INSTITUTE FOR DEFENSE ANALYSES

Handbook on Statistical Design & Analysis Techniques for Modeling & Simulation Validation

Heather Wojton, Project Leader

Kelly M. Avery
Laura J. Freeman
Samuel H. Parry
Gregory S. Whittier
Thomas H. Johnson
Andrew C. Flack

February 2019

Approved for public release.
Distribution is unlimited.

IDA Document NS D-10455

Log: H 2019-000044

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

Approved for public release; distribution is unlimited.

The Institute for Defense Analyses is a non-profit corporation that operates three federally funded research and development centers to provide objective analyses of national security issues, particularly those requiring scientific and technical expertise, and conduct related research on other national challenges.

About This Publication
This work was conducted by the Institute for Defense Analyses (IDA) under contract HQ0034-14-D-0001, Task BD-9-2299(90), "Test Science Applications," for the Office of the Director, Operational Test and Evaluation. The views, opinions, and findings should not be construed as representing the official position of either the Department of Defense or the sponsoring organization

Acknowledgments
The IDA Technical Review Committee was chaired by Mr. Robert R. Soule and consisted of Denise J. Edwards, Douglas A. Peek, Jason P. Sheldon, Robert M. Hueckstaedt, Sabrina Lyn Hiner Dimassimo, and William G. Gardner from the Operational Evaluation Division, and Nathan Platt and James M. Gilmore from the Systems Evaluation Division. Thanks also to Chris Henry, Jane Pinelis, Dave Beyrodt, Stargel Doane, Doug Ray, and Simone Youngblood for their valuable contributions to this project.

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

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

INSTITUTE FOR DEFENSE ANALYSES

IDA Document NS D-10455

Handbook on Statistical Design & Analysis Techniques for Modeling & Simulation Validation

Heather Wojton, Project Leader

Kelly M. Avery
Laura J. Freeman
Samuel H. Parry
Gregory S. Whittier
Thomas H. Johnson
Andrew C. Flack

Approved for public release; distribution is unlimited.

Approved for public release; distribution is unlimited.

Executive Summary

Evaluations of system operational effectiveness, suitability, and survivability increasingly rely on models to supplement live testing. In order for these supplements to be valuable, testers must understand how well the models represent the systems or processes they simulate. This means testers must quantify the uncertainty in the representation and understand the impact of that uncertainty.

Two broad categories of uncertainties (statistical and knowledge) are of central importance to test and evaluation (T&E), particularly as testers try to extrapolate the model output and live test data into predictions of performance in combat. The validation process should include parametric analyses and a comparison of simulation output to live data to support quantification of statistical uncertainty. However, qualitative and non-statistical techniques may also be required to compare the future hypothetical combat environment with the non-quantitative portions of the validation referent.

A model's intended use stipulates the fidelity and type of model necessary. Intended use is based on the model's overarching purpose (model hierarchy); the quantities of interest for evaluation (response variables); the range of input conditions for the model; the range of input conditions over which experimental data (live testing) can be collected; and the acceptability criteria (typically stated as an allowable difference between the model and live data). Determining which uncertainty and how much uncertainty matters for a modeled system or process requires detailed understanding of the model, the system under test, and the model's specific intended use

The goal of this handbook is to aid the T&E community in developing test strategies that support data-driven model validation and uncertainty quantification. Chapter 2 of the handbook discusses the overarching steps of the verification, validation, and accreditation (VV&A) process as it relates to operational testing. Chapter 3 describes analytical methods for analyzing the simulation itself, making comparisons with the live data, and quantifying the associated uncertainties. Chapter 4 outlines design of experiment techniques and their application to both live and simulation environments.

Process

The purpose of this chapter is to outline in detail the VV&A process for using models and simulations to augment operational test and evaluation (OT&E). It also provides general principles for achieving a meaningful accreditation.

Approved for public release; distribution is unlimited.

Once the T&E community has determined that a model or simulation is required to support operational evaluation, then the VV&A process can commence. This process is comprised of nine steps:

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.

The successful implementation of this process is contingent on the tester, modeler, and user communities working together and communicating early and often. The VV&A strategy, including the associated acceptability criteria, has routinely been developed too late (or not at all) for operational evaluations. This practice creates unacceptable risk that the delivered models will not support their intended use for operational evaluations, or that the intended use will need to be significantly limited from that originally planned.

Statistical concepts and methodologies can be incorporated throughout this VV&A process. Table 1 correlates statistical ideas to the appropriate steps in the V&V process.

Table 1. Correlating Statistical Concepts to the VV&A Process

VV&A Information	Process Step	Statistical Concepts
M&S Requirements and Intended Use	1 – 4	Response variables Factor space Stochastic vs. deterministic models
Data Requirements	5	Design for Computer Experiments (Chapter 3) Classical Design of Experiments (Chapter 3)
Iterate Model-Test-Model	6	Design for Computer Experiments (Chapter 4) Classical Design of Experiments (Chapter 4) Variation Analysis & Statistical Emulation (Chapter 3) Comparison to M&S runs (Chapter 3) Calibration (Chapter 4)
Verification Analysis	7	Parametric Analysis (Chapter 3)
Validation Analysis	8	Parametric Analysis (Chapter 3) Comparison of live and M&S data (Chapter 3)
Quantify Uncertainty	8 – 9	Hypothesis Testing and Interval Estimation (Chapter 3)

Analysis

Statistical analyses can and should inform VV&A decision makers by providing information about model performance across the input space and by identifying risk areas. In addition to discussing inherent limitations of the model in terms of knowledge uncertainty, a good validation analysis should characterize the performance of the model across the operational space and quantify the statistical uncertainty associated with that performance. Inappropriate analysis techniques, such as those that average or roll up information despite the data being collected in distinct operational conditions, can lead to incorrect conclusions.

Evaluations should ultimately be based on both statistical and operational knowledge. Subject matter expertise is critical for answering the question, "Do the identified statistical differences actually make a practical difference?" Accreditation reports should use all relevant information to discuss what the model is useful for and what it is not.

After thoroughly exploring and visualizing the data, testers should evaluate the simulation on its own (to include sensitivity analysis and statistical emulators) and compare simulation results with the live test data (external validation). Sensitivity analysis can refer to either large changes in the inputs to build parametric emulators (parametric analysis) or small changes in inputs to look for bugs in code and reasonable perturbations in outputs. In this context, sensitivity analysis is used to determine how different values of an input variable affect a particular simulation output variable. Statistical emulators, also known as meta-models, use simulation output from across a set of conditions (ideally a designed experiment) to build a statistical model. An emulator can be used to estimate uncertainty and predict the output of the simulation at both tested and untested conditions.

The most appropriate method for statistically comparing live data and simulated output will depend on a variety of circumstances. There is no one-size-fits-all solution. In some cases, it may be useful or necessary to apply multiple techniques in order to fully understand and describe the strengths and weaknesses of the model. Generally speaking,

statistical validation techniques that account for possible effects due to factors are preferred over one-sample averages or roll-ups across conditions. If a designed experiment was executed in the live environment, an appropriate statistical modeling approach should be used to comparing the live data and simulated output. In all cases, even if no factors are identified and a one-sample approach is taken, it is crucial that the uncertainty about the difference between live data and simulated output be quantified. Hypothesis tests and confidence intervals are simple ways to quantify statistical uncertainty. Table 2 shows our recommended validation analysis methods based on response variable distribution, factor structure, and sample size for live testing.

Table 2. Recommended Analysis Methods

		Recommended Method by Sample Size		
Distribution	**Factors**	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

Design

Design of Experiments (DOE) provides a defensible strategy for selecting data from live testing and simulation experiments to support validation needs. DOE characterizes the relationship between the factors (inputs) and the response variable (output) of a process and is a common technique for planning, executing, and analyzing both developmental and operational tests.

The most powerful validation analysis techniques require coordination between the designs of the physical and computer experiments. Classical DOE (used for live tests) and computer experiments provide the building blocks for conducting validation. Even though they are often grouped into "DOE" as a whole, their design principles, algorithms for

generating designs, and corresponding analysis techniques can be quite different. Understanding these differences is crucial to understanding validation experiments.

A robust validation strategy will include a combination of classical DOE and computer experiment techniques. Computer experiments employ space-filling designs, which cover the model input domain. Classical design can be used for selecting model runs for replication and for matching points to live tests. When combining simulation designs and classical designs into the overall validation process, there are numerous reasonable implementations. The following process is one potential execution of the hybrid approach that has worked in practice. It assumes the simulation is available before live tests. In this case, the validation process might proceed as follows:

1. Conduct a computer experiment on all model input variables.
2. Add replicates to a subset of the simulation runs for Monte Carlo variation analysis.
3. Conduct Monte Carlo variation analysis.
4. Conduct parametric analysis. Evaluate the simulation experiment using emulator/interpolator.
5. Determine important factors and areas for investigation for live testing.
6. Design live tests using classical DOE, record all other relevant variables not included in the design, and include replicates if feasible.
7. Run live test design in the simulator, set additional factors at values realized during live tests, and include replications if simulation is non-deterministic.

This approach allows for complete coverage across the simulation space, estimates experimental error for both the simulation and live tests if replicates are included, and provides a strategy for direct matching between simulation and live test points.

Ultimately, the best design for the simulation experiment depends on the analytical goal and the nature of the simulation and the data it produces. Statistical designs should support both comparison with live data and exploration of the model space itself, including conducting sensitivity analyses and building emulators. For completely deterministic simulations, space-filling designs are the recommended approach for both comparison and model exploration. On the other end of the spectrum, for highly stochastic models, classical designs are the recommended approach for both goals. For simulations in the middle, a hybrid approach is recommended. In this case, a space-filling approach can be useful for building an emulator, but replicates are also needed to characterize Monte Carlo variation. Table 3 below summarizes these recommendations.

Table 3. Simulation* Design Recommendations

	Recommended Method by Validation Goal	
Level of Randomness	Compare to Live Data	Explore Model Space
None (Deterministic)	Hybrid Design	Space Filling
Low (E.g., Physics-based with calibration factors)	Classical	Hybrid Design
High (E.g., Effects-based, Human-in the-loop)	Classical with Replications	Classical with Replications

***The recommended strategy for live data is classical DOE**

Handbook on
Statistical Design & Analysis Techniques for
Modeling & Simulation Validation

February 2019

Approved for public release; distribution is unlimited.

Approved for public release; distribution is unlimited.

Contents

Approved for public release; distribution is unlimited.

1. Introduction

A. Models and Simulation in Test and Evaluation

This handbook focuses on methods for data-driven validation to supplement the vast existing literature for Verification, Validation, and Accreditation (VV&A) and the emerging references on uncertainty quantification (UQ)[1]. The goal of this handbook is to aid the test and evaluation (T&E) community in developing test strategies that support model validation (both external validation and parametric analysis) and statistical UQ.

In T&E, the validation process generally includes comparison with quantitative test results from live testing. However, while operational testing is meant to emulate a hypothetical combat environment as closely as possible, there are often still gaps and imperfections in the tests' representation of reality. Therefore, qualitative and non-statistical techniques may be required to compare the hypothetical combat environment with the non-quantitative portions of the validation referent.

Director, Operational Test and Evaluation (DOT&E) has noted the usefulness of models, simulations, and stimulators in planning, executing, and evaluating operational tests. In a June 2002 memo[2] to the Operational Test Agencies (OTAs), DOT&E discussed modeling and simulation (M&S) as a data source supporting core T&E processes. [Note: This handbook often uses "model" for brevity to refer to a model, simulation, or stimulator.]

In recent years, evaluations of operational effectiveness, suitability, and survivability increasingly rely on models to supplement live testing. In order for these supplements to be valuable, we must understand how well the models represent the systems or processes they simulate. This means we must quantify the uncertainty in the representation and understand the impact of that uncertainty.

---

[1] National Academy of Sciences Report (ISBN 978-0-309-25634-6), "Assessing the Reliability of Complex Models: Mathematical and Statistical Foundations of Verification, Validation, and Uncertainty Quantification (VVUQ)," 2012.

[2] Models and Simulations, 2002, http://www.dote.osd.mil/guidance.html

In 2016 and 2017, DOT&E noted shortfalls in the analytical tools used to validate DoD models[3,4]. A 2016 guidance memo from DOT&E requires a statistically-based quantitative method be used to compare "live data" to "simulation output" when models are used to support operational tests or evaluations. The guidance requires that test planning documents capture 1) the quantities on which comparisons will be made (response variables), 2) [a/the] range of conditions for comparison, 3) a plan for collecting data from live testing and simulations, 4) analysis of statistical risk, and 5) the validation methodology.

In 2017 DOT&E published a follow-on clarification memo emphasizing that validation not only required a sound experimental design for the live data (ideally matched with the model data), but also a sound design strategy for covering the model domain. This clarification is consistent with the National Research Council[5] observation that validation activities can be separated into two general categories 1) external validation (i.e., comparison to live test data), and 2) parametric analysis[6] (i.e., investigation of model outcomes across the model input domain).

DOT&E also maintains a website devoted to M&S resources.[7] It hosts all relevant memos, some case studies highlighting best practices, and other useful references, including the current and future editions of this handbook. Another resource is the Test Science website,[8] which contains tools and material on rigorous test and evaluation methodologies, including many techniques introduced in this handbook.

B. Components of Verification, Validation, and Accreditation

The Defense M&S Coordination Office's (MSCO's) *M&S VV&A Recommended Practices Guide*[9] provides a conceptual framework for describing VV&A for DoD applications. The National Research Council provides a similar depiction of the same process. Figure 1 and Figure 2 show the two conceptual processes.

Both processes generally proceed as follows:

---

[3] DOT&E Guidance Memorandum, "Guidance on the Validation of Models and Simulation used in Operational Testing and Live Fire Assessments," March 14, 2016.

[4] DOT&E Guidance Memorandum, "Clarification on Guidance on the Validation of Models and Simulation used in Operational Testing and Live Fire Assessments," January 19, 2017.

[5] National Academy of Sciences Report (ISBN 0-309-06551-8), "Statistics, Testing, and Defense Acquisition, New Approaches and Methodological Improvements," 1998.

[6] Parametric analysis includes and is sometimes used interchangeably with sensitivity analysis.

[7] https://extranet.dote.osd.mil/ms/ (requires CAC to access)

[8] https://testscience.org/

[9] https://vva.msco.mil

- **Model development**: The existing body of knowledge (depicted as a conceptual/mathematical model and sometimes referred to as the development referent) is transformed into a computerized or computational model.
- **Verification**: The computerized or computational model is tested against the developer's intent to verify that the transformation was successful.
- **Validation**: The verified model is compared with another, likely overlapping, body of knowledge (the validation referent).
  Note: The validation referent is referred to as the "Problem Entity" or the "True, Physical System" in Figures 1 and 2.
- **Accreditation**: The decision to use the model with any necessary caveats or restrictions based on the information provided in the verification and validation (V&V) process.

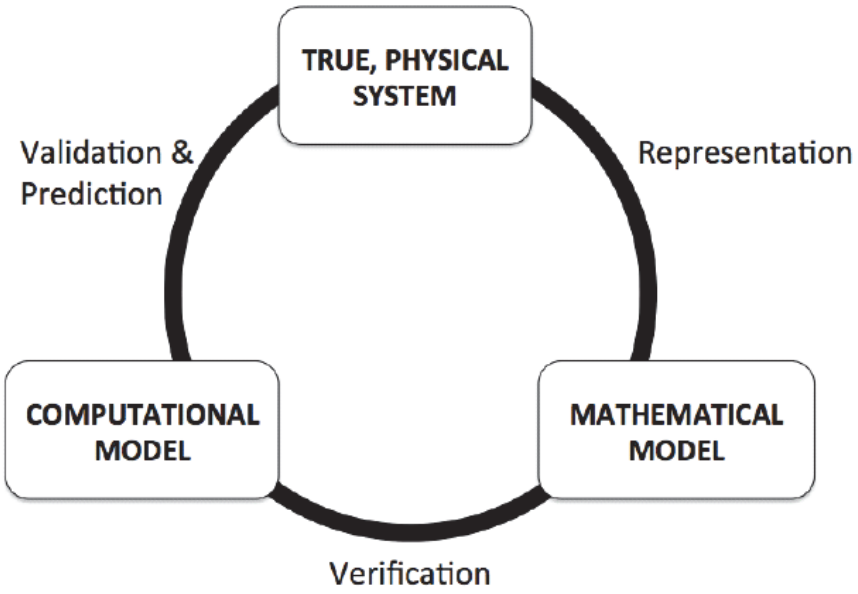

Figure 1. Conceptual Relationships for Verification & Validation Relationships From NRC, Adapted From AIAA 1998[5 above].

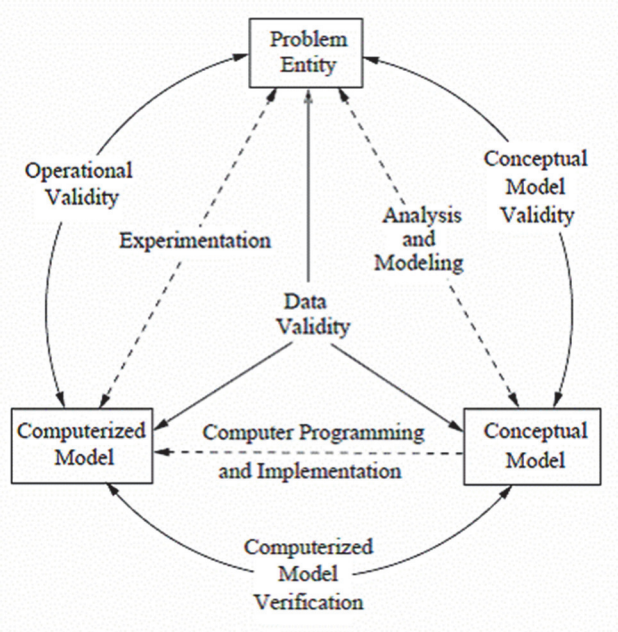

Figure 2. Conceptual Framework for Modeling and Simulation[10]

Typically, this process is iterative, where the model is updated based on validation data collected from the "true physical system" and the V&V process is repeated. It is important to highlight that verification and validation are not yes/no answers; rather they involve the quantitative characterization of differences in quantities of interest across a range of input conditions.[5 above]

The MSCO *Recommended Practices Guide*[11] includes a taxonomy of validation techniques, which describes statistical techniques as well as less formal methods such as face validation[12]. This handbook provides additional statistical techniques beyond those identified in the MSCO guide, as well as guidance as to why a practitioner would select a given technique. These two guides together provide a solid foundation from which the validation process should begin.

For use within the DoD, after models or simulations have been verified and validated they must be accredited, which is the official determination that a model or simulation is acceptable for an intended purpose. Accreditation is a decision based on the information provided in the verification and validation process. Model accreditation for T&E requires the collection of data and the demonstration of acceptable uncertainty for specific attributes (acceptability criteria). Understanding the acceptability criteria is required to develop a

[10] From Modeling and Simulation Coordination Office, https://vva.msco.mil

[11] https://vva.msco.mil/files/ST05_VV_Techniques.pdf

[12] A subjective assessment of the extent to which a model represents the concept it purports to represent. Face validity refers to the transparency or relevance of a test as it appears to test participants.

test program. T&E acceptability criteria should specify the allowable differences on the quantities of interest (response variables) between the model and live testing. Knowing those allowable differences allows testers to plan a test program that supports model validation with adequate data for accreditation decisions. This is done by designing a test that can determine if the test outcomes are statistically different from the model outcomes across relevant input conditions. T&E stakeholders (Program Office, testers, and oversight organizations) should agree on the acceptability criteria early in the VV&A process to ensure the VV&A process supports the needs of all stakeholders.

The decision to accredit a model ties to intended use, which may vary across different organizations and by phase in the acquisition process. Different organizations, based on their independent reviews of the V&V process, may have different views on whether that information supports accrediting the model. The handbook provides an overview of how to incorporate data-driven validation strategies into the existing test processes, thereby supporting quantitative inputs into the accreditation decision by any of the stakeholders.

C. Uncertainty Quantification

Historical DoD V&V processes have not formally acknowledged UQ as a critical aspect of using models for evaluations. Recent research and access to better mathematical tools make it possible to quantify uncertainty from both models and live data. A quantitative validation should aim to accurately convey the uncertainty in any generalizations from models and data.

The National Research Council Report[5 above] defined UQ as "the process of quantifying uncertainties associated with model calibrations of true, physical quantities of interest, with the goals of accounting for all sources of uncertainty and quantifying the contributions of specific sources to the overall uncertainty." They refer to VVUQ [Verification, Validation and Uncertainty Quantification], which emphasizes UQ as a critical aspect of the V&V process.

