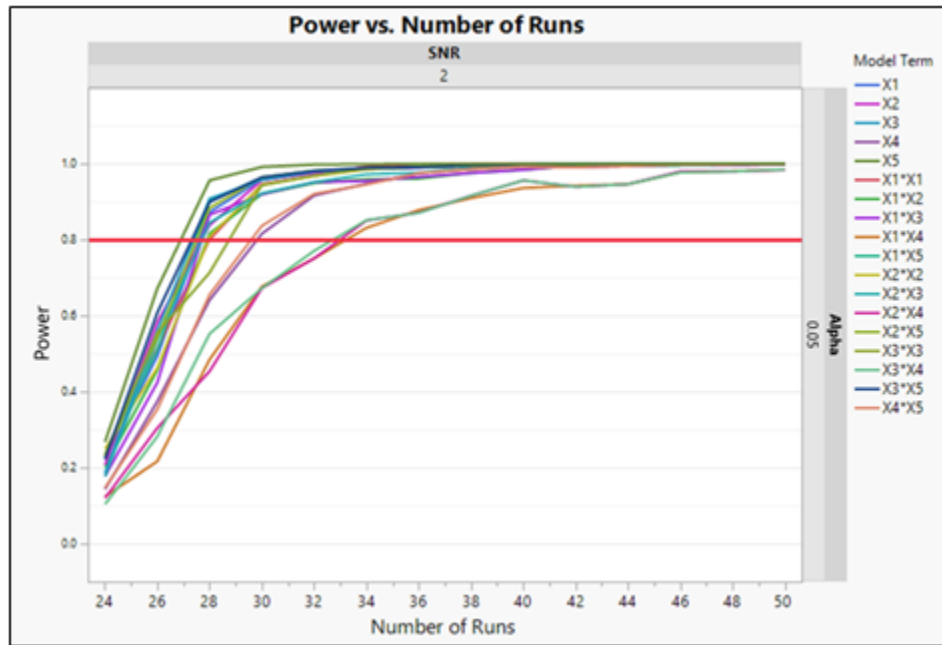


**Figure 3-4. Survivability Testing Example of an Infrared Countermeasure System**

Assuming that the experiment requires two levels for Target Range, Target Altitude, Ozone, and Target, and three levels for Threat, 162 runs would be required to consider all possible combinations. However, using DOE methods and balancing power and confidence number reduces the total number of runs to approximately 34 if 80 percent power is desired for all main effects and first-order interactions (i.e., interactions between the main effects); see Figure 3-5. Similarly, confidence over 90 percent for all terms can be achieved with only 38 total runs<sup>3</sup>. This shows the strength of statistical test design: The program can not only reduce the total number of tests needed to characterize the system but also quantify the risks associated with that reduction.

<sup>3</sup> These calculations are based on effect size that is double the standard deviation (noise). The reasonableness of the effect size assumed is part of the iterative V&V process.

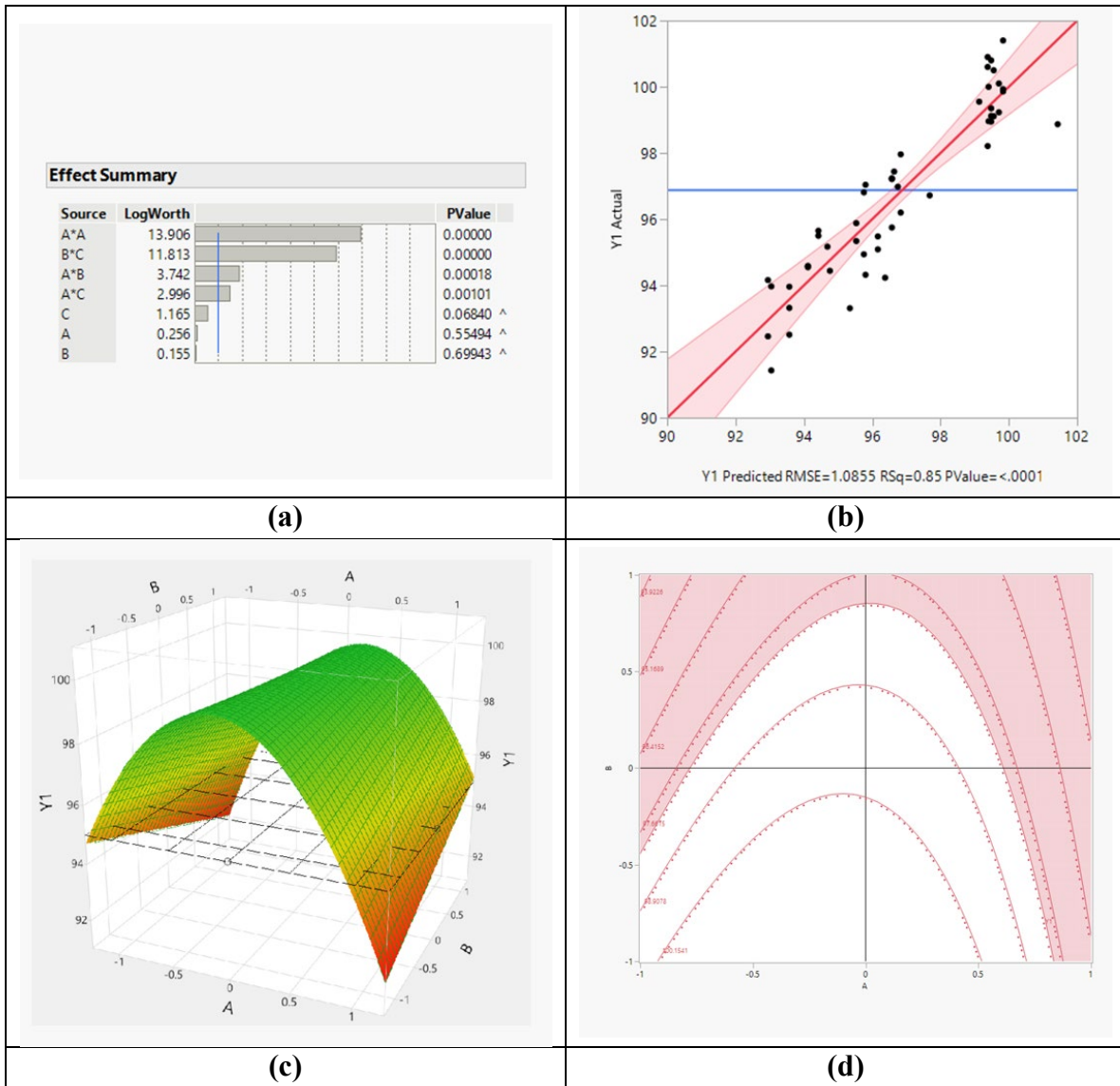




Note: The JEDIS tool used here is available at the IDA Test Science Website (<https://testscience.org>).

**Figure 3-5. Power Calculations for Model Terms (Main Effects and All First-Order Interactions) Versus the Number of Runs for the Test Design**

As shown in Figure 3-6, based on the 34-run test design, analysis can identify what factors and interactions are most important (Figure 3-6a), build a predictive model based on these significant factors (Figure 3-6b), create a response surface that describes performance across the factor space (Figure 3-6c), and then use this model to identify areas of concern via contour plots (Figure 3-6d). This type of analysis aids in comparing simulated results with real-world observations obtained via DOE. Discrepancies between simulated and experimental data may highlight areas for model refinement or improvement.



**Figure 3-6. Notional Example Analysis from the Results of Using a Designed Experiment**

#### 3.1.6 Useful Software and Resources

Common software for constructing designed experiments for physical systems includes JMP (licensed software), Design-Expert (licensed software), and R (free programming language). Some commonly used DOE R packages are skpr, AlgDesign, and FrF2, but there are many more. A comprehensive list can be found on the CRAN Task View: DOE and Analysis of Experimental Data Website.

### 3.2 Design of Computer Experiments

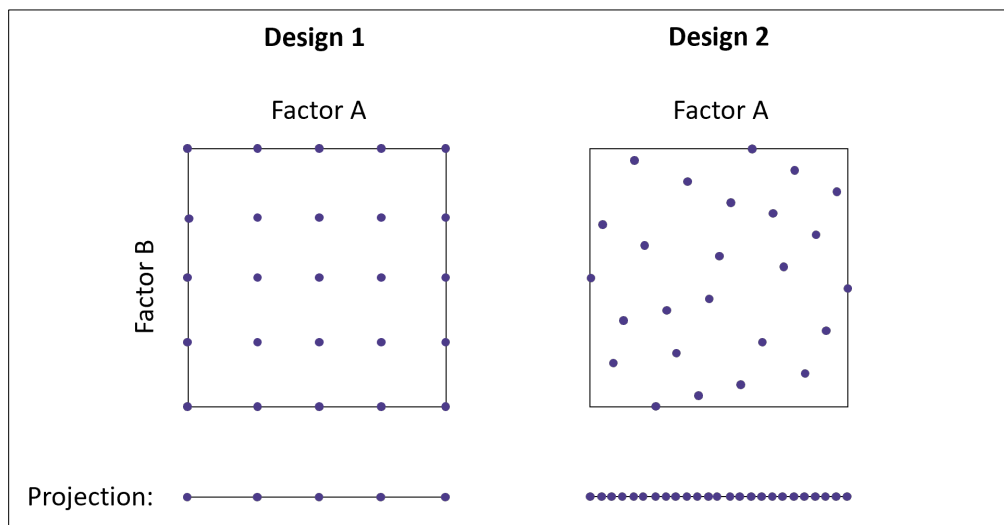
Computer experiments are a class of designs that were developed to efficiently cover a large input space for computer simulations with deterministic (or nearly deterministic) outcomes (e.g.,

they produce the same response for a given set of inputs). Test designs for VV&A of deterministic systems can capitalize on the deterministic nature of the models using highly efficient computer experiments that sparsely cover the space.

M&S is often quick to run, especially compared with the time it would take to run a live experiment. This relatively quick runtime can allow for many more runs to be performed than is typically seen in live test designs. M&S can also feature a complex response surface, for instance, nonlinear response curves due to complex physical phenomena that are not amenable to linear modeling of the M&S responses.

These properties of M&S lead to a few common characteristics of designs for computer experiments used to accomplish M&S VV&A. First, deterministic responses mean no replication of runs is needed, unlike designs for physical experiments. Instead, designs for computer experiments focus on covering the entire factor space of the M&S to obtain responses that allow nonlinear response surfaces to be fully characterized. For this reason, designs for computer experiments are often called space-filling designs.

Space-filling designs do not replicate any points, but they do try to avoid replication of responses should some factors prove to not influence the response. In other words, projecting the design into a smaller factor space should yield a design that also does not contain response replicates. Figure 3-7 pictures two possible designs that both fill the space; however, Design 1 contains replicates when projected, meaning it could prove to be a waste of resources if one factor does not influence the response, whereas Design 2 still contains no replicates upon projection. Designs such as Design 2 are preferred for computer experiments.



**Figure 3-7. Projection of Space-Filling Designs**

### 3.2.1 Purpose of Space-Filling Designs in VV&A

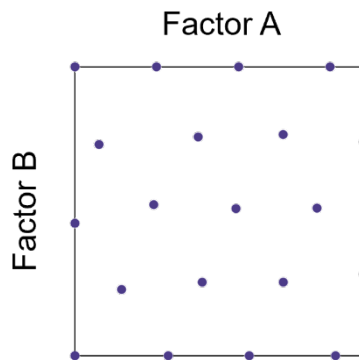
Space-filling designs should be used with specific analysis techniques in mind to satisfy specified goals of VV&A. These techniques are discussed in Section 4 and could include UQ, metamodel building, and sensitivity analysis. These analysis methods could meet a variety of objectives, such as to investigate the effects of a multitude of variables, to characterize system performance, or to compare model and referent data.

### 3.2.2 Space-Filling Designs

Many different methods exist for constructing space-filling designs. The following subsections present a subset of possible design types. The designs presented have good space-filling properties and are likely best suited for VV&A. The trade-offs between these designs will be discussed.

#### 3.2.2.1 Maximin Designs

Maximin designs aim to maximize the minimum distance between any two points in the design. Figure 3-8 shows an example of a maximin design with two design factors. Maximin designs are also known as sphere-packing designs because they effectively fill the space with as many spheres as possible in an optimal arrangement, with design points placed at the center of each sphere. Physical examples of sphere packing include optimally filling a baking sheet with cookies or filling a bucket with golf balls.

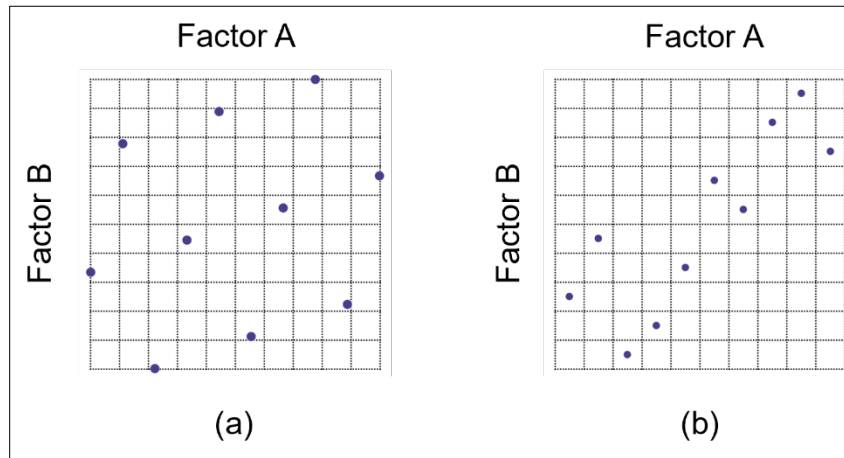


**Figure 3-8. Maximin Design**

Maximin designs are ideal for spreading out points as much as possible, and they cover the edges and corners of the space more than alternative designs. However, as seen in Figure 3-8, maximin designs can tend to follow a more ordered structure, where design points cluster together when projected down to a smaller set of factors. This tendency makes maximin designs less ideal when it is likely that some factors may not have an effect.

### 3.2.2.2 Latin Hypercube Designs

Latin hypercube (LH) designs are constructed to ensure each factor individually is well covered, with even spacing between factor levels. Figure 3-9 shows two different LH designs with 10 points. To construct a design with 10 points, a grid is created with 10 rows and 10 columns. Then points are assigned to grid cells so that each row and each column contain exactly one point. Within each cell, the design point may be centered, placed at random, or placed at an optimized location (Cioppa and Lucas 2007).



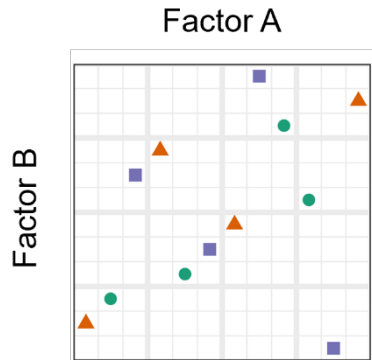
**Figure 3-9. Latin Hypercube Designs: (a) Maximin Optimized and (b) Highly Correlated Design**

Many possible designs satisfy the basic properties of an LH design, but not all are considered to have good space-filling properties. For example, in Figure 3-9, design (b) has one design point in each row and each column, but it is missing large areas of the design space, and the factors are highly correlated, meaning the design only tends to test high Factor A values when Factor B is also high. Therefore, instead of randomly generating a single LH design, it is advantageous to adopt some optimality criterion—for example, maximin—to generate a design that satisfies LH properties and has generally good space-filling properties (Damblin et al. 2013). Design (a) in Figure 3-9 uses a maximin criterion to choose design points, meaning points are placed as far apart from each other as possible while maintaining LH properties.

### 3.2.2.3 Sliced Latin Hypercube Designs

A major limitation of the maximin and LH designs discussed thus far is that they cannot be applied when the M&S contains categorical factors because categorical factors have no “distance” that can be calculated and cannot be subdivided into intervals. Sliced LH designs adapt the LH to accommodate categorical factors. Figure 3-10 shows an example of a sliced LH design with three different levels of a categorical factor (green circle, orange triangle, and purple square). The design is created such that each categorical “slice” is an LH design, and all slices overlaid are also an LH design. Figure 3-10 shows a major 4x4 grid, where each column and row

contain one point for each categorical factor, and an overall 12x12 grid, where each column and row contain one point. This structure ensures that points will not be replicated should the categorical factor not impact the response.

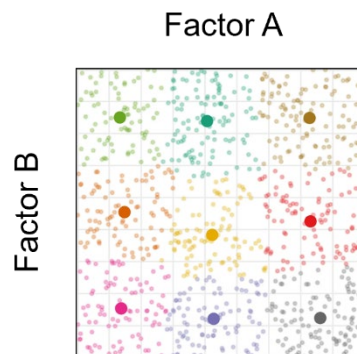


Source: Adapted from IDA Document NS D-21562 (Wojton et al. 2021)

**Figure 3-10. Sliced Latin Hypercube Design**

#### 3.2.2.4 Fast Flexible Filling Designs

Fast flexible filling (FFF) designs are another common design type chosen when categorical factors are present. The design is generated by randomly generating a large number of points, randomly distributed across factor levels, and then clustering those points into a number of clusters that equals the specified number of runs. Typically, design points are placed at the centroid of each cluster. Figure 3-11 shows an FFF design with nine points generated by clustering many points in a two-factor continuous space.



Source: Adapted from IDA Document NS D-21562 (Wojton et al. 2021)

**Figure 3-11. Fast Flexible Filling Design Generated from Clustering**

When categorical factors are present, design points are balanced across the total number of combinations of categorical factor levels (see Fast Flexible Filling Design Details on the JMP Statistical Discovery Website). Another advantage of FFF designs is that they can be constructed when there are constraints on which factor levels or combinations can contain design points. For

example, if designing a test for a system with a concept of operations that specifies that the system will not be used at high speeds and low altitudes, a constraint (e.g., linear inequality) can be defined so that no design points will be created for the disallowed combinations of conditions.

### 3.2.2.5 MaxPro Designs

Space-filling designs may optimize some mathematical function defining what a “good” space-filling design should do. The minimum distance between points in the design (optimized by maximin designs, which maximize this minimum distance) is a simple example of such a mathematical criterion, and the IDA Document, “Space Filling Designs for Modeling & Simulation Validation” (Wojton et al. 2021), includes many such mathematical criteria that a design may optimize. The MaxPro (meaning “maximum projection”) criterion function encourages spreading out points just as the minimum distance between points does, but it also rewards designs that retain good spread between points even if factors were removed; the minimum distance between design points lacks this feature. Additionally, the criterion handles categorical factors well.

JMP provides FFF designs that optimize the MaxPro criterion. The IDA Paper, “Metamodeling Techniques for Verification and Validation of Modeling and Simulation Data” (Haman and Miller 2022), discusses how sliced LH designs can use the MaxPro criterion for moderate to large numbers of categorical factors. Analysts should consider using the MaxPro criterion because it generates designs for robustness against dropping factors, spreads points throughout the design space, and can handle the presence of many categorical factors.

### 3.2.2.6 Design Comparison

Table 3-2 summarizes the five design types discussed above. These designs have generally good properties, but many other designs can be used beyond those discussed in this guidebook. The IDA Document, “Space Filling Designs for Modeling & Simulation Validation” (Wojton et al. 2021) discusses a wider array of design options and additional evaluation criteria.

**Table 3-2. Comparison of Five Space-Filling Design Types**

Design Type	Description	Handles Categorical Factors?	Handles Design Region Constraints?
Maximin	Maximizes distance between pairs of points.	No	No
Latin Hypercube (Maximin)	Maximizes distance between points while maintaining even spacing between factor levels.	No	No



Design Type	Description	Handles Categorical Factors?	Handles Design Region Constraints?
Sliced Latin Hypercube	Maintains even spacing between factor levels; contains no replicates projected across categorical factors.	Yes	No
Fast Flexible Filling	Design points are chosen by clustering many points generated across the design region.	Yes	Yes
MaxPro	Spreads points throughout the design space while also accounting for the possibility of dropping factors.	Yes	No

### 3.2.3 Determining Sample Size

Although M&S is typically not as resource restrained as physical experiments, it is important to carefully choose the sample size to ensure that enough information can be gained without wasting resources. The sample size should adequately cover the complexity of the space. Often the complexity of the space is not well understood, however, because of things like nested statements in software. A simple rule of thumb is that a space-filling design should contain at least 10 times the number of factors or variables being tested (Loeppky et al. 2009). This usually allows for sufficient modeling with techniques such as GP modeling (see Section 4.4.5). If the response surface is expected to be complex (the model outputs change rapidly with input changes), the number of points per factor should be increased to be able to model a more complex curve. If categorical factors are present, this rule of thumb may underestimate the number of runs required. Instead, choose a sample size of 10 times the number of continuous factors times the number of categorical combinations (Wojton et al. 2021). Sample size planning for stochastic M&S should include other methods for calculating sample size (including power analyses or sample size needed to obtain a desired margin of error). All these rules are conditional on correctly understanding the complexity of the space in continuous and categorical factors, so they should be thought of as a starting point that needs to be evaluated as part of the V&V process.

### 3.2.4 Useful Software and Resources

Software for constructing computer experiments includes JMP (licensed software) and R (free programming language). Of the designs discussed, JMP can build maximin (called sphere packing in JMP), maximin LH, and FFF designs. JMP also has built-in tools for modeling space-filling designs. In R, several packages are available that can be used to generate space-filling designs: “MaxPro,” “lhs” (Latin hypercube samples), “maximin,” “SLHD” (sliced Latin hypercube designs), and “DiceDesign” (maximin and LH). Additional packages are available for modeling.



For a more detailed discussion of design construction, evaluation criteria, and additional design types, see the IDA Document, “Space Filling Designs for Modeling & Simulation Validation” (Wojton et al. 2021).

### 3.3 Methods for Comparing Data from Physical Experiments and M&S

As discussed in Section 2 of this guidebook, VV&A of M&S entails combining and comparing data from live testing (physical experiments) and M&S outputs to determine whether the M&S provides a representation of system behavior valid for the purposes of T&E. Statistically based methods of design and analysis, as described previously in Section 3, exist for determining the M&S outputs needed (e.g., the space-filling designs described in Section 3.2.2), determining the live test data needed (e.g., DOE, Section 3.1), and analyzing whether the differences between the two sets of data are acceptable. This section of the guidebook elaborates on several steps in the overall nine-step approach to VV&A cited in Section 2.1 (Wojton et al. 2019) and the seven-step DOE process (Section 3.1) to illustrate how data from live testing and M&S can be combined and compared. It should also be noted that in addition to comparisons with live test data, M&S validation can and should use information from other referents for comparison, as discussed in Section 4 of this guidebook.