This handbook focuses on methods for quantifying uncertainties on the differences between models and live data. However, there are many possible sources of uncertainty that should be considered in a model V&V process. They include:

- Model input uncertainty (often specified in the model by a probability distribution)
- Model parameter (i.e., coefficient) uncertainty,
- Model inadequacy: models are approximations and have discrepancies from reality that are a source of uncertainty
- Experimental uncertainty: includes measurement error in live data and uncertainty due to unmeasured input conditions (e.g., nuisance variables)
- Interpolation/extrapolation uncertainty: areas where data cannot be collected.

Traditional UQ has focused on statistical methods for quantifying uncertainty based on data and statistical models (i.e., focused on experimental, interpolation, and extrapolation uncertainty). Recent advances in the field of UQ have enabled development of a more general theory on UQ that synergizes statistics, applied mathematics, and the relevant domain sciences (i.e., the model and/or simulation) to quantify uncertainties that are too complex to capture solely based on sampling methods.[13] For a more detailed understanding of uncertainty quantification, see Smith 2013.[14]

The many sources of uncertainty can be identified by two major categories of uncertainty in M&S: statistical and knowledge uncertainty.[15] Understanding the type of uncertainties that exist is a key consideration when developing the verification and validation plan and deciding how to allocate resources.

Statistical uncertainty captures information on the quantity of data and the variability of the data that were collected under a certain set of conditions. It cannot be reduced or eliminated through improvements in models, but can be reduced by collecting more data. Stochastic models have statistical variations, as do test data. Many statistical analysis techniques are available for quantifying statistical uncertainty. The verification and validation strategy should identify data collection requirements to achieve acceptable statistical uncertainty.

Knowledge uncertainty reflects data inaccuracy that is independent of sampling; in other words, collecting more of the same data does not reduce knowledge uncertainty. Knowledge uncertainty can only be reduced by improving our knowledge of the conceptual model (e.g., improved intelligence on threats) or by incorporating more proven theories into the models (e.g., the tactical code from systems).

Uncertainty is always defined relative to a particular statement. The nature of that statement guides how we might reduce the uncertainty associated with it. For instance, we might ask about the mean range at which the F-22 is detected by an F-16. We could reduce statistical uncertainty by repeated observation (increased sampling). However, if instead we wish to understand the mean range at which the F-22 is detected by an enemy fighter (that is unavailable to test against), the approach is likely different. We would be more likely to reduce the overall uncertainty not by repeating trials with surrogate aircraft, but

---

[13] Adapted from Ralph Smith, DATAWorks Presentation 2018, available at https://dataworks2018.testscience.org/wp-content/uploads/sites/8/2018/03/DATAWorks2018_Smith_Part1.pdf

[14] Smith, R. C. (2013). Uncertainty quantification: theory, implementation, and applications (Vol. 12). Siam.

[15] Formally known as aleatoric and epistemic uncertainty, respectively

by gathering more intelligence on the enemy radar capabilities (a reduction in knowledge uncertainty).

Both categories of uncertainties (statistical and knowledge) are of central importance to T&E, particularly as we try to extrapolate our model and live test data into predictions of performance in combat. Validation reports should convey both types of uncertainty to the decision maker carefully and completely. Ignoring sources of uncertainty that are difficult to quantify does the decision maker a disservice and could result in poor decisions.

D. Model Intended Use: How Good is Good Enough?

The first step in determining whether a specific model is good enough to support a particular test program is identifying how the model will be used to support the test program. Different model uses require different levels of fidelity. Common intended uses for models in systems engineering and T&E include:

- Refining system designs and evaluating system design tradeoffs for meeting performance requirements
- Designing tests to focus on critical information, which could include important test factors, performance transition boundaries, or areas of unknown performance
- Identifying edges of the operational mission space
- Characterizing performance across the operational space
- Extrapolating performance into conditions where live testing is constrained by safety or environmental considerations
- Bolstering conclusions from live testing by interpolating through gaps in live tests
- Improving operational understanding of results from live testing
- Creating operationally relevant test environments by using stimulators to augment live test conditions.

A model's intended use stipulates the fidelity and type of model necessary based on the model's overarching purpose (model hierarchy); the quantities of interest for evaluation (response variables); the range of input conditions for the model; the range of input conditions over which experimental data (live testing) can be collected; and the acceptability criteria (typically stated as an allowable difference between the model and live data). Determining which uncertainty and how much uncertainty matters for a modeled system or process[16] requires detailed understanding of the model, the system under test, and the model's specific intended use.

[16] Simuland

1. Considerations for Model Hierarchy

Once we have determined our model and defined the intended use, our next step is to assess the required fidelity based on the model hierarchy. Figure 3 shows a notional hierarchy of models frequently used in DoD decision making. T&E often employs models that span all of these levels. Operational test and evaluation often focuses on engagement- and mission-level models, but physics-based and engineering-level models often provide context to the evaluation. Campaign-level models may be used to extend the results of T&E into campaign-level analyses.

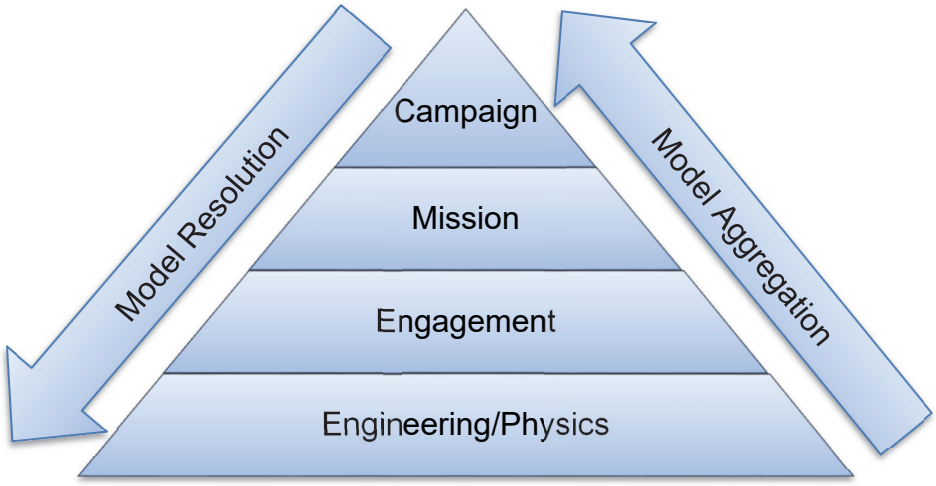

Figure 1. Notional Model Hierarchy

Examples of engineering models include computer aided design (CAD) drawings and finite element models of platforms. Examples of engagement models include MOSAIC [Modeling System for Advanced Investigation of Countermeasures], which investigates the effectiveness of countermeasures against different surface-to-air threats, and ESAMS [Enhanced Surface-to-Air Missile Simulation], which models the probability of successful engagements by surface missiles. A popular mission-level model is SUPPRESSOR, which models air combat missions and reflects both the platforms in the mission and the tactics employed to assess survivability. An example of a campaign-level model is the Logistics Composite Model (L-COM), which can model sortie generation rates from a new class of carrier over a multi-day campaign.

Our expectations for validation must reflect the resolution and level of aggregation of the model. For example, mission-level models may aggregate multiple engagement models and their corresponding uncertainties, as well as current tactics, techniques, and procedures (TTPs). Therefore, mission-level models will likely have higher uncertainty when compared to a single engineering model. That uncertainty must be accounted for in the validation strategy. Techniques include replicating points in the model and live testing,

such that outcomes can be compared based on means and variance. These design techniques are discussed in more detail in Chapter 4.

Similarly, some operational evaluations may require the integration of numerous individual models to represent the system or process being modeled. End-to-end testing that includes all modeled systems during live testing can be resource intensive, making the determination of statistical uncertainty from test data alone challenging. For systems-of-systems, the V&V plan should consider determining statistical uncertainty at the individual system level and quantifying the uncertainty of the modeled systems-of-systems using analytical methods (e.g., propagation of errors or Monte Carlo). A smaller quantity of end-to-end live testing can then be used, with subject matter expert (SME) evaluation, to validate that the interaction of the individual models is qualitatively accurate.

2. Considerations for the Quantity of Interest

Once we've established where the model falls on the hierarchy, we need to determine a response variable that characterizes the quantity of interest and corresponds to models location on hierarchy. A model should be designed to characterize the appropriate quantities of interest. For example, if a mission-level model uses engagement-level models to generate a quantity of interest, then it might be appropriate to compare both mission-level quantities and engagement-level quantities. If, on the other hand, the mission-level model assumes a fixed value for an engagement-level quantity, or draws one from a probability distribution based on historical data, the engagement-level quantity will no longer be a meaningful variable to consider for validation.

Choosing an appropriate response variable (quantity of interest) is a critical step towards achieving successful accreditation. Mission success or force exchange response variables are noisy and unlikely to generalize, making them poor choices for a validation response variable. However, validating based on a more specific performance mechanism that affects mission success may generalize when transitioning from the model, to the test, to the combat environment, making it a better validation response variable.

For example, instead of modeling F-22 mission success, a more appropriate response variable for model validation might be the ability of the F-22 to achieve specific tasks using its stealth and weapons as it faces the kinds of threats seen in operational testing (e.g., range for positive identification for a specific threat). Response variable selection must be correct in order to maximize the likelihood of correctly validating or invalidating a particular model.

3. Considerations for the Input Conditions

Almost as important as selecting appropriate response variables for model validation is identifying the correct input conditions to define the model validation scope. SMEs for

the system under test can identify the input conditions that are important to system performance. Operational SMEs can identify the input conditions that matter operationally. For example, SMEs on a radar system may identify that certain factors on a target are important to radar detection (e.g., target speed, target maneuverability, target's radar cross section), and an operational SME might add that terrain matters. A scope of the validation plan would span all of these input conditions. However, if a model lacks an input condition (e.g., radar models do not consider terrain), that becomes a limitation to the model use.

Analysis using other models (e.g., code-on-code comparison) or data collected from similar systems is another useful information source for determining the scope of the validation. A sensitivity analysis that makes small perturbations in the input variables can identify how much deviation in the input conditions will affect system performance. This is useful for determining whether the input variable should be considered as part of the validation strategy requiring data collection in live testing. For example, if a particular input causes little variation in output that might warrant removing it from the validation plan. It can also inform whether acceptability criteria should change as a function of input conditions.

We may find that the acceptability criteria for a specific intended use is unachievable due to the resources necessary for model development or limited knowledge on a particular input condition (e.g., threat system). In these cases, the T&E community should consider the practicality of the intended use. For example, if insufficient data exist on a specific threat platform to validate a target as that specific threat (e.g., we cannot show that the F-18 is representative of a particular foreign threat), the V&V plan could validate the target as being representative of a meaningful category of a platform type (e.g., a fourth generation fighter jet including representative radar cross sections, jamming capability, etc.). The model can still be useful even if it cannot be validated against a specific threat.

E. Insufficiency of Statistical Techniques Alone

In Chapters 3 and 4 we will consider statistical design and analysis methods that provide a characterization of quantitative differences between model data and reference data (often from live testing). These methods can also be used to estimate uncertainty based on unexplained variance in model and live test outcomes. The statistical techniques discussed are not a complete method for answering the accreditation assessment question: "Is the model or simulation believable enough to be used?"[17] As previously discussed, statistical estimates of uncertainty do not account for all aspects of uncertainty.

[17] https://vva.msco.mil

Additionally, the certainty with which we can generalize (e.g., the combat conditions match the test conditions, which match simulation conditions) is dependent on the degree of similarity between the contexts. For this reason, operational testing includes operationally representative test articles, operating units, threats, and combat environments. M&S will never replace all of these aspects of operational testing, which is why the accreditation decision should include statements of what the model is useful for and the associated limitations.

For operational test and evaluation (OT&E), the modeled system is often the combat system in some future combat scenario that we cannot observe. By necessity, the validation referent must include a considerable theoretical component as a basis for evaluating the model and sometimes that theory exists only implicitly in the intuition of SMEs. For example, intelligence SMEs develop future threat representations. Often the nature of the comparison is highly multi-dimensional and qualitative to a degree that is extremely difficult (if not impossible) to capture completely as differences in quantities of interest.

In cases where explicit observations and theories do not provide a comprehensive and sufficiently reliable description of the reality represented in the simulation, information from theory, other simulations, and SMEs may have to serve as all or part of the validation referent.[18] Even in cases where a predominantly qualitative assessment is necessary, validation approaches and results should be documented in an organized fashion, and information should be elicited from SMEs in as scientific and unbiased a manner as possible.

F. Summary

This handbook is intended to be an introductory guide to applying statistical design and analysis techniques to model validation in the context of T&E. A key assumption in this handbook is that we acknowledge that the test team has input both on what model runs are conducted and what test (validation referent) data should be collected to support validation and the quantification of statistical uncertainty. Therefore, in addition to focusing on statistical techniques for comparing the two outcomes, we also devote a chapter to design of computer experiments and design of experiments for live testing.

The remainder of the handbook provides more detail on how statistical design and analysis techniques can fit into and improve the overall VV&A process. Chapter 2 discusses the overarching steps of the VV&A process as it relates to operational testing. Chapter 3 describes analytical methods for analyzing the simulation itself, making comparisons with the live data, and quantifying the associated uncertainties. Chapter 4

[18] https://vva.msco.mil/Special_Topics/Validation_Referent/val_ref-pr.pdf

outlines design of experimental techniques and their application to both live and simulation environments.

The handbook is not meant to be a guide for all VV&A; it is not exhaustive, and it does not contain all of the most advanced, cutting-edge statistical techniques. The field of UQ is rapidly expanding to include tools for model calibration and for combining model outcomes with live testing to quantify their combined uncertainty.[1920] Those topics are not covered in this handbook.

[19] Oberkampf, W. L., & Roy, C. J. (2010). Verification and validation in scientific computing. Cambridge University Press.

[20] Smith, R. C. (2013). Uncertainty quantification: theory, implementation, and applications (Vol. 12). Siam.

2. Process

A. Introduction

This chapter outlines in detail the VV&A process for using models and simulations to augment operational test and evaluation (OT&E). It also provides general principles for achieving a meaningful accreditation. The overarching process steps for VV&A for OT&E include:

- Identify the intended use of the proposed model and/or simulation
  - Intended use is more than just the purpose (e.g., extrapolate beyond live tests); it also includes the appropriate model hierarchy, quantities of interest, modeling input scope, acceptability criteria, and the testing scope.
- Determine the data requirements to develop and validate the model
- Iterate the Model-Test-Model loop until desired model fidelity is achieved
- Verify that the simulation accurately represents the intended conceptual model (**verification**)
- Diagnose and quantify the uncertainty in the comparison of real-world data and model data/results (**validation**)
  - For simulations that combine multiple system or component models:
    - Assess the impact of the cumulative uncertainty
    - Assess if the interactions between individual models within a simulation are representative of real-world interactions.
- Identify the acceptability of the model or simulation for the intended use (**accreditation**)
  - Does the model provide sufficient representation for the entirety, portions, or none of the intended use statement?

B. The Need for Early and Continued Participation

It is essential that the testing, modeling, and operational user communities work together to develop a VV&A strategy as soon as there is an indication that modeling or simulation will be used as part of the system evaluation. A comprehensive VV&A strategy will specify the intended use of the model, identify the specific models and their required fidelity (through identification of acceptability criteria), and estimate the real-word data needed for validation. The VV&A strategy forms the foundation for resource planning and for estimating timelines to achieve development and validation that support test strategies and test execution.

The VV&A strategy, including the associated acceptability criteria, has routinely been developed too late (or not at all) for operational evaluations. This practice creates unacceptable risk that the delivered models will not support their intended use for operational evaluations, or that the intended use will need to be significantly limited from that originally planned. Model development programs, like all development programs, are susceptible to cost and schedule overruns, technical complications, and budget reductions. Therefore, decisions will likely be required that include deviation from planned development, deviation from delivery schedule, or reassignment of resources. Known accreditation criteria and an understanding of their impact on the model intended for use in operational testing must inform these decisions. A formal interface between the modelers and the testers is a necessary part of the decision process. To avoid these pitfalls, testers should engage with the modeling community early and often.

C. Critical Process Elements for Including Modeling or Simulation in Operational Test Evaluation

1. Model or Simulation Requirements

A critical first step for including modeling or simulation in operational test (OT) evaluation is to figure out whether or not a model is needed, and if so, what it is needed for. In the process of developing test strategies for system evaluation, the test communities[21] may identify a need to supplement live data with information from models and/or simulations due to limitations with live test events. If the need for modeling exists within the test design, the test community should develop a general concept of the model that supports its test requirements. The requirements should include the intended use statement(s), which in turn should include the primary measures that will be assessed (response variables) and model input scope (threat platform models, threat weapon models, environment types, energy propagation models, etc.), including all associated attributes that the model needs to consider. From this general concept, the test community can review existing models to determine if they satisfy the testing need. Based on this research, the test community should then propose to the program manager the use of an available model or the need to develop a new model. The OT representative of the test community should document their requirements for models to support program manager development of models. The requirements for models should be identified as early as possible in the test planning process.

[21] Includes developmental and operational testers, program office, and contractors; commonly called the Integrated Test Team (ITT) or Working Integrated Product Team (WIPT),) depending on the Service.

2. Feasibility of Proposed Model or Simulation

Once testers have determined the model requirements and stakeholders have proposed the use or development of a model, the program manager should assess the feasibility of the proposed modeling solution. They should identify the level of maturity of any existing models or simulations, and, if no models currently exist, the Program Office should investigate the feasibility of developing a model that will meet the identified requirements. The Program Office's assessment should include estimates for cost and schedule required for development and V&V of the proposed model or simulation. The assessment may also include alternative models or simulations that would meet or nearly meet the identified requirements. If the development timeline will not support the scheduled OT or the development cost is prohibitive, the assessment may also identify what limited capabilities can be delivered to support the OT. If the proposed modeling solution is feasible and expected to support OT requirements then it should be included in the test strategy documented in the Test and Evaluation Master Plan (TEMP).

3. Verification and Validation Plan and Accreditation Plan

The VV&A planning process commences once the determination is made to include models or simulations in the test strategy. The planning for VV&A, to include development of acceptability criteria and estimation of live testing data requirements, should be completed as soon as possible. Model feasibility should be re-evaluated after detailed acceptability criteria and validation data requirements are developed. The VV&A process is the focus of the discussion in the remainder of this document. The OTA, or other test organizations involved at the time, should document in an Accreditation Plan the acceptability criteria, including the allowable uncertainty and the corresponding derived data requirement for model runs and live testing. The test data requirements should be included in the TEMP. The program manager or model manager should develop a Verification and Validation Plan based on the approved OTA Accreditation Plan.

D. Verification, Validation, and Accreditation Process

The VV&A process is composed of nine steps. Steps 1-5 of this generalized process should be completed as early as possible in order to estimate resource requirements, inform model development, and reassess feasibility of the intended modeling solution. The test community collectively must decide on the outcomes from Steps 1-5. These steps should be revisited as model development proceeds. The model developer completes Step 6 with oversight from the program manager and model users (for operational evaluations the model users are the OT&E and live fire test and evaluation community). Steps 7 and 8 are completed by the V&V agent, typically appointed by the program manager, with oversight from the model users. The accreditation authorities complete Step 9. For most programs of record, there are multiple accreditation authorities, including the OTA and DOT&E.

1. **Develop intended use statement**
   - Define the system under test.
   - Detail how the model or simulation will be used to support the test objectives. Include the specific test objective(s) and the domains under which a test objective(s) is being assessed.
   - Identify known capabilities and limitations of the model or simulation and their potential impact on the test objective(s).
   - Use Steps 2-5 below to articulate the scope of the intended use.
2. **Identify the response variables or measures.**
   - The response variables are the quantities of interest that are outputs of the model. They provide the basis of comparison between the model and live testing. Because they are satisfying different objectives, the response variables used for validation may not be exactly the same as those used to evaluate system performance, but there is typically overlap. While OT metrics are often at the mission-level, testers may also want to collect lower-level performance metrics to help evaluate the model. In some cases, response variables for validation can directly contribute to the assessment of operational effectiveness, suitability, or survivability.
     - For example, a radar system or sonar system may use detection range as a response variable that is operationally meaningful in assessing system effectiveness.
3. **Determine the factors that are expected to affect the response variable(s) or that are required for operational evaluation.**
   - The list of factors (i.e., independent variables) should be comprehensive of both model scope and test scope. SMEs on the system under test, operational users, and knowledge of the proposed model are essential in determining the factors.
     - For example, for radar detection range, the list of factors could include target radar cross section (frequency dependent, elevation angle dependent, bearing dependent), target speed, and target maneuvers.
   - From the comprehensive list of factors, identify the factors that can be represented in the proposed model, the necessary subset for operational evaluation, and the factors that will be evaluated during the V&V.
   - Identify model-specific factors that should be considered during the V&V. These factors or conditions, due to the nature of models, may not reflect reality. Categories include:
     - Accuracy in representation
       - For example, are environmental conditions such as air quality, moisture content, etc. represented in the model? If so, does their representation match (or can it match) reality from the day live testing is conducted?

- Model-to-model interactions or behaviors
  - For example, for a fire control system, do the radar model and firing model interact as they do in reality?
- Human factors considerations, to include tactics, rules of engagement, operator experience, etc. that could affect performance.

4. Determine the acceptability criteria.

- For each response variable, determine the difference between simulation and reality that can be accepted. The acceptability criteria should consider the context of making a decision based on the model. One simple question we can ask is whether we would accept the data generated by the simulation as evidence that the system is ineffective. Sensitivity analysis using lower fidelity models, analysis of data from legacy systems with similar capability, and SME assessment are all useful inputs to the acceptability criteria.
- The acceptability criteria could be constant or a function of specific factors.
  - For example, detection range must be within one nautical mile for threats with low radar cross section and detection range must be within 500 feet for threats with a large radar cross section.

5. Estimate the quantity of data that will be required to assess the uncertainty within the acceptability criteria.

- Design the simulation experiment (see Chapter 4) to span the model factor space (input scope).
- Design the live test (see Chapter 4) to span the live test factor space and ensure that there are matching points to the model runs to the extent possible. Ensure the live testing provides sufficient data by estimating prediction variance and/or calculating statistical power (see Chapter 3 for definition) for a detectable difference equivalent to the acceptability criteria.

6. Iterate the Model-Test-Model (MTM) loop until desired model fidelity is achieved.

- This is the development stage of the model. Model development is best paired with the sequential testing of the system. A commonly useful approach is one of spiraling complexity, where early, immature versions of a model are initially tested and validated against simple, benign live environments, and live tests get progressively more complex and realistic as the model matures.
- The MTM paradigm is a process where prior model runs and live tests can be used to inform the next step in model development.
- The test community should remain engaged throughout the model development process.
- The MTM loop follows these steps:

  - Design the simulation experiment based on the factors and response variables selected.
  - Review the outputs of the simulation experiment with SMEs and compare to any existing data, including data from similar models or legacy test data.
  - Design the live testing based on the same factors and response variables that will be used to compare to the model. These test designs should progress sequentially as the program iterates through the developmental and operational test continuum.
  - Compare and quantify differences between test data and the model outputs until the acceptability criteria are achieved.
- Note that the above steps are an over-simplification and must be tailored to the situation. If OT occurs before the final desired model fidelity is achieved, the MTM process may need to stop, and the model in its current state and fidelity can be used to inform OT evaluation. Of course in this case, testers must carefully caveat what the model is good for and what it is not, and the original intended use may need to be updated. Synchronizing model development and testing schedules is challenging and is one of the risks in using models for OT evaluation. Testers may not know if the model is adequate for their needs until OT.
- In cases where the MTM process must be interrupted to support accreditation for OT, that does not necessarily mean the model is frozen and the MTM process is over. There can often be a short-term V&V process to support accreditation for a specific purpose, as well as a long-term (or spiral) V&V process that continues the model fidelity improvement cycle.
- Care should be taken to avoid simply tuning the model to the data. Since models are frequently designed to fill the gaps and extrapolate outside of the regime of the available data, naively tuning models to the data provides a false impression that the model is valid, and thus that interpolation and extrapolation are also acceptable.
- When models are not updated directly because of testing, a Bayesian framework can provide an empirical methodology for combining previous model runs, live test results, and new results from either models or live tests.

7. Verify that the final instance of the simulation accurately represents the intended conceptual model (verification process).

- The developer should provide evidence that the model or simulation accurately represents the intended conceptual model.

- Methods for verification should include document review, code review, and an initial parametric analysis to check for bugs. An important aspect of the code review is to ensure that the model correctly propagates uncertainty internally.

8. Determine differences between the model and real-world data for acceptability criteria of each response variable using appropriate statistical methods (validation process).

- Use statistical models to estimate the uncertainty of each of the response variables as a function of the appropriate factors.
- For any uncertainty that exceeds the acceptability criteria of a response variable, the impact of this uncertainty needs to be further assessed against the intended use. Are there aspects of the intended use that are unaffected by this uncertainty?
    - For example, does the model meet the intended use with specific conditions or environments excluded?
- For collections of models where uncertainty is quantified at the level of the individual models, provide an assessment of the cumulative effect of the individual models when assessing a response variable that uses a combination of system models.
- Assess if the interaction between models is representative of real-world interaction. This assessment may need to be qualitative and rely on the judgement of system SMEs due to limited (in both quantity and variance) live test data for highly complex systems.

9. Identify the acceptability of the model or simulation for the intended use.

- The V&V process should provide a recommendation on the acceptability of the models and any associated limitations to support the model's intended uses for OT. The range of acceptability includes usable without limitation, usable with specified limitations, and not usable for the intended use.
- The accreditation authority (e.g., program manager, OTA, DOT&E) will then formalize the acceptability of the model for OT with an accreditation letter.

E. Statistical Analysis

Many statistical design and analysis methods are available for quantifying differences between data sets. Different statistical analysis methods include different assumptions and limitations that need to be considered when identifying the best method for the specific model. Many of these methods are detailed in Chapters 3 and 4. Table 1 correlates statistical concepts and methods to the appropriate steps in the V&V process.

Table 1. Correlating Statistical Concepts to the VV&A Process

VV&A Information	Process Step	Statistical Concepts
M&S Requirements and Intended Use	1 – 4	Response variables Factor space Stochastic vs. deterministic models
Data Requirements	5	Design for Computer Experiments (Chapter 3) Classical Design of Experiments (Chapter 3)
Iterate Model-Test-Model	6	Design for Computer Experiments (Chapter 4) Classical Design of Experiments (Chapter 4) Variation Analysis & Statistical Emulation (Chapter 3) Comparison to M&S runs (Chapter 3) Calibration (Chapter 4)
Verification Analysis	7	Parametric Analysis (Chapter 3)
Validation Analysis	8	Parametric Analysis (Chapter 3) Comparison of live and M&S data (Chapter 3)
Quantify Uncertainty	8 – 9	Hypothesis Testing and Interval Estimation (Chapter 3)

3. Analysis

Statistical analyses can and should inform VV&A decision makers by providing information about model performance across the input space and by identifying risk areas. In addition to discussing inherent limitations of the model in terms of knowledge uncertainty, a good validation analysis should characterize the performance of the model across the operational space and quantify the statistical uncertainty associated with that performance. Inappropriate analysis techniques, such as those that average or roll up information despite the data being collected in distinct operational conditions, can lead to incorrect conclusions.

Evaluations should ultimately be based on both statistical and operational knowledge. Subject matter expertise is critical for answering the question, "Do the identified statistical differences actually make a practical difference?" Accreditation reports should use all relevant information (SME and statistical) to discuss what the model is and isn't useful for.

Before diving into statistical testing or modeling, it is always a good idea to explore the data visually in order to "get a lay of the land" and develop an intuition about the data and the analytical results. Gaining an understanding of the shape of the data and basic trends across the factor space will help practitioners decide which analysis technique is appropriate.

After thoroughly exploring and visualizing the data, testers should evaluate the simulation on its own (to include sensitivity analysis) and compare simulation results with the live test data (external validation). The following sections (A, B, and C) address each of these goals, respectively, and provide recommendations for how to accomplish them. An example is provided in Section D, and additional toy examples and analysis tools are provided in the Appendix.

A. Exploratory Data Analysis

Exploratory data analysis (EDA) is an approach to analyzing data sets for the purpose of summarizing their main characteristics, typically via graphical methods. EDA on its own is not sufficient for VV&A, but it can be a helpful and informative first step.

The primary goal of EDA is to develop an understanding of the data. Ultimately, this is a creative and iterative process; there are no formal rules.[22] Oftentimes EDA begins with a basic question, then visualizations are performed to attempt to answer that question, and then, based on this new information, the question is refined or a new question is developed. Two common types of questions that are usually useful for making discoveries are:

- What type of variation occurs within my variables?
- What type of covariation, or relationship, occurs between my variables?

Visualizing distributions is a great way to begin to answer the first question. Depending on whether your outcome variable is categorical or continuous, this might mean creating bar graphs/dot plots or histograms/boxplots/density plots. When studying these charts you might notice which values are the most common, which values are rare or missing, and whether any interesting patterns occur. These observations, along with the overall shape of the data, may drive the type of statistical analysis techniques to employ later in the validation process. Figure 4 is an example of a histogram showing the range (horizontal displacement) of 100 projectile weapon fires. We notice an approximately symmetric distribution of data, centered somewhere around 615 or 620 m, with about 20 m of variation on either side of that center.

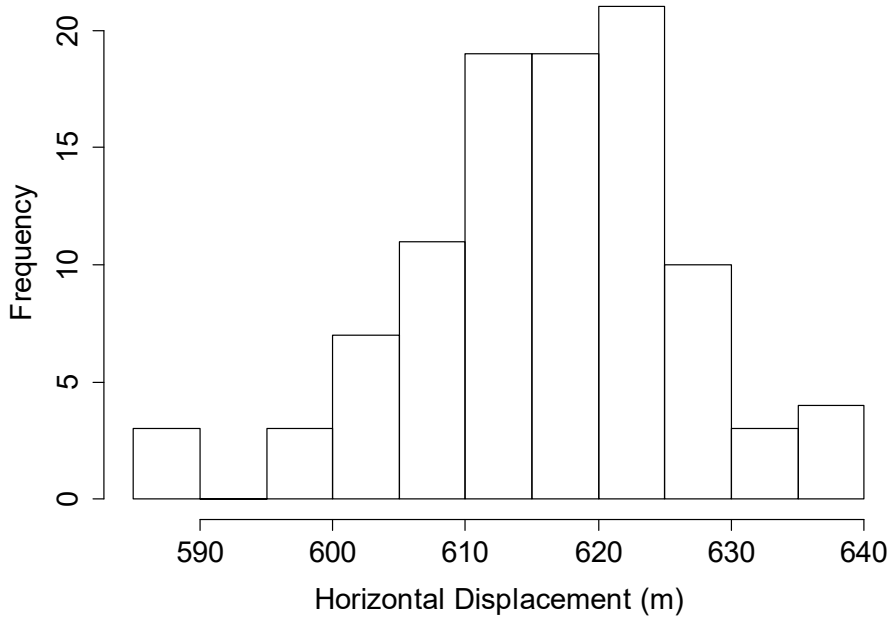

Figure 4. Example of a Histogram to Explore the Variation of Continuous Data

[22] Hadley Wickham and Garrett Grolemund. 2017. R for Data Science: Import, Tidy, Transform, Visualize, and Model Data (1st ed.). O'Reilly Media, Inc.

A good way to study the covariation question before doing formal analysis is to plot your outcome variables across the different factor levels in your data. Again, the best plotting technique depends on the nature of the response and explanatory variables. The relationship between two continuous variables can be visualized using a scatterplot, while a continuous outcome across a categorical factor can be visualized with side-by-side boxplots or histograms. To visualize two or more categorical variables, colored bar graphs or heat maps may be useful. Figure 5 is an example of side-by-side boxplots comparing the horizontal displacement of multiple types of projectiles. We can clearly see differences between the projectiles, with projectile D having an especially larger horizontal displacement than the other projectiles.

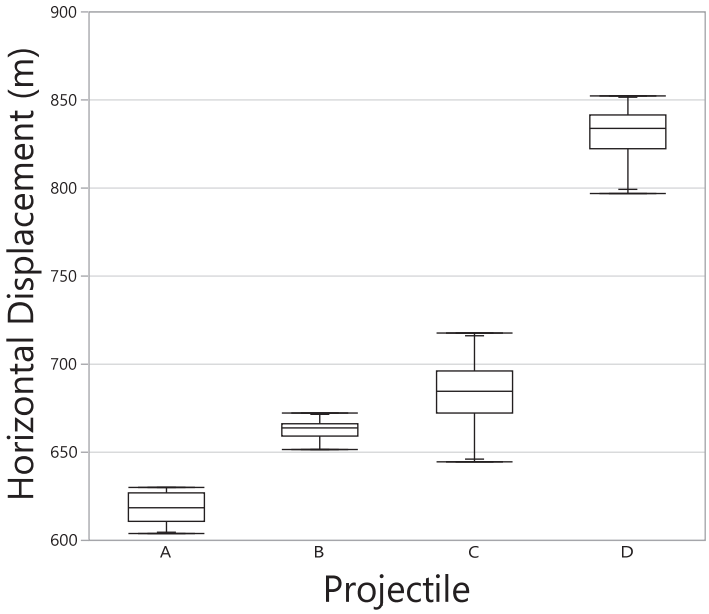

Figure 5. Example of Side-By-Side Boxplots to Explore the Covariation of Continuous Data Across a Categorical Factor

Analysts conducting an EDA at the beginning of the VV&A process may only have initial simulation data on which they plan to perform parametric analysis as part of the verification and early validation processes. As the VV&A process progresses, analysts will have both simulation data and live data to explore and analyze.

Any patterns or relationships found during EDA can help guide the analysis process. Having a basic understanding of what the live data look like, what the simulation data look like, and how each data set behaves across the factor space means that practitioners already know some of what to expect from an analysis, such as regression modeling, and can prioritize certain interesting analysis questions instead of starting from scratch.

B. Analysis of Simulation Data

A critical part of verifying and validating a model is to understand and evaluate the model's behavior. Ideally, this should be done before live data are collected to compare to the model as well as iteratively thereafter, each time the model is updated.

Rather than relying on a standard set of reference cases or SME selection of points, testers should explore the simulation space systematically using design of experiments (DOE) techniques, as discussed in Chapter 4. If the simulation is deterministic[23], design of computer experiments methodologies and associated statistical models are most useful; if the simulation is stochastic[24], classical DOE techniques are appropriate (see Chapter 4 for examples of each).

Two broad types of analysis can and should be performed on simulation data: variation analysis and statistical emulation.

1. Variation Analysis

Variation analysis includes concepts like sensitivity analysis and Monte Carlo analysis. The specific techniques that are appropriate depend on whether the simulation is deterministic. Under constant input conditions, the output of a deterministic simulation will not vary, so there is no randomness to characterize. However, if the model under evaluation is non-deterministic (e.g., human-in-the-loop simulations, hardware-in-the-loop simulations, discrete event simulation, or random variable draws hard-coded into an otherwise deterministic simulation), then an analysis of the Monte Carlo variation inherent in the model is warranted.

A Monte Carlo analysis can entail running the model many times under the same input conditions to determine the range of possible outputs under those conditions. How many repetitions depends on how much variation exists and how costly and time-consuming the model is to run, but could be anywhere from a few times to hundreds of times per condition. Ideally, the analyst should obtain a distribution of output values under each set of input conditions, such that center and spread can be characterized. If the set of input conditions is unreasonably large, using a designed experiment can help cover the simulation space efficiently without necessarily testing every possible combination of inputs.

Sensitivity analysis is appropriate for both deterministic and non-deterministic simulations. Sensitivity analysis can refer to either larger changes in the inputs to build

[23] A deterministic model will always produce the same output from a given starting condition or initial state.

[24] Stochastic models possess some inherent randomness. The same set of parameter values and initial conditions will lead to an ensemble of different outputs.

parametric emulators (termed parametric analysis earlier) or small changes in inputs to look for bugs in code and reasonable perturbations in outputs. In this context, sensitivity analysis is used to determine how different values of an input variable affect a particular simulation output variable. For example, a lethality modeler might want to test how changing the range to target affects the probability of kill, with all other conditions held constant. In this case, multiple discrete values that span the spectrum of possible use cases for range to target should be input into the simulation and the resulting probability of kill should be recorded each time. If the simulation is stochastic, this process may have to be repeated many times in order to characterize patterns fully.

Variation analyses have multiple purposes. Both Monte Carlo analysis and sensitivity analysis can be used to test for model robustness. Typically, small changes to an input variable should produce small, predictable, and reasonable changes to the output variable. In addition, Monte Carlo variation so large that it overpowers variation due to changing controlled inputs is probably not desired. Any randomness in a model should closely match the random variation seen in the real world after controlling for as many other sources of variation as possible.

Sensitivity analysis is also useful for increasing understanding of relationships between inputs and outputs in a model and for finding errors by encountering unexpected relationships. Subject matter expertise can be a critical piece of this analysis, because SMEs can tell whether an observed pattern is reasonable and expected.

Finally, variation analyses can be used to identify risk areas, support uncertainty reduction, and inform live testing. By identifying model inputs that cause significant uncertainty in the output, testers can focus attention on those areas and strive to reduce that uncertainty and increase robustness by obtaining real data in those conditions and/or updating the model appropriately.

2. Statistical Emulation

Statistical emulators, also known as meta-models, use simulation output from across a set of conditions (ideally a designed experiment) to build a statistical model. An emulator, similar to any type of empirical model fit, is used to predict the output of the simulation at both tested and untested conditions. If the emulator is robust enough, it can also serve as a surrogate for the model or simulation itself, which can save time and cut costs by avoiding the need to re-run the simulation repeatedly. Standard statistical model validation techniques, such as goodness-of-fit statistics and outlier analysis, should be used to ensure that the statistical emulator is a reasonable representation of the data on which it was built.

As discussed in Chapter 4, building an emulator first requires constructing a designed experiment. The properties of the simulation drive the choice of DOE (classical vs. computer experiments), which in turn drives the choice of statistical model that becomes

the emulator. If the simulation is stochastic and outputs a binary response variable, a suitable emulator might be a logistic regression model, while if the simulation is deterministic and nonlinear, a Kriging model (or other interpolator; See Chapter 4) is more appropriate.

Statistical emulators can accomplish some of the same things as variation analyses, but in a more integrated way. Statistical emulators characterize relationships between input variables and simulation output just as a sensitivity analysis does, but can also uncover more complex relationships, such as higher-order interactions.

Emulators also support prediction and uncertainty quantification across the simulation space, even in conditions that were not explicitly run in the simulation. Typically, if some sort of space-filling design was performed in the simulation, interpolation between those points via an emulator is acceptable, so long as the associated uncertainty intervals are provided with the predictions. Extrapolation to areas just outside of tested regions is sometimes reasonable as well, especially provided those points at the extremes of the tested region can be shown to adequately match reality.

Finally, evaluation of statistical emulators can help support the choice of live test points. For this reason, programs should make every effort to develop and mature their model as early as possible so that it may inform their live test planning. If testers and modelers are confident that certain input variables should not significantly affect the simulation output, those variables might be able to be dropped from future live test designs. On the other hand, if predictions under other conditions have large uncertainty intervals, or the emulator produces unexpected or counterintuitive results, that may be cause for including those factors in a live test so the patterns can be better understood and the model can be appropriately updated.

C. Comparing Physical and Simulation Data

Once the properties of the simulation have been characterized, the next goal of validation is to detect and quantify differences between the live test data and simulation output (external validation). These differences could be in terms of mean (or any measure of center), variance (or any measure of spread), or distribution more generally. Typically, at least as much simulation data can be obtained as live data. Figure 6 compares the distribution of several dozen simulation data points (blue curve) to a handful of live data points (red lines) under the same conditions. In the top panel, the two data sets appear to match quite well in terms of center and spread. In the center panel, the mean of the live data is noticeably higher than that of the simulation data. In the bottom panel, the spread of the live data is much smaller than that of the simulation data. While these differences are easy to see by eye, statistical techniques can be crucial to detecting such differences in the presence of multiple factors and interactions or noisy data.