Identify response variables. These are the outputs of the M&S and results of live testing of interest for assessing the behavior of the model and performance of the system. In this case, the chosen measures should be sufficient to satisfy *both purposes* (assessing model behavior and assessing system performance), as well as sufficient to provide the basis for comparing live test results and M&S outputs. There will likely be overlap and uniqueness between and among the two sets of responses. Subject matter expertise is used to select the responses.

Determine the factors that will affect the response variables. These are the conditions under which the system will be used (such as threat characteristics, weather) that will affect its performance and will be present in live testing, as well as the inputs that will affect the behavior of M&S outputs. The factors chosen should, again, be sufficient for *both purposes* (effects on performance obtained during live testing and model outputs). There will be overlaps and unique entries among and between the two sets of factors (e.g., both the results of live testing and M&S outputs will be affected by threat characteristics). Subject matter expertise is used to select the factors.

Determine the acceptability criteria. These are the differences between each response variable common to the live testing and M&S outputs that are acceptable for using the M&S to evaluate the system’s performance (e.g., using the M&S results to conclude that the system’s performance is inadequate). These criteria can be response specific and will be determined using subject matter expertise, performance of other analogous systems, and outputs of other M&S.

Design the M&S experiment. Depending on the level of randomness in the model, as shown in Table 3-3, use the methods described in Sections 3.1 or 3.2 to design the M&S experiment to cover the M&S factors and to generate test points that can be matched to live test data.

Design the live test. Using the approach described in Section 3.1, design the live test to cover the live testing factors, ensuring that there are live test points that match outputs generated from the M&S experiment as much as is feasible. Use estimates of prediction variance and/or statistical power for a detectable difference consistent with the acceptability criteria to determine the sufficiency of the data to be obtained from live testing.

Use statistical methods to determine whether differences between live test data and M&S outputs are acceptable. Use statistical methods and models, such as those described in Section 4 of this guidebook, to “estimate the uncertainty of each of the response variables as a function of the appropriate factors” (Wojton et al. 2019). When any uncertainty exceeds the acceptability criterion for a given response variable, apply subject matter expertise to judge whether the exceedance is significant for the intended use of the M&S.

Given sufficient resources and time, the steps described above can be repeated iteratively until the desired model fidelity is achieved as M&S development and system development proceed, model complexity increases, and system complexity and maturity increase.

An example of a hybrid design approach combining M&S experimental design with DOE for live testing along the lines of the approach discussed above is described in Chapter 4, Sections C and D of the IDA Handbook (Wojton et al. 2019).

**Table 3-3. Simulation Design Recommendations**

Level of Randomness	Recommended Method by Validation Goal	
	Compare to Live Data	Explore Model Space
<b>None</b> (Deterministic)	Hybrid Design	Space Filling
<b>Low</b> (e.g., physics-based with calibration factors)	Classical	Hybrid Design
<b>High</b> (e.g., effects-based, human-in-the-loop)	Classical with Replications	Classical with Replications

Source: IDA Handbook (Wojton et al. 2019)

## 4 Analysis

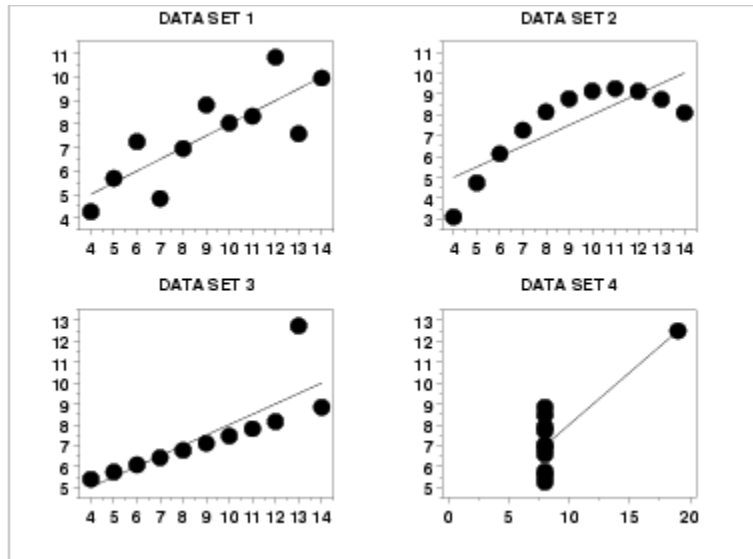
Section 4 provides guidance on the application of statistical techniques to data analysis within the context of M&S for T&E. This section outlines best practices for analyzing M&S results, highlighting the role of statistical methods in validating models and ensuring their practical significance in T&E. It aims to equip testers, systems engineers, analysts, and PMs with the necessary tools to analyze M&S outputs effectively, focusing on VV&A processes and using statistical methods to communicate M&S results. By applying statistical techniques, stakeholders can quantify uncertainties, compare M&S results with live test data, and make informed decisions based on comprehensive data analysis.

### 4.1 Exploratory Data Analysis

Exploratory data analysis (EDA) is a data analysis philosophy used most often to discover new information about a data set. Whether it is a data set about which nothing is known or one to which a few other analysis methods have been applied, EDA is used to discover the unknown. The first step in performing EDA is to formulate a question that needs to be answered about the data: Are there patterns that are not obvious? Are patterns that are obvious not quite what they seem?

It is important to think of EDA not as a technique or set of methods to use; it is a mindset, a way of thinking about data. Data analysis techniques such as regression and inferential statistics have prescribed inputs, processes, and outputs. EDA serves in contrast to these methods. There are no prescribed techniques, only those that analysts think might reveal information about what is not otherwise already known. EDA allows the analysts to let the data determine the appropriate techniques to utilize. Graphical representations are often employed in EDA as they are efficient to create and easy to interpret. Scatter plots, box plots, and histograms are only a few of the graphical analysis techniques that can shed light on data.

Figure 4-1 is a simple example from the National Institute of Standards and Technology (NIST) Web-based Engineering Statistics Handbook (<https://www.itl.nist.gov/div898/handbook/>) that shows how EDA is used to interpret a set of data analyses.



Source: NIST Engineering Statistics Handbook

**Figure 4-1. Anscombe's Quartet: A Set of Four Data Sets with Identical Means, Variance, R-Squared, Correlations, and Linear Regression Lines**

Each of the data sets indicates a unique behavior between the two variables. If any of the possible data points were filtered, the actual behavior of the entire data set might not be observed.

EDA allows outliers to be identified, such as in Data Sets 3 and 4, which should be investigated further. If the outliers are caused by data collection or data entry errors, it may be possible to repeat the collection or fix the entries; if neither option is feasible, then the data should be removed before proceeding with further analysis. However, if the outliers are valid data, then they may indicate real phenomena that should be modeled. EDA allows informed decisions to be made about whether to exclude data before proceeding with modeling. Note that even when outliers are removed from a modeling analysis, they are still part of the data set and should be included in the overall analysis in support of the V&V.

The NIST Engineering Statistics Handbook further recommends the following basic set of four techniques to conduct a new EDA. Although these techniques are recommended, they are not mandatory. Simplified examples of how to use these techniques are provided in subsequent subsections of this guidebook.

- Run-sequence plot
- Lag plot
- Histogram
- Normal probability plot

## 4. Analysis

This set of four plots provides a substantial amount of information pertaining to a data set, able to cover a wide variety of questions and discovery. According to NIST, these four plots are excellent for testing the four assumptions of randomness, fixed distribution, fixed location, and fixed variation. For example, if the run-sequence plot has a vertical spread that is the same over the entire plot, then the fixed variation assumption holds. If the lag plot is structureless, then the randomness assumption holds. If the histogram is bell-shaped, the underlying distribution is symmetric. If the normal probability plot is linear, then the underlying distribution is approximately normal. NIST provides a detailed discussion for the proper understanding of data when testing indicates that any of the four assumptions are invalid.

As mentioned earlier in this section, the first step in EDA is to formulate the set of questions that need to be answered. Once the set of questions is chosen, they must be ranked by order of importance. Some analytic techniques are meant for specific purposes, whereas others are best suited to answer many general questions. The following sections of the NIST Engineering Statistics Handbook provide additional information: Section 1.3.3. provides a variety of graphical analysis methods, with descriptions as well as instructions for their use; Section 1.3.5. provides quantitative analysis methods; Section 1.3.6. explains probability distribution methods; and Section 1.4.2. provides in-depth examples about how these various techniques can be used with EDA in mind.

### 4.1.1 Run-Sequence Plot

A run-sequence plot is a plot of the data against the order it was observed. The run-sequence plot can be constructed via a scatter plot, with the order of the observations intact, as shown in Figure 4-2.



**Figure 4-2. Run-Sequence Plot Showing an Overall Trend in Observations Along With Two Outliers**

The plot shows a gradual decline in the observations, with two seeming outliers. These two points are similar in that they are roughly the same number and somewhat follow the same declining behavior as the rest of the data set. Based on this run-sequence plot, it might be decided that these two points are not outliers but instead phenomena that should be investigated.

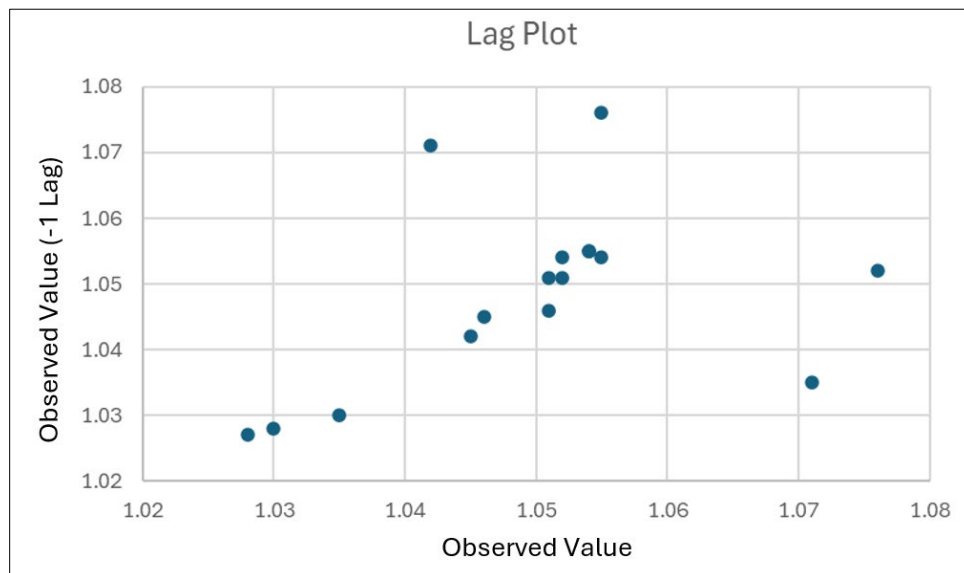
#### 4.1.2 Lag Plot

Similar to a run-sequence plot, a lag plot examines the structure of data. In a lag plot, the observations are plotted on the y-axis against the indexed observations based on the chosen amount of lag. For example, the data set in Table 4-1 pairs each observation with the next observation in the set.

**Table 4-1. Data Set for Lag Plot Example**

Data	1.052	1.054	1.055	1.054	1.055	1.076	1.052	1.051	1.051	1.046	1.045	1.042	1.071	1.035	1.03	1.028	1.027
-1 Lag	1.054	1.055	1.054	1.055	1.076	1.052	1.051	1.051	1.046	1.045	1.042	1.071	1.035	1.03	1.028	1.027	0

With this data set, a scatter plot can be created to show a unique perspective; see Figure 4-3.

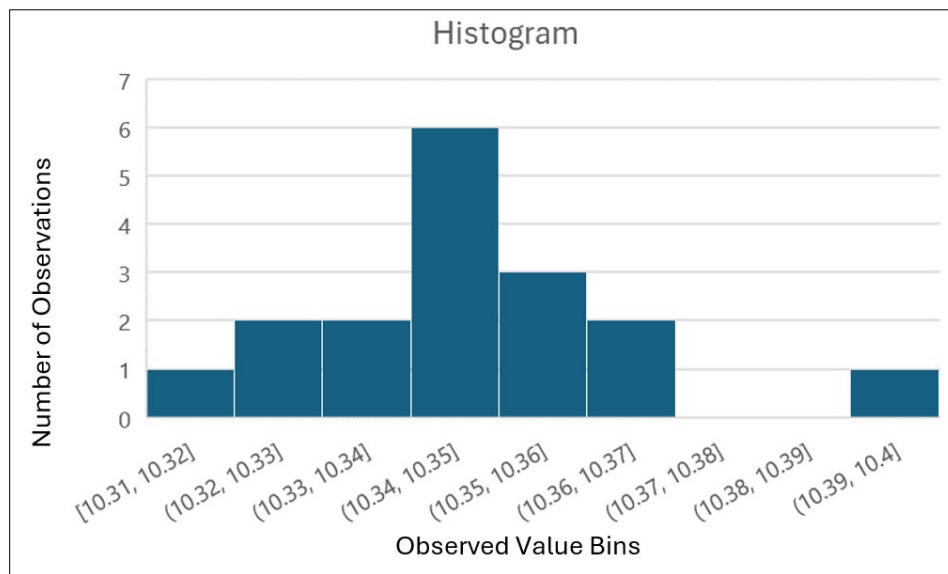


**Figure 4-3. Lag Plot of Notional Data Showing a Linear Relationship, Indicating Autocorrelation**

Lag plots show autocorrelation. If the data were truly random, a high variability, or scatter, in the lag plot would be expected. What is seen in the example is more of a linear relationship that indicates a low amount of randomness.

### 4.1.3 Histogram

The purpose of the histogram is to analyze the distribution of a data set. The example in Figure 4-4 is a data set for observed transmission times of a network. Plotting the observations shows that most of the transmissions are arriving at or near the theoretical value of 10.35 seconds. There is moderate variability between 10.33 and 10.37 seconds, with a seeming outlier at 10.4 seconds. Based on an understanding of the different types of distributions, this data set could be classified as normally distributed. There are many other studied types of distributions, including binomial, uniform, Poisson, and Weibull. The shape of the distribution can provide a wealth of information, including potential probability and cumulative distribution functions (CDFs).



**Figure 4-4: Histogram Showing Distribution of Notional Data**

### 4.1.4 Normal Probability Plot

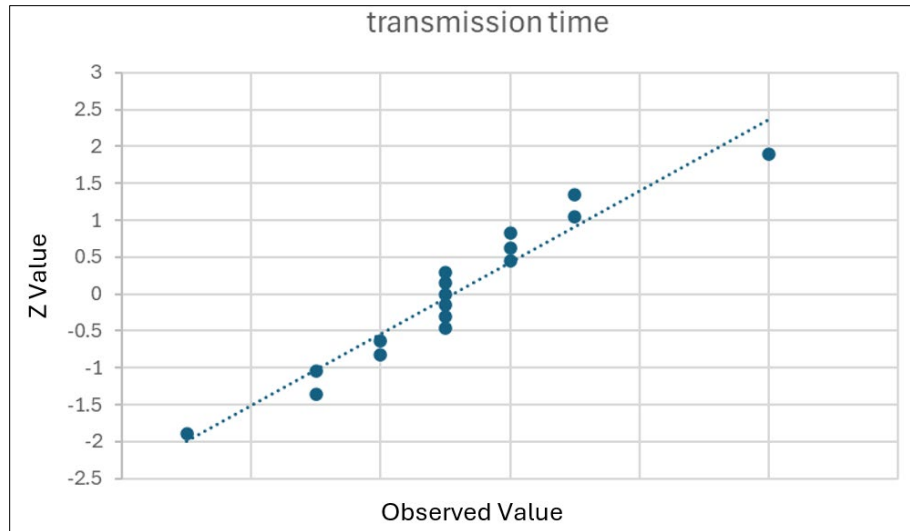
The normal probability plot is a technique used to determine whether a data set is normally distributed. There are many different methods to create a normal probability plot; the example below is one method. Taking the network transmission times from the previous example, the first step is to reorder the raw values in ascending order, as shown in the second column of Table 4-2. Then each value is ranked (the Number column); the cumulative probability of each value is calculated; and the corresponding z-value for each probability is determined. The cumulative probabilities are calculated as  $(i - 0.5)/n$ , where  $i$  is the rank of the value (Number column) and  $n$  is the total number of observations (17 in this case). For this example, the z-values were determined via the NORMSINV function within Microsoft Excel, which was applied to the Cumulative Probability column.

**Table 4-2. Data Table for Normal Probability Plot Example**

Transmission Time	Ordered	Number	Cumulative Probability	Z-Value
10.31	10.31	1	0.029411765	-1.88950996
10.35	10.33	2	0.088235294	-1.35170224
10.34	10.33	3	0.147058824	-1.049131398
10.35	10.34	4	0.205882353	-0.820792088
10.4	10.34	5	0.264705882	-0.628904218
10.35	10.35	6	0.323529412	-0.457851931
10.33	10.35	7	0.382352941	-0.29930691
10.34	10.35	8	0.441176471	-0.14798711
10.35	10.35	9	0.5	0
10.35	10.35	10	0.558823529	0.14798711
10.36	10.35	11	0.617647059	0.29930691
10.37	10.36	12	0.676470588	0.457851931
10.36	10.36	13	0.735294118	0.628904218
10.35	10.36	14	0.794117647	0.820792088
10.36	10.37	15	0.852941176	1.049131398
10.33	10.37	16	0.911764706	1.35170224
10.37	10.4	17	0.970588235	1.88950996

With the data table complete, the results can be plotted on a scatter plot; see Figure 4-5. The x-axis should be the set of ordered data values (column 2 in this example) and the y-axis the z-values. A trend line is added to better expose the desired effect. Ideally, this plot would show a perfect alignment of values along the trendline, or a perfect linear relationship. With only 17 values, this becomes a bit difficult, but the linear behavior can be observed. Additional data points would smooth out the appearance.





**Figure 4-5. Normal Probability Plot**

This method can be applied to other types of distributions as well. No matter the type of distribution, the plot shows the linearity of the ordered values to the statistical expectation.

## 4.2 Statistical Methods for Comparing Physical and M&S Data

Statistical methods provide a rigorous framework for comparing data from physical experiments with outputs from simulations. These methods enable analysts to assess the validity and reliability of models, ensuring that they accurately represent real-world scenarios. By quantifying differences and uncertainties, statistical techniques help in identifying discrepancies and guiding improvements in model fidelity. This process not only enhances the credibility of M&S but also supports informed decision making, ultimately leading to more effective and efficient system development and evaluation.

### 4.2.1 Basics of Hypothesis Testing

Hypothesis tests can be used to determine whether the data support a hypothesis, for example, “the model and the referent data do not have the same mean.” When creating the hypothesis, a program should consider what difference is of practical consequence.

Hypothesis tests always have at least two hypotheses: the null hypothesis and an alternative hypothesis. The specific null/alternative hypotheses will vary slightly depending on the test applied. When applied to model validation, however, the hypotheses typically follow the structure below.

Null Hypothesis ( $H_0$ ): The model and the referent agree.

Alternative Hypothesis ( $H_1$ ): The model and the referent disagree.

The null hypothesis is what is assumed to be true, and the alternative hypothesis can be supported by collecting and analyzing data. The outcome of a hypothesis test is to either reject the null hypothesis and conclude that the alternative hypothesis is true or fail to reject the null hypothesis due to lack of evidence. The decision of whether to reject the null is based on the comparison of a test statistic calculated from data to a critical value based on the significance level ( $\alpha$ ), or equivalently of the calculated p-value to the significance level. The p-value is the probability of obtaining data at least as extreme as the observed data, given the null hypothesis is true, and the significance level is the acceptable rate of rejecting the null hypothesis if it were true (commonly set to 0.05). If the p-value is smaller than the significance level, the result of the hypothesis test is statistically significant.

If the model and the referent agree, the model can be considered valid, and if the model and referent disagree, the model cannot be considered valid. A key thing to highlight is that using the hypotheses above, it is impossible to prove that the model and referent agree; this is because the null hypothesis is assumed to be true. Rather, the conclusion would be that the model and referent data do not indicate evidence of disagreement. The hypotheses above can be used to prove that the model and referent disagree if disagreement is indicated by the data.

Some types of hypothesis tests swap the null and alternative hypotheses listed above, enabling the test to prove agreement between a model and referent. These are called equivalence tests and are described further in the STAT Center of Excellence (COE) Report, “Equivalence Testing” (Ramert and Westphal 2020).

### 4.2.2 Two-Sample T-Test

The two-sample t-test is a simple hypothesis test that can be used to determine whether a statistically significant difference exists between the model mean,  $\mu_m$ , and referent mean,  $\mu_r$ . This test assumes that the data drawn from the model and referent are normally distributed. The hypotheses are below. This setup is commonly called a two-sided hypothesis test because the model mean could be higher or lower than the referent mean, indicating that the model and referent disagree.

Null Hypothesis ( $H_0$ ):  $\mu_m = \mu_r$

Alternative Hypothesis ( $H_1$ ):  $\mu_m \neq \mu_r$

## 4. Analysis

Evidence for disagreement between the model and referent means is captured by the test statistic  $T$  given in Equation (4-1), where  $\bar{x}_m$  is the model sample mean,  $\bar{x}_r$  is the referent sample mean,  $s_m^2$  is the model sample variance,  $s_r^2$  is the referent sample variance,  $N_m$  is the model sample size, and  $N_r$  is the referent sample size.

$$T = \frac{\bar{x}_m - \bar{x}_r}{\sqrt{s_m^2/N_m + s_r^2/N_r}} \quad (4-1)$$

To determine whether the evidence is sufficient to prove the alternative hypothesis at the chosen significance level  $\alpha$ , a critical value must be calculated. In this case, the critical value is computed using the Student's t-distribution, which gives the probability of observing different test statistics assuming that the model and referent have the same mean. If there is a low probability (a low p-value) of observing a test statistic as extreme or more extreme than the one calculated using Equation (4-1), there is sufficient evidence to conclude that the model and referent means are different. The critical value,  $t_{1-\alpha/2, \nu}$ , for a two-sided hypothesis test is calculated at a  $1 - \alpha/2$  confidence level using software or lookup tables for a t-distribution with degrees of freedom  $\nu$  given in Equation (4-2).

$$\nu = \frac{(s_m^2/N_m + s_r^2/N_r)^2}{(s_m^2/N_m)^2/(N_m - 1) + (s_r^2/N_r)^2/(N_r - 1)} \quad (4-2)$$

If  $|T| > t_{1-\alpha/2, \nu}$ , then there is sufficient evidence to reject the null hypothesis, resulting in the conclusion that the model and referent means are not the same and the model is not valid. Otherwise, the conclusion would be that the data do not indicate evidence of disagreement between the means. The t-test can be implemented in R using the `t.test` function.

The disadvantage of a t-test is that it only accounts for differences in the mean rather than differences between distributions in general. Additionally, it only applies to a single set of factor conditions and cannot account for factor changes. Section 4.2.3 discusses the ANOVA hypothesis test, which accounts for multiple factors, and Section 4.2.4 discusses goodness-of-fit tests, which compare two distributions.

### 4.2.3 Analysis of Variance

ANOVA is a hypothesis test that underlies statistical regression. ANOVA can be used on a model or referent data alone when creating a linear regression model to understand which factors influence outputs. However, ANOVA can also be applied to a combined regression model built

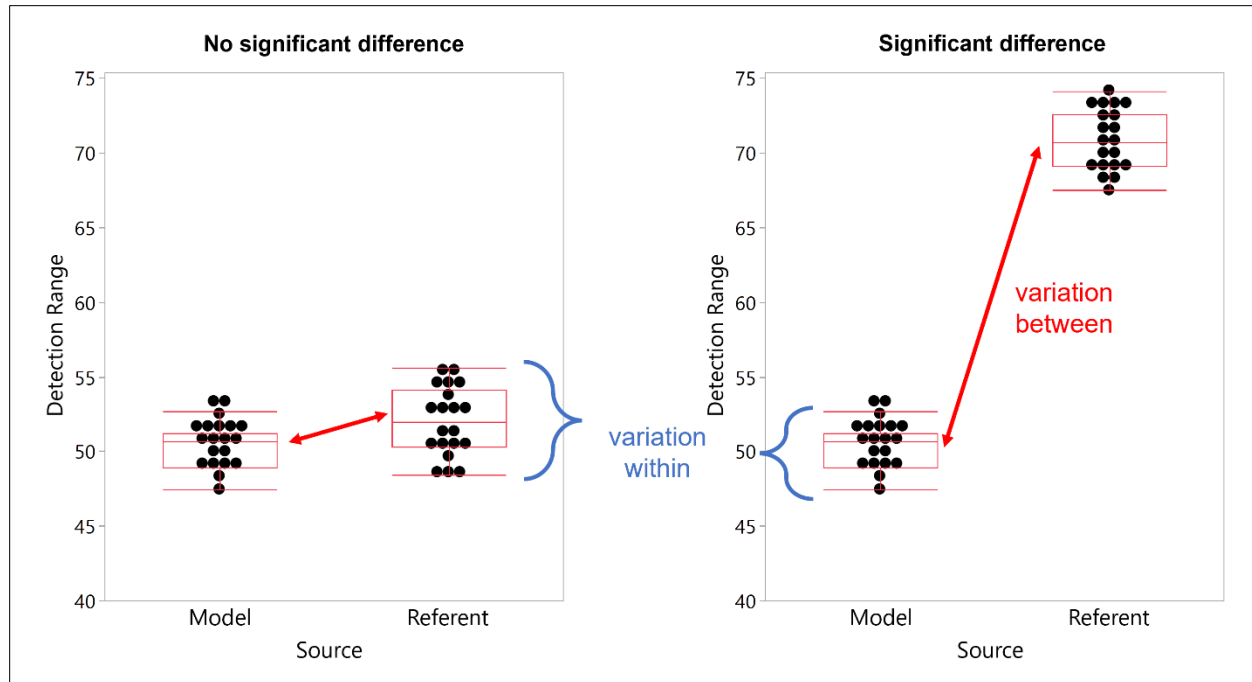
from both model and referent data to find statistically significant differences between model and referent data across a range of factors.

ANOVA requires several assumptions: independent observations, constant variance of error terms, and normally distributed error terms. For a discussion of checking model assumptions, see the STAT COE Report, “Model Building Process Part 1: Checking Model Assumptions V 1.1” (Burke 2017). Equation (4-3) gives the general form for a linear regression model with main effects (single factor effects) and interaction effects between factors, where  $y$  is the response,  $\beta_0$  is intercept,  $k$  is the number of factors,  $\beta_j$  is the main effect model parameter for the  $j^{\text{th}}$  factor,  $\beta_{ij}$  is the model parameter for the interaction between factor  $i$  and factor  $j$ ,  $x_i$  is the  $i^{\text{th}}$  independent variables or factors, and  $\varepsilon$  represents the unknown error. Linear regression models are flexible and may also contain quadratic terms (e.g.,  $x_1^2$ ), or even higher-order terms (e.g.,  $x_1^2 x_2$ ).

$$y = \beta_0 + \sum_{j=1}^k \beta_j x_j + \sum_{j=1}^k \sum_{i < j} \beta_{ij} x_i x_j + \varepsilon \quad (4-3)$$

For model validation, in addition to operationally relevant factors, the data source (model/referent) can be included as a factor in the model, as well as interactions with the source factor. For each term included in the model, the null hypothesis is that the term does not have an effect on the response, and the alternative is that the term does have an effect on the response.

To determine statistical significance, an F statistic is calculated from the ratio of variation between groups to the variation within groups. This is illustrated conceptually for a single factor case in Figure 4-6. The figure shows notional data observed for radar detection range from two sources, a model and referent, for two different cases, with boxplots showing the distributions of each data set. In the left graph, the ratio of variation between the model and referent groups to the variation within the groups is small, and there is substantial overlap between the two groups, so the evidence does not indicate a significant difference. In the right graph, however, the ratio of variation between the model and referent groups to the variation within the groups is large, and the two groups are well separated, so the source term is considered statistically significant. Therefore, the model could be considered valid in the left case but not in the right case.



**Figure 4-6. ANOVA Illustration for a Single Factor**

The concept shown in Figure 4-6 can be extended for multiple model terms, allowing the difference between the model and referent to be assessed across a range of factors. For the mathematical details for performing ANOVA, see the STAT COE Report, “Understanding Analysis of Variance: Best Practice” (Natoli 2017).

The statistical software JMP has easy-to-use tools for performing regression and ANOVA. The R programming language can also be used.

A key consideration when performing regression/ANOVA is the correlation of design factors. To correctly attribute an effect to a factor, factors should have zero or low correlation. Using designed experiments (see Section 3.1) is an effective way to reduce correlation between factors. Correlation can also be introduced when combining unbalanced data sets from a model and referents. Section 4.4.1 discusses strategies for combating issues with unbalanced data sets.

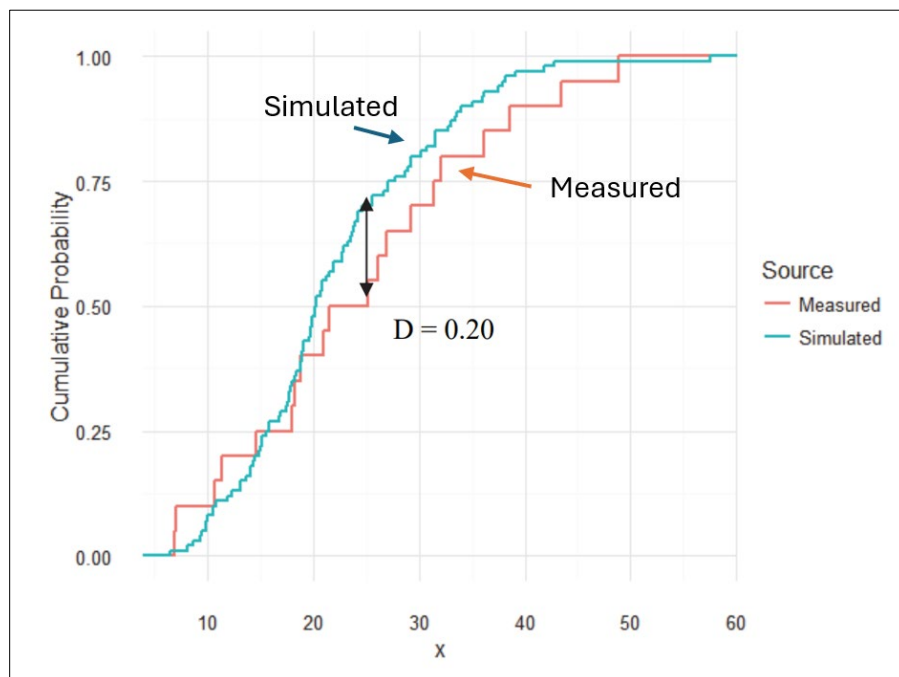
#### 4.2.4 Goodness-of-Fit Tests

Goodness-of-fit tests compare two samples to each other to determine whether they came from the same distribution. When applied to model validation, goodness-of-fit tests compare model data to referent data to determine whether they have the same distribution, establishing evidence that the model agrees with referent data. Unlike ANOVA, goodness-of-fit tests typically apply to a unique combination of factors (or when no factors are present) rather than applying over a

range of many factors. Section 4.2.5 discusses how these analysis techniques can be extended across multiple factor settings.

#### 4.2.4.1 Nonparametric Kolmogorov-Smirnov Test

The nonparametric or two-sample Kolmogorov-Smirnov (KS) test can be used to compare model and referent data with continuous responses at a unique combination of factors to determine whether the two data sets come from the same or different underlying, unknown distributions. The model and referent data are compared using their empirical CDF. Figure 4-7 shows examples of CDFs for simulated and referent (measured) data. The empirical CDF plots data points in a sample against their percentile. For example, 50 percent of the sample data is above or below the median value of a sample, indicated by where the CDFs reach 0.5 cumulative probability in Figure 4-7. Typically, simulated data sets have a higher number of samples due to the relative ease and speed of simulating data. This results in the smoother CDF seen for the simulated data shown in Figure 4-7.



Source: IDA Handbook (Wojton et al. 2019)

**Figure 4-7. Cumulative Probability Functions for Measured and Simulated Data**

Figure 4-7 visualizes how the test statistic  $D$  is calculated, which is used to determine whether there is sufficient evidence that the model and referent disagree. The formula for calculating  $D_{m,n}$  is given in Equation (4-4), where  $m$  is the number of model samples,  $n$  is the number of referent samples,  $F_{\text{model}}(x)$  is the CDF of the model,  $F_{\text{referent}}(x)$  is the CDF of the referent, and  $x$  is the response being compared. This equation simply states that  $D$  is equal to the maximum difference between the model and referent CDFs for any value of  $x$ .

$$D_{m,n} = \max_x |F_{\text{model}}(x) - F_{\text{referent}}(x)| \quad (4-4)$$

The test statistic  $D_{m,n}$  is compared to a critical value  $D_\alpha$ , where  $\alpha$  is the significance level. If  $D_{m,n} > D_\alpha$ , the difference between the model and referent is statistically significant, meaning the data indicate that there is sufficient evidence that the model and referent disagree, and the null hypothesis is rejected. Equation (4-5) gives the equation to calculate  $D_\alpha$  for large sample sizes, where the value of  $c(\alpha)$  is given by Table 4-3 for commonly chosen significance levels.

$$D_\alpha = c(\alpha) \sqrt{\frac{m+n}{m * n}} \quad (4-5)$$

**Table 4-3. Values of  $c(\alpha)$  for Different Significance Levels**

$\alpha$	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001
$c(\alpha)$	1.073	1.138	1.224	1.358	1.48	1.628	1.731	1.949

The KS test is excellent for comparing entire continuous distributions rather than just two means (as in a t-test). However, the KS test can be used to demonstrate model validity only at a single input combination at a time and requires a higher number of data points in both model and referent data to establish the shape of their distributions. The KS test can be implemented in R using the `ks.test` function, which can be used to calculate the p-value (compared against the significance level to determine statistical significance).

#### 4.2.4.2 Fisher's Exact Test

Fisher's exact test can be used to compare model and referent data that have a categorical response (e.g., pass/fail, hit/miss) at a unique combination of factors (or when no factors are present). Fisher's exact test is similar to the chi-squared test but is preferred for smaller sample sizes, such as those often encountered in DoD. Model and referent data with a binary response can be summarized in a contingency table such as in Table 4-4.

Table 4-4. Example of a Contingency Table

	Successes	Failures	Row Sums
Model	$a = 8$	$b = 14$	$a + b = 22$
Referent	$c = 1$	$d = 3$	$c + d = 4$
Column Sums	$a + c = 9$	$b + d = 17$	$n = 26$

Source: IDA Handbook (Wojton et al. 2019)

Table 4-4 shows that the model and referent both had low but slightly different rates of success ( $8/14 = 0.37$  for the model and  $1/3 = 0.33$  for the referent). Fisher's exact test can be used to determine whether any calculated difference in the rate of success is statistically significant: in other words, if there is enough evidence that the model and referent distributions disagree. To perform the hypothesis test, assume that the model and referent are equally likely to succeed, then calculate the probability of observing the data *or data more extreme* conditional on the marginal totals in Table 4-4 (conditional on total success/failures and data counts for model and referent). If the probability is small, there is a low chance of obtaining the observed data, meaning it is unlikely that the model and referent have the same probability of success. This probability of the observed data is calculated using Equation (4-6), where  $a$ ,  $b$ ,  $c$ ,  $d$ , and  $n$  are the counts drawn from Table 4-4, assuming fixed margins.

$$p = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!} \quad (4-6)$$

For Table 4-4,  $p = 0.4094$ . However, to calculate the probability of observing the data *or data more extreme*, the conditional probability of observing other possible  $a$ ,  $b$ ,  $c$ , and  $d$  values that have the marginal totals must also be calculated, as shown below in the same arrangement as Table 4-4.

9	13
0	4

$$p = 0.1592$$

7	15
2	2

$$p = 0.3275$$

6	16
3	1

$$p = 0.0955$$

5	17
4	0

$$p = 0.0084$$

The probabilities calculated above indicate that the observed data (Table 4-4) are not improbable compared to other arrangements that would give the same marginal totals. To calculate the p-value for the two-sided significance test (to be compared to significance level  $\alpha$ ), all probabilities *smaller than or equal to* the probability calculated for the observed data are summed with Equation (4-6) ( $\leq 0.4094$ ). For this example, the p-value equals  $0.4094 + 0.1592 + 0.3275 + 0.0955 + 0.0084 = 1$ . For a significant level of  $\alpha = 0.05$ , it can be



concluded that the difference between the model and referent is not statistically significant because  $1 > 0.05$ , and therefore it can be concluded that the evidence does not indicate the model is invalid.

Fisher's exact test can be implemented in R using the `fisher.test` function.

#### 4.2.5 Fisher's Combined Probability Test

Fisher's combined probability test is a "meta-analysis" technique that can be used to combine the results of several independent hypothesis tests with the same hypotheses. Fisher's combined probability test can be used to combine results from tests such as the t-test, KS test, and Fisher's exact test discussed in Sections 4.2.2, 4.2.4.1, and 4.2.4.2, respectively. For example, a goodness-of-fit test could be conducted individually for different unique factor combinations, then combined using a Fisher's combined probability test to determine whether there is evidence of model and referent disagreement overall across the operational space.

The test statistic  $\chi^2_{2k}$  for Fisher's combined probability test can be calculated using Equation (4-7), where  $k$  is the total number of hypothesis tests and  $p_i$  is the p-value for the  $i^{\text{th}}$  hypothesis test.

$$\chi^2_{2k} = -2 \sum_{i=1}^k \ln p_i \quad (4-7)$$

This test statistic follows a chi-squared distribution with  $2k$  degrees of freedom, which can be used to calculate either the critical value for comparison to the test statistic or the p-value for the overall hypothesis test.

If the overall p-value is small and indicates that the model and referent do not agree, investigate which conditions had small individual p-values. It is possible that the model predictions are invalid under some conditions rather than invalid across the entire space. Regression techniques are powerful for understanding these factor effects and should also be used when validating across an operational space.

#### 4.2.6 Summary of Statistical Analysis Methods for Comparing Physical and M&S Data

The hypothesis tests most appropriate for analyzing the live test and M&S data are determined by the data and their characteristics as described in previous sections. Examples of appropriate hypothesis tests and other statistical methods include the following:

#### 4. Analysis

- The t-test (or log t-test for transforming skewed data) is used to determine whether the means of two sets of data (or transformed data) are statistically different/independent.
- Fisher's combined probability test is used to combine the results of multiple difference/independence tests associated with the same overall null hypothesis.
- Fisher's exact test is used to statistically analyze contingency tables comprising categorical data.
- The nonparametric KS test is used to assess whether two sets of data arise from the same distribution.
- Regression models are used to assess what factors affect responses, with the factors/terms in the model composing the sets of hypotheses. Lognormal regression is performed on the logarithm of skewed, nonnegative responses.
- Emulation and prediction compare live test data as they become available to the prediction interval (PI) of a metamodel of the M&S (see Sections **Error! Reference source not found.** and **Error! Reference source not found.**).

Furthermore, it may be necessary to apply several methods to fully understand and characterize any given set of data. "There is no one-size-fits-all solution" (Wojton et al. 2019). Table 4-5 summarizes which analysis methods are appropriate based on the data characteristics and sample size.

**Table 4-5. Statistical Analysis Methods According to Distribution of Responses and Amount of Live Test Data**

Distribution	Factors	Recommended Method by Sample Size		
		Small	Medium	Large
Skewed (Lognormal)	Univariate	Fisher's Combined	Log t-test Fisher's Combined Non-parametric K-S	Log t-test Fisher's Combined Non-parametric K-S
	Distributed	Log t-test Fisher's Combined Non-parametric K-S	Non-parametric K-S	Non-parametric K-S
	Designed Experiment	Log-Normal Regression Emulation & Prediction	Log-Normal Regression Emulation & Prediction	Log-Normal Regression Emulation & Prediction
Symmetric (Normal)	Univariate	Fisher's Combined	t-test Fisher's Combined Non-parametric K-S	t-test Fisher's Combined Non-parametric K-S
	Distributed	t-test Fisher's Combined Non-parametric K-S	Non-parametric K-S	Non-parametric K-S
	Designed Experiment	Regression Emulation & Prediction	Regression Emulation & Prediction	Regression Emulation & Prediction
Binary	Univariate	Fisher's Exact	Fisher's Exact	Fisher's Exact
	Distributed	Logistic Regression	Logistic Regression	Logistic Regression
	Designed Experiment	Logistic Regression	Logistic Regression	Logistic Regression

Source: IDA Handbook (Wojton et al. 2019)

### 4.3 Methods for Uncertainty Quantification

As stated in Section 2.2.1, there are generally two broad categories of uncertainty: statistical and knowledge (Wojton et al. 2019).<sup>4</sup> Statistical uncertainty is due to inherently random effects; can be better characterized, but not reduced, by accumulating more samples (i.e., more replicates); and is generally characterized by a probability distribution. An example of statistical uncertainty is measurement error. In contrast, knowledge uncertainty is due to a lack of information and thus can be reduced by collecting data, which provides new information, but generally cannot be

<sup>4</sup> Aleatory and epistemic are commonly used terms for statistical and knowledge uncertainty, respectively.

characterized by a probability distribution.<sup>5</sup> An example of knowledge uncertainty is the unknown performance capabilities of a foreign weapon system due to limited intelligence data.

Intervals are a key way to quantify statistical uncertainty around an estimate, and different types of intervals quantify different kinds of uncertainty. Statistical intervals quantify statistical uncertainty, and there are several types of statistical intervals (Ortiz and Truett 2015):

- A **confidence interval**<sup>6</sup> (CI) is calculated from sample data, and it provides a range of values where the true population parameter likely resides. An example of a question that a CI can answer is “What is the expected performance of my system at a specific condition?”
- A **prediction interval** (PI) is an estimated range of values in which future observations will fall at a specified level given what has already been observed. An example of a question that a PI can answer is “What is the predicted performance of my system at a specific condition?” PIs are wider than CIs but often more useful when using M&S to predict performance of the real-world system.
- A **tolerance interval** (TI) is a range of values where X percent (specified by the user) of the population should fall. An example of a question that a TI can answer is “Will 99 percent of my observations fall under the threshold specification at least 95 percent of the time?” Although TIs are not widely used in DoD T&E, they are arguably more appropriate than CIs for many requirements.

In addition to choosing which of the three types of intervals is appropriate, a choice must also be made between a one-sided or two-sided interval. A one-sided interval provides either an upper or lower bound on the estimate, and it is appropriate when the quantity of interest must be either above or below a single threshold. A two-sided interval, on the other hand, brackets the estimate with both an upper and lower bound.

There are many ways to calculate the intervals described above, and choosing the appropriate method depends on the characteristics of the data. Therefore, EDA (see Section 4.1) is a crucial prerequisite. For example, there are closed-form solutions for all three types of intervals for normally distributed data. Prediction and tolerance intervals, however, are highly sensitive to the

---

<sup>5</sup> Knowledge uncertainty is typically represented as an interval with no associated probability density function (PDF). However, it may be represented as a PDF that reflects the subject matter expert’s degree of belief (Cortes et al. 2021). Methods for eliciting subjective probability functions are outside the scope of this guidebook; an introduction to the topic can be found in the book *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis* (Morgan and Henrion 1990).

<sup>6</sup> The credible interval is the Bayesian counterpart to the frequentist confidence interval. For an introduction to Bayesian statistics, see the book *Statistical Rethinking: A Bayesian Course with Examples in R and Stan* (McElreath 2020).

normality assumption. If the normality assumption is not satisfied, then it may be possible to transform the response such that the transformed response is normal. It is hard to interpret intervals, however, for transformed variables (Cortes and Ortiz 2020). Interactive tools for calculating CIs for variables that follow an exponential, binomial, or lognormal distribution can be found on the IDA Test Science Website (<https://testscience.org/interactive-tools>).

Alternatively, distribution-free methods should be considered; see the book *An Introduction to the Bootstrap* (Efron and Tibshirani 1994) for more information on distribution-free methods.

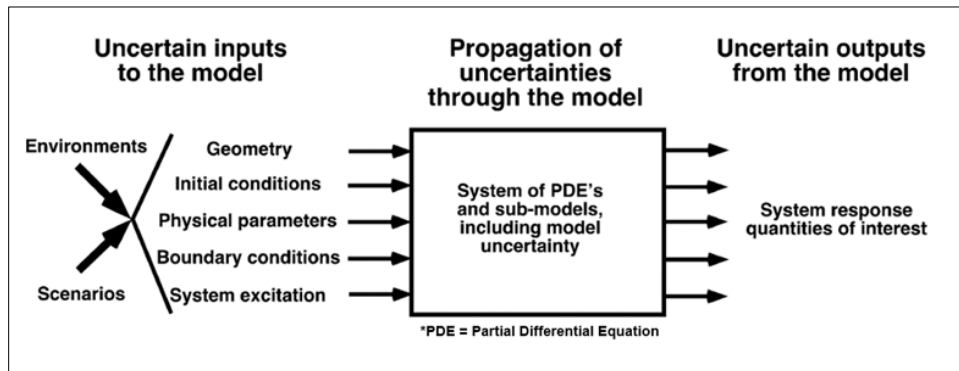
There are many sources of uncertainty in M&S, such as model inputs, simplifying model assumptions, and numerical errors (e.g., round-off error) (Roy and Oberkampf 2011). Thus, M&S can have both statistical and knowledge uncertainties. Instead of relying solely on point estimates (e.g., the average miss distance was 10 feet), decision makers should expect M&S V&V results and model predictions to also include information about the uncertainty around point estimates (e.g., the average miss distance was 10 feet with a 95 percent CI of [7 feet, 13 feet]). Uncertainty intervals are a convenient way to quantitatively convey that information. See Section **Error! Reference source not found.** for more information about when to use different types of uncertainty intervals.