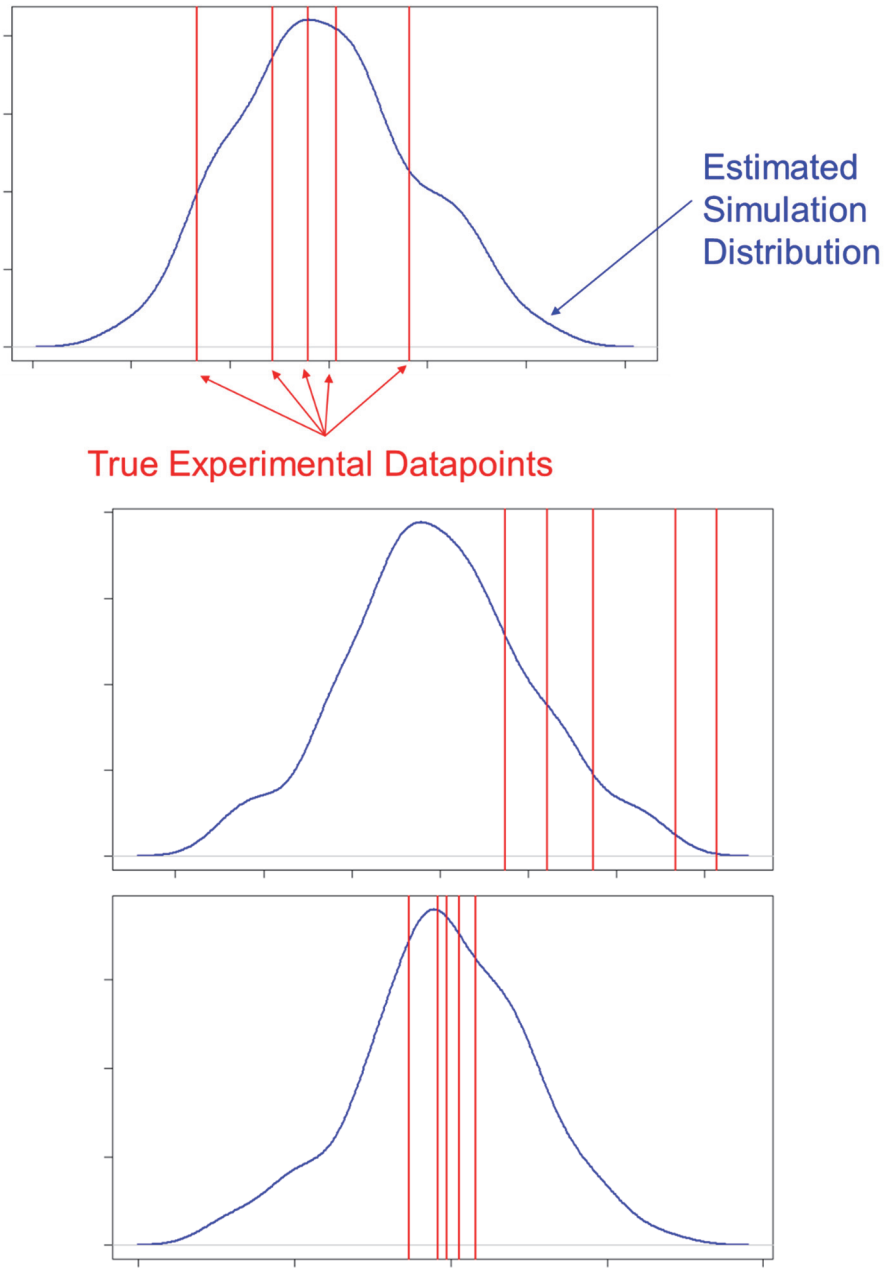

Figure 6. Comparison of Live Data to Simulation Output Where Data Sets Match (Top Panel), are Different in Center (Middle Panel), and are Different in Spread (Bottom Panel)

In general, statistical validation techniques that account for possible effects due to factors are preferred over one-sample averages or roll-ups across conditions. Especially if a designed experiment was executed in the live environment, an appropriate statistical modeling approach for comparing the live data and simulated output should be taken. In all cases, even if no factors are identified and a one-sample approach is taken, it is crucial that the uncertainty about the difference between live data and simulated output be

quantified. Hypothesis tests and confidence intervals are simple ways to quantify uncertainty.

The most appropriate method for statistically comparing live data and simulated output will depend on a variety of circumstances. There is no one-size-fits-all solution. In some cases, it may be useful or necessary to apply multiple techniques in order to fully understand and describe the strengths and weaknesses of the model.

The distribution of the response variable, or metric of interest, will inherently drive the techniques that can be used. For example, binary (pass/fail) data should not be analyzed in the same way as continuous data, such as miss distance or detection time. One should perform EDA prior to employing any specific validation techniques in order to determine which distribution or distribution category is most appropriate for the data.

As mentioned earlier, the presence or absence of factors will also influence which methods are preferred. Applying a simple hypothesis test on the means, for example, to a data set obtained across multiple diverse conditions should be avoided, as it will cover up any potential differences in how the simulation performs across the operational space. A one-sample or overall goodness-of-fit test may be an insightful first step in the statistical validation process, but should not be the only analysis performed if factors are present.

Collecting small amounts of live data limits the application of certain statistical techniques. In such cases, approaches that involve building a statistical model with the live data may not be feasible. However, the simulation data can still be modeled to determine how well it can predict live outcomes, and quantitative comparisons can still be made.

1. Hypothesis Testing

At the core of nearly every statistical technique is a set of hypotheses. The null hypothesis (H_0) is typically one of no change or no difference, while the alternative hypothesis (H_1) is the change or difference one is interested in detecting via test. In the context of external model validation, these hypotheses are typically framed as differences (or lack thereof) between live and simulation data for a particular response variable. Suppose we are interested in comparing the mean miss distance of an air-to-air missile produced via simulation with that of live test outcomes under identical conditions. Our statistical hypotheses in this case might be:

H_0: The mean miss distance of the simulation is *the same as* the live shots and

H_1: The mean miss distance of the simulation is *different than* the live shots.

Collected data is then used as evidence to either reject the null hypothesis (which indicates a significant difference between live and simulation data) or fail to reject the null hypothesis. The example above has only one set of hypotheses and a simple technique such as the t-test (see Analysis Appendix) is sufficient for analyses. However, the

statistical technique(s) used to analyze the data should be able to address all hypotheses of interest. Hypotheses become more numerous and more complex as more questions are asked and thus more advanced statistical techniques are employed. For example, in regression modeling we want to know if there are any changes in the response variable caused by any factor (or interaction) in the model. In this case, there are as many sets of hypotheses as there are terms in the model.

As presented in the hypotheses above, the live and simulation data are assumed to match, unless the data provide evidence otherwise (a hypothesis test set up in this manner is often referred to as an "acceptance test" in the literature). This logic may seem backwards and could imply that collecting little or even no data means the simulation "passes." This is definitely not the argument here. In conjunction with this or any set of hypotheses, analysts must conduct power analyses (see discussion below) to ensure that the data they plan to collect will provide enough information to meaningfully evaluate the results against the previously determined acceptability criteria. In the absence of a sufficient power, the results of the hypothesis test are meaningless.

Another way to set up the hypotheses is to set the null hypothesis to be that the two data sources do not agree, unless the data show that they do (known as a "rejection test"). Setting up validation hypotheses as a rejection test is certainly not wrong and, given sufficient resources, is arguably the more rigorous and defensible method. However, rejection testing requires much more data to show model agreement, and testers almost never have the sample sizes necessary to support a powerful rejection test. So ultimately, using the hypotheses shown above (null of model and live data agreement) may not be ideal, but, provided the test is adequately powered, it is certainty better than not analyzing the data statistically at all.

Ultimately, uncertainty quantification and the magnitude of detected differences[25] matter more, and are more interpretable, than the outcome of a hypothesis test. The amount of data required to address certain hypotheses depends on the magnitude of the difference we are trying to detect. Performing hypothesis tests on small quantities of data will enable us to uncover large differences between live and simulation data, as those are easier to detect. However, in order to detect smaller differences, more data are needed.

To ensure that planned data collection will produce meaningful results, practitioners should calculate the statistical power for a detectable difference equivalent to the acceptability criteria. Power is the probability of correctly rejecting the null hypothesis. Similarly, confidence is the probability of correctly failing to reject the null hypothesis. Suppose that in the air-to-air missile example, a difference of 10 feet or more is operationally important, meaning that if the simulation prediction is within 10 feet of the

[25] Effect size

actual live hit, this will have minimal effect on lethality predictions. The test should be sized to ensure that, if miss distances are actually 10 or more feet different, then the t-test (see Analysis Appendix) will reject the null hypothesis. Otherwise, the uncertainty about the estimated difference will be too large to draw any meaningful conclusions about the acceptability criteria. In other words, a confidence interval about the difference may span both 0 and 10. Figure 7 conceptually outlines the meaning of statistical power and confidence, and associated risks, in the context of this example.

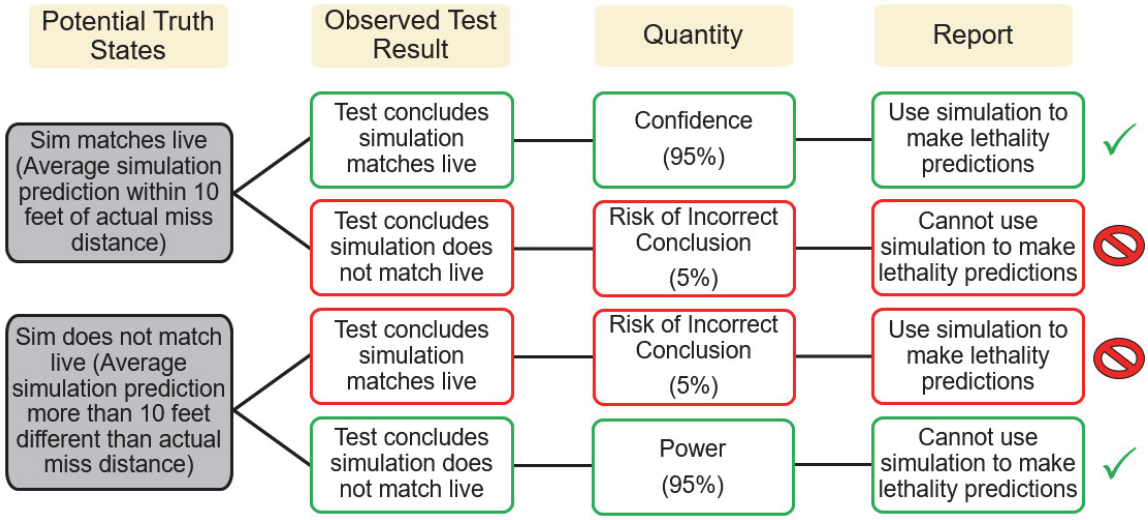

Figure 7. Power and Confidence Flow Chart

2. Interval Estimation

While UQ should include all sources of uncertainty, interval estimation is one way to quantify uncertainty in *statistical* results. Rather than simply reporting a point estimate (e.g., the mean difference in miss distance is 10 feet), it is useful to calculate a range of plausible values for that parameter (e.g., the mean difference in miss distance is between 6 feet and 14 feet, with 80% confidence). Wider intervals indicate more uncertainty about an estimate, and thus more risk.

There are several types of intervals, each of which can be useful depending on the context. A confidence interval is a range of values likely to contain a particular population parameter, such as the mean. In cases where there are simulation requirements or accreditation criteria focused on specific parameters, confidence intervals about estimates of those parameters are useful to report.

Perhaps more often useful in the context of validation are prediction intervals. A prediction interval is a range of values likely to contain a future individual data point. Prediction intervals take into account the scatter of the data in addition to the uncertainty in knowing the value of the population parameter. Thus, prediction intervals are always wider than confidence intervals. Because simulations are often used to estimate

performance in untested regions of the operational space, prediction intervals are an intuitive way to convey the uncertainty associated with these predictions. Figure 8 demonstrates these two most common types of interval. The blue band is the 95% confidence interval on the mean miss distance, while the orange band is the 95% prediction interval.

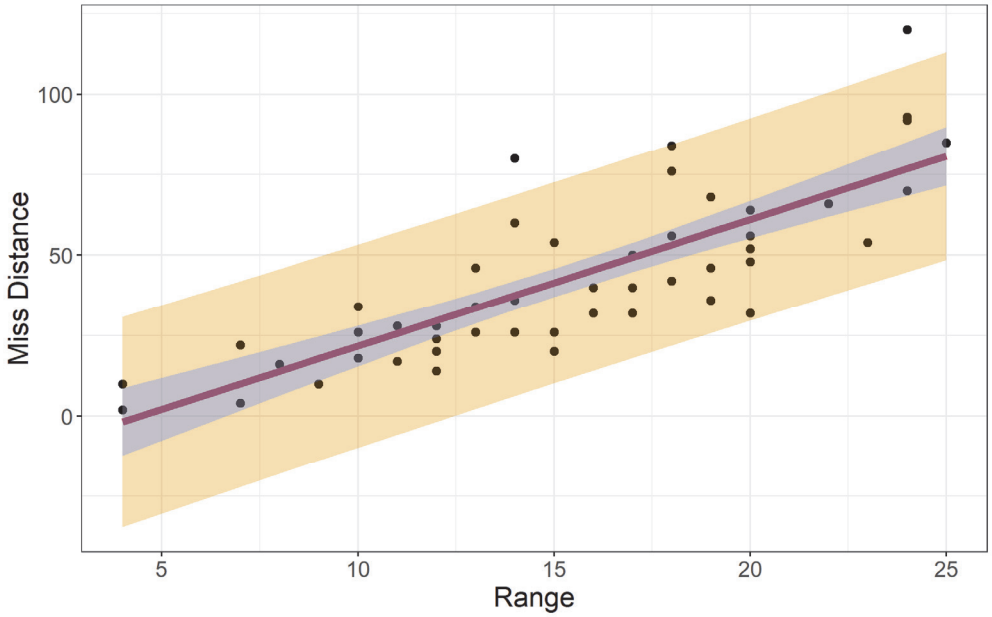

Figure 8. Confidence and Prediction Intervals

Finally, a tolerance interval is a range of values within which some proportion of the population lies, with a certain level of confidence. These intervals can be useful when dealing with requirements phrased in terms of percentiles (e.g., 95% of miss distances must be below 20 feet).

Any kind of quantitative validation result should include an uncertainty interval, even if statistical significance is also reported. While significance[26] is potentially useful, the width of an uncertainty interval is more interpretable and ultimately more important for both accreditation decision makers and users of the model.

3. Literature Gaps

The research on model validation in the academic literature is vast[27][28], and some of these techniques are described in detail later in this handbook. However, previous work

[26] Statistical significance as determined by the p-value(s) resulting from a hypothesis test

[27] Jack PC. Kleijnen, "Verification and validation of simulation models," European journal of operational research 82.1, pp. 145-162, 1995.

[28] S. Y. Harmon and Simone M. Youngblood, "A proposed model for simulation validation process maturity," The Journal of Defense Modeling and Simulation 2.4, pp. 179-190, 2005.

tends to ignore the limited live data problem that is prevalent in OT, and much of the academic research is focused on improving the predictive capabilities of computational models that are highly deterministic in nature. Most of the models used in operational testing are quite stochastic. Even if a deterministic model is employed for OT purposes, the goal of validation is typically not to improve prediction, but to identify regions of the operational space where the simulation adequately matches the live data, and regions where it might not. While updating or calibrating the model to improve prediction would certainly be beneficial, this task presents a slew of additional challenges that will not be discussed in this handbook.

The following example illustrates the importance of applying the appropriate statistical techniques while comparing live and simulation data. See the Appendix for a more detailed description of the analytical methods, R code to implement these methods, and additional recommendations based on a Monte Carlo simulation study.

D. Example

Consider a test of an air-to-ground missile. A key metric of interest is miss distance. Factors that might affect miss distance include:

- The specific aircraft platform from which the missile is launched (A or B)
- The altitude of the aircraft (Low, Medium, or High)
- Whether or not countermeasures were in use (Yes or No).

Only 12 missiles are available for live testing. One test was conducted in each of the 12 conditions (2 platforms x 3 altitudes x countermeasures on/off). This is called a full factorial design (see Chapter 4 for formal definition). For each condition executed during testing, 100 replicates were conducted in the simulation.

An initial statistical look at the data might involve testing whether the mean miss distance from the live data is equal to that of the simulation data. Figure 9 depicts the mean miss distance (with 80% confidence interval) from the simulation runs (blue) and the live runs (red). The t-test (see Analysis Appendix) does not reject the null hypothesis, meaning that the mean miss distance of the live data is not statistically different than the mean miss distance of the simulation data. However, this test ignores the variance of the data and does not account for the test conditions in the design.

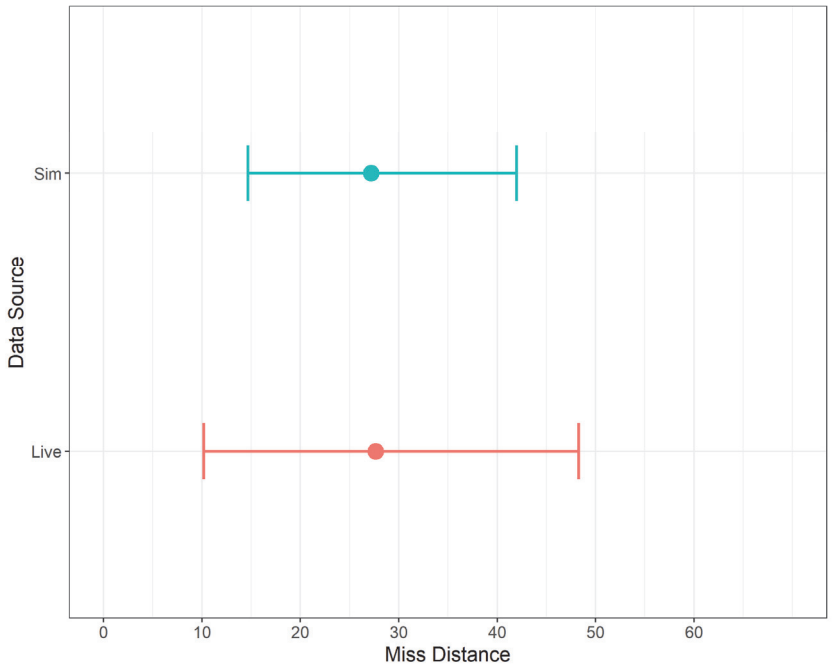

Figure 9. Mean Miss Distance (With 80 Percent Confidence Interval) for Live and Simulation Runs

Looking into the data further might include comparing the overall miss distance distribution of the live data to that of the simulation data. Figure 10 shows the cumulative distribution of miss distance for live data points (in red) and simulation (blue). Note that the stair step effect is caused by the fact that there are only 12 live data points. In statistical distribution tests, the variance of the data is considered in addition to the central tendency, but test conditions are still completely ignored. The Kolmogorov-Smirnov test (see Analysis Appendix) fails to reject the null hypothesis that the two data sets come from the same distribution.

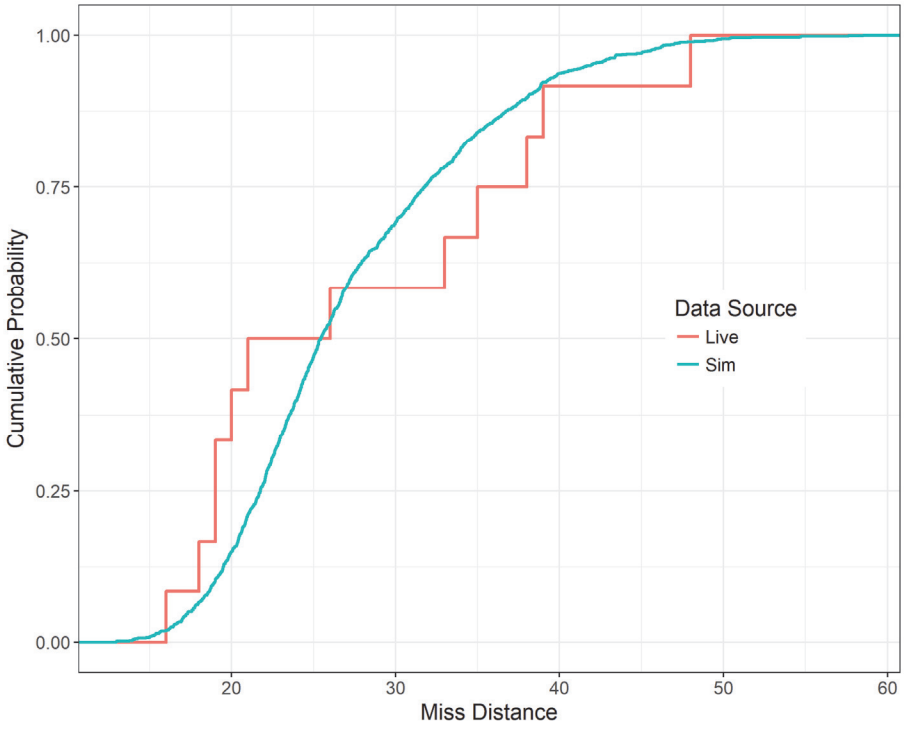

Figure 10. Cumulative Distribution of Miss Distance for Live and Simulation Runs

Fisher's Combined Probability Test (see Analysis Appendix) takes a stove-piped look at each of the 12 conditions and then attempts to summarize these individual findings using a global hypothesis test. Figure 11 shows the distribution of simulation miss distances under each condition (blue histogram) overlaid with the live miss distance result in that condition (green line). The results of this test are more ambiguous. Notice, for instance, that the green lines always fall towards the right end of the distribution when countermeasures is "Yes." A uniformity test on the p-tail values[29] passes, but Fisher's original statistic finds borderline evidence[30] of extreme behavior. These results give us a good hint that something is going on with the conditions in our design. The simulation matches the live testing better in some conditions than others.

[29] A p-tail is the proportion of simulated miss distances greater than the observed live miss distance in each condition. If the simulation data and the live data are statistically equivalent, the distribution of p-tails will be uniform.