Roy and Oberkampf (2011) developed a comprehensive framework for UQ in scientific computing, which is summarized in their article, “A Comprehensive Framework for Verification, Validation, and Uncertainty Quantification in Scientific Computing.” As shown in Figure 4-8, the first step in the UQ process is to assess sources of uncertainty. This entails identifying and quantifying all potential sources of uncertainty in the model. Each source of uncertainty should be categorized as statistical, knowledge, or mixed uncertainty. Then a mathematical structure can be assigned (e.g., probability function, interval), and numerical values can be determined for each source of uncertainty. Consider each of the following categories when identifying sources of uncertainty:

- **Model inputs:** This category includes parameters used in the model and data from the description of the surroundings. Examples include model parameters, initial conditions, and boundary conditions.
- **Numerical approximation:** Most scientific models use approximate numerical solutions. Examples include round-off error, discretization error, and convergence error.
- **Model form:** This is due to approximate or imprecise representation of the underlying physical, biological, economic, or social processes (Smith 2013). Examples are problem dependent.

The number of potential sources of uncertainty can be very large, especially for complex models. Inputs that have little to no uncertainty or that have little to no effect on the variability of *all*

responses of interest may be treated as deterministic. This determination, however, must be justified and based on strong evidence such as a sensitivity analysis.



Source: (Roy and Oberkampf 2011)

**Figure 4-8. Propagation of Input Uncertainties to Obtain Output Uncertainties**

The second step is to propagate uncertainties through the model. Statistical and knowledge uncertainties should be treated independently because they characterize different types of uncertainty. Monte Carlo sampling is often the simplest approach. For inputs with statistical uncertainty, this entails randomly choosing a number between 0 and 1 and then using the inverse CDF to choose the sample for the input parameter. For inputs with knowledge uncertainty, there is not an associated probability distribution function but rather a range of possible values. In most practical applications, there will be a mix of both types of uncertainties. In these circumstances, a nested sampling approach is recommended. For each sample from the knowledge uncertainties, compute and propagate the statistical uncertainties as previously described. Repeat across the intervals of the knowledge uncertainties. The result will be an ensemble of CDFs. This ensemble is then used to construct a probability box (i.e., p-box) to express both the statistical and knowledge uncertainty without confounding them, as illustrated in Figure 4-9. There are two ways to interpret the p-box. First, for a given probability value, there is a predicted interval-value range of the response. Likewise, for a given response value, there is a range of probabilities for which the value will occur. In other words, there is no single value of probability that describes the uncertainty given the present state of knowledge.

Note that the Monte Carlo sampling approach requires a large number of samples to characterize the uncertainty, especially for low-probability events. Time may be a limiting factor if the simulation takes a very long time to run; in these situations, alternative methods should be considered. Although beyond the scope of this guidebook, some alternatives include polynomial chaos expansion (Xiu 2010), stochastic collocation (Xiu 2015), low-rank tensor approximations (Konakli and Sudret 2016) for statistical uncertainty and interval analysis (Jones et al. 1998), and Dempster-Shafer evidence theory (Li et al. 2010) for knowledge uncertainty.

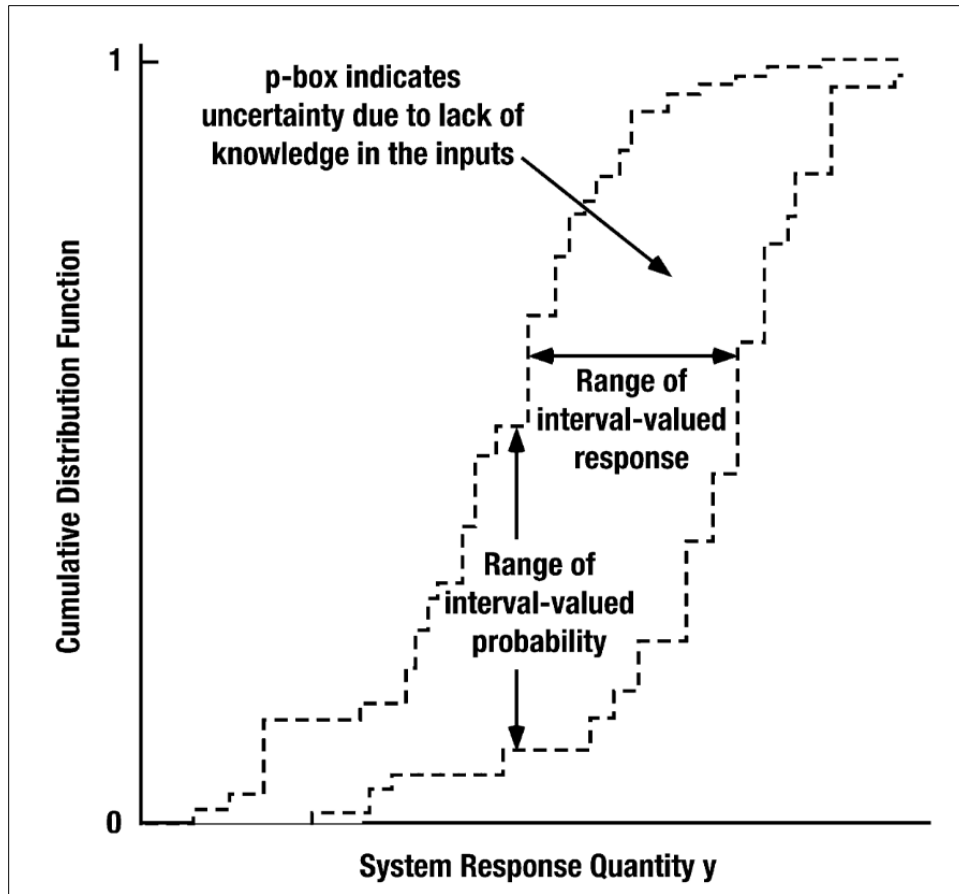
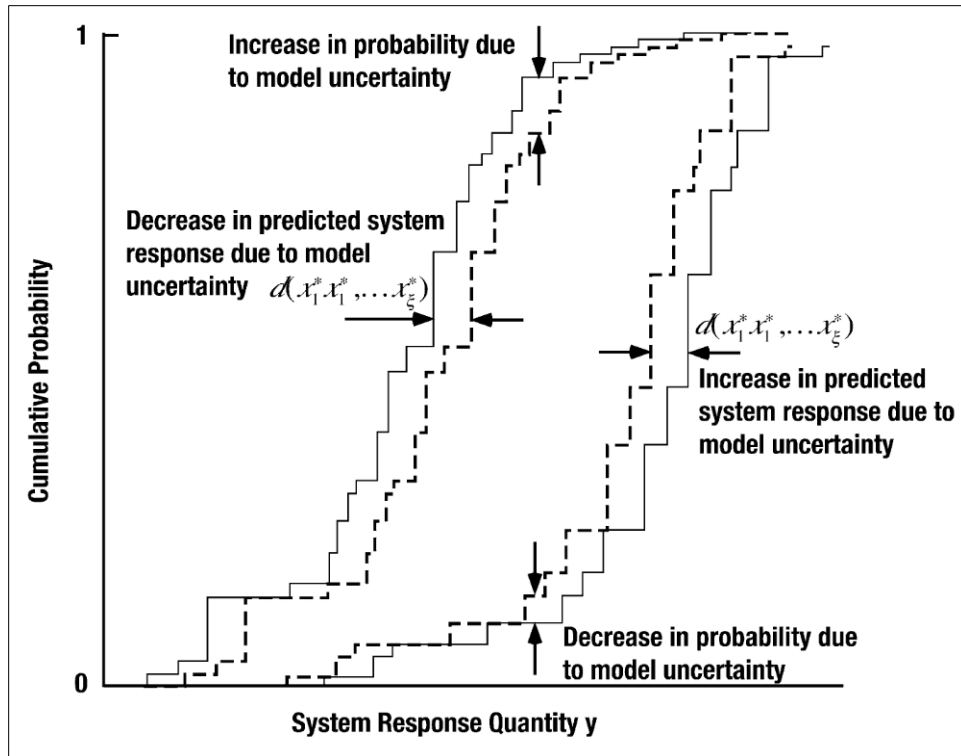


Figure 4-9. Example P-Box

The final step is to determine the total uncertainty in the model outputs. Start with the p-box that was generated from propagating the model input parameters. Next, append the model form uncertainty from the validation metric<sup>7</sup> to the sides of the p-box, which captures the increased knowledge uncertainty of the response; see Figure 4-10. Numerical approximation errors are also treated as knowledge uncertainties, which are appended to the sides of the p-box. The result is a p-box that explicitly accounts for and segregates uncertainties due to model inputs, numerical approximations, and model form errors.

<sup>7</sup> In this example, the validation metric is the smallest area between the p-box and the experimental CDF.



Source: (Roy and Oberkampf 2011)

**Figure 4-10. Increase in Predictive Uncertainty Due to the Addition of Model Form Uncertainty**

UQ does not indicate whether a model’s predictions are “right” or “true.” Instead, UQ indicates that if the validity of a model is accepted for an intended use, then the validity of the conclusions drawn from the model must be accepted up to the degree suggested by the UQ analysis. UQ plays 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.

#### 4.3.1 Useful Software and Resources

The Sandia National Laboratories Dakota project (<https://dakota.sandia.gov/>) provides software for UQ with sampling, reliability, stochastic expansion, and epistemic methods. Dakota versions 5.0 and newer are available under a GNU Lesser General Public License. Extensive user resources—including manuals, videos, and runnable examples—are provided on the Dakota Project Website.

#### 4.3.2 Example

The following example is taken from the STAT COE Report, “Using Statistical Intervals to Assess System Performance” (Ortiz and Truett 2015) and illustrates the statistical intervals previously described. Note that all data presented in this example are notional and used for demonstrative purposes only.



#### 4. Analysis

This case study uses a generic example of a designed experiment applied to assess an MWS as shown in Figure 4-11. An MWS works in conjunction with a CM tracker to defeat guided seeker threats to aircraft. The MWS acts as a cueing system by detecting, declaring, and eventually handing off a potential threat to the CM tracker. The goal of the analysis is to assess various performance measures and help make a determination on the suitability of the MWS. This example focuses on the “time to handoff” performance measure, which has a threshold requirement to be less than 500 milliseconds.



Source: ITT Defense

**Figure 4-11. Missile Warning System**

MWS handoff capabilities and timelines vary according to threat type, engagement slant range, atmospheric conditions, clutter level, and platform flight profile.

For simplicity, the designed experiment will consider only one threat type and will vary the following factors at a high and low level (+1, -1 in coded units, respectively):

- Altitude
- Range

#### 4. Analysis

- Aircraft Speed
- Clutter

The following  $2^4$  design (with 6 center points) shown in Table 4 was created and executed for the MWS. The performance measure of interest (i.e., response) is time to handoff and is shown in the last column of Table 4.

**Table 4-6.  $2^4$  Design for MWS Test**

Run	A:Altitude	B:Range	C:Aircraft Speed	D:Clutter	Time to Handoff (ms)
1	-1	-1	-1	Low	352.63
2	1	-1	-1	Low	386.31
3	-1	1	-1	Low	385.61
4	1	1	-1	Low	518.39
5	-1	-1	1	Low	326.29
6	1	-1	1	Low	375.43
7	-1	1	1	Low	358.07
8	1	1	1	Low	489.84
9	-1	-1	-1	High	394.13
10	1	-1	-1	High	391.74
11	-1	1	-1	High	431.25
12	1	1	-1	High	499.41
13	-1	-1	1	High	373.54
14	1	-1	1	High	367.18
15	-1	1	1	High	422.40
16	1	1	1	High	485.20
17	0	0	0	Low	397.37
18	0	0	0	High	415.79
19	0	0	0	Low	402.17
20	0	0	0	High	412.67
21	0	0	0	Low	401.35
22	0	0	0	High	417.09

The MWS program wishes to demonstrate that the time to handoff will not exceed 500 milliseconds throughout the operational region as defined by the factors and levels. A more statistically precise statement would be that the program wants to show, with 95 percent confidence, that the probability of success ( $P_s$ ) is at least 99 percent. That is,

$$Ps(\text{time to handoff} < 500\text{ms}) \geq 0.99$$

at any point within the design space. Run number 4 (highlighted in red in Table 4) already demonstrates that the MWS can exceed 500 milliseconds under certain conditions.

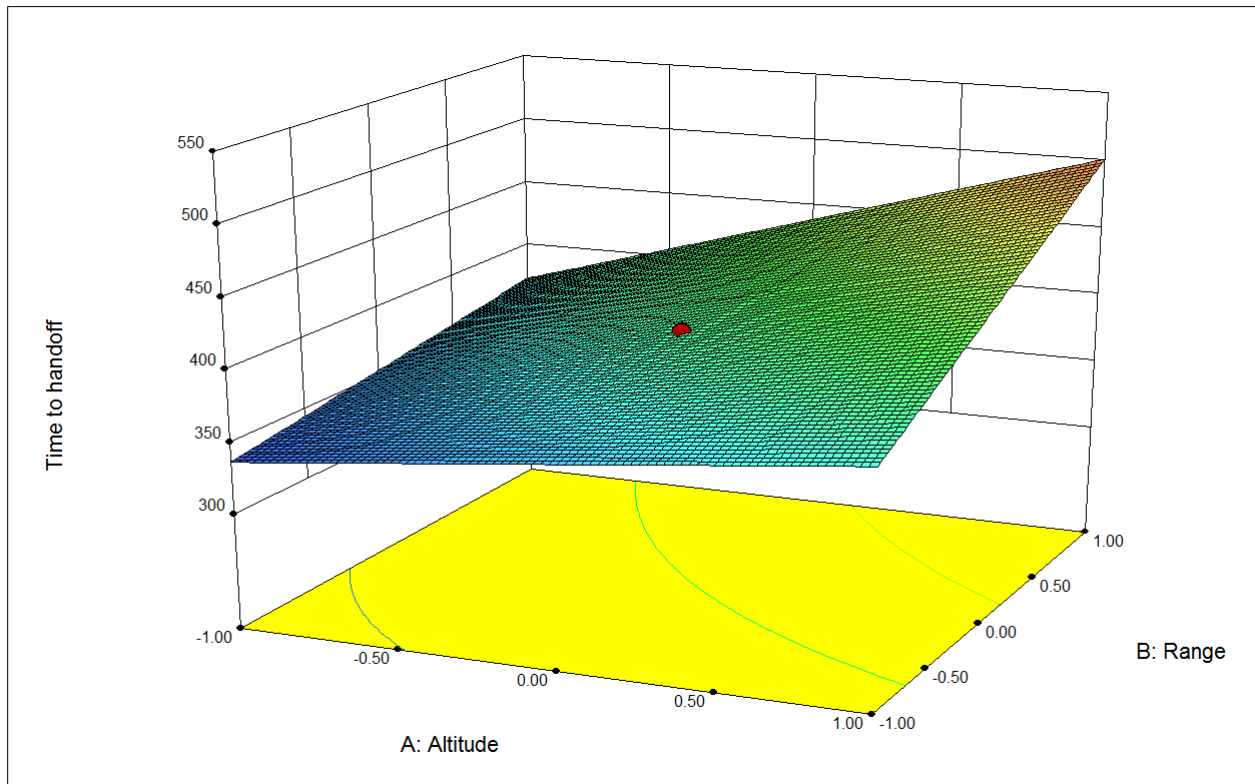
The following regression model was created based on the collected data and can be used to predict time-to-handoff performance (in milliseconds) across the design space:

#### 4. Analysis

$$\text{Time to handoff} = 409.27 + 29.35A + 38.93B - 10.09C + 9.86D + 20.09AB - 14.07AD$$

where A is altitude, B is range, C is aircraft speed, and D is clutter.

The regression model is represented graphically in Figure 4-12 with aircraft speed and clutter held constant. This surface plot clearly shows the relationships between the response (time to handoff) and two of its input factors (altitude and range). The figure shows that as range and altitude increase, the MWS takes longer to hand off. The model allows for interpolation within the design space, thus allowing for prediction of untested scenarios.



Source: STAT COE-Report-04-2015 (Ortiz and Truett 2015)

**Figure 4-12. 3-D Graphical Representation of the Regression Model Developed Using DOE for MWS Example**

For this MWS case study, the parameter of interest is the mean time to handoff. The basic form for a statistical interval for the mean is as follows:

$$\bar{y} \pm c_{(level,n)} * s$$

where

- $\bar{y}$  is the sample mean.
- $s$  is the standard error.

## 4. Analysis

- $n$  is the sample size.
- $c_{(level,n)}$  is a critical value that changes depending on the interval type and the specified confidence level.

### 4.3.2.1 Confidence Interval

For the MWS example, suppose the null and alternate hypotheses are as follows:

$$H_0: \mu_{time} \geq 500$$

$$H_1: \mu_{time} < 500$$

Note that  $\mu_{time}$  represents the population mean time to handoff. In this example, a one-sided interval is constructed with the pre-specified 95 percent confidence level. The 95 percent upper confidence bound for each test condition is shown in Table 4-7. As highlighted in red, the 95 percent upper confidence bound for Runs 4 and 12 exceeds 500 milliseconds. There is not sufficient evidence to reject the null hypothesis. (i.e., the true mean of the population could be more than 500 milliseconds when altitude and range are both at the high level and aircraft speed is at the low level).

**Table 4-7. Calculated Upper Confidence Intervals for MWS Example**

Run	A:Altitude	B:Range	C:Aircraft Speed	D:Clutter	Time to Handoff (ms)	95% CI high
1	-1	-1	-1	Low	352.63	353.11
2	1	-1	-1	Low	386.31	399.77
3	-1	1	-1	Low	385.61	390.80
4	1	1	-1	Low	518.39	517.82
5	-1	-1	1	Low	326.29	332.92
6	1	-1	1	Low	375.43	379.58
7	-1	1	1	Low	358.07	370.61
8	1	1	1	Low	489.84	497.63
9	-1	-1	-1	High	394.13	400.98
10	1	-1	-1	High	391.74	391.35
11	-1	1	-1	High	431.25	438.66
12	1	1	-1	High	499.41	509.39
13	-1	-1	1	High	373.54	380.79
14	1	-1	1	High	367.18	371.16
15	-1	1	1	High	422.40	418.47
16	1	1	1	High	485.20	489.20
17	0	0	0	Low	397.37	380.79
18	0	0	0	High	415.79	421.91
19	0	0	0	Low	402.17	402.19
20	0	0	0	High	412.67	421.91
21	0	0	0	Low	401.35	402.19
22	0	0	0	High	417.09	421.91

Source: STAT COE-Report-04-2015 (Ortiz and Truett 2015)

#### 4.3.2.2 Prediction Interval

PIs encompass the variation in both the estimation and the response. Therefore, PIs tend to be wider than CIs. Table 4 shows the results for the MWS case study with a new column that provides the 95 percent PIs. As the table shows, Runs 4, 8, and 12 all have a value greater than 500 milliseconds (highlighted in red in Table 4). These results suggest that although the true mean could be less than 500 milliseconds at Run 8, the response variation can show values that exceed 500 milliseconds.

**Table 4-8. Calculated Upper Prediction Intervals for MWS Example**

Run	A:Altitude	B:Range	C:Aircraft Speed	D:Clutter	Time to Handoff (ms)	95% CI high	95% PI High
1	-1	-1	-1	Low	352.63	353.11	358.19
2	1	-1	-1	Low	386.31	399.77	404.85
3	-1	1	-1	Low	385.61	390.80	395.88
4	1	1	-1	Low	518.39	517.82	522.90
5	-1	-1	1	Low	326.29	332.92	338.00
6	1	-1	1	Low	375.43	379.58	384.67
7	-1	1	1	Low	358.07	370.61	375.69
8	1	1	1	Low	489.84	497.63	502.71
9	-1	-1	-1	High	394.13	400.98	406.06
10	1	-1	-1	High	391.74	391.35	396.43
11	-1	1	-1	High	431.25	438.66	443.74
12	1	1	-1	High	499.41	509.39	514.48
13	-1	-1	1	High	373.54	380.79	385.87
14	1	-1	1	High	367.18	371.16	376.24
15	-1	1	1	High	422.40	418.47	423.56
16	1	1	1	High	485.20	489.20	494.29
17	0	0	0	Low	397.37	380.79	409.07
18	0	0	0	High	415.79	421.91	428.79
19	0	0	0	Low	402.17	402.19	409.07
20	0	0	0	High	412.67	421.91	428.79
21	0	0	0	Low	401.35	402.19	409.07
22	0	0	0	High	417.09	421.91	428.79

Source: STAT COE-Report-04-2015 (Ortiz and Truett 2015)

#### 4.3.2.3 Tolerance Interval

A column for the 99 percent population/95 percent TI for the MWS case study data has been added in Table 4-9. Runs 4, 8, 12, and 16 fail to meet the requirement that time to handoff is less than 500 milliseconds at least 99 percent of the time. Note that neither the CI nor the PI calculations were able to address this requirement directly. The TI is the only interval that indicates what scenario will result in failures more than 1 percent of the time. However, the TI does not directly provide an estimate for  $P_s$ . To get an estimate for  $P_s$ , the inverse of the TI needs to be found, as shown in the last column of Table 4-9.

**Table 4-9. Calculated Upper Tolerance Interval for MWS Example**

Run	A:Altitude	B:Range	C:Aircraft Speed	D:Clutter	Time to Handoff (ms)	95% CI high	95% PI High	99% population 95% TI high	% Below Spec
1	-1	-1	-1	Low	352.63	353.11	358.19	369.73	>99%
2	1	-1	-1	Low	386.31	399.77	404.85	416.39	>99%
3	-1	1	-1	Low	385.61	390.80	395.88	407.42	>99%
4	1	1	-1	Low	518.39	517.82	522.90	534.44	<0.1%
5	-1	-1	1	Low	326.29	332.92	338.00	349.54	>99%
6	1	-1	1	Low	375.43	379.58	384.67	396.21	>99%
7	-1	1	1	Low	358.07	370.61	375.69	387.23	>99%
8	1	1	1	Low	489.84	497.63	502.71	514.25	78.9%
9	-1	-1	-1	High	394.13	400.98	406.06	417.60	>99%
10	1	-1	-1	High	391.74	391.35	396.43	407.97	>99%
11	-1	1	-1	High	431.25	438.66	443.74	455.28	>99%
12	1	1	-1	High	499.41	509.39	514.48	526.02	11.1%
13	-1	-1	1	High	373.54	380.79	385.87	397.41	>99%
14	1	-1	1	High	367.18	371.16	376.24	387.78	>99%
15	-1	1	1	High	422.40	418.47	423.56	435.10	>99%
16	1	1	1	High	485.20	489.20	494.29	505.83	97.1%
17	0	0	0	Low	397.37	380.79	409.07	397.41	>99%
18	0	0	0	High	415.79	421.91	428.79	439.26	>99%
19	0	0	0	Low	402.17	402.19	409.07	419.54	>99%
20	0	0	0	High	412.67	421.91	428.79	439.26	>99%
21	0	0	0	Low	401.35	402.19	409.07	419.54	>99%
22	0	0	0	High	417.09	421.91	428.79	439.26	>99%

Source: STAT COE-Report-04-2015 (Ortiz and Truett 2015)

## 4.4 Advanced Methods and Additional Techniques

This section of the guidebook covers advanced techniques that may be applicable for some tests. Project managers should explore and understand their potential applications and use the guidance within this section to determine the contexts in which these techniques are most beneficial. Through informed application, these advanced techniques can significantly elevate the quality and impact of T&E processes. By employing statistical techniques early in the process, practitioners can gain critical insights into the data's underlying structure, identify potential anomalies, and ensure that the data is suitable for further analysis. This preliminary assessment helps to mitigate risks associated with data quality issues and lays a solid foundation for subsequent analyses. Additionally, careful planning of statistical analyses ensures that the chosen methods are appropriate for the research questions and data characteristics, optimizing resource use and enhancing the credibility of the findings.

### 4.4.1 Unbalanced Data

The proper approach to regression or other statistical analysis for validating M&S depends on sample sizes—in particular, on the amounts of live versus simulation data. When both sets of data are balanced, a single regression model combining the two, for example as the differences

between matched data pairs, can be appropriate. However, if there are much more M&S data than live test data (unbalanced data sets), confounding correlations can make it difficult or impossible to determine whether statistically significant differences exist between the live and simulation data. When the data sets are unbalanced, at least two approaches can be used:

- Bootstrapping (or resampling), in which an overall balanced data set is created by randomly sampling with replacement to generate a smaller subset from the larger M&S data set and combining those resampled simulation data with the live test data (Efron and Tibshirani 1994). Regression or other statistical analysis can be performed on the balanced data set, and the procedure can be repeated many times, generating distributions of test, model coefficients, and predictions (Wojton et al. 2019; Wojton and Avery 2019). This approach can eliminate confounding and allow statistical testing of the significance of interactions among factors in a regression model. In particular, assuming that the model has a live versus M&S source for purposes of validation, bootstrapping can be used to analyze whether statistically significant differences exist between the live and M&S data (the sources) and other regression factors (i.e., whether the M&S and live data agree in some circumstances but not others).
- Constructing two regression models, one for the live data and one for the M&S data, and testing statistically for differences between the coefficients composing each of the two models. This approach should be straightforward for live test data collected using plans developed by application of DOE, as a statistical model of the testing and its results underlies that approach.

### 4.4.2 Sequential Analysis

This section of the guidebook is a synopsis of the IDA Document, “A Review of Sequential Analysis” (Wojton et al. 2020), with material drawn from additional sources.

In sequential analysis, the characteristics of the data and their amount undergoing statistical analysis can change and be adjusted as results are obtained. Thus, sequential analysis offers the possibility that defensible conclusions can be obtained more quickly using lesser resources than would be the case if test plans and resources are predetermined at the outset. Sequential analysis comprises sequential testing and sequential estimation, in which the amount of data used can change as information accrues, and sequential design, in which the characteristics of the data being collected can also change. These three categories are not distinct in that the approaches and methods they employ can overlap.

### Sequential Testing

Sequential testing involves testing hypotheses as data are collected and as the sample size, which is not predetermined at the outset of the analysis, increases. After each test, decisions are made about whether more data should be collected. The sequential probability ratio test (SPRT) can be regarded as the origin of this approach (Wald 1945).<sup>8</sup> Using the SPRT, data are obtained one datum at a time. With the addition of each datum, statistical testing is performed to decide to accept the null hypothesis, to reject the null hypothesis, or to collect additional data. Using the SPRT can often require lesser collection of data than nonsequential testing in which the amount of data needed is fixed at the outset (Wojton et al. 2020). The SPRT also deals with Type I errors (incorrect rejection of the null hypothesis) and Type II errors (incorrect acceptance of the null hypothesis) on par with nonsequential testing (Wald 1945).

The potential savings offered by sequential testing is illustrated by analysis of a system's failure rate using the SPRT to determine whether its failure rate  $p = p_1 > p_0$ . The null hypothesis is that the failure rate is  $p_0$ , and the alternative is that it is greater and equal to at least  $p_1$ . In this particular case, as shown in Figure 4-13, after 11 runs and 4 failures, use of the SPRT indicates the null hypothesis should be rejected in favor of the alternative. The required number of runs for a nonsequential exact binomial test is 50. See Appendix A of the IDA Document, "Case Study on Applying Sequential Analyses in Operational Testing" (Medlin et al. 2021), for additional details.

---

<sup>8</sup> Abraham Wald developed the SPRT while a member of the Statistical Research Group at Columbia University during World War II. The Navy was interested in determining the probability an anti-aircraft gun could hit a directly approaching dive bomber and desired to do so using the minimum test resources necessary (Stein 1945).





Source: IDA Document NS D-32904 (Medlin et al. 2021)

**Figure 4-13. SPRT Applied to Analyze a System's Failure Rate**

### Sequential Estimation

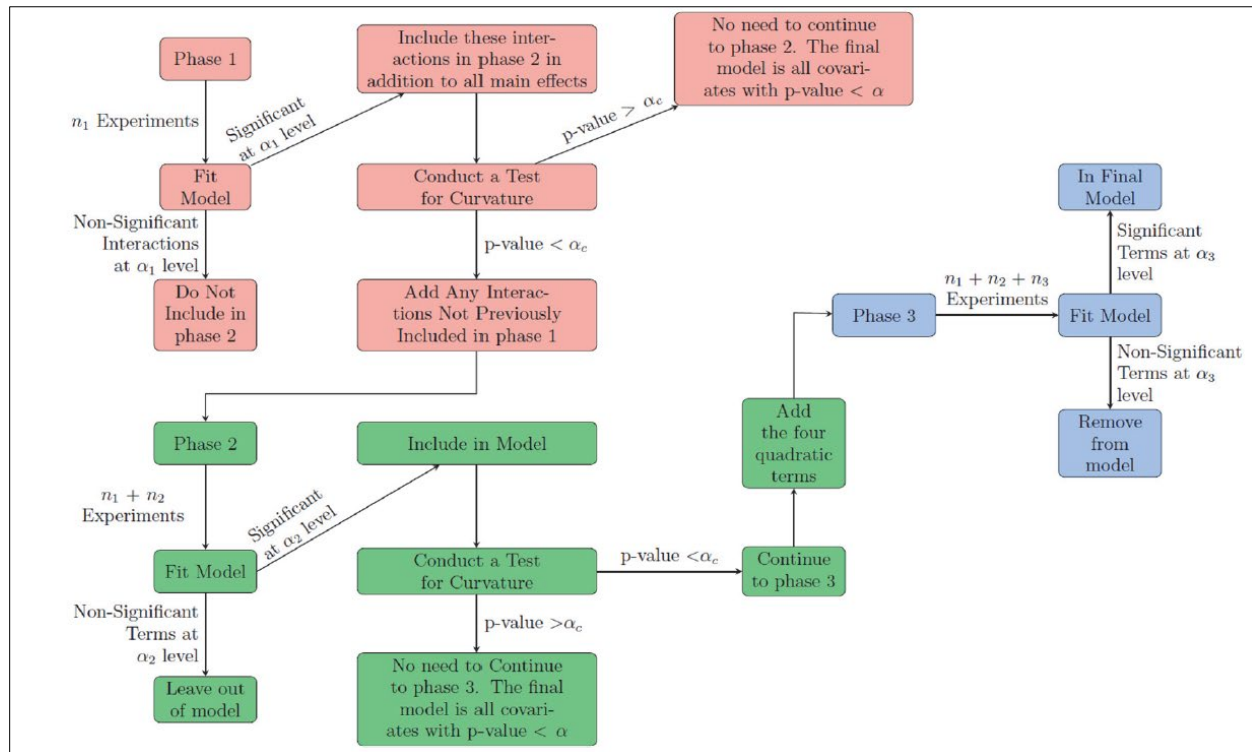
Sequential estimation involves calculating statistical quantities or intervals where the amount of data used is determined based on the accruing information as the calculations proceed. Examples of sequential estimation include:

- Inverse binomial sampling “which can estimate the log-likelihood of an entire data set efficiently and without bias” and with minimum variance (van Opheusden et al. 2020).
- Double sampling, in which one sample is used to estimate the variance and a second sample, sized based on the first (and not necessarily nested within the first sample), is used to estimate the mean (Stein 1945; Hidiroglou 2001). This approach enables “calculation of a confidence interval of fixed width that is independent of the mean” (Wojton et al. 2020).

### Sequential Design

Sequential design involves applying DOE to plan and conduct a series of tests in which the factors used and the characteristics and amount of data collected during each test are determined based on the information gained as the testing is conducted. Although not commonly described as sequential, DOE can be used in this manner, in which the results of prior testing are used to

inform and design subsequent testing. An example of sequential application of DOE would be an approach in which the number of tests/experiments and the factors and interactions/terms included in each experimental design/model are determined for each additional phase of testing based on the significance (i.e., p-values of significance tests) of the interactions/terms found upon analysis of the results from the prior phase; see Figure 4-14. See Appendix B of the IDA Document, “Case Study on Applying Sequential Analyses in Operational Testing” (Medlin et al. 2021) for additional details.



Source: IDA Document NS D-32904 (Medlin et al. 2021)

**Figure 4-14. Example of Implementation of Sequential DOE**

These sequential analysis approaches can be used when planning and conducting validation of M&S. For example, sequential design can be used to plan the live testing that generates data for comparison with M&S responses. And sequential testing and estimation can be used to decide whether enough data, both from M&S and live testing, have been obtained to accept or reject the null hypothesis that the M&S matches the live data and should be validated. The use of these methods offers the possibility that the amount of both live test data and simulation results needed for validation will be less than if the validation effort is planned and conducted nonsequentially.

#### 4.4.3 Predicted Probability Validation

This section of the guidebook comprises a synopsis of the IDA Document, “Predicted Probabilities Validation” (Haman et al. 2022).

Predicted probability validation (PPV) provides a means to statistically compare binary results from testing with probabilities predicted by M&S (Haman et al. 2022). Examples of the application of PPV include the following: comparing M&S predictions of the probability of critical damage to a system's components by a kinetic threat to live binary test results of whether damage was or was not observed; or comparing binary test results of the success or failure of executing elements of a kill chain to M&S predictions of the probability of success for each element of the kill chain.

PPV data comprise M&S-predicted probabilities  $p_i$  ranging from 0 to 1 and paired with corresponding binary test outcomes  $y_i$  of 0 or 1:  $(p_i, y_i), i = 1, 2, \dots, n$ . As shown in Table 4-10, the PPV methods used to compare the M&S and live test outcomes involve calibration, discrimination, and overall performance:

- Calibration methods assess the extent to which probabilities predicted by the M&S are too high or too low compared to the binary test data. For example, if M&S predicts the probability of success to be 40 percent, then 20 test runs should have 8 successes.
- Discrimination methods assess whether the relative relationship of test outcomes matches the predicted probabilities, for example, whether the number of successes observed in live testing is higher for larger M&S-predicted probabilities and vice versa.
- Overall performance methods assess calibration and discrimination simultaneously.

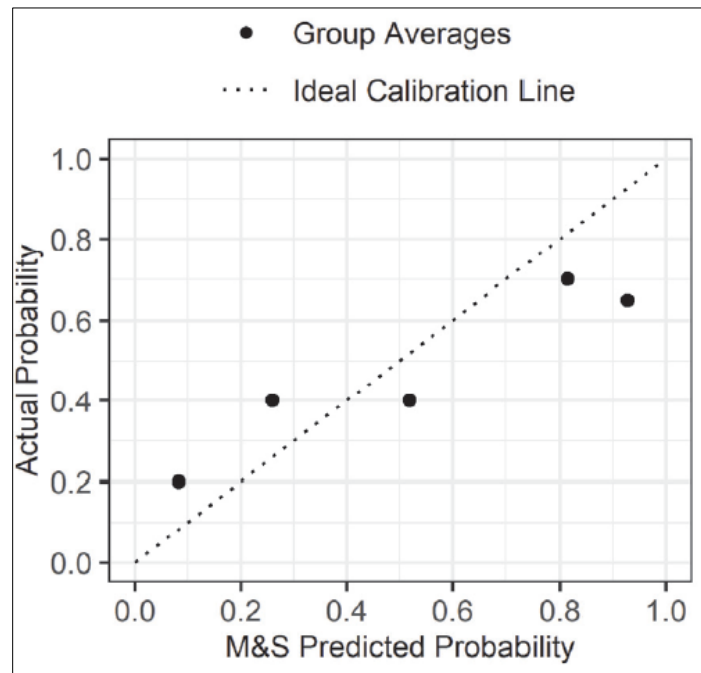
**Table 4-10. PPV Methods for Comparing M&S Probabilities with Binary Test Results**

Type	Metrics	Graphical Techniques
Calibration	<ul style="list-style-type: none"> <li>• Harrell's U index</li> <li>• Intercept, slope</li> <li>• <math>E_{max}</math>, <math>E_{90}</math>, <math>E_{avg}</math></li> <li>• Spiegelhalter's statistic</li> </ul>	<ul style="list-style-type: none"> <li>• Group averages</li> <li>• Parametric and nonparametric calibration curves</li> </ul>
Discrimination	<ul style="list-style-type: none"> <li>• Harrell's D index</li> <li>• Somers' <math>D_{xy}</math> index</li> <li>• C index (AUC)</li> </ul>	<ul style="list-style-type: none"> <li>• Histograms of binary test outcomes</li> </ul>
Overall Performance	<ul style="list-style-type: none"> <li>• Harrell's Q index</li> <li>• Nagelkerke's <math>R^2</math></li> <li>• Brier score</li> </ul>	

Source: IDA Document D-33156 (Haman et al. 2022)

### Group Average Scatter Plots

This method divides the live data into equal-sized groups and calculates the average M&S-generated probability and average binary test result within each group. The results are then plotted and compared with the ideal calibration line for which the two would be equal; see the notional example in Figure 4-15.



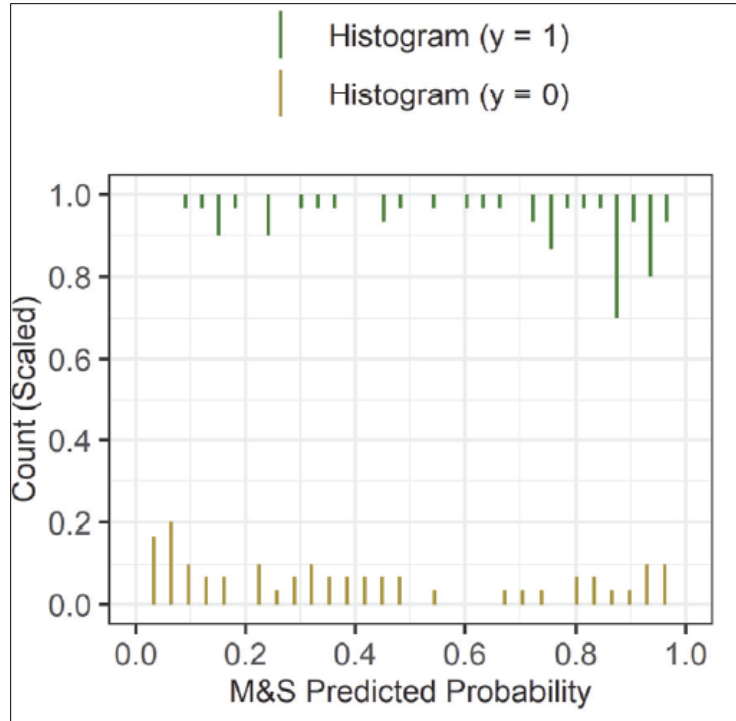
Source: IDA Document D-33156 (Haman et al. 2022)

**Figure 4-15. Example of a Group Average Plot**

### Histograms

Plotting the frequency of binary successes ( $y = 1$ ) and failures ( $y = 0$ ) associated with each live data group at the top and bottom of a histogram can provide useful information; see the notional example in Figure 4-16 showing good discrimination:

- Good discrimination is indicated by the longest bars at the top clustering around predicted probability of 1 and the longest bars at the bottom clustering around 0.
- Perfect discrimination is indicated if all the bottom bars are to the left of all the top bars.
- Bad discrimination is indicated if the top and bottom bars are all about the same height and spread evenly across the predicted probabilities.



Source: IDA Document D-33156 (Haman et al. 2022)

**Figure 4-16. Example of PPV Histogram**

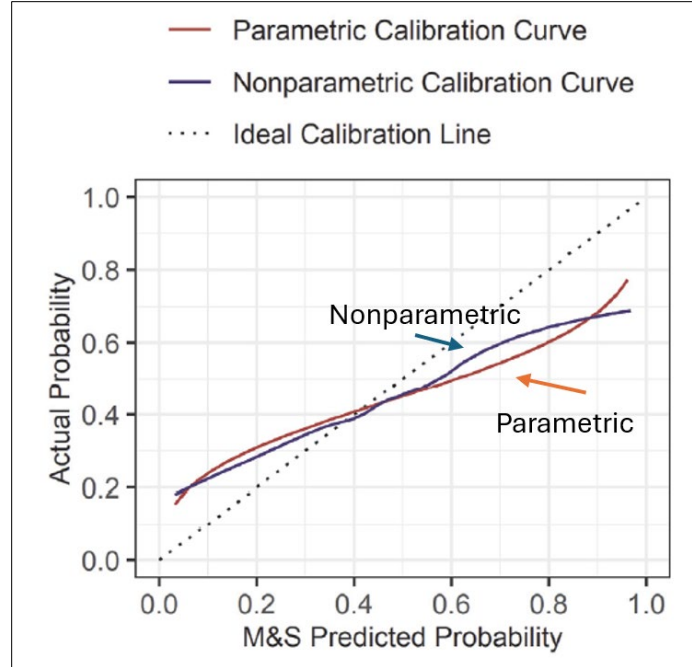
### Calibration Curves

A parametric calibration curve employs locally estimated scatterplot smoothing to plot the binary test data as a function of the probabilities predicted by the simulation (Cleveland 1979). A nonparametric calibration curve employs logistic regression to model the actual probability of success,  $\tilde{p}$ , as a function of the logit of the probability predicted by the simulation,  $p$ :<sup>9</sup>

$$\text{logit}(\tilde{p}) = y_0 + y_1 \text{logit}(p)$$

with  $y_1$  being the slope of the curve and  $y_0$  the intercept. These calibration curves are then compared to the ideal calibration line for which the predicted probability matches the actual probability; see the notional example in Figure 4-17. In this example, the slope is less than 1, and the M&S-generated low probabilities are too small, and the M&S-generated high probabilities are too large. When the slope is greater than 1, the M&S-generated low probabilities are too large, and the M&S-generated high probabilities are too small.

<sup>9</sup>  $\text{logit}(z) = \log\left(\frac{z}{1-z}\right)$



Source: IDA Document D-33156 (Haman et al. 2022)

**Figure 4-17. Example of PPV Calibration Curves**

### Harrell's U Index

Harrell's U index measures calibration by comparing the likelihood of the test data ( $n$  observed outcomes) given the M&S-predicted probabilities (represented using  $L_{01}$ , the corresponding negative log-likelihood) with the likelihood of the test data given the actual probabilities (represented using  $L_{ab}$ , the corresponding negative log-likelihood):<sup>10</sup>

$$U = (L_{01} - L_{ab})/n$$

$$L_{01} = -2 \sum_{i=1}^n (y_i \log(p_i) + (1 - y_i) \log(1 - p_i))$$

$$L_{ab} = -2 \sum_{i=1}^n (y_i \log(\tilde{p}_i) + (1 - y_i) \log(1 - \tilde{p}_i))$$

When  $U$  is a large positive number,  $L_{ab}$  is more likely than  $L_{01}$ , indicating poor calibration. When  $U$  is a large negative number,  $L_{01}$  is more likely, indicating good calibration.

<sup>10</sup> Maximizing likelihood is equivalent to minimizing the corresponding negative log-likelihood.

### Error Indices

Error indices measure the difference between the nonparametric calibration curve and the ideal calibration line.  $E_{max}$ ,  $E_{90}$ ,  $E_{avg}$  are, respectively, the maximum, 90th percentile, and average vertical distances between the nonparametric calibration line and the ideal calibration curve taken at increments along the horizontal axis. The smaller their values, the better the calibration.

### Spiegelhalter's Statistic

Spiegelhalter's statistic tests a null hypothesis that the M&S is perfectly calibrated. The p-value for this statistic,  $S_p$ , is computed according to the journal article, "Probabilistic Prediction in Patient Management and Clinical Trials" (Spiegelhalter 1986):

$$S_z = \frac{\sum_{i=1}^n (y_i - p_i)(1 - 2p_i)}{\sum_{i=1}^n (1 - 2p_i)^2 p_i (1 - p_i)}$$

$$S_p = 2\Phi(-|S_z|)$$

where  $\Phi(x) = \text{Prob}(Z \leq x)$  is the standard normal CDF.<sup>11</sup>

If  $S_p$  is sufficiently small, the null hypothesis is rejected, indicating the M&S is not perfectly calibrated.

### Somers' $D_{xy}$ Index

Somers'  $D_{xy}$  index assesses the extent to which all pairs of binary test data and M&S-predicted probabilities are concordant (ordered properly) or discordant (not properly ordered). Pairs of data in which failure ( $y = 0$ ) is associated with a low predicted probability and success ( $y = 1$ ) is associated with a high predicted probability are concordant and conversely:

$$D_{xy} = \frac{N_C - N_D}{N_T}$$

where  $N_C$  is the number of concordant pairs,  $N_D$  is the number of discordant pairs, and  $N_T$  is the total number of pairs. If all the pairs are concordant,  $D_{xy} = 1$ , and the M&S is perfectly discriminating.