[30] A p-value between .05 and .10

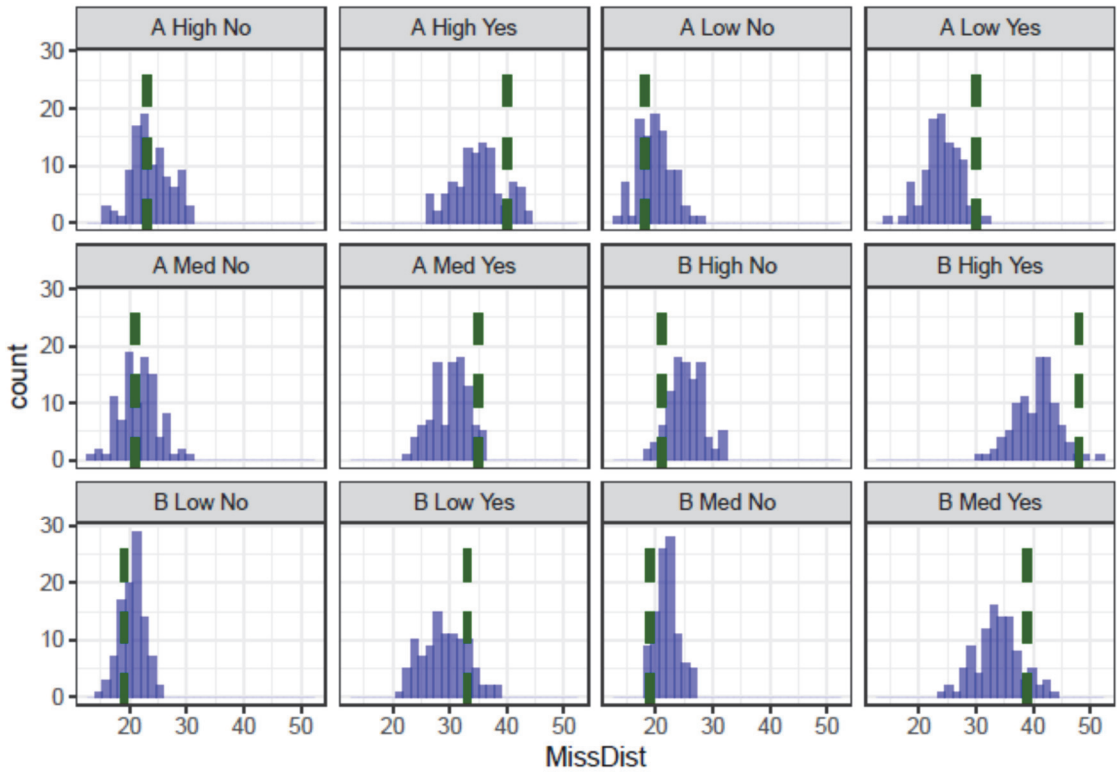

Figure 11. Histogram of Simulated Miss Distances in Each Condition (Blue) With Live Result Overlaid (Green); Labels in Gray Indicate the Factor Settings (Platform, Altitude, and Countermeasures)

Regression analyses can formally account for all factor settings (and their interactions) simultaneously in order to pinpoint which specific conditions are driving differences in performance. In the context of this example, the primary goal is to determine whether actual (live) miss distances and estimated (simulation) miss distances are consistent across the factor space. In cases where every condition for which simulation data was collected also has corresponding live data and vice versa (matched pairs), regression can be performed on the matched *differences* in miss distance between live and simulation. Otherwise, using the raw miss distances is appropriate, so long as the regression includes a "source" term (an indicator for whether the data point came from live or sim) and at minimum all two-way interactions with that source term.

Sample size is an important consideration for regression validation analysis. While putting all of the data (live and all simulation replicates) into one single regression model works well for balanced designs, if there are substantially more simulation runs per live run, as in this case, the correlation structure of model terms becomes a serious issue. Figure 12 shows the correlation map for all main effects and two-way interactions in the model. Notice that some main effects are confounded with the two-way interaction terms with source, which are the exact terms we are most interested in for the purposes of validation.

Interaction terms reveal that the simulation may reflect reality in some conditions but not others.

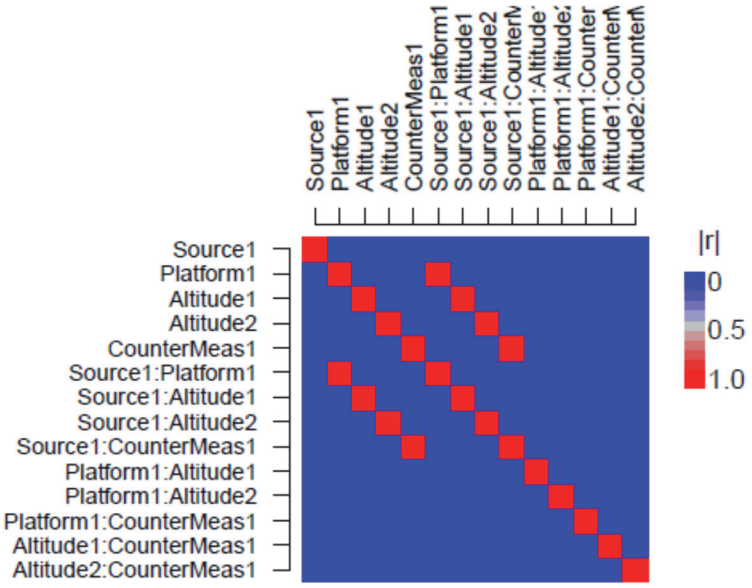

Figure 12. Correlation Map Using all the Live and Simulation Data in an Unbalanced Regression Model

One possible solution to this correlation problem is to take a resampling approach. The idea is to create a balanced data set by sampling with replacement from the simulation data set (in this case we're only choosing one simulation run per condition so replacement is irrelevant), combining that resampled data set with the live data set, performing a regression, saving the results, and repeating thousands of times. This bootstrapped[31] regression approach 1) controls for differences in sample size between live and sim and 2) mathematically accounts for two-way (or higher) interactions between factors in the test design. Figure 13 depicts the (now perfect) correlation structure for the balanced bootstrapped regression model. Another way to avoid the correlation problem is to construct two separate regression models (one for live and one for simulation data) and statistically test for differences in the coefficients from each model.

[31] See Efron, B. (1981). "Nonparametric estimates of standard error: The jackknife, the bootstrap and other methods"

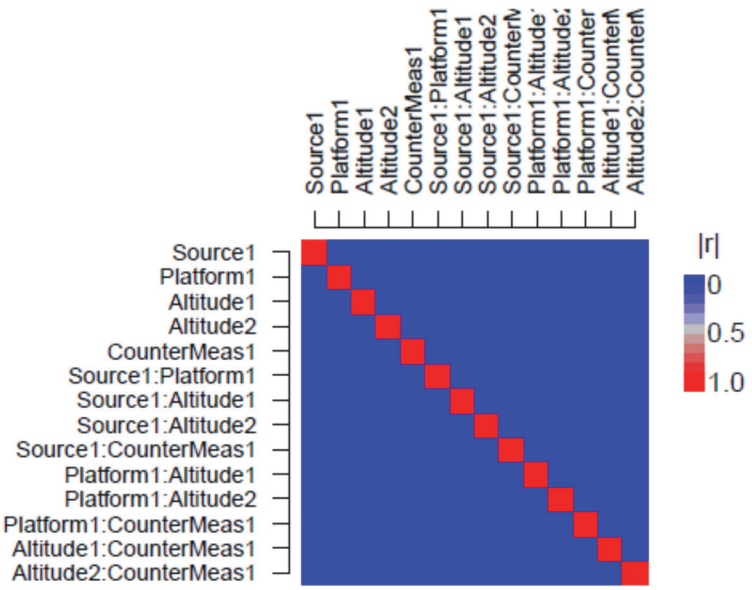

Figure 13. Correlation Map Using Bootstrapped Live and Simulation Data in a Balanced Regression Model

The terms we care about the most for validation are the terms involving source. While the main effect for source is not significant, the interaction between source and countermeasure is. One way we can visualize these effects is by plotting differences in model predictions between live and sim and calculating 80% confidence intervals[32] on those differences. If these intervals fall entirely on one side of zero, there is evidence of a statistically significant difference. Based on Figure 14, we see that the simulation underestimates miss distance in the presence of countermeasures, overestimates miss distance for Platform B with no countermeasures, and performs well for the Platform A no countermeasures case.

[32] Confidence interval calculated from the bootstrapped distribution of predictions

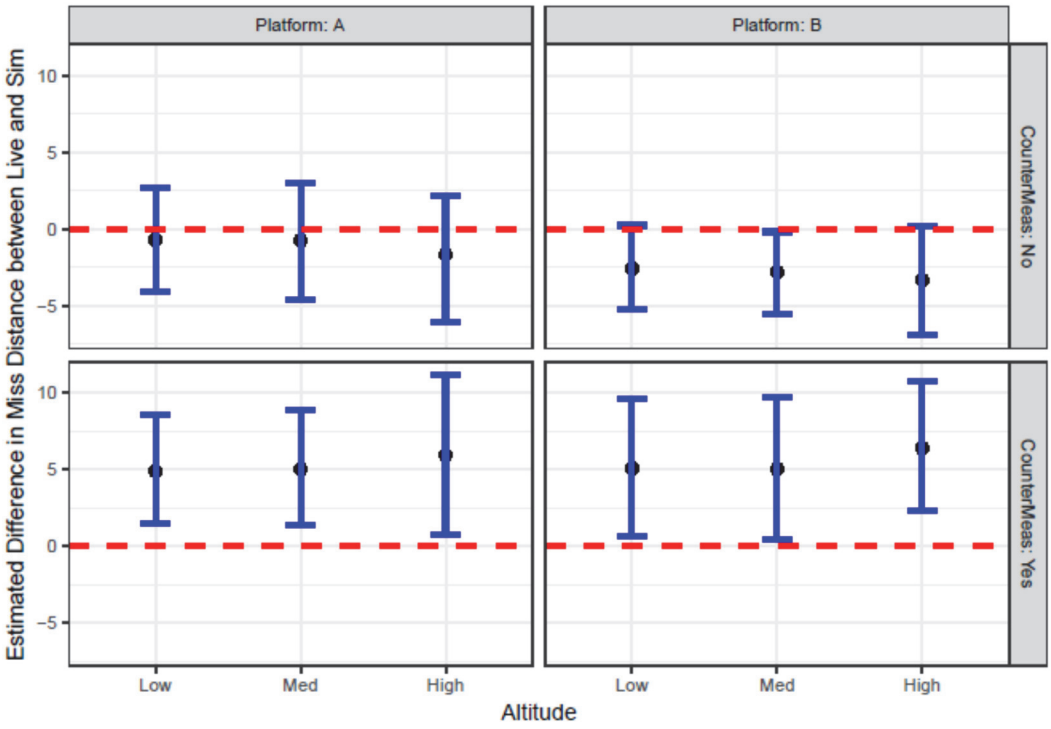

Figure 14. Estimated Differences in Miss Distance Between Live and Sim, With 80% Bootstrapped Confidence Intervals

This example highlights the importance of regression-based approaches that consider all factors simultaneously, which in turn requires early planning to ensure matched conditions. Roll-up results and univariate or distribution tests fail to identify key differences between the live and simulation data. Including interaction terms is key to uncovering sources of variation and providing a more accurate and nuanced validation.

Another way to avoid the correlation problem is to construct two separate regression models (one for live and one for simulation data) and statistically test for differences in the coefficients from each model. However, this test is not trivial to implement, and the matched points approach described above is more powerful.

E. Recommended Methods

The most appropriate method for statistically comparing live data and simulated output will depend on a variety of conditions. We used Monte Carlo power simulations under various settings to develop the methods recommended below. See the Appendix for details on the methodology. In the Monte Carlo simulations , we varied the structure of factors, distribution of the response variable, and sample size of the validation referent.

Table 2 shows the recommended techniques in each category. In the cases where there are multiple methods per cell, more than one test is required to have high power to

test differences in both mean and variance. In addition, some tests are more sensitive to cases where the live test data are more variable than the simulation data, while others perform better in the reverse case, where the simulation data are more variable than the live data. Thus, it may be necessary to use up to three techniques depending on the goals of the validation study.

This table is not intended to include every possible appropriate technique or prohibit the use of any method. There are entire classes of methodologies, such as statistical process control, time series techniques, and Bayesian analyses, that were not included in this study but may perform well in certain contexts. See the Analysis Appendix for more details on each of the methods identified in Table 2, as well as examples and R code.

Table 2. Recommended Validation Analysis Methods Based on Response Distribution and Sample Size for Live Testing (Validation Referent)

		Recommended Method by Sample Size		
Distribution	**Factors**	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

The following descriptions provide the background for interpreting Table 2.

Distribution of the response variable:

- Skewed – data generated from the lognormal distribution
- Symmetric – data generated from the normal distribution
- Binary – data generated from the binomial distribution

Structure of factors:

- Univariate – no varying factors across the collected data
- Distributed level effects – factors significantly affect the mean, but there is no underlying designed experiment. The difference between simulation and test varies across factor levels. The amount of variation across factor levels is represented by a distribution (hence the name, distributed level effects).
- Designed experiment – factors in the underlying designed experiment determine the difference between the live and simulation data, with significant factor mean effects.

Validation Referent Data size:

- Small – 2-5 (continuous data) / 20 (binary data)
- Moderate – 6-10 (continuous data) / 40 (binary data)
- Large – 11-20+ (continuous data) / 100+ (binary data)

4. Design

Design of Experiments (DOE)[33] provides a defensible strategy for selecting data from live testing (validation referent) and simulation experiments to support validation needs. DOE is a common technique for planning, executing, and analyzing both developmental and operational tests. DOE characterizes the relationship between the factors (inputs) and the response variable (output) of a process. For example, in live fire testing, the "process" could be a projectile fired at armor, where the response variable is the penetration depth of the projectile, the factor is armor thickness, and the levels of that factor are thin and thick.

As we saw in the Chapter 3 example, the most powerful validation analysis techniques (e.g., regression analyses) require some degree of coordination between the designs of the physical and simulation experiments. DOE for physical experiments, referred to here as "classical DOE", was developed in the 1920s by Ronald Fisher and is typically taught in introductory DOE courses. DOE for simulation experiments, often referred to as computer experiments, was developed in the late 1970s and early 1980s and is still actively researched today. Classical DOE and computer experiments provide the building blocks for conducting a validation and, even though they are often grouped into "DOE" as a whole, their design principles, algorithms for generating designs, and corresponding analysis techniques can be quite different. Understanding these differences is crucial to understanding validation experiments.

A challenge for analysts is that popular textbooks for classical DOE and computer experiments do not explicitly lay out the necessary steps to tackle validation. This chapter will look at how an analyst can design the physical and computer experiments to facilitate validation. It will examine how designs support the ensuing analysis, and provide an illustrative example implementing different design types in a program. The subsequent sections review Classical DOE (section A) and computer experiments (section B) and highlight aspects that are crucial to validation. Section C introduces hybrid designs, with an associated example in Section D.

[33] Douglas C. Montgomery, "Design and Analysis of Experiments," John Wiley & Sons, 1990.

A. Design of Physical Experiments

1. Distinguishing Features

Physical experiments played a dominant role for the first 70 to 80 years of the long history of DOE,[34]. DOE started with agriculture in its founding days (Fisher 1926[35]), moved to process industries with the development of Response Surface Methodologies (Box 1952[36]), and then to manufacturing (Taguchi 1986[37]). Classical DOE is well suited to handle the distinguishing features of a physical experiment. These features include:

- Having stochastic response variables.
- Being suitable for evaluation with linear regression.
- Emphasizing design robustness over optimality.
- Having few significant factors and few levels per factor.

Live Tests generally have stochastic response variables, that is, re-running an experimental trial with the same factor settings will usually give different observations. Assumptions about the structure of the variance in these observations enables linear models (empirical models) to systematically separate important factor effects from background noise.

This non-deterministic nature of physical experiments is why Classical DOE emphasizes design robustness over optimality[38]. Robustness, here, refers to the capability of the experimental design and its associated analysis to withstand complications that arise in physical testing. Complications include outliers, missing data, nuisance errors, and violations of the statistical assumptions.

Replication, randomization, and blocking are fundamental principles in physical experimentation. Replication is repeating at least some of the trials in the experiment in order to estimate the experimental uncertainty. Randomization refers to the practice of running the trials in the experiment in random order to minimize systematic variation and to provide a valid estimate of uncertainty. Blocking is a technique to prevent the variability

---

[34] C. J. Wu, Post-Fisherian experimentation: from physical to virtual, Journal of the American Statistical Association 110 (2015) 612-620.

[35] R. Fisher, On the capillary forces in an ideal soil; correction of formulae given by W.B. Haines, The Journal of Agricultural Science 16 (1926) 492-505.

[36] G. E. P. Box, Multi-factor designs of first order, Biometrika 39 (1952) 49-57

[37] G. Taguchi, Introduction to quality engineering: designing quality into products and processes, 1986.

[38] R. H. Myers, Response surface methodology, current status, and future directions, Journal of Quality Technology 31 (1999) 30.

from known nuisance sources, which are variables that affect the test outcome but are not of interest for evaluation purposes, from increasing the experimental error[39].

Another distinguishing feature of physical experiments is that, in general, only a small subset of the factors that are part of the design will actually be statistically significant. The goal of a screening test designs is to determine the few factors that have the largest impact on the outcome of the test. On the other hand, characterization test designs are used to better understand how the response variable(s) change across the subset of important factors.

2. Common Classical Design Methods

Figure 15 visualizes three common experimental designs. The left panel shows a factorial experiment, the center panel shows a fractional factorial experiment, and the right panel shows a central composite design. The number of factors in the designs are visualized by the number of dimensions in the figure. Since the designs in Figure 15 are cubic, they each have three factors.

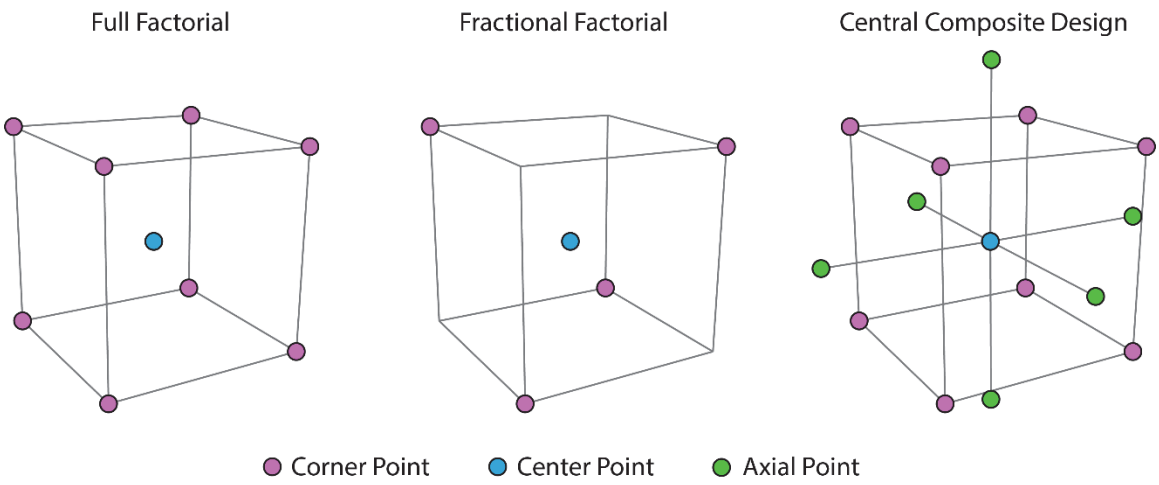

Figure 15. Classical Experimental Designs

Factorial and fractional factorial experiments are some of the most popular types of physical experiments. They are examples of screening experiments that place test points for each factor at a low and high setting. For the three-factor full factorial experiment this results in eight total runs. The blue "center point" is often added to factorial designs to check for curvature in the response, because two points creates a line and three points a curve. Factorial experiments simultaneously vary factor settings from run to run. This

[39] R. T. Johnson, G. T. Hutto, J. R. Simpson, D. C. Montgomery, Designed experiments for the defense community, Quality Engineering 24 (2012) 60-79.

enables the estimation of both main effects and interactions, which is not possible when only one factor is varied at a time.

Two-level factorial experiment designs are highly efficient (in that they only use two levels of each factor) and informative (in that they allow for the estimation of all possible interactions). However, they are potentially prohibitively costly as they grow in size by powers of two as additional factors are added. The design in Figure 15 includes only nine points: a three-factor full factorial design with a center point. A best practice for increasing the design robustness is to include replications. If replicating all nine points is prohibitively expensive, replicating only the center point, or replicating a fraction of the full factorial can provide additional design strength for less expense.

Fractional factorial designs, shown in the center panel of Figure 15, are efficient screening designs. They are similar to full factorials in that factors typically each have two levels, but different because they only include a fraction of the number of unique design points, in this case a half, resulting in four total runs. This design is termed a "half-fraction". Smaller fractions exist for experiments with more than three factors. These designs are efficient, robust, and can accommodate many factors in a relatively few number of runs, while still simultaneously varying factor setting. For example, a one-eighth fraction of a five-factor full factorial reduces the test size from 64 runs to eight runs, but the design can no longer estimate the interaction effects, only the main effects. Fractional factorials trade off the ability to estimate some interaction effects for reductions in test points: generally, fewer points equals fewer estimable interactions.