---

<sup>11</sup> See the Cumulative Distribution Function of the Standard Normal Distribution in the NIST Engineering Statistics Handbook (<https://www.itl.nist.gov/div898/handbook/eda/section3/eda3671.htm>).

Consider the notional data displayed in Table 4-11 in which  $N_C = 2 \times 8 + 2 \times 5 + 6 \times 5 = 56$ ;  $N_D = 1 \times 6 + 1 \times 1 + 8 \times 1 = 15$ ;  $N_T = (2 + 6 + 1) \times (1 + 8 + 5) = 126$ ;  $D_{xy} = 0.33$ .

**Table 4-11. Notional Pairs of Observed Binary Data and Predicted Probabilities**

p (row) / y (column)	0.30	0.60	0.90
0	2	6	1
1	1	8	5

### C Index

The  $C$  index is computed using the  $D_{xy}$  index:  $C = \frac{D_{xy} + 1}{2}$ . When  $C = 1$ , the M&S is perfectly discriminating; when  $C = 0.5$ , the M&S predictions are random.

### Harrell's D Index

Harrell's  $D$  index assesses discrimination by comparing the mean of the  $n$  binary test outcomes,  $\bar{y}$ , to the M&S-predicted probabilities (Harrell and Lee 1990). The index is computed using the difference between the negative log-likelihoods of the mean binary outcome and the M&S predictions:

$$L_{a0} = -2 \sum_{i=1}^n (y_i \log(\bar{y}) + (1 - y_i) \log(1 - \bar{y}))$$

$$D = (L_{a0} - L_{01})/n$$

If  $D$  is positive,  $L_{01}$  is more likely than  $L_{a0}$ , and the M&S discrimination is good. If  $D$  is negative,  $L_{a0}$  is more likely than  $L_{01}$ , and discrimination is poor.

### Harrell's Q Index

Harrell's  $Q$  index assesses overall performance by combining the  $D$  and  $U$  indices described above:  $Q = D - U$ . Recall that positive values of  $D$  mean discrimination is good, and positive values of  $U$  mean calibration is poor. Therefore, when  $Q$  is positive, the quality of the discrimination is greater than the shortcomings in calibration, meaning overall performance is acceptable, if not good. If  $Q$  is negative, overall performance of the M&S is poor.



### Nagelkerke's $R^2$ Index

Nagelkerke's  $R^2$  index assesses overall performance by measuring predictive strength (Nagelkerke 1991):

$$R^2 = \frac{1 - \exp\left(\frac{L_{01} - L_{ao}}{n}\right)}{1 - \exp\left(-\frac{L_{ao}}{n}\right)}$$

Good overall performance is indicated by large positive values, poor performance by small or negative values.

### Brier Score

The Brier score assesses overall performance by computing the mean squared error between the binary test data and the M&S-predicted probabilities:

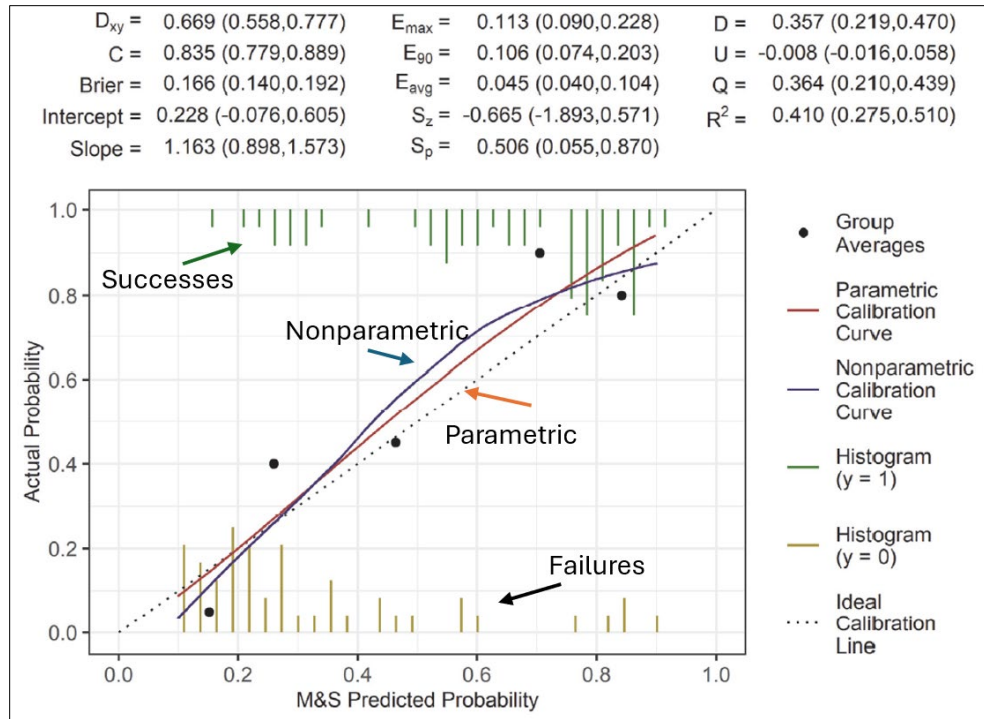
$$\text{Brier} = \frac{1}{n} \sum_{i=1}^n (p_i - y_i)^2$$

Small score values indicate good performance. Score values greater than 0.25 mean the predicted probabilities are no better at matching the binary test data than flipping a coin.

### Summary of Metrics Indicating Good Calibration, Discrimination, and Overall Performance

When the group averages cluster fairly well around the ideal calibration curve; the slope of the parametric calibration curve is close to 1 and its intercept is close to 0; top histogram bars cluster around predicted probabilities near 1 and bottom histogram bars cluster around 0; and all the other metrics discussed above are within their favorable ranges for either discrimination, calibration, or overall performance, then both calibration and discrimination are good, as is overall performance; see Figure 4-18. Note that bootstrapping can be used to estimate CIs for all the metrics of calibration, discrimination, and overall performance discussed above; again, see Figure 4-18. Additional examples are provided in the IDA Document, “Predicted Probabilities Validation” (Haman et al. 2022).

## 4. Analysis



Source: IDA Document D-33156 (Haman et al. 2022)

**Figure 4-18. Example of Metrics with Confidence Intervals Indicating Good Calibration, Good Discrimination, and Good Overall Performance**

### 4.4.4 Time Series Analysis and Functional Data Analysis

Sometimes the measure of interest from a simulation is a time series or a function. Consider a simulation that models the flight path of a projectile through the atmosphere. The data in this case is the relative position of the object in space at a particular time. The objective of validation in this case is to evaluate how well the simulation compares to discrete live event observations of the actual system, such as a flight test. The fact that the response or measure of interest is not a value but rather a series of values presents a unique set of challenges. The following subsections present, at a high level, some approaches that have been used for these occasions and discuss some of the advantages and disadvantages of these approaches.

It is important to emphasize that a rigorous validation approach should help determine how well the M&S represents the system under test. The approach should be in line with an acceptability criterion that is based on a system performance specification. The approach should also help quantify test risk (both power and confidence).

#### 4.4.4.1 Approach: Functional Data Analysis

Functional data analysis (FDA) is a branch of statistics that deals with information represented by functions, curves, or shapes (Ramsay and Silverman 2005). This type of data analysis is used

when data is collected over a continuous range of points rather than at discrete intervals. FDA is particularly useful in fields such as medicine, finance, and environmental science, where observations are often recorded over a period of time. The main goal of FDA is to explore and model the underlying dynamics of the data, focusing on the structural features of the data, such as peaks, troughs, and overall shape.

FDA can be used to compare two time series by considering each time series as a single function or curve. The steps involved in comparing two time series using FDA are as follows:

1. **Data Collection:** Collect the time series data to be compared. This data could be any type of data that is recorded over a period.
2. **Data Preprocessing:** Clean the data by removing any outliers or noise. This step is crucial as it can significantly impact the results of the analysis.
3. **Data Representation:** Represent each time series as a smooth function. This can be done using various techniques such as smoothing splines, Fourier series, or wavelets. This step transforms the discrete time series data into continuous functions.
4. **Functional Principal Component Analysis (FPCA):** Perform FPCA to reduce the dimensionality of the functional data and to identify the main modes of variation between the curves. This step helps in understanding the key differences between the two time series.
5. **Functional Hypothesis Testing:** After identifying the main modes of variation, perform hypothesis testing to determine whether the differences between the two time series are statistically significant. This step involves comparing the mean functions of the two time series and testing whether they are equal.
6. **Functional Regression:** If one time series is dependent on the other, use functional regression to model this relationship. This step involves using one time series as the predictor variable and the other as the response variable.
7. **Interpretation of Results:** Finally, interpret the results of the analysis. This could involve identifying key differences between the two time series, determining whether one time series can predict the other, or determining whether the differences between the two time series are statistically significant. Remember, the specific steps and techniques used can vary depending on the nature of the time series data and the specific goals of the analysis.

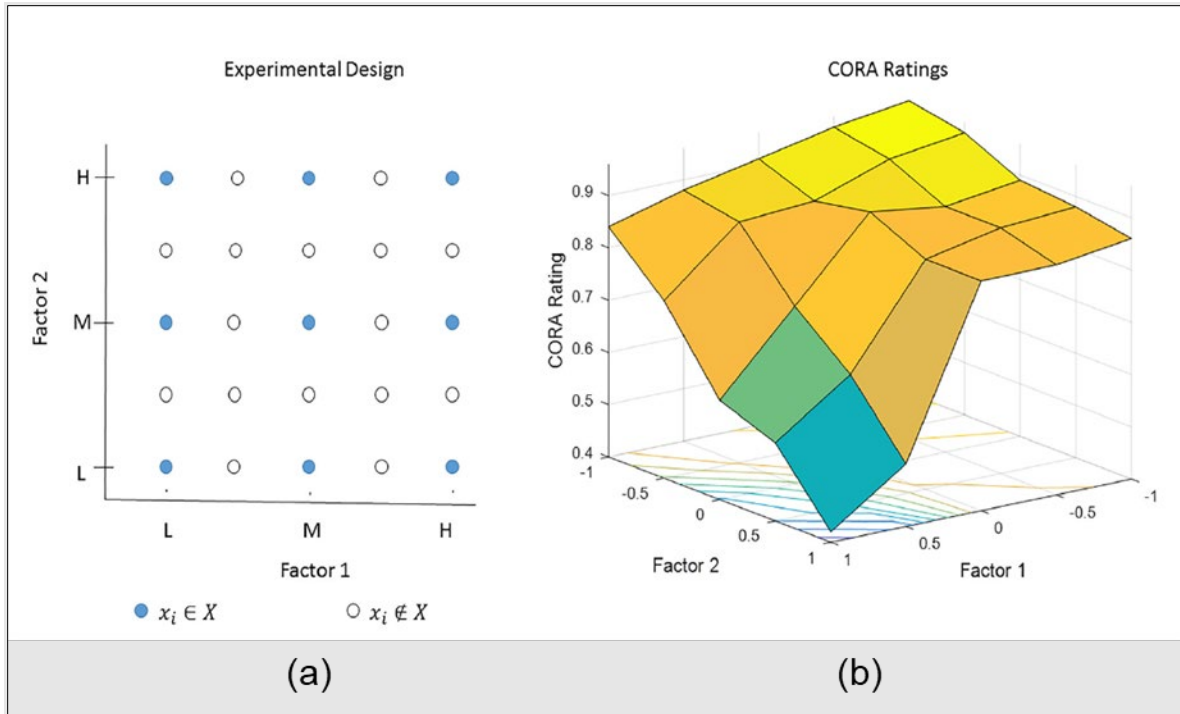
A coding example of this process and its steps is available in the book *Functional Data Analysis* (Ramsay and Silverman 2005).

Using FDA for time series comparison offers several benefits. The approach captures the entire shape of the time series, not just summary statistics. It accounts for time warping and misalignments and is robust to noise. FDA also enhances interpretability with a rich set of tools for similarity measurement and visualization. These tools can help determine how well the M&S represents the system under test. Using FDA also presents some challenges. The approach can be computationally expensive for large data sets, and choosing the most appropriate similarity measure depends on the specific characteristics of the data.

### 4.4.4.2 Example of Using Functional Data Analysis and Design of Experiments

An important objective of any validation method is that it should aid in improving the M&S of the sensor. Ultimately, the goal of these large-scale simulations is to continuously become less and less wrong with their predictions. A validation approach should not only score the M&S on its ability to predict performance accurately but should aid in identifying where in the operational space the M&S does not fit as well, and it should help identify what factors contribute to these discrepancies. Using such a validation approach will lead to a more thorough assessment of how valuable the M&S is and where future resources, research, and testing need to be focused.

An excellent example of a more rigorous validation methodology is presented in the journal article, “Model Validation of Functional Responses Across Experimental Regions Using Functional Regression Extensions to the CORA Objective Rating System” (Storm et al. 2017). The article presents a methodology specifically devoted to the validation of dynamical responses of M&S, such as a time series. The approach combines DOE and FDA to assess the predictive capability of a simulation across the operational space. DOE is used to sample the operational space efficiently. FDA is used to model the system response (a time series). The correlation analysis (CORA) objective rating system, a validation metric, is used to provide a measure of disagreement between live and simulated results anywhere within the factor space. The CORA rating varies from 0 to 1; values near 1 represent little discrepancy between M&S and live data; see Figure 4-19.



Source: (Storm et al. 2017)

**Figure 4-19. Example Output from the CORA Validation Methodology**

Figure 4-19(a) is the experimental design; blue points represent tested scenarios, and clear points are estimated points based on an empirical prediction model. Figure 4-19(b) is the CORA rating across Factors 1 and 2. The rating varies from 0 to 1, with values near 1 representing little discrepancy between M&S and live data. Again, using a more rigorous validation approach such as the one presented above will lead to a better understanding of the M&S predictive capabilities and will aid in root cause analysis.

#### 4.4.4.3 Approach: Other Validation Metrics in the Literature

There are several validation metrics or statistical approaches for time series data in the literature that are more informative and allow for a more robust assessment of the M&S. For example, the Sandia National Laboratories Report, “Validation Metrics for Deterministic and Probabilistic Data” (Maupin and Swiler 2017), compares metrics to validate simulation models where test cases produce data over a period. These validation metrics quantify the difference between live and simulated observations. The report provides guidance in selecting a validation metric based on the quantity of interest and the characteristics of the observed and simulated data sets. The report groups validation metrics into three data types:

- Type 1: Live and simulated responses are treated as point values (deterministic data with no uncertainty or noise).

#### 4. Analysis

- Type 2: Live responses are treated as stochastic (noisy) values.
- Type 3: Both live and simulated data are treated as noisy data.

Table 4-12 summarizes the validation metrics presented in the Sandia Report and the types of data for which they are applicable.

**Table 4-12. Validation Metrics and the Types of Data to Which They Are Targeted**

Metric	Type 1	Type 2	Type 3
Root Mean Square	X		
Minkowski Distance	X		
Normalized Euclidean Metric		X	
Mahalanobis Distance		X	
Kullback-Leibler Divergence			X
Symmetrized Divergence			X
Jensen-Shannon Divergence			X
Hellinger Metric			X
Kolmogorov-Smirnov Test		X	X
Total Variation Distance			X
Simple Cross-Correlation	X		
Normalized Cross-Correlation	X		
Normalized Zero-Mean Sum of Squared Distances	X		
Moravec Correlation	X		
Index of Agreement	X		
Sprague-Geers Metric	X		

Source: Sandia Report SAND2016-1421 (Maupin and Swiler 2017)

The Sandia Report also highlights signal processing validation metrics. These metrics address high-frequency and phased-based data. In general, these metrics calculate the correlation between simulated and live results and can identify trends in the differences between the results. Some validation methods mentioned include normalized cross-correlation, Moravec correlation, and Sprague-Geers metric. These signal processing validation metrics are meant to be applied to Type 1 data. Analysis with Type 2 or 3 data can be done; however, the uncertainty information would not be utilized. Further research on how best to incorporate uncertainty is needed.

##### 4.4.4.4 Legacy Approaches

Some legacy approaches are not recommended for analysis.

### **Summarizing the Time Series**

An initial approach when dealing with time series or functional data could be to try to reduce the dimensions of the problem by looking at and comparing summary statistics. This approach fails to capture the entire shape of the time series, leading to very limited insight and possibly an incorrect conclusion as to whether the simulated results match the live data. Therefore, more sophisticated approaches are needed to capture temporal patterns.

### **Using M&S to Bound the Live Data**

One common approach is to use the simulation itself to establish the bounds of the live data. This validation approach compares bounds created by a Monte Carlo generated time series of a validation metric from a simulation against a single, discrete live event observation of the actual system.

However, using the M&S to bound the live data does not aid in improving the M&S by identifying the factors affecting prediction performance, does not provide information on the areas in the operational space where prediction is poor, and gives no insight into the uncertainties that exist in the system.

#### **4.4.4.5 Summary of Approaches**

Currently, there appears to be no consensus on what the best approach is when comparing live and simulated time series data. The appropriate approach is dependent on the specific characteristics of the data and the amount of data that can be obtained from live and simulated tests. This section presented a high-level overview of approaches that have been discussed in the literature and highlighted their pros and cons. It is important to remember that, ideally, a validation approach should inform the user not only about how well the M&S represents the system under test but also about the conditions under which the simulation does not perform well. The approach should also help quantify test risk (e.g., power, confidence) to quantify the credibility of the analysis and aid the decision makers.

### **4.4.5 Metamodel Techniques for Analysis of M&S Data**

This section of the guidebook is a synopsis of introductory and overview material in Sections 4 and 5 of the IDA Paper, “Metamodeling Techniques for Verification and Validation of Modeling and Simulation Data” (Haman and Miller 2022).

As noted previously, because metamodels summarize M&S responses, they can be used to assess M&S outputs relatively quickly and efficiently across the range of factors and inputs affecting a system’s performance. Metamodels facilitate the comparison of M&S responses with live data;

the understanding of system performance under conditions that are impossible to create for live testing; the discovery of potentially problematic M&S outputs; and the quantification of uncertainty in M&S predictions.

### Building Metamodels

The appropriate techniques to build metamodels depend on whether M&S responses are deterministic (no randomness in responses) or stochastic (varying responses for the same inputs), as well as whether the responses are discrete or continuous:

- Nearest neighbor or decision tree interpolations are best suited for M&S generating deterministic, discrete responses.
- GP interpolations are best suited for M&S generating deterministic, continuous responses.
- GAMs are best suited for M&S generating stochastic responses, either discrete or continuous.

A *nearest neighbor interpolator* makes predictions identical to the observed output closest to the point at which a prediction is needed. This interpolation is not based on parametric statistical theory. Rather, it assumes responses vary continuously without large variation.

A *decision tree* comprises multiple sequential nodes with branching paths. The path branch followed from each node to the next is determined by the answer to a true or false question. The predicted M&S response is provided at the termination of the path determined by the sequence of answers.

Examples of the implementation of both techniques are provided in the IDA Paper (Haman and Miller 2022).

GP interpolation assumes that because the unobserved M&S responses are not known with certainty, they, and their uncertainty, can be predicted by the likely values of a random function. GPs are determined by the mean function,  $\mu(t)$ , usually taken to be zero, and the two-parameter covariance kernel  $K(s, t)$ , which determines the behavior of the metamodel between any two observed M&S responses. There are many possible covariance kernels; the choice of kernel for any particular problem determines the quality of the GP metamodel. Often,  $K(s, t) = K(r), r = \|s - t\|$ .

Although other, more complex kernels (e.g., the Matérn kernel (Xing 2015)) are available, the Gaussian kernel is often used for GP interpolation:



#### 4. Analysis

$$K(r, \tau, \lambda) = \tau^2 e^{-\frac{r^2}{\lambda}}$$

Large  $\lambda$  means two nearby points have similar values; small  $\lambda$  means less similarity between nearby points. Approaches for fitting a GP interpolator (i.e., choosing kernel parameters) and quantifying uncertainty are discussed in more detail in the IDA Paper (Haman and Miller 2022).

A *GAM* is a statistical model that makes general (i.e., not strictly linear) assumptions about the relationship between factors and the average values of response variables, the latter of which are assumed to follow a non-normal distribution. Building a GAM requires making decisions on the functional form relating factors to responses and their interactions. Fitting the GAM requires choosing parameters that determine how data are smoothed and what the characteristics of the smoothing relationships should be (e.g., cyclical, increasing, decreasing). As an example, consider a GAM in which the response,  $y$ , depends on three factors  $x_1, x_2, x_3$  according to unknown functions  $f_1, f_2, f_3$ , with an intercept  $\alpha$  and noise  $\varepsilon$ :

$$y = \alpha + f_1(x_1) + f_2(x_2) + f_3(x_3) + \varepsilon$$

To build the GAM, the three functions, as well as the intercept and noise terms, must be estimated. Examples of how to do so are provided in the IDA Paper (Haman and Miller 2022).

#### Evaluating Metamodels

The appropriate techniques to use to evaluate whether the metamodel, once built, fits M&S output well enough to be useful must consider the metamodel's ability to match both observed outputs (in-sample outputs) and hypothetical unobserved outputs (out-of-sample outputs). Output splitting should be employed that uses only some of the M&S responses to build the metamodel, and other output data not used to build the model are used to evaluate its predictive abilities. Generally, three sets of data will be needed: a training set for building the model; a screening set that augments the training set, as needed; and an evaluation set for determining how well the metamodel can predict out-of-sample responses. DOE can be used to plan for collecting the M&S responses composing the three data sets.

Calibration methods and metrics analogous to those discussed subsequently for predictive probability validation can be used to evaluate the performance of the metamodel. If the metamodel's performance is the same when evaluated against the training, screening, and evaluation sets, the metamodel can be used. If not, the metamodel may be overfitting (i.e., predicting the training set in the larger data sets) or underfitting (i.e., making imprecise predictions).

#### 4.4.6 Model Validation Levels

A Model Validation Level (MVL) is an objective, automatable metric scored from 0 to 9 that quantifies how much trust can be placed in the results of a model to represent the real world (Stafford et al. 2024). The MVL framework is recommended to provide a quick, interpretable measure of model validity and identify areas for increasing model trust. In addition, MVLs are recommended as a metric for quantitatively tracking model validity over time with changes in referent data, the scope of intended use, and/or the model itself. This can help enable a fast pace in a DE environment. Because the MVL framework is generalized to be applicable to a wide range of models, additional methods (such as those discussed in this guidebook) that are tailored to the specific validation scenario should be used with MVLs to provide additional insight.