The right panel in Figure 15 shows a central composite design, a type of response surface design. Response surface designs are a collection of designs that seek to characterize the test outcome with more granularity, by adding additional levels to the factors in the test design. They are used to locate where in the design space (or under what conditions) responses are maximized and minimized, and to characterize a system's performance across a variety of conditions. This often involves the inclusion of higher order effects (e.g., quadratic model terms) that can estimate curves instead of monotonic linear effects.

Another broad class of classical designs are optimal designs, which are defined in terms of some optimality criteria. They are algorithmically generated based on a researcher-specified model and a fixed sample size (the fixed sample size must exceed the number of model terms). Optimal designs are useful when an external reason (cost, safety, etc.) constrains the number of test points and precludes a factorial design. They also provide flexibility for designing tests when the region of interest is irregularly shaped or includes impossible combinations of factors (e.g., disallowed combinations). Several methods exist for optimizing the test point coverage in optimal designs; these include, but

are not limited to, D-, I-, and G-optimal criteria[40]. D-optimal designs minimize the overall variance of the parameter estimates while also not letting the covariance between the parameter estimates get too large. I-optimal designs minimize the prediction variance. G-optimal designs minimize the *maximum* prediction error over the design space rather than the average prediction error. While many other design criteria have been proposed in the literature, D- and I- and G-optimal designs are perhaps the most popular and are available options in most statistical software packages. Optimal designs can be constructed for varying complexities of model surfaces, but the robustness of the design is dependent on specifying the right model prior to running the experiment, which can be challenging.

3. Analysis Methods for Classical Designs

Linear regression modeling is the most widely used statistical method for analyzing Classical DOE and physical testing. Linear models are appropriate for well-behaved, linear response surfaces. Consider an experiment with a response variable and two factors. Figure 16 shows examples of linear model fits to the predicted response for a main effects only model (left), a main effects and interactions model (center), and a response surface model (right). Note that the model is termed linear because it is linear in the model coefficients and not in the factors. Therefore, a linear model can contain quadratic terms or higher order polynomials, providing flexible surfaces illustrated in Figure 16.

Figure 16. Different linear model surfaces

The fractional factorial design from Figure 15 allows analysts to estimate the main effects model (left panel) in Figure 16. The full factorial design from Figure 15 allows analysts to estimate the main effects and interactions model (center panel) in Figure 16. The central composite design from Figure 15 allows analysts to estimate the response surface (right panel) in Figure 16. Generally, estimations of main effects models are

[40] See Myers, R. H., & Montgomery, D. C. (1995). Response surface methodology: process and product optimization using designed experiments (Vol. 4, pp. 156-179). New York: Wiley.

supported by full-factorial, fractional-factorial, or optimal designs. Response surface designs are overkill for main effects models (no need to estimate curvature).

4. Considerations for Validation

The reasons classical DOE is appropriate for physical testing also apply to validation. The wide spectrum of validation techniques all require an estimate of the noise in the physical response variable. Simple hypothesis tests require this, as do regression based techniques. A system's physical response is considered the gold standard representation of reality, so designing a test and using validation methods that characterize and reduce noise in the response is crucial. Classical DOE principles accurately estimate and even reduce this noise, making the comparison to simulation more powerful.

Classical DOE also provides an efficient strategy for selecting points for validation between the model and the live test. Using a classical DOE approach allows for a comparison between live testing and model outcomes across all of the factors considered in the design. This provides increased flexibility for validation, because it is entirely possible that the model may reflect test outcomes under only a subset of the region of interest. Classical DOE allows these differences to be characterized across the domain of the design.

Classical design is also useful for conducting parametric analyses on the model outcomes, which as discussed previously are an important aspect of verification and validation. Parametric analyses quantify the rate at which the response variable changes as the factor settings change. This could be useful for identifying the most or least important factors from a computer experiment to look at in live testing.

B. Computer Experiments

1. Distinguishing Features

Much of the preliminary work on computer experiments can be traced to the 1980s. Sacks, Welch, and Mitchell identified specific challenges in computer experimentation and were the first to apply a type of interpolator model, called a Kriging model, to the analysis of simulation experiments[41]. Their research catalyzed the invention of new experimental designs and modeling improvements that form the contents of modern computer experiment textbooks (see for example Santner (2003)[42], Fang (2005)[43], and Kleijnen

[41] J. Sacks, W. J. Welch, T. J. Mitchell, H. P. Wynn, Design and analysis of computer experiments, Statistical science (1989) 409-423.

[42] T.J. Santner, W.J. Thomas, W. J. Williams, W.I. Notz, The design and analysis of computer experiments, Springer, 2003.

[43] K.-T. Fang, R. Li, A. Sudjianto, Design and modeling for computer experiments, CRC Press, 2005.

(2008)[44]). This section briefly reviews the most important assumptions for computer experiments, which include:

- Having deterministic response variables.
- Emphasizing coverage optimality over robustness.
- Having numerous significant factors and levels per factor.
- Being suitable for evaluation with interpolators.

In general, the development of computer experiments assumes the response variable from simulation is deterministic. That is, repeated samples with fixed factor settings produce the same output. Therefore, replication is not included as part of computer experiments, which provides the opportunity for more levels within each factor. This is an important assumption to highlight because not all models used in the DoD T&E community are deterministic and thus the direct application of a computer experiment may not be appropriate. Computer experiments also tend to assume that factors are generally easy to set and often it is possible to run a large number of trials (i.e., >1000).

Another contrast between physical and simulation experiments is the relative importance of predictions versus understanding the most important factors. Computer experiments emphasize covering the input space and using interpolators for predictions. Computer experiment designs are generated to fill in the space to facilitate robust interpolations. This is especially useful in cases where the outcome is expected to be highly non-linear and not easily approximated by low order polynomial approximations. This space-filling objective contrasts with classical experiments, where designs are generated to obtain the best estimate of the model coefficients, which is best accomplished for a main effects model by pushing points to the edge of the space.

2. Common Design Techniques

Without having to worry about random variation, replication, or most practical considerations about hard-to-change factors, design points in simulation experiments are free to spread out across the design space. Space-filling designs are the general class of experimental designs for computer experiments. Space-filling designs maximize the likelihood of identifying problems with the model or simulation by allowing for the discovery of local maxima/minima or non-linearity (e.g., divide by zero in the code). Figure 17 displays four different space-filling designs: uniform grid, maximin, fast flexible

[44] J. P. Kleijnen, Design and analysis of simulation experiments, volume 20, Springer, 2008.

filling[45], and Latin hypercube. The bottom of the figure displays the distribution of these designs' points projected onto a horizontal axis.

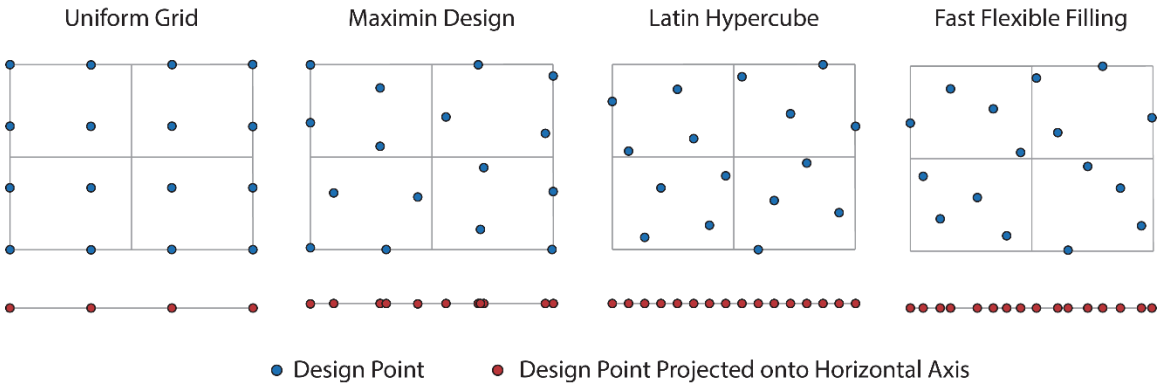

Figure 17. Space-Filling Designs

The simplest way to fill the design space is to create a uniform grid. For example, the first panel in Figure 17 creates a uniform grid with four unique levels per factor. Though uniform grids achieve good spacing, a disadvantage is that the number of design points increases exponentially as the number of factors and levels increases.

The second panel in Figure 17 contains a maximin design, which offers a solution to this design point problem by using an algorithm that optimizes design point spacing for any given number of design points, and for any given number of factors. The maximin design is also known as a sphere-packing design due to the nature of the spacing of its design points.

A disadvantage of maximin (and related minimax) designs is that their projections onto subspaces are unevenly distributed. For example, in the projection of the maximin design onto the horizontal axis in Figure 17, the sixteen design points reduce to eleven [46]distinct points. This means that if the vertical-axis variable does not affect the output, then five runs (or points) of the design would be wasted because they do not provide any additional information over the other eleven runs.

The third panel in Figure 17 contains a Latin hypercube design, which provides a solution to the projection problem by ensuring even projection to smaller subspaces. For example, the projection of the Latin hypercube in Figure 17 onto the x-axis results in sixteen evenly spaced design points. Similarly, projecting the design points to the y-axis would result in sixteen unique levels. Latin hypercube designs maintain good spacing and

[45] Lekivetz, R., & Jones, B. (2015). Fast flexible space-filling designs for nonrectangular regions. Quality and Reliability Engineering International, 31(5), 829-837.

[46] Note points 7,8,9 counting left to right are grouped closely together.

flexibility, like maximin designs, and gain good projection, making them most popular space-filling design for computer experiments with continuous inputs.

The fourth panel in Figure 17 contains a fast flexible filling design, which is the most versatile type of space filling design. It is the only space filling design that allows for categorical variables and is therefore extremely useful in defense applications. Fast flexible filling designs are generated by a simple algorithmic process: 1) randomly generate n>>N data points, where N is the final desired design size, 2) cluster the n data points into N clusters using your favorite clustering technique, and 3) use each cluster to form a design point using a summary statistic, for example the centroid of the cluster. The specifics of the resulting design and the speed at which it can be generated vary based on the clustering method and summary statistic selected.

3. Analysis Methods for Simulation Designs

Computer experiments tend not to use linear regression modeling for two reasons. First, with no random variation in the response, the full complexity of the true response surface can emerge. The response surface may include non-linearity and multiple local maxima or minima. Linear models have difficulties capturing this complexity. Second, for simulation experiments with a deterministic response, it is desirable that the model perfectly fits through all the data. The reason for this is that error between the model fit and the data can only be interpreted as a poorly fitted model. By contrast, when using linear models in physical experiments, one would expect the model to fit imperfectly due to the random variation in the response, and would even have cause for concern if the fit were perfect.[47]

Computer experiments often include thousands of runs. Therefore, one of the goals for analyzing computer experiments is simply to understand how outcomes change across all of the model inputs. The analysis of computer experiments can also inform if there are any unexpected irregularities in the output of the executed code (supporting verification), and the determination of where live test points might most benefit the validation process.

The computer experiments literature focuses on using interpolators to analyze computer experiment outcomes. Interpolators are a class of models that connect observed points with linear or smooth curves. They provide flexibility in fitting nonlinear, global response surfaces, and they are the method of choice for analyzing space-filling designs. Figure 18 contrasts a single predictor linear model (left panel), which is not an interpolator, with two interpolators (middle and right panels). Note that the Kriging model with nugget effect can accommodate a non-deterministic response variable by including an estimate of pure error as part of the estimate of the prediction variance, as shown in the right panel of

[47] Perfect fit might indicate that the model is overfit, which makes the model specific to a particular set of data., reducing generalizability, potentially leading to failures in predicting future observations.

Figure 18. However, this would require a computer experiment that was designed to estimate pure error (i.e., included replicates).

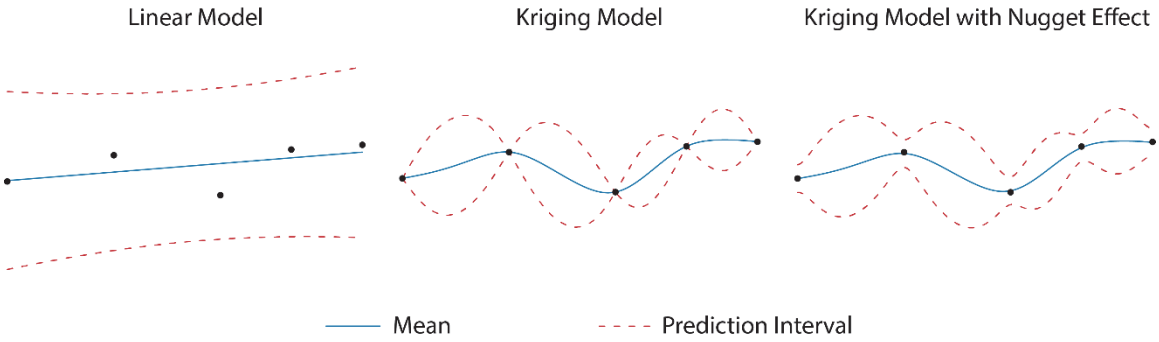

Figure 18. Kriging Model Compared to Linear Model

All interpolation algorithms estimate the response value at a given location as a weighted sum of data values at surrounding locations. This procedure allows for a more perfect fit to the data compared to linear regression. Kriging (also referred to as Gaussian process prediction) is a Bayesian technique that assumes samples come from a normal distribution, and covariance between any two points governs the uncertainty. The advantage of the Kriging model, shown in the middle panel of Figure 18, is its ability to fit the data perfectly. A benefit of the Kriging model over other interpolators is that it accommodates the estimation of prediction variance on the mean.

A Kriging model fit to the simulation output is often called an emulator (or meta-model, among other names). Similar to any other type of model fit, an emulator is used to predict the output of the simulation and can serve as surrogate for conducting additional analyses that require re-running the simulation. This helps cut costs, especially when the simulation is expensive to run.

4. Considerations for Validation

Simulation experiments generated by space-filling designs provide the ability to check simulation code under a wide variety of input conditions. They also focus on the ability to predict simulation outcomes under conditions between design points. These characteristics are good for finding bugs in the code, identifying areas for potential investigation during live testing, and understanding anticipated outcomes from live tests. However, they are sub-optimal from a validation perspective because physical and computer designs generated by different algorithms generally do not contain matching points, which prohibits a straightforward comparison between the response variables from the model and from live testing at any given combination of input variables. Additionally, if the computer experiment includes more variables, they must be accounted for when comparing to live testing.

One approach to validating simulations is to use a computer experiment design to investigate the simulation space. This simulation design can be used to conduct a global sensitivity analysis on the simulation output to determine the relative importance of the each of the input factors. It may be appropriate to omit the factors that have little or no impact on the response variable in designing the physical tests. However, this assumption should be checked ideally with screening experiments in early physical testing and all variables should be recorded during testing. On the other hand, the important factors from the simulation should be strategically set in physical experiments.

Factors can be divided into those that can be controlled in live tests and those that are controllable inputs only for simulation experiments. Factors that are controllable only in the simulation experiment are often referred to as calibration factors. The simplest strategy is to set the calibration factors at fixed nominal values according to subject matter expert opinion or values measured in test (the computer experiment literature refers to this process as calibration). Another commonly used strategy in defense testing is to run Monte Carlo iterations across all of the calibration factors and use a summary statistic (e.g., mean) as the simulation output. Once the calibration factors are accounted for, one can proceed with validation using statistical comparison techniques that ignore the calibration factors. Advanced analysis techniques that simultaneously include calibration factors and then make comparisons to live data are not considered in this handbook.

C. Hybrid Design Approaches

A robust validation strategy will include a combination of classical DOE and computer experiment techniques. A space-filling design covers the model input domain and a classical design can be used for selecting model runs for replication and for matching points to live tests. When combining simulation designs and classical designs into the overall validation process, there are numerous reasonable implementations. The following process is one potential execution of the hybrid approach that has worked in practice. It assumes the simulation is available before live tests. In this case, the validation process might proceed as follows:

1. Conduct a simulation experiment on all model input variables (e.g., using fast flexible filling design).
2. Add replicates to a subset of the simulation experiment points for Monte Carlo variation analysis.
3. Conduct Monte Carlo variation analysis.
4. Conduct parametric analysis. Evaluate the simulation experiment using emulator/interpolator (e.g., Kriging or linear model).
5. Determine important factors, areas for investigation for live testing.

6. Design live tests using classical DOE, record all "calibration" variables, include replicates if feasible.
7. Run live test design in the simulator, set calibration factors at realized values during live tests, and include replications if simulation is non-deterministic.

This approach allows for complete coverage across the simulation space, estimates experimental error for both the simulation and live tests if replicates are included, and provides a strategy for direct matching between simulation and live test points. However, in gaining all of those benefits, it loses its optimality for any single objective.

Additionally, while this approach makes sense for a quick-to-execute simulation with relatively small Monte Carlo variation, it may not make sense for a highly stochastic simulation or one with a long execution time. In cases where simulation experiments take a long time or cost a lot to execute, matching classical designs with the live testing may provide the best approach. See the last section of this chapter for recommendations on which designs to use for specific types of simulations.

D. Hybrid Approach Example

Reconsider our notional test of an air-to-ground missile. We previously identified that the key response variable is miss distance and the factors that might affect miss distance that were controlled during testing included:

- The specific aircraft platform from which the missile is launched (A or B)
- The altitude of the aircraft (Low, Medium, or High)
- Whether or not countermeasures were in use (Yes or No)

A full factorial design was used to design a 12-shot test in each of the 12 conditions (2 platforms x 3 altitudes x countermeasure on/off). For each condition executed during test, 100 replicates were conducted in the simulation. However, there are other factors a simulation experiment could include, which may be difficult or expensive to include in live testing, such as:

- Off-boresight angle (0-180 degrees)
- Release range (5 – 30 nmi)
- Aircraft velocity (Mach 0.5 – 2.0)
- Environment clutter (none, partial, full)
- Wind speed (random normal)
- Release point stability derivatives (random normal based on release conditions).

Clearly some of the factors could also be controlled in the actual live tests (e.g., off-boresight angle, release range, aircraft velocity), albeit not as precisely as in the simulation experiments. Other factors cannot be directly controlled (e.g., environmental clutter, wind speed), although the choice of test day and mission setup will influence their values.

Release point stability derivatives are bounded by the launch acceptability conditions, but not precisely controlled. All of these factors can be input directly into the simulation with the exception of wind speed (random draw) and release point stability derivatives (random draw based on other inputs). The wind speed and release point stability derivatives are included in the simulation as Monte Carlo variables and sampled from a normal distribution, introducing a small amount of stochastic variation in the simulation.

As this is a notional example, we generated our own highly notional miss-distance simulation model (MDSM) that we wish to validate. The model we use for the simulation outcomes is:

$$MissDist = \frac{Altitude + Boresight + RandomNormal\ (5,1)}{4} + \begin{Bmatrix} 0.75, & if\ PF = A \\ 0, & if\ PF = B \end{Bmatrix} + \begin{Bmatrix} 1.25, & if\ CM = Yes \\ 0, & if\ CM = No \end{Bmatrix}$$

where PF is the platform, CM is the countermeasures, and the random normal represents the Monte Carlo for the stability derivatives and wind speed. Using this "simulation" we can run our simulation design.

The first step in our hybrid approach is to generate a computer experiment. We used JMP's Fast Flexible Filling space-filling design to generate a 1000-point computer experiment across six factors: Platform, Countermeasures, Off-boresight, Range, Velocity, and Clutter. Figure 19 shows the design matrix visually using pairwise scatter plots of the 1000 points. The 1000 points provides good coverage across the input space in two-dimensional space.