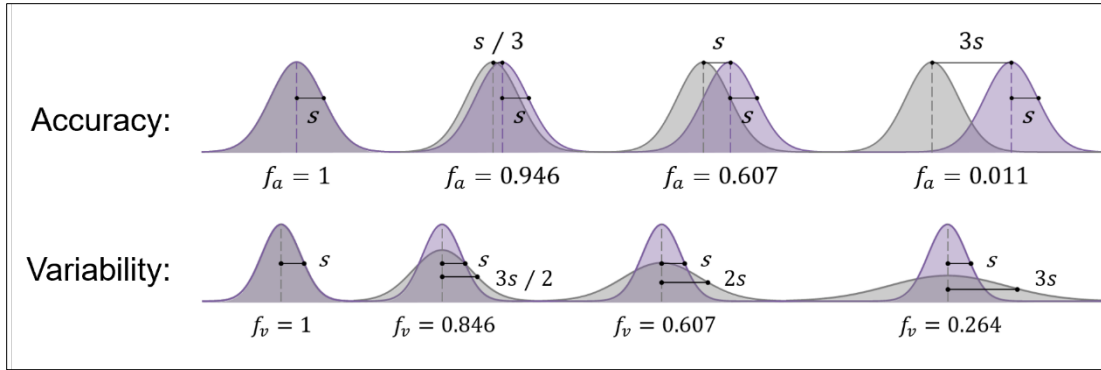
The MVL framework determines model validity based on three pillars: fidelity, referent authority, and scope. Fidelity is the level of consistency between a model and a referent, defined in the three dimensions of accuracy, repeatability, and resolution; referent authority is the strength of credibility of a referent's claim to be a high-fidelity representation of reality; and scope is the set of inputs, outputs, assumptions, and limitations representing the mission-relevant system parameters, environmental conditions, constraints, and requirements and their allowable values. Together, these pillars ensure that the model and referent data agree, the referent is trustworthy, and the validity is evaluated over the entire intended use. Each pillar is addressed by the MVL framework using quantitative metrics, including a fidelity metric, referent authority level, and scope coverage metric.

The fidelity metric is calculated at each unique input combination and is the product of an accuracy metric, which rates similarity in mean behavior between the model and referent, and a variability metric, which rates similarity in variability, where variability comprises both repeatability and resolution. Zero indicates no fidelity and one indicates perfect fidelity between the model and referent. The fidelity metric is given in Equation (4-8), where  $\bar{x}_m$  is the model response mean,  $\bar{x}_r$  is the referent mean,  $s_m^*$  is the model variability, and  $s_r^*$  is the referent variability. The variability,  $s^*$ , is defined in Equation (4-9), where  $s$  is the sample standard deviation and  $\delta$  is the resolution. Figure 4-20 illustrates the accuracy and variability metric for different cases.

$$f = f_a f_v = e^{-\frac{1}{2} \left( \frac{\bar{x}_m - \bar{x}_r}{s_r^*} \right)^2} e^{-\frac{(s_m^* - s_r^*)^2}{s_m^* s_r^*}} \quad (4-8)$$

$$s^* = \sqrt{s^2 + \frac{\delta^2}{12}} \quad (4-9)$$

#### 4. Analysis



Source: Adapted from the STAT COE Paper (Stafford et al. 2024)

**Figure 4-20. Visualizing the Accuracy and Variability Components of the Fidelity Metric**

The referent authority level of referent data is determined according to , which is derived from Technology Readiness Levels. In , referents that are more representative of real-world operational conditions receive a higher authority level. The MVL framework also enables multiple referents to be combined to validate the model against the body of available data instead of just one referent.

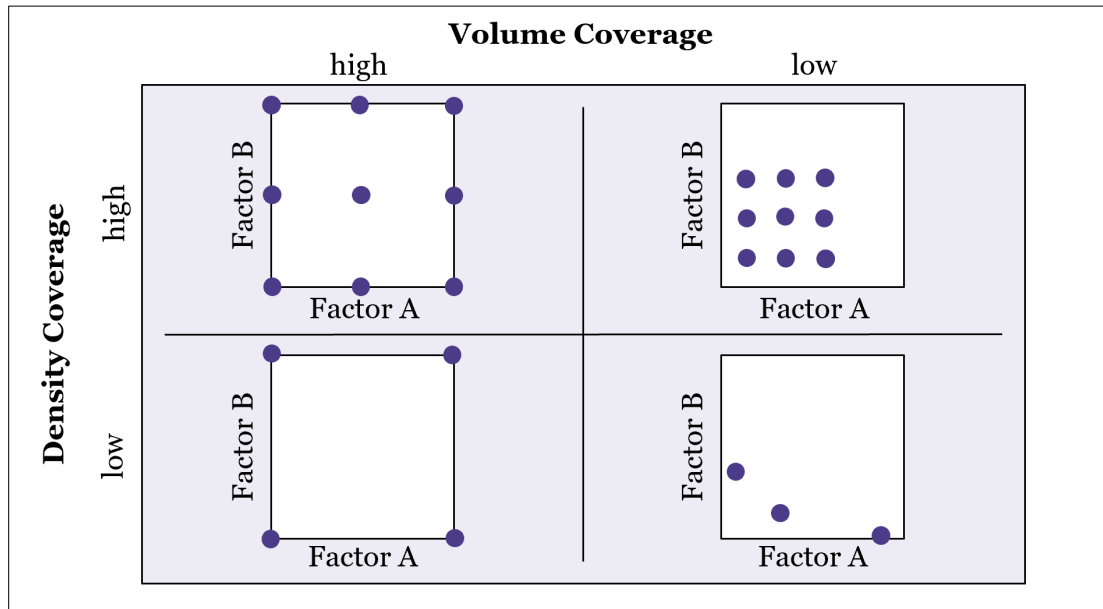
**Table 4-13. Referent Authority Level Scale**

Authority Level	Relevant Referent
1	Subject Matter Expert Judgment
2	First Principles/Physics Predictions
3	Component Lab Test Data
4	Integrated Component Lab Test Data
5	Lab-Scale System Test Data
6	Hardware-in-the-Loop and Software-in-the-Loop Data
7	Prototype Field Test Data
8	Live System Test Data
9	Operational Real-World Data

Source: STAT COE Paper (Stafford et al. 2024)

Scope coverage is assessed using the scope metric score between 0 and 1. The coverage metric is the product of  $C_v$ , the volume coverage metric, and  $C_D$ , the density coverage metric. These metrics are conceptually pictured in Figure 4-21, with mathematical formulations provided in the

STAT COE Paper, “Model Validation Levels: Methods and Implementation” (Stafford et al. 2024). Alternate formulations allow scope coverage to be calculated when not all factors are continuous.



Source: STAT COE Paper (Stafford et al. 2024)

**Figure 4-21. Volume Coverage Versus Density Coverage for Continuous Factors**

The metrics assessing fidelity, referent authority, and scope coverage are then quantitatively combined across different factor combinations to produce a continuously scaled 0 to 9 score for the model. The MVL is interpreted by mapping the score to the authority ranking in . For example, an MVL of 5.1 would indicate that the model is slightly more trustworthy than lab-scale system test data. The MVL can be no higher than the level of the most authoritative referent since model authority is inherited from the referent(s) used to validate it. Depending on the model’s intended use, different MVLs may be acceptable for different use cases (e.g., to plan testing versus to make an operational assessment of the system). In addition to the MVL, many lower-level scores can be calculated, such as average fidelity, average authority level, and coverage, to further understand model validity and identify areas for increasing model trust.

For further details on the MVL framework, see the STAT COE Paper (Stafford et al. 2024). An R package is available upon request (contact STAT COE at [AFIT.ENS.STATCOE@us.af.mil](mailto:AFIT.ENS.STATCOE@us.af.mil)) that automates the MVL calculation given model data, referent data, and the scope of intended use (Provost et al. 2024; Jones et al. 2024).

## 5 Conclusion

This guidebook provides project leadership and T&E practitioners with the background to understand when M&S can be helpful in T&E and why a rigorous VV&A is needed before using M&S output to support program decisions. This guidebook is not prescriptive; rather, it provides recommendations and best practices for how to accomplish a rigorous VV&A based on STAT. Where practical, this guidebook provides examples of the recommended techniques. However, this guidebook is not intended to serve as the definitive, exhaustive reference for all possible VV&A methods. Therefore, an extensive list of references is provided for practitioners to get further information about the recommendations made in this guidebook.

M&S should be viewed as a tool for improving T&E and therefore systems engineering and capability development. The composability of M&S compared to physical articles, especially combined with model-based engineering, may provide significant value. By focusing on creating validated and accredited M&S to support T&E, DoD can implement more Agile processes, as recommended in the GAO Report, “Leading Practices: Iterative Cycles Enable Rapid Delivery of Complex, Innovative Products,” and engage in a true campaign of learning in which T&E creates reusable data and models to support decision making throughout the capability life cycle.

The Office of the Director, Developmental Test, Evaluation, and Assessments (DTE&A) and the Office of the DOT&E intend to update this guidebook on a regular basis. Future versions will include more examples, including software code that can be downloaded. Additional topics planned for future updates include the following:

- Detailed verification methods.
- Additive models.
- GP models.
- Sensitivity analysis.
- Surrogate methods for UQ.
- Updated training information as the Defense Acquisition University (DAU) releases more credentials and courses.
- Subjective probability distributions elicited from SMEs.
- Federated model uncertainty propagation.

## Appendix A: Catalog of Popular Experiment Designs

This appendix outlines common classes of DOE designs.

### A1. Screening Designs

full factorial design. This class of design consists of two or more factors with a discrete number of levels that produces runs for all possible combinations of those levels across all factors.

Sometimes these designs are also referred to as  $2^k$ , where  $k$  represents the number of factors, and 2 represents the number of levels. For example, consider a full factorial design that consists of five factors with two levels each. The number of possible combinations in this type of design is  $32 (2^5)$ .

fractional factorial design. This design is constructed by systematically choosing a fraction of runs from a full factorial design to start the process of efficiently determining the relationship between the factors and responses. Fractional factorial designs are described mathematically as  $2^{k-p}_{Res}$  where  $k$  refers to the number of factors,  $1/2^p$  refers to the fraction, 2 represents the number of levels, and  $Res$  refers to the ability of the design to reveal main effects and 2FIs. For example, a  $2^{5-1}$  is a  $1/2$  fraction of the  $2^5$  full factorial that requires only 16 runs.

optimal design. This class of design is a good option whenever it is inadequate to use classical designs. The design is generated from computer algorithms that could be optimal with respect to a single criterion for a specified statistical model, but they could be suboptimal according to another criterion. The designs are model dependent and may require a model that the user may not have. The efficiency of these designs depends on the number of factors, the number of points, and the maximum standard error for prediction over the design space. Typically, the best design for an application is the design with the highest optimality efficiency. The designs have designations corresponding to the letters of the alphabet. The most popular are the D-optimal design, which is mostly used for screening, and the I- (or G-) optimal design, which is used for response surface. Other designs include the A-, V-, and E-optimal designs:

- D-optimal: Minimizes the generalized variance of the model regression coefficients.
- G-optimal: Minimizes the maximum scaled prediction variance over the design region.
- I-optimal: Minimizes the average scaled prediction variance over the design region.
- A-optimal: Minimizes the average variance of the regression coefficients.
- V-optimal: Minimizes the average prediction variance over a set of specific  $m$  points of interest in the design region.
- E-optimal: Maximizes the minimum eigenvalue of the information matrix.

definitive screening design (DSD). DSD is a class of three-level screening designs for numeric  $k > 5$  that provide estimates of the main effects that are uncorrelated with 2FIs and pure quadratic terms. Because they are three-level designs, the quadratic effects are estimable (as discussed later in this section). DSDs require  $2^{k+1}$  runs. 2FIs are only partially confounded with other 2FIs as opposed to other screening designs in which 2FIs are completely confounded with other 2FIs. Pure quadratic effects are not completely confounded with interactions.

Oehlert and Whitcomb minimum run Resolution IV (MR Res IV) design. This is a class of economical designs that offers savings over the  $2^{k-p}$  fractional factorial designs. Like all minimum run designs, MR Res IV designs are extremely sensitive to missing data.

general factorial design. This is a class of design where there are  $a$  levels of factor A,  $b$  levels of factor B, and  $c$  levels of factor C to produce a total number of  $a*b*c$  combinations.

## A2. Response Surface Designs

central composite design (CCD). CCD is a five-level design that consists of a combination of  $2^k$  factorial design or  $2^{k-p}$  fractional factorial design, center points, and axial points located a distance  $\alpha$  from the center of the design. The  $2^k$  or  $2^{k-p}$  designs provide for estimating first-order effects, 2FIs, and lack of fit if there is replication. The center points provide for estimating pure error, determining the presence of curvature, and making a more uniform estimation of the prediction variance. The axial points provide for estimating second-order effects. The position of the axial points produces designs for spherical regions or designs for cuboidal regions of space. The CCD is the workhorse of response surface methodology.

Box-Behnken design (BBD). BBD is a three-level design for fitting response surfaces. The construction technique relies on using balanced incomplete block designs and  $2^k$  factorial designs. The design avoids the corners of the design space in favor of edge points located at the mid-level ( $x_i = 0$ ) of the factor levels, which results in poor estimation at the factorial point locations. Thus, a BBD is more useful for situations in which there is no interest in predicting at the factorial points of the cube. Similar to a CCD, replicated runs at the center points permit a more uniform estimation of the prediction variance over the design space. Some BBDs with four or fewer factors are more economical than their equivalent CCDs. Practitioners often associate the BBD with cuboidal regions because of its cubic appearance; however, the BBD is a spherical design. BBDs require sufficient center points to improve their prediction accuracy.

$3^k$  general factorial design. This is a three-level factorial design that allows for the estimation of curvature. However, this class of design is generally inadequate because it requires a large sample size.

space-filling design. This is a class of design that aims at filling the test space with as few gaps as possible. These designs are suitable for computer simulations and other situations in which the response is deterministic.

Draper-Lin design. This is a minimal point design for  $k = 3, 5, 6, 7$ , and 10. These designs share a common structure with other small composite designs.

Koshal design. This is a class of design that uses exactly the same number of runs as the number of model coefficients to be estimated.

### **A3. Small Response Surface Designs**

There are situations in which scarce resources—funds, time, material, manpower, equipment—make it impractical to allow the use of traditional designs for fitting second-order models, especially when the number of factors  $k$  is high. For those situations, small or saturated response surface designs could be attractive because they are more economical. Some of the most popular small or saturated response surface designs and their features are listed below.

Hoke design. This saturated, nonorthogonal, second-order design is based on irregular fractions of partially balanced  $3^k$  factorial designs (for  $k > 2$ ). Hoke designs are suitable for cuboidal regions. For a small number of factors ( $k < 7$ ), some of the designs are near-saturated and permit estimating pure error and lack of fit. Except for the quadratic terms, the best Hoke designs (for  $2 < k < 10$ ) are comparable to BBDs and Hartley small composite designs based on the determinant and the trace of the information matrix. Hoke designs have better prediction performance and more similar sample size than Hartley small composite designs.

Doehlert uniform shell design. Uniform shell designs (for  $k < 11$ ) have an equally spaced distribution of points lying on concentric spherical shells.

Roquemore hybrid design. This is a set of saturated or near-saturated second-order rotatable or near-rotatable designs for  $k = 3, 4, 5$ , and 6. Roquemore hybrid designs are competitive with CCDs based on the scaled prediction variance criteria.

Box and Draper minimum point design. This class of design fits main effects plus interaction models that are optimal for  $k = 2$  and 3 but not optimal for  $k > 3$ .

Hartley small composite second-order design. This design is based on the idea that the cube portion of the composite design can be as low as a Resolution III fraction if the 2FIs are not aliased with other 2FIs.



Westlake cuboidal design. This is a second-order cuboidal design for  $k = 5, 7$ , and  $9$  in  $22, 40$ , and  $62$  runs that expands the idea from Hartley.

Rechtschaffner dummy factor design. This design adds a dummy factor, whose main effect and 2FIs are assumed to be zero, to saturated fractions of  $2^k$  and  $3^k$  factorial Resolution V designs to increase the degrees of freedom for estimating pure error and preserving the balanced structure of the design.

Pesotchinsky minimum point D-optimal design. This design is for  $k < 8$ .

Lucas saturated D-optimum composite design. This saturated D-optimum composite design uses a subset of points from the saturated Resolution V of the Rechtschaffner dummy factor design.

Mitchell and Bayne exchange algorithm design. This design uses an algorithm to find a  $k$ -run design that maximizes the information matrix  $|X'X|$  given the number of factors  $k$ , a specified model, and a set of candidate points.

Notz minimal point asymptotic D-efficiency design. This is a class of  $3^k$  (for  $k < 7$ ) minimal point second-order design with asymptotic D-efficiency of 1 relative to the number of factors  $k$  and the number of minimal points  $q$ .

Draper near-saturated design. This cuboidal, near-saturated design uses the Plackett-Burman design for the factorial portion of Hartley's small composite design, which is a compromise between a saturated small composite design and a CCD allowing degrees of freedom for estimating lack of fit.

Morris augmented-pairs design. This is a class of three-level design constructed by combining the levels of every pair of points in a two-level first-order design to form the third level.

Oehlert and Whitcomb minimum run Resolution V (MR Res V) design. MR Res V is a class of equireplicated, irregular fractions of  $2^k$  designs constructed using the D-optimality criterion algorithm. This design provides Resolution V designs in fewer runs when regular fraction designs contain significantly more degrees of freedom than are needed to estimate the model up to 2FIs. Judged only on a D-optimality criterion, these designs are more efficient than many other types.

Haines' San Cristobal design. This design fits a quadratic response surface for  $k$  factors that are restricted to nonnegative levels.

Gilmour subset design. This is a class of three-level response surface design obtained by using subsets of  $2^k$  factorial designs at levels of  $-1$  and  $1$  for each combination of  $k$  factors while holding the other  $q - k$  factors at their middle level.

## Appendix B: Policy and Guidance for M&S in T&E

Table B-1 lists DoD policies and guidance that describe the use of M&S as part of the defense acquisition process. The 5000-series of policy documents pertaining to the Defense Acquisition System are directly relevant to this guidebook. These documents outline the principles for T&E within the Defense Acquisition System and how M&S can be used to inform programmatic decisions. DoDI 5000.61 is a key document that describes the VV&A process. MIL-STD-3022 is another core document that establishes a template for VV&A documents along with a framework for data sharing across stakeholders. Each of the policies describes roles and responsibilities but provides minimal detail about how to accomplish the overall program goals of using M&S to support system development.

**Table B-1. Top-Level DoD T&E Policy and Guidance**

Policy and Guidance	Title	Effective Date	Description
DoDD 5000.01	"The Defense Acquisition System"	September 9, 2020 July 28, 2022 (Change 1)	<ul style="list-style-type: none"> <li>• Provides guiding principles for the management of capability acquisitions.</li> <li>• Provides policies to support the Defense Acquisition System.</li> </ul>
DoDI 5000.02	"Operation of the Adaptive Acquisition Framework"	January 23, 2020 June 8, 2022 (Change 1)	<ul style="list-style-type: none"> <li>• Establishes policy and prescribes procedures for managing acquisition programs.</li> <li>• Assigns acquisition program management responsibilities.</li> <li>• Describes the responsibilities of principal acquisition officials and the purpose and key characteristics of the acquisition pathways.</li> <li>• Restructures defense acquisition guidance to improve process effectiveness and implement the Adaptive Acquisition Framework.</li> </ul>

Policy and Guidance	Title	Effective Date	Description
DoDI 5000.61	“DoD Modeling and Simulation Verification, Validation, and Accreditation”	September 17, 2024	<ul style="list-style-type: none"> <li>Establishes policy, assigns responsibilities, and prescribes procedures for the VV&amp;A of models, simulations, distributed simulations, and associated data.</li> <li>Establishes the basis for credible M&amp;S across DoD.</li> </ul>
DoDI 5000.70	“Management of DoD Modeling and Simulation (M&S) Activities”	May 10, 2012 October 15, 2018 (Change 3)	<ul style="list-style-type: none"> <li>Implements DoDD 5000.59, which was canceled and incorporated into DoDI 5000.97.</li> <li>Assigns responsibilities for the DoD M&amp;S Steering Committee.</li> <li>Establishes the Director, DoD M&amp;S Coordination Office.</li> <li>Extends discovery metadata policy to key DoD M&amp;S tools, data, services, data assets, models, and simulations.</li> </ul>
DoDI 5000.88	“Engineering of Defense Systems”	November 18, 2020	Establishes policy, assigns responsibilities, and provides procedures to implement engineering of defense systems.
DoDI 5000.97	“Digital Engineering”	December 21, 2023	Establishes policy, assigns responsibilities, and provides procedures for implementing and using DE in the development and sustainment of defense systems.

Policy and Guidance	Title	Effective Date	Description
DoDI 5000.98	“Operational Test and Evaluation and Live Fire Test and Evaluation”	December 9, 2024	<ul style="list-style-type: none"> <li>Establishes policy, assigns responsibilities, and prescribes procedures for OT&amp;E and LFT&amp;E of DoD systems and services acquired via the Defense Acquisition System or via other nonstandard acquisition systems.</li> <li>Supersedes information regarding OT&amp;E and LFT&amp;E located in DoDI 5000.89.</li> </ul>
DoDM 5000.100	“Test and Evaluation Master Plans and Test and Evaluation Strategies”	December 9, 2024	Implements policy, assigns responsibilities, and provides procedures for developing OT&E and LFT&E input to the TEMP, a T&E strategy, or an equivalent product for DoD systems and services acquired via the Defense Acquisition System or via other nonstandard acquisition systems.
DoDM 5000.102	“Modeling and Simulation Verification, Validation, and Accreditation for Operational Test and Evaluation and Live Fire Test and Evaluation”	December 9, 2024	Implements policy, assigns responsibilities, and provides procedures for VV&A of M&S tools critical to meeting the OT&E and LFT&E objectives of DoD systems and services acquired via the Defense Acquisition System or via other nonstandard acquisition systems.
MIL-STD-3022	“Documentation of Verification, Validation, and Accreditation (VV&A) for Models and Simulations”	January 28, 2008 April 5, 2012 (Change 1)	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.

Policy and Guidance	Title	Effective Date	Description
M&S Enterprise Core Document	“Verification, Validation, and Accreditation (VV&A) Recommended Practices Guide (RPG)”	Current edition available at <a href="https://www.cto.mil/sea/vva_rpg/">https://www.cto.mil/sea/vva_rpg/</a>	Facilitates the application of DoD directives, policies, and guidelines to promote effective and efficient VV&A processes for the full spectrum of M&S products employed in DoD.
DOT&E TEMP Guidebook	“Director, Operational Test and Evaluation (DOT&E) Test and Evaluation Master Plan (TEMP) Guidebook”	January 19, 2017	Provides guidance for the content of the TEMP.
DoD M&S Glossary	“Department of Defense Modeling and Simulation (M&S) Glossary”	October 1, 2011	Provides definitions for common M&S terminology.