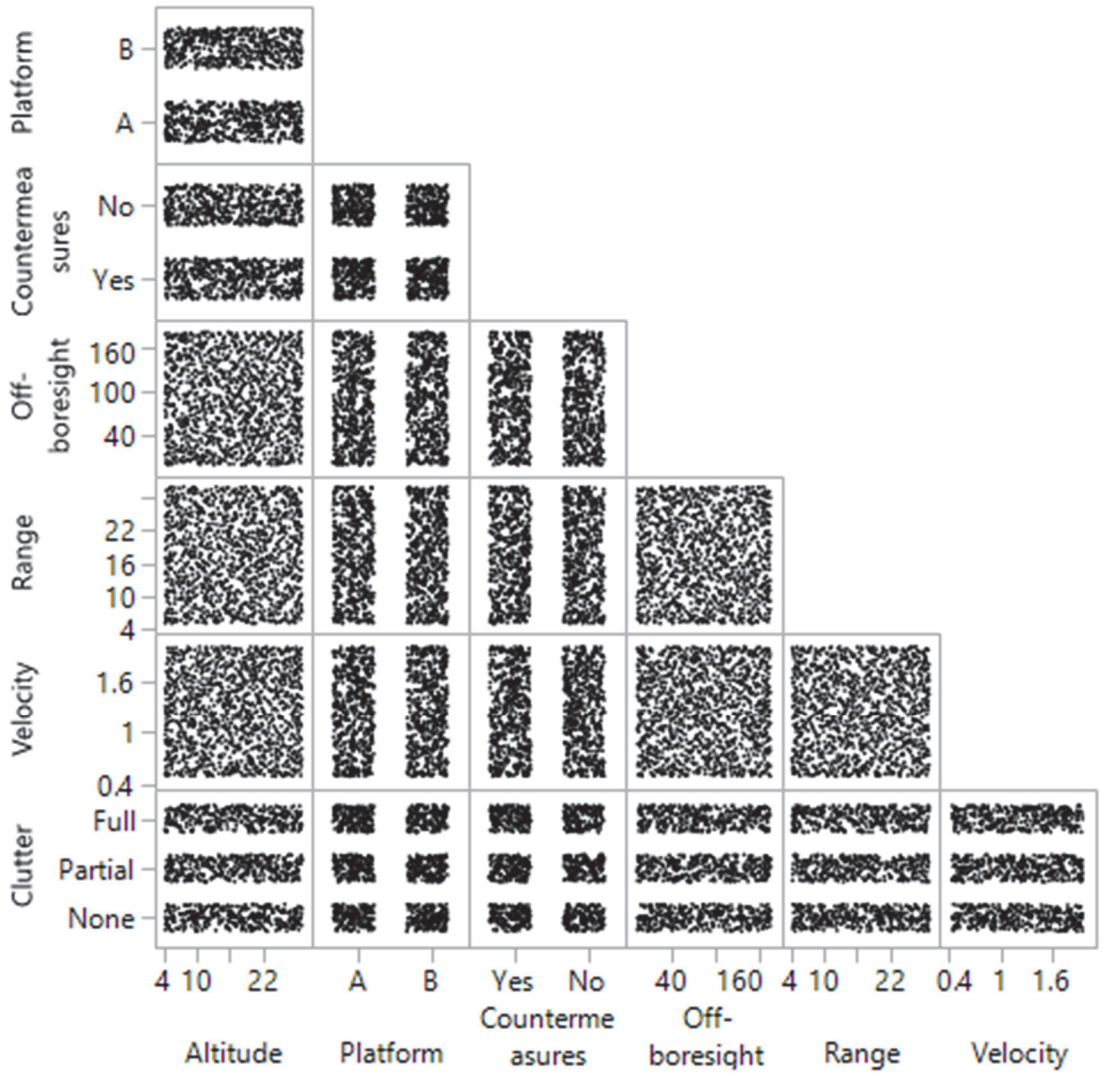

Figure 19. Scatter Plots Illustrating 1000-point Fast Flexible Filling Design

Figure 20 shows the resulting miss distances for the 1000 simulation runs.

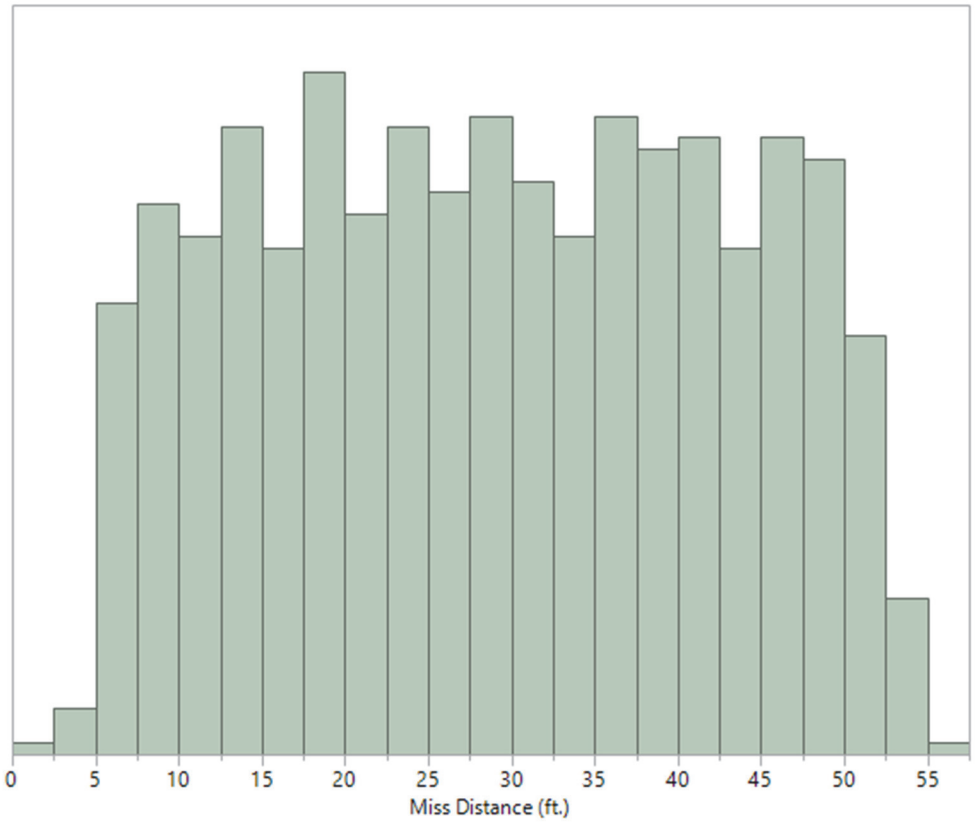

Figure 20. Histogram of Miss Distance

Conducting an exploratory data analysis across all of the potential factors shows that altitude, platform, and countermeasures only reveal slight differences in miss distance, while off-boresight angle is the primary cause of larger miss distances. Note that this makes sense because our hypothetical model has coefficients of similar magnitude for all of the factors, but the input values for off-boresight angles are an order of magnitude greater than other values (e.g., altitude). Figures 21 and 22 contrast boxplots for each platform-countermeasures combination for various altitudes and off-boresight ranges. This provides initial input towards the appropriate analysis and important information that should be validated by live tests.

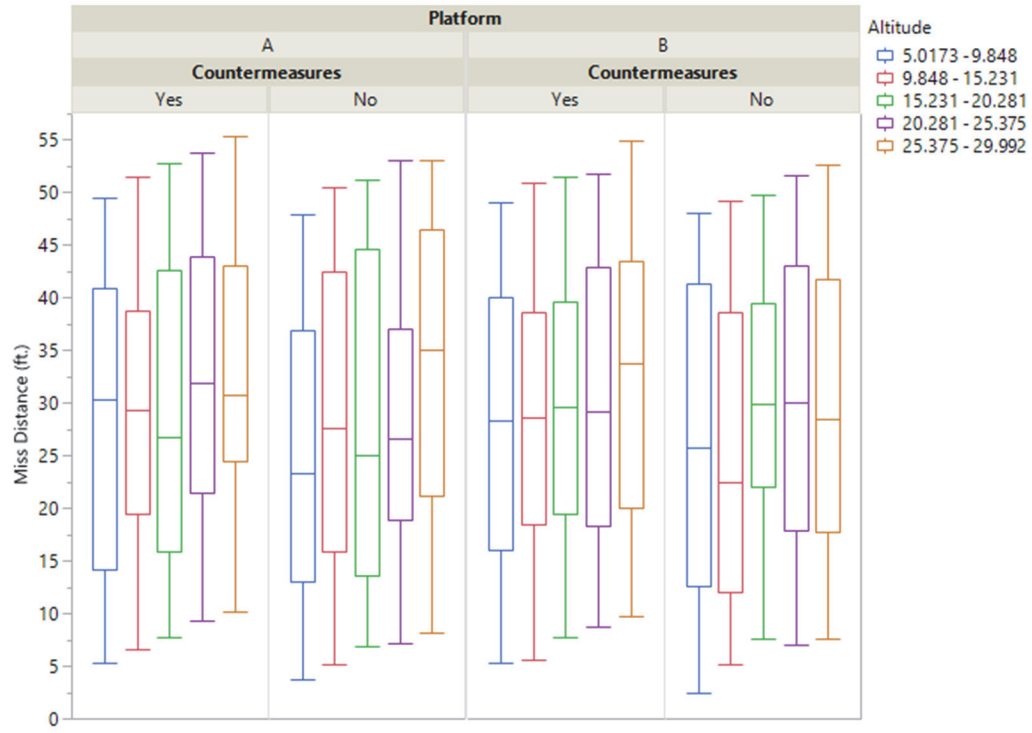

Figure 21. Boxplots of Miss Distance by Platform, Countermeasures, and Altitude

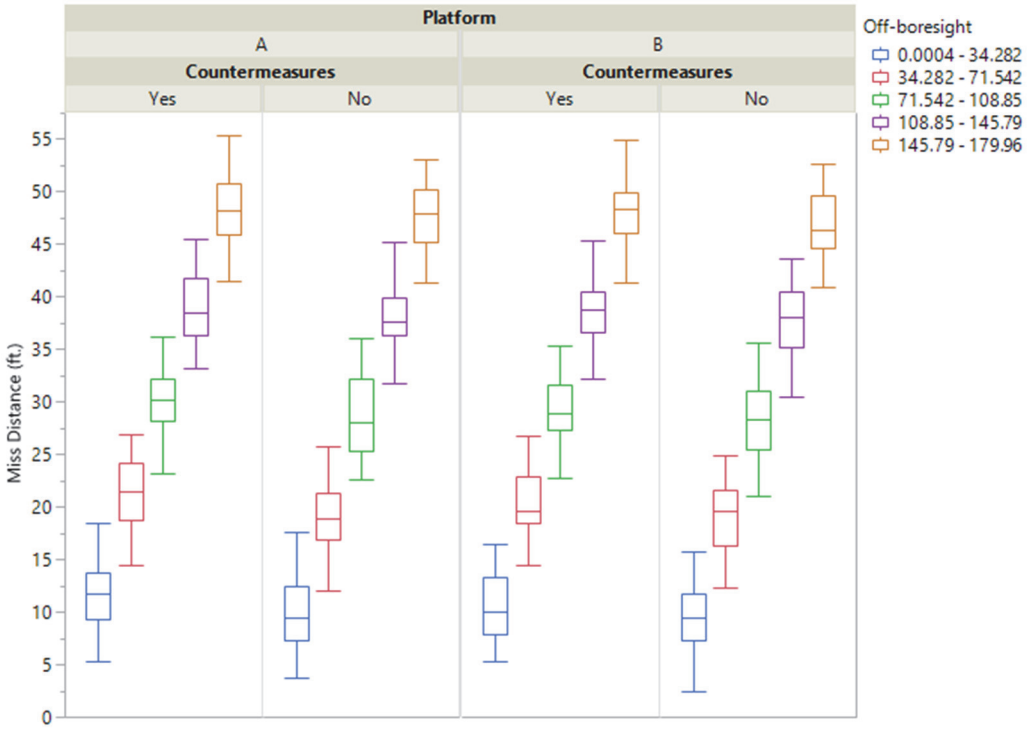

Figure 22. Boxplots of Miss Distance by Platform, Countermeasures, and Off-Boresight

Following the exploratory data analysis of the simulation outcomes, we run a Kriging model and a linear model regression analysis. Figure 23 shows the Kriging model for the two continuous factors for Platform B with countermeasures. Separate Kriging models are fit for each of the categorical factors. Visually, one can identify that larger off-boresight angles and higher altitudes appear to increase miss distance.

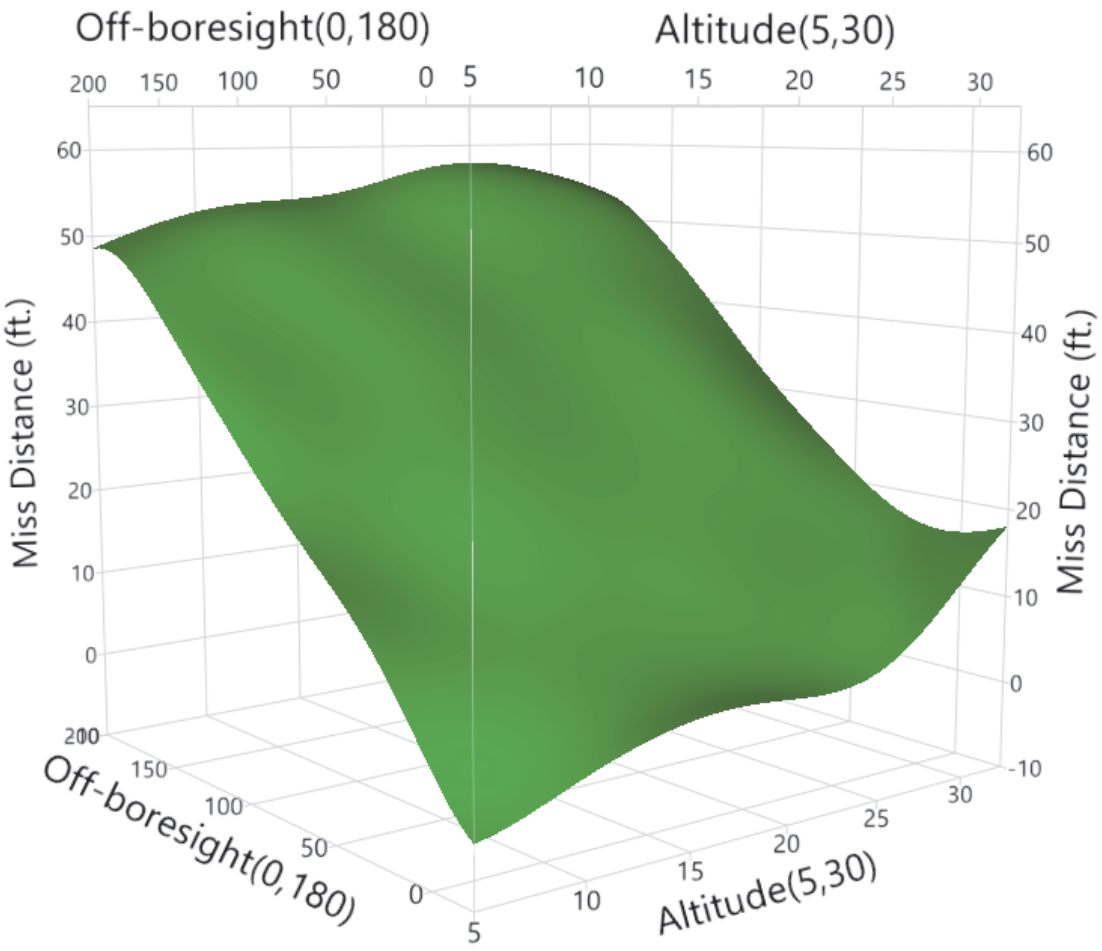

Figure 23. Gaussian Process Model for Platform B, With Countermeasures.

Figure 24 shows the linear model contour plot. Contrasting the linear model fit with the Kriging model provides an intuitive understanding of when each type of model is appropriate. The small curves in the Gaussian process fit reflect the random variation in this extremely simple model, but would be useful in capturing anomalous behavior, or points of interest for a more complex simulation. The linear model more clearly identifies the significant factors and allows for all of the factors (both continuous and categorical) to be analyzed in one analysis framework. Table 4 shows that the linear model easily

identifies the four factors we programmed into our simulation (off-boresight angle, altitude, countermeasures, and platform) as statistically significant (P-value near 0)[48].

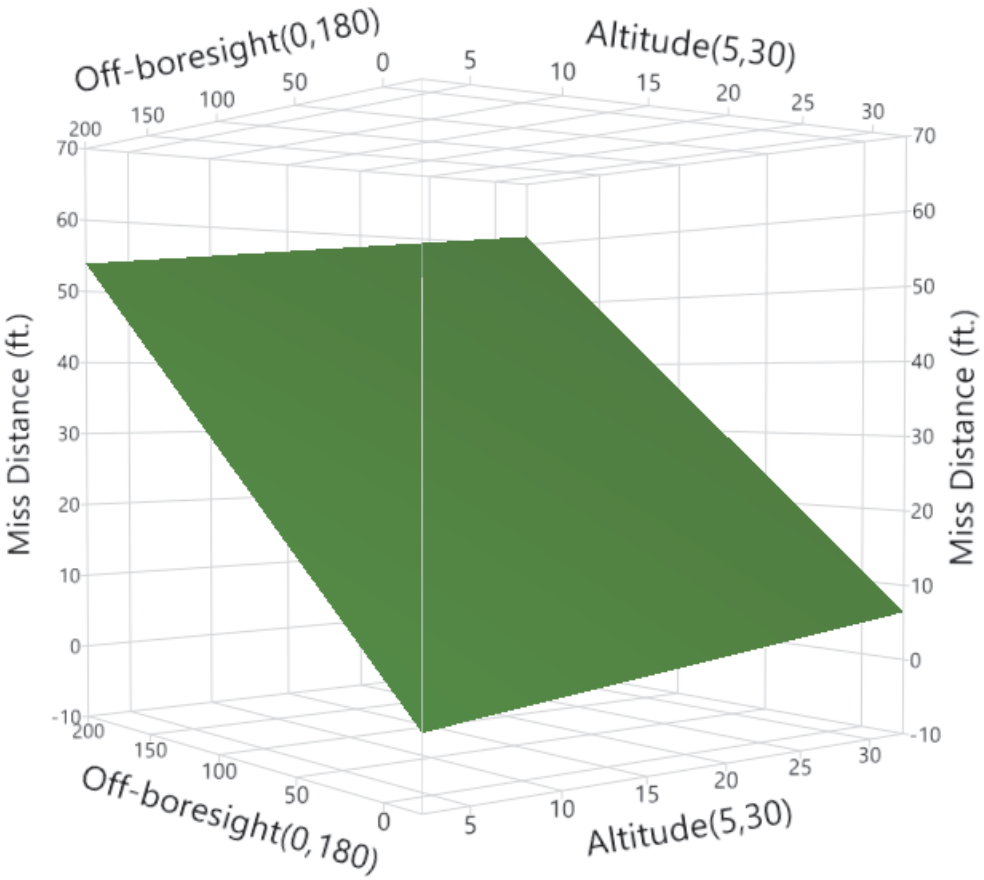

Figure 24. Gaussian Process Model for Platform B, With Countermeasures.

Table 4. Linear Model Significant Factors

Factor	Model Terms	P-value
Altitude(5,30)	1	<.0001*
Platform	1	<.0001*
Countermeasures	1	<.0001*
Off-boresight(0,180)	1	<.0001*
Range(5,30)	1	0.1804
Velocity(0.5,2)	1	0.4165
Clutter	2	0.8316

[48] Note the p-value is the probability that if the factor truly had no effect on miss-distance, then these data or data resulting in a larger effect would have been observed. A small value (typically ≤ 0.05) indicates strong evidence against the null hypothesis and we conclude the factor is statistically significant.

Now we have some interesting information to consider in our test planning process. The model predicts that off-boresight angle has a large effect. With that piece of information, we can plan a more informed test design than the example from Chapter 3. Based on the results of this simulation study we generate a new test design, this time one with four factors – adding off-boresight angle. Figure 25 shows the new test design. It is a 12-point D-optimal design generated to support the estimation of main-effects for platform, countermeasures, altitude, and off-boresight. It also includes points for estimating quadratic terms for off-boresight and altitude, and the interactions terms *platform * countermeasures* and *countermeasures * off-boresight*.

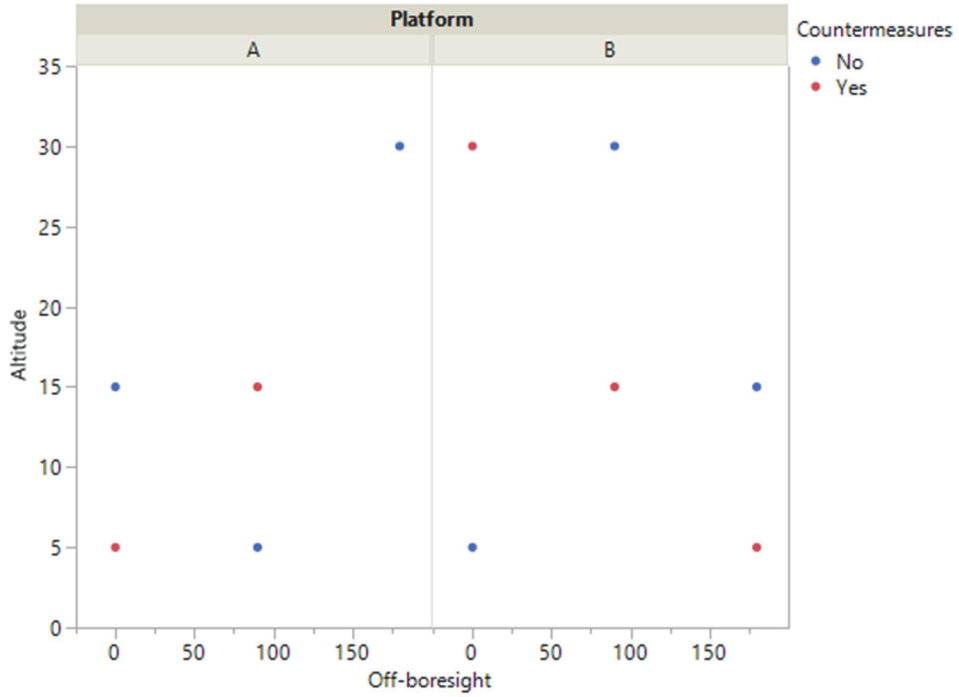

Figure 25. Twelve-point D-optimal Classic Design for Live Testing

The next step would be to run these 12 points in live testing and replicate them in the simulation, matching calibration variables (i.e., range, velocity, clutter) to the values observed during live tests. The validation analysis illustrated in the example in Chapter 3 would then be applied to this new dataset.