Table B-2 lists Service-level policies and Office of the Secretary of Defense (OSD) guidance for implementing M&S in acquisition programs. Most of these instructions are focused on how the VV&A process is implemented within each Service. These documents are a good reference for programs to further understand Service-level requirements for M&S in T&E. The Service guidance documents listed in Table B-2 recommend using statistical analysis for VV&A, but they typically do not provide specific examples of how to apply analytical techniques.

**Table B-2. Service and OSD Policy and Guidance on VV&A of M&S for T&E**

Policy and Guidance	Title	Effective Date
Department of the Army Pamphlet 5-11	“Verification, Validation, and Accreditation of Army Models and Simulations”	September 30, 1999
Army Regulation 5-11	“Management of Army Modeling and Simulation”	May 30, 2014
SECNAVINST 5200.46	“Department of the Navy Modeling, Simulation, Verification, Validation, and Accreditation Management”	March 7, 2019
OPTEVFORINST 5000.1D	“Use of Modeling and Simulation in Operational Test”	March 8, 2022
OPNAVINST 3960.15B	“Validation of Navy Threat Simulators, Targets and Digital Threat Models and Simulations”	July 27, 2017
Air Force Instruction 16-1001	“Verification, Validation and Accreditation (VV&A)”	April 29, 2020
Air Force Instruction 16-1005	“Modeling & Simulation Management”	June 23, 2016

Policy and Guidance	Title	Effective Date
OSD M&S Guidance for the Acquisition Workforce	"Modeling and Simulation Guidance for the Acquisition Workforce"	October 2008

## Appendix C: Training

Refer to Service-level policy and guidance in Appendix B for workforce development requirements. In addition, the following DAU credentials and courses are recommended because they contain content relevant to M&S development, testing, and use in T&E:

- CTST 001: Applying Scientific Test & Analysis Techniques Credential (in development)
- CTST 002: Cyber T&E Fundamentals Credential (in development)
- CLE 023: Modeling and Simulation in Test and Evaluation
- CLE 084: Models, Simulations, and Digital Engineering
- CENG 001: Digital Engineering for DoD Consumers Credential

For more information, see the DAU Website (<https://www.dau.edu/>).

STAT methods should underpin the M&S VV&A strategy, and these topics are covered in more detail in Sections 2.2.1 and 4 of this guidebook. The Air Force Institute of Technology STAT COE provides short classes on some of these topics; course offerings can be found at <https://www.afit.edu/STAT/page.cfm?page=1093>. IDA offers online courses and interactive tools for test design and analysis on the IDA Test Science Website (<https://testscience.org/>).

## Acronyms

2FI	two-factor interaction
ANOVA	analysis of variance
BBD	Box-Behnken design
CCD	central composite design
CDF	cumulative distribution function
CI	confidence interval
CM	countermeasure
COE	center of excellence
CORA	correlation analysis
DAU	Defense Acquisition University
DE	digital engineering
DoD	Department of Defense
DoDD	DoD directive
DoDI	DoD instruction
DoDM	DoD manual
DOE	design of experiments
DOT&E	Director, Operational Test and Evaluation
DSD	definitive screening design
DT	developmental test
DT&E	developmental test and evaluation
DTE&A	Developmental Test, Evaluation, and Assessments
EDA	exploratory data analysis
FDA	functional data analysis
FFF	fast flexible filling
FPCA	functional principal component analysis
GAM	generalized additive model
GAO	Government Accountability Office
GP	Gaussian process



## Acronyms

IDA	Institute for Defense Analyses
IDSK	Integrated Decision Support Key
IPO	input-process-output
IR	infrared
KS	Kolmogorov-Smirnov
LFT&E	live fire test and evaluation
LH	Latin hypercube
M&S	modeling and simulation
MBSE	model-based systems engineering
MIL-STD	Military Standard
MR Res IV	minimum run Resolution IV
MR Res V	minimum run Resolution V
MSWG	Modeling and Simulation Working Group
MVL	Model Validation Level
MWS	missile warning system
NIST	National Institute of Standards and Technology
OPNAVINST	Office of the Chief of Naval Operations instruction
OPTEVFORINST	Operational Test and Evaluation Force instruction
OSD	Office of the Secretary of Defense
OT&E	operational test and evaluation
OUSD(R&E)	Office of the Under Secretary of Defense for Research and Engineering
PDF	probability density function
PI	prediction interval
PM	program manager
PPV	predicted probability validation
SECNAVINST	Secretary of the Navy instruction
SME	subject matter expert
SNR	signal-to-noise ratio
SPRT	sequential probability ratio test

## Acronyms

STAT	scientific test and analysis techniques
T&E	test and evaluation
TEMP	Test and Evaluation Master Plan
TI	tolerance interval
UQ	uncertainty quantification
V&V	verification and validation
VV&A	verification, validation, and accreditation

## References

- Air Force Instruction 16-1001, “Verification, Validation and Accreditation (VV&A),” April 29, 2020.
- Air Force Instruction 16-1005, “Modeling & Simulation Management,” June 23, 2016.
- Army Regulation 5-11, “Management of Army Modeling and Simulation,” May 30, 2014.
- Baty, R.S., et al. “Enhanced Surface-to-Air Missile Simulation (ESAMS) Computer Program - Analyst Manual, Basic Methodology.” Flight Dynamics Laboratory, Air Force Wright Aeronautical Laboratories, Air Force Systems Command, Wright-Patterson Air Force Base, 1988.
- Beling, Peter, Barry Horowitz, and Tom McDermott. “WRT-1022: Developmental Test and Evaluation (DTE&A) and Cyberattack Resilient Systems.” Final Technical Report: SERC-2021-TR-015 (v2). Systems Engineering Research Center, September 14, 2021.
- Boehm, Barry. “WRT 1016: Reducing Total Ownership Cost (TOC) and Schedule.” Final Technical Report SERC-2021-TR-009. Systems Engineering Research Center, January 17, 2021.
- Box, George E. P. “Science and Statistics.” *Journal of the American Statistical Association* 71(356): 791–799, 1976.
- Boyle, Edward. “LCOM Explained.” AFHRL Technical Paper 90-58. Logistics and Human Factors Division, Air Force Human Resources Laboratory, Wright-Patterson Air Force Base, July 1990.
- Burke, Sarah. “Model Building Process Part 1: Checking Model Assumptions V 1.1.” STAT COE-Report-09a-2017. Scientific Test and Analysis Techniques Center of Excellence, October 24, 2017, as amended.
- Cioppa, Thomas M., and Thomas W. Lucas. “Efficient Nearly Orthogonal and Space-Filling Latin Hypercubes.” *Technometrics* 49(1): 44–55, 2007.
- Cleveland, William S. “Robust Locally Weighted Regression and Smoothing Scatterplots.” *Journal of the American Statistical Association* 74(368): 829–836, 1979.
- Coleman, David E., and Douglas C. Montgomery. “A Systematic Approach to Planning for a Designed Industrial Experiment.” *Technometrics* 35(1): 1–12, 1993.
- Cortes, Luis A., and Francisco Ortiz. “Modern Test Design and Analysis Playbook.” MITRE Technical Report MTR 200387. The MITRE Corporation, August 19, 2020.
- Cortes, Luis A., Melissa Kim Wong, Angel Cortes-Morales, David Wells, and Jason Daly. “Advance M&S in Acquisition T&E.” MITRE Technical Report MTR210454. The MITRE Corporation, November 19, 2021.
- Damblin, Guillaume, Mathieu Couplet, and Bertrand Iooss. “Numerical Studies of Space-Filling Designs: Optimization of Latin Hypercube Samples and Subprojection Properties.” *Journal of Simulation* 7(4): 276–289, 2013.  
<https://doi.org/10.1057/jos.2013.16>
- Department of Defense Modeling and Simulation (M&S) Glossary. Modeling and Simulation Coordination Office, October 1, 2011.

## References

- Department of the Army Pamphlet 5-11, “Verification, Validation, and Accreditation of Army Models and Simulations,” September 30, 1999.
- Director, Operational Test and Evaluation (DOT&E) Test and Evaluation Master Plan (TEMP) Guidebook, Version 3.1, January 19, 2017.
- DoD Directive 5000.01, “The Defense Acquisition System,” September 9, 2020, as amended.
- DoD Instruction 5000.02, “Operation of the Adaptive Acquisition Framework,” January 23, 2020, as amended.
- DoD Instruction 5000.61, “DoD Modeling and Simulation Verification, Validation, and Accreditation,” September 17, 2024.
- DoD Instruction 5000.70, “Management of DoD Modeling and Simulation (M&S) Activities,” May 10, 2012, as amended.
- DoD Instruction 5000.88, “Engineering of Defense Systems,” November 18, 2020.
- DoD Instruction 5000.89, “Test and Evaluation,” November 19, 2020.
- DoD Instruction 5000.97, “Digital Engineering,” December 21, 2023.
- DoD Instruction 5000.98, “Operational Test and Evaluation and Live Fire Test and Evaluation,” December 9, 2024.
- DoD Instruction 5000.DT, “Developmental Test and Evaluation,” currently under development by OUSD(R&E)/DTE&A to be approved by the USD(R&E) and published on the DoD Issuances Website.
- DoD Manual 5000.100, “Test and Evaluation Master Plans and Test and Evaluation Strategies,” December 9, 2024.
- DoD Manual 5000.102, “Modeling and Simulation Verification, Validation, and Accreditation for Operational Test and Evaluation and Live Fire Test and Evaluation,” December 9, 2024.
- Efron, Bradley, and Robert J. Tibshirani. *An Introduction to the Bootstrap*. 1st ed. Chapman and Hall/CRC, 1994.
- Fang, Kai-Tai, Runze Li, and Agus Sudjianto. *Design and Modeling for Computer Experiments*. 1st ed. Chapman and Hall/CRC, 2005.
- Guidelines for Modelling and Simulation (M&S) Use Risk Identification, Analysis, and Mitigation. STO-TR-MSG-139. North Atlantic Treaty Organization Science and Technology Organization, September 2021.
- Haman, John T., Thomas H. Johnson, David Grimm, Kerry Walzl, and Lindsey Butler. “Predicted Probabilities Validation.” IDA Document D-33156. Institute for Defense Analyses, 2022.
- Haman, John T., and Curtis G. Miller. “Metamodeling Techniques for Verification and Validation of Modeling and Simulation Data.” IDA Paper P-33230. Institute for Defense Analyses, September 2022.
- Harrell, F. E., and K. L. Lee. “Using Logistic Model Calibration to Assess the Quality of Probability Predictions.” Division of Biometry, Duke University Medical Center, 1990.

- Hidiroglou, M. A. “Double Sampling.” *Survey Methodology* 27(2): 143–154, 2001.
- Johnson, Rachel T., Gregory T. Hutto, James R. Simpson, and Douglas C. Montgomery. “Designed Experiments for the Defense Community.” *Quality Engineering* 24(1): 60–79, 2012.
- Jones, Donald R., Matthias Schonlau, and William J. Welch. “Efficient Global Optimization of Expensive Black-Box Functions.” *Journal of Global Optimization* 13(4): 455–492, 1998.
- Jones, Nicholas, Kyle Provost, and Corinne Stafford. “Model Validation Level (MVL) R Tool User Guide.” Scientific Test and Analysis Techniques Center of Excellence, 2024.
- Kleijnen, Jack P.C. *Design and Analysis of Simulation Experiments*. 1st ed. Springer New York, NY, 2008.
- Konakli, Katerina, and Bruno Sudret. “Global Sensitivity Analysis Using Low-Rank Tensor Approximations.” *Reliability Engineering and System Safety* 156(3): 64–83, 2016.
- Krasner, Jerry. “How Product Development Organizations Can Achieve Long-Term Cost Savings Using Model-Based Systems Engineering (MBSE).” *Embedded Market Forecasters*, October 2015.
- Leading Practices: Iterative Cycles Enable Rapid Delivery of Complex, Innovative Products. GAO-23-106222. Government Accountability Office, July 2023.
- Li, Jiaming, Suhuai Luo, and Jesse S. Jin. “Sensor Data Fusion for Accurate Cloud Presence Prediction Using Dempster-Shafer Evidence Theory.” *Sensors* 10(10): 9384–9396, 2010.
- Loeppky, Jason L., Jerome Sacks, and William J. Welch. “Choosing the Sample Size of a Computer Experiment: A Practical Guide.” *Technometrics* 51(4): 366–376, 2009.
- Maupin, Kathryn A., and Laura P. Swiler. “Validation Metrics for Deterministic and Probabilistic Data.” Sandia Report SAND2016-1421. Sandia National Laboratories, January 2017.
- McDermott, Tom, and Eileen Van Aken. “Summary Report, Task Order WRT-1001: Digital Engineering Metrics.” Supporting Technical Report SERC-2020-SR-003. Systems Engineering Research Center, June 8, 2020.
- McElreath, Richard. *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. 2nd ed. Chapman and Hall/CRC, 2020.
- McGinnity, Kelly. “The Importance of M&S in Operational Testing and the Need for Rigorous Validation.” IDA Document NS D-5807. Institute for Defense Analyses, April 2016.
- Medlin, Rebecca M., Monica L. Ahrens, Keyla Pagán-Rivera, and John W. Dennis. “Case Study on Applying Sequential Analyses in Operational Testing.” IDA Document NS D-32904. Institute for Defense Analyses, December 2021.
- Military Standard MIL-STD-3022, “Documentation of Verification, Validation, and Accreditation (VV&A) for Models and Simulations,” January 28, 2008, as amended.
- Modeling and Simulation Guidance for the Acquisition Workforce. Version 1.01. Office of the Deputy Under Secretary of Defense for Acquisition and Technology, Systems and Software Engineering, Developmental Test and Evaluation, October 2008.

- Montgomery, Douglas C. *Design and Analysis of Experiments*. 8th ed. John Wiley & Sons, Inc., 2013.
- Morgan, Millet Granger, and Max Henrion. *Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis*. Cambridge University Press, 1990.
- Nagelkerke, N. J. D. “A Note on a General Definition of the Coefficient of Determination.” *Biometrika* 78(3): 691-692, 1991.
- Natoli, Cory. “Understanding Analysis of Variance: Best Practice.” STAT COE-Report-29-2017. Scientific Test and Analysis Techniques Center of Excellence, December 21, 2017, as amended.
- Office of the Chief of Naval Operations Instruction 3960.15B, “Validation of Navy Threat Simulators, Targets and Digital Threat Models and Simulations,” July 27, 2017.
- Ogilvie, John F. “A Monte-Carlo Approach to Error Propagation.” *Computers and Chemistry* 8(3): 205–207, 1984.
- Operational Test and Evaluation Force Instruction 5000.1D, “Use of Modeling and Simulation in Operational Test,” March 8, 2022.
- Ortiz, Francisco, and Lenny Truett. “Using Statistical Intervals to Assess System Performance.” STAT COE-Report-04-2015. Scientific Test and Analysis Techniques Center of Excellence, April 17, 2015, as amended.  
[https://www.afit.edu/stat/statcoe\\_files/Using%20Statistical%20Intervals%20to%20Assess%20System%20Performance%20V2.pdf](https://www.afit.edu/stat/statcoe_files/Using%20Statistical%20Intervals%20to%20Assess%20System%20Performance%20V2.pdf)
- Provost, Kyle, Corinne Stafford, and Nicholas Jones. MVL R Tool. Scientific Test and Analysis Techniques Center of Excellence, 2024.
- Ramert, Aaron, and Emily Westphal. “Equivalence Testing.” STAT COE-Report-12-2020. Scientific Test and Analysis Techniques Center of Excellence, August 28, 2020.
- Ramsay, J. O., and B. W. Silverman. *Functional Data Analysis*. 2nd ed. Springer New York, NY, 2005.
- Roy, Christopher J., and William L. Oberkampf. “A Comprehensive Framework for Verification, Validation, and Uncertainty Quantification in Scientific Computing.” *Computer Methods in Applied Mechanics and Engineering* 200(25–28): 2131–2144, 2011.
- Santner, Thomas J., Brian J. Williams, and William I. Notz. 1st ed. *The Design and Analysis of Computer Experiments*. Springer New York, NY, 2003.
- Secretary of the Navy Instruction 5200.46, “Department of the Navy Modeling, Simulation, Verification, Validation, and Accreditation Management,” March 7, 2019.
- Smith, Ralph C. *Uncertainty Quantification: Theory, Implementation, and Applications*. Society for Industrial and Applied Mathematics (SIAM), 2013.
- Spiegelhalter, David J. “Probabilistic Prediction in Patient Management and Clinical Trials.” *Statistics in Medicine* 5(5): 421–433, 1986.

- Stafford, Corinne, Kyle Provost, and Nicholas Jones. “Model Validation Levels: Methods and Implementation.” Scientific Test and Analysis Techniques Center of Excellence, February 2024.  
[https://www.afit.edu/docs/Model%20Validation%20Levels\\_Methods%20and%20Implementation.pdf](https://www.afit.edu/docs/Model%20Validation%20Levels_Methods%20and%20Implementation.pdf)
- Standard for Verification and Validation in Computational Solid Mechanics. ASME V&V 10-2019. The American Society of Mechanical Engineers, 2020.  
<https://www.asme.org/codes-standards/find-codes-standards/standard-for-verification-and-validation-in-computational-solid-mechanics>
- Stein, Charles. “A Two-Sample Test for a Linear Hypothesis Whose Power Is Independent of the Variance.” *The Annals of Mathematical Statistics* 16(3): 243–258, 1945.
- Storm, Scott M., Raymond R. Hill, Joseph J. Pignatiello, G. Geoffrey Vining, and Edward D. White. “Model Validation of Functional Responses Across Experimental Regions Using Functional Regression Extensions to the CORA Objective Rating System.” *Journal of Verification, Validation and Uncertainty Quantification* 2(4): 041004, 2017.
- Thomas, Dean, and Rebecca Dickinson. “Validating the PRA Testbed Using a Statistically Rigorous Approach.” IDA Document NS D-5445. Institute for Defense Analyses, 2015.
- van Opheusden, Bas, Luigi Acerbi, and Wei Ji Ma. “Unbiased and Efficient Log-Likelihood Estimation with Inverse Binomial Sampling.” *PLoS Computational Biology* 16(12), 2020.
- Verification, Validation, and Accreditation (VV&A) Recommended Practices Guide (RPG). Office of the Under Secretary of Defense for Research and Engineering, Systems Engineering and Architecture, current edition.  
[https://www.cto.mil/sea/vva\\_rpg/](https://www.cto.mil/sea/vva_rpg/)
- Wald, Abraham. “Sequential Tests of Statistical Hypotheses.” *The Annals of Mathematical Statistics* 16(2): 117–186, 1945.
- Weapon Systems Annual Assessment: Knowledge Gaps Pose Risks to Sustaining Recent Positive Trends. GAO-18-360SP. Government Accountability Office, April 2018.
- Whitehurst, Robert, Jane Phipps, and Victor Kowalenko. “A Programmer’s Reference to the Suppressor Simulation System.” AR-010-154, DSTO-GD-0130. Defence Science and Technology Organisation (DSTO) Aeronautical and Maritime Research Laboratory (Commonwealth of Australia), February 1997.
- Wojton, Heather, and Kelly Avery. “Statistical Design & Analysis Challenges in Defense Testing.” IDA Non-Standard Document NS D-9225. Institute for Defense Analyses, January 2019.
- Wojton, Heather, Kelly Avery, Han Yi, and Curtis Miller. “Space Filling Designs for Modeling & Simulation Validation.” IDA Document NS D-21562. Institute for Defense Analyses, June 2021.
- Wojton, Heather, Kelly M. Avery, Laura J. Freeman, Samuel H. Parry, Gregory S. Whittier, Thomas H. Johnson, and Andrew C. Flack, “Handbook on Statistical Design & Analysis Techniques for Modeling & Simulation Validation.” IDA Document NS D-10455. Institute for Defense Analyses, February 2019.

- Wojton, Heather, Rebecca Medlin, John Dennis, Keyla Pagan-Rivera, and Leonard Wilkins. “A Review of Sequential Analysis.” IDA Document NS D-20487. Institute for Defense Analyses, December 2020.
- Xing, Eric P. “Advanced Gaussian Processes.” Lecture Notes, Course 10-708: Probabilistic Graphical Models. Carnegie Mellon University School of Computer Science, Spring 2015. [https://www.cs.cmu.edu/~epxing/Class/10708-15/notes/10708\\_scribe\\_lecture21.pdf](https://www.cs.cmu.edu/~epxing/Class/10708-15/notes/10708_scribe_lecture21.pdf)
- Xiu, Dongbin. *Numerical Methods for Stochastic Computations: A Spectral Method Approach*. Princeton University Press, 2010.
- Xiu, Dongbin. “Stochastic Collocation Methods: A Survey.” In *Handbook of Uncertainty Quantification*, edited by Roger Ghanem, David Higdon, and Houman Owhadi. Springer, Cham, pp. 1–18, 2015.

## Websites

- CRAN Task View: Design of Experiments (DoE) & Analysis of Experimental Data.  
<https://CRAN.R-project.org/view=ExperimentalDesign>
- DAU Website.  
<https://www.dau.edu/>
- DoD Issuances.  
<https://www.esd.whs.mil/DD/DoD-Issuances/>
- IDA Test Science.  
<https://testscience.org/>
- JMP Statistical Discovery, Fast Flexible Filling Design Details.  
<https://www.jmp.com/support/help/en/18.0/index.shtml#page/jmp/fast-flexible-filling-design-details.shtml#>
- NIST Engineering Statistics Handbook.  
<https://www.itl.nist.gov/div898/handbook/>
- Sandia National Laboratories Dakota Project.  
<https://dakota.sandia.gov/>
- STAT COE.  
<https://www.afit.edu/STAT/>
- VV&A Recommended Practices Guide.  
[https://www.cto.mil/sea/vva\\_rpg/](https://www.cto.mil/sea/vva_rpg/)



## **Modeling and Simulation for Developmental Test and Evaluation Guidebook**

Office of the Director, Developmental Test, Evaluation, and Assessments  
Office of the Under Secretary of Defense for Research and Engineering  
3030 Defense Pentagon  
Washington, DC 20301  
[osd.r-e.comm@mail.mil](mailto:osd.r-e.comm@mail.mil)  
<https://www.cto.mil/dtea>

Office of the Director, Operational Test and Evaluation  
1700 Defense Pentagon  
Washington, DC 20301-1700  
<https://www.dote.osd.mil/>

Distribution Statement A. Approved for public release. Distribution is unlimited.