This illustrative example shows that by using a combination of computer experiments and classical designs we can ensure sufficient data for validation, customize the design of live tests to emphasize information learned from the simulation, and maximize the probability that we identify potential problems in the model by spanning the model input domain efficiently.

E. Recommended Designs

At each stage of the model-test-model cycle, designed experiments should be used for both the live test and model runs to ensure that an adequate comparison can be made at the desired sensitivity. Deciding how much live data are needed will ultimately depend on how much uncertainty in the difference between live data and simulated outcome is acceptable across the factor space, which will in turn depend on the intended use of the model or simulation, and how much validation testing has been done previously.

Live test plans should always employ classical experimental designs and include replication on a subset of the design if possible. Various statistical measures of merit can help determine an adequate design and sample size. Power and confidence are directly related to the amount of uncertainty in comparisons. Increasing the power of the hypothesis test comparing the model with live testing leads to smaller confidence interval widths in the analysis.

Figure 26 contrasts conceptually two strategies for selecting points from models and live testing. A key element of this selection is what points to conduct in both live and model-based testing. Ideally, the live design should encompass the simulation design (left panel) so that comparisons made between the two are interpolations instead of extrapolations. However, this is often not feasible due to practical constraints that exist in live testing. In these cases, illustrated in the right panel, the domain of the live testing should span the maximum possible domain of the simulation experiment and regions of extrapolation should be clearly identified in the validation limitations.

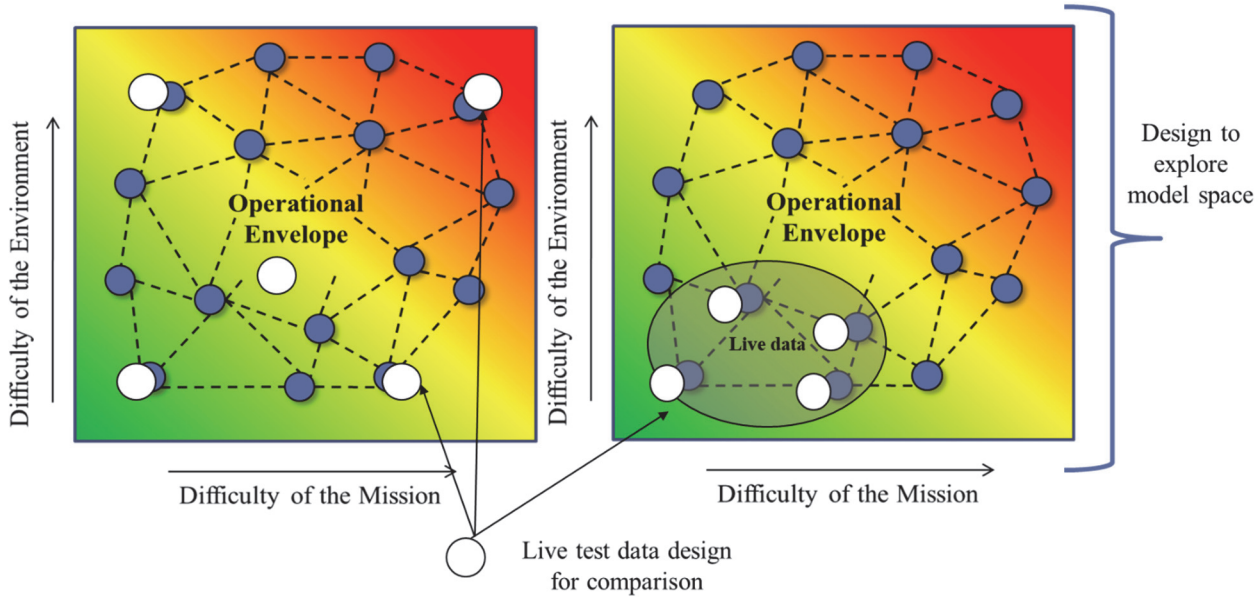

Figure 26. Notional Picture of Data Selection Across M&S and Live Testing

Direct matching of points between the live testing and simulation provides the best validation strategy in all cases. If the subset of matched points are based on a classical designed experiment, the validation strategy can include regression analysis, which is the most powerful validation approach for characterizing differences across a range of conditions. If the subset of matched points do not come from a designed experiment, statistical comparison is still possible but requires aggregation of the comparison across conditions using a less powerful approach. If data cannot be matched directly, emulation and prediction is the best approach for validation.

The best design for the simulation experiment depends on the analytical goal and the nature of the simulation and the data it produces. Statistical designs should support both comparison with live data and exploration of the model space itself, including conducting sensitivity analyses and building emulators. For completely deterministic simulations, such as most finite element models, space-filling designs are the recommended approach for both comparison and model exploration. On the other end of the spectrum, highly stochastic models, such as effects-based models, operator-in-the-loop simulations, or system-of-system models, classical designs are the recommended approach for both goals. For simulations somewhere in the middle in terms of randomness, such as a physics-based model with some built in Monte Carlo (random draw) input variables, a hybrid approach is recommended. In this case, a space-filling approach can be useful for building an emulator, but replicates are also needed to characterize Monte Carlo variation. Table 5 below summarizes these recommendations.

Table 5. Simulation* Design Recommendations

	Recommended Method by Validation Goal	
Level of Randomness	Compare to Live Data	Explore Model Space
None (Deterministic)	Hybrid Design	Space Filling
Low (E.g., Physics-based with calibration factors)	Classical	Hybrid Design
High (E.g., Effects-based, Human-in the-loop)	Classical with Replications	Classical with Replications

***The recommended strategy for live data is classical DOE.**

Approved for public release; distribution is unlimited.

5. Conclusions

Computer models and simulations are critical sources of information for developing, testing, and evaluating military systems. In the development of new systems, systems engineers and developmental testers may use engineering- and engagement-level models to refine system design and evaluate design tradeoffs in meeting performance requirements. Once a system design is finalized, engagement- and mission-level models can be used to design live tests to answer informed questions about operational effectiveness, suitability, or survivability. The models can be useful in bolstering conclusions from live testing by filling in gaps in knowledge, identifying edges of the operating space, and providing limited abilities to extrapolate findings. They are also useful in shaping operational test environments, and can provide test environments that cannot be generated by live test assets alone.

The validity of the information provided by models, and therefore the trustworthiness of the evaluations that use models, depends on a rigorous verification and validation process. In the past, validation processes have lacked sufficient statistical rigor to support the quantification of uncertainty in accreditation decisions. This handbook provides the foundational design and analysis tools that are required to support a statistical comparison of simulation and live data. These techniques are essential for being able to quantify statistical uncertainty in models used for evaluation of the effectiveness, suitability, and survivability of military systems.

Advances in computational power have allowed both greater fidelity and more extensive use of such models. Almost every complex military system has a corresponding model that simulates its performance in the field. In response, the DoD needs defensible practices for validating these models. The focus of this handbook was to provide an overview of the process, the required data, and the core statistical analysis techniques that are useful for validating the use of computer models for operational test and evaluation. It is not exhaustive and does not include the most recent developments in model calibration and uncertainty quantification, which could also benefit the DoD. It provides a first step forward towards a sound methodology for data-driven validation.

Approved for public release; distribution is unlimited.

Analysis Appendix

As discussed in Chapter 3, Monte Carlo power simulations were used to make the analysis recommendations. Monte Carlo simulation uses iterations of random draws from a probability distribution of interest to estimate uncertainty. The three Monte Carlo settings varied are listed below:

Distribution of the response variable

- Skewed – data generated from the lognormal distribution
- Symmetric – data generated from the normal distribution
- Binary – data generated from the binomial distribution

Structure of factors

- Univariate – no varying factors across the collected data
- Distributed Level Effects – factors significantly affect the mean, but no underlying designed experiment. The difference between simulation and test varies across factor levels. The amount of variation across factor levels is represented by a distribution (hence, the name distributed level effects).
- Designed experiment – factors in the underlying designed experiment determine the difference between the live and simulation data with significant factor mean effects

Validation Referent Data Size

- Small – 2-5 (continuous data) / 20 (binary data)
- Moderate – 6-10 (continuous data) / 40 (binary data)
- Large – 11-20+ (continuous data) / 100+ (binary data)

Candidates for appropriate statistical methodologies to use in each of the above settings came from standard statistical literature[49][50][51]. Similar tests were chosen as candidates in the skewed and symmetric data cases. These tests include:

- t-test (and log transformed t-test in the skewed case)

[49] Montgomery, Douglas C. Design and analysis of experiments. John Wiley & Sons, 2017.

[50] Hollander, Myles, Douglas A. Wolfe, and Eric Chicken. Nonparametric statistical methods. Vol. 751. John Wiley & Sons, 2013.

[51] Agresti, Alan. Categorical data analysis. Vol. 482. John Wiley & Sons, 2003.

- Parametric Kolmogorov-Smirnov (KS) test – the simulation and live data are assumed to be from the same distribution and are together compared to the true CDF of the appropriate distribution family with parameters estimated from the sample
- Non-parametric 2-sample KS test – the simulation and live data compared to each other (can scale and center each factor level and perform the KS test on transformed data to account for factor effects)
- Anderson-Darling (AD) test (or log transformed AD test in the skewed case)
- Empirical test 1 – find the median of the live data and find the (smaller) empirical quantile of that value in the simulation data
- Empirical test 2 – find the more extreme of the 25th or 75th quantile of the live data (as opposed to the median previously)
- Fisher's Combined Probability Test – find the empirical quantile of each data point in the live data as compared to the simulated data and then combine the p-tails via Fisher's combined probability
- Combined test 1 – combine results from the scaled non-parametric KS test and Fisher's Combined Probability Test
- Combined test 2 – combine results from the t-test, scaled non-parametric KS test, and Fisher's Combined Probability Test
- Regression (linear in the symmetric case and lognormal in the skewed case) – traditional application using all the data from both live and simulation
- Matched data size regression – data sizes between live and simulation are first matched (via a resampling scheme)
- Emulation and prediction – simulation data is used to create prediction intervals and live data is evaluated against those intervals

In the case of binary data, we must consider a different set of techniques for comparison. For the purposes of the Monte Carlo simulation, the parameter p (from the binomial probability distribution) is set to .5 for the simulation data set and allowed to vary in the live data. The methodologies considered include:

- Fisher's Exact test
- Chi-squared test (equivalent to test for difference in proportions)
- Permutation test – randomly assign labels of the data as simulation vs. live and calculate an odds ratio in each simulated case. Then the odds ratio of the true data is empirically compared to the distribution of odds ratios.
- Logistic regression on all the data
- Matched data size logistic regression
- Emulation and prediction – create a confidence interval for each factor level and then use bootstrapping of the live data to obtain samples to compare coverage of that interval.

In order to compare results across these techniques meaningfully, a type I error[52] correction was empirically computed individually for each test. For each combination of sample size and number of factor levels, a p-value cutoff value was computed that most closely yielded the desired type I error rate in the overall test, i.e., the type I error's quantile of the observed rejection p-values. Then, when a p-value is computed via a test (in the original way), it is multiplied by the ratio of the desired level and that cutoff value. By doing this, the p-value obtained by this procedure can be compared directly to the originally intended type I error rate, e.g., .2. Thus, the user does not need to consider the type I error rate in active use: the p-values obtained will be comparable to familiar values. The benefit of performing this type 1 error correction is shown in Figure A-1.

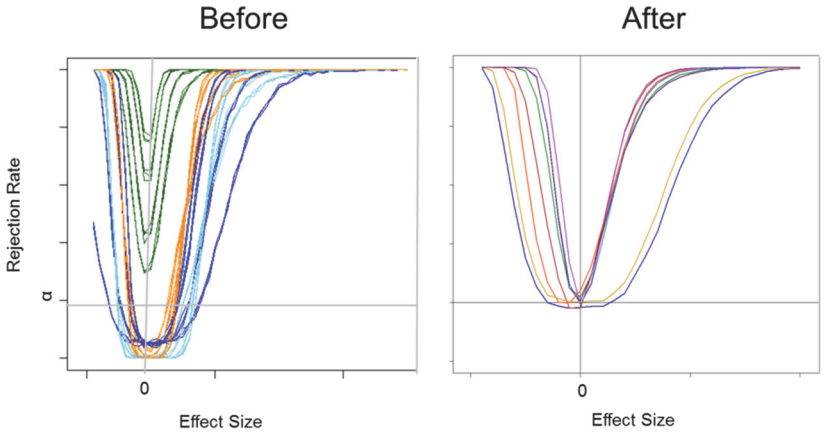

Figure A-1. Power Curves Before and After Type 1 Error Correction

1. Univariate Techniques

One overarching class of statistical methods is the hypothesis test on a single parameter. Such tests compare a certain metric of interest, such as the mean or the variance, of one data set with that of another data set. Examples of methods in this class include the t-test, the F-test for equality of variances, empirical median or quantile tests, and several nonparametric procedures, such as the Wilcoxon signed rank and rank sum tests[53]. These tests are aggregate tests in that they are not designed to account for factor differences. Overall, these techniques can be useful for quickly invalidating a model, but since they only capture one dimension of the data, they typically are not sufficient for a comprehensive statistical comparison between model output and live tests.

---

[52] Type I error is the probability of rejecting a true null hypothesis and is typically set a priori when conducting statistical tests

[53] George E. P. Box, William G. Hunter and J. Stuart Hunter, "Statistics for experimenters," 2nd edition, John Wiley & Sons, 2005.

a. T-test

Example

Consider an operational test of a truck in which testers collect miles per gallon data during multiple runs of a single specific operational mission type. The Gas Mileage Simulator (GMS) lab conducts dozens of simulations under the same mission scenario and environment as the live test. Gas mileage data is symmetric (approximately normal) both in the lab and in real life. As part of the validation process of GMS, testers are interested in determining whether the average miles per gallon from the simulation is statistically the same as in the live environment for this particular mission.

Description of Method

The t-test is a parametric statistical hypothesis test used to compare the means of two data sets. This test is only valid if the data is approximately normal and observations are independent of one another. The test statistic is as follows:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

where $\bar{x}_1$ is the mean of the simulated data, $\bar{x}_2$ is the mean of the live data, $s_1$is the standard deviation of the simulated data, $s_2$is the standard deviation of the live data, n_1 is the sample size of the simulated data set, and n_2 is the sample size of the live data set. This statistic is then compared to the critical value of a t-distribution with degrees of freedom $\frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1-1) + (s_2^2/n_2)^2/(n_2-1)}$ in order to determine whether the observed difference in means is statistically significant.

Characterization & Limitations of Method

The t-test is one of the most powerful tools for detecting differences in means, provided the parametric assumptions are met. However, the test is limited in the sense that it ONLY determines changes in means. The t-test is not designed to check for differences in variance, nor does it account for any possible factor effects.

As a side note, if the data is lognormally distributed a t-test can be performed on the log-transformed data using the same procedure as above.

As shown in the summary recommendations table in Chapter 3, the t-test performs well compared to other methods when the goal is to detect a change in mean, the sample

size is moderate or large, and the data is symmetric. If the data is skewed, the log t-test performs well under the same conditions.

Interpretation of Results in Context of Example

Returning to the truck miles per gallon example, suppose 15 runs of the mission were conducted during the live operational test and produced a mean miles per gallon of about 17.3 and a standard deviation of about 5.8. Fifty runs of the mission were conducted in the GMS lab and produced a mean miles per gallon of around 24 and a standard deviation of 7.1. The test statistic in this case would be:

$$t = \frac{26 - 20.5}{\sqrt{\frac{6.6^2}{50} + \frac{8^2}{15}}} \approx 2.43$$

The associated p-value for this test is less than .01. Thus, in this hypothetical example, the mean miles per gallon from the simulation is statistically different from the mean miles per gallon from the live data at the .01 significance level. A visual representation of this difference is shown in Figure A-2. The small dots are the individual data points, the large dots represent the mean of each group, and the error bars represent a 95% confidence interval about those means. These intervals quantify the statistical uncertainty in the results.

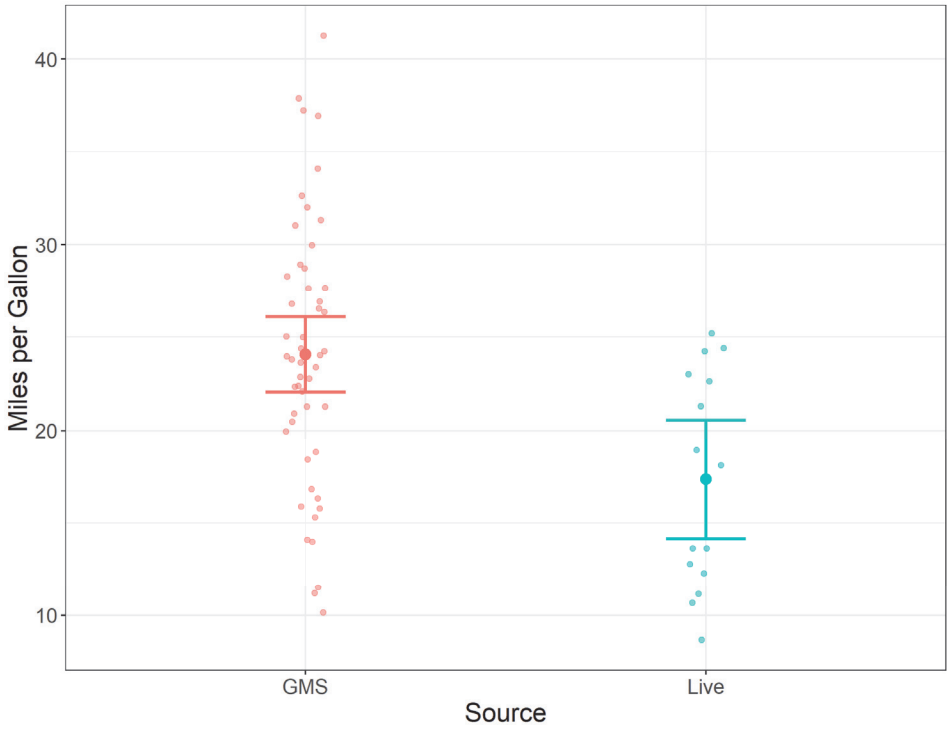

Figure A-2. Box Plots of Miles Per Gallon for GMS Verse Live Data

Prior to the test, testers and subject matter experts decided on an acceptability criterion of 5 miles per gallon. In other words, a difference in mean miles per gallon between live and simulation of 5 or more would be operationally significant. In this case the raw difference in mean miles per gallon is 7.7. Considering both the statistical and the practical results, testers concluded that the simulation did not adequately represent the real world and did not accredit the GMS.

R code

If *x* is the simulated data set and *y* is the live data set, the analysis for comparing the means of the two samples in R would be:

```
t.test(x,y)
```

2. Distribution Techniques

A second class of methods compare the entire distribution, or shape, of a data set to that of another data set, rather than focusing on a particular summary statistic such as the mean or variance. Examples of statistical tests that employ distributions include the Kolmogorov Smirnov (KS) Test, Anderson Darling[54], and Fisher's Combined Probability Test[55]. Methods that worked well for many settings in our simulation study were the non-parametric KS test, and Fisher's Combined Probability Test.

a. Kolmogorov Smirnov Test

Example

Consider an operational test of a stealthy air vehicle in which testers collect detection range data during multiple runs of a single specific operational mission type. Because live testing is costly, the B 5000 X lab conducts many simulations under similar mission scenarios and environments as the live test. In order to validate the simulation, the testers would like to determine whether the distribution of the simulated data and the distribution of the live data are the same.

Description of Method

The two-sample KS test is a non-parametric test used to determine whether two samples are drawn from populations with the same distribution.

---

[54] M.A. Stephens, "EDF Statistics for Goodness of Fit and Some Comparisons," Journal of the American Statistical Association 69.347, pp. 730-737, 1974.

[55] Ronald A. Fisher, "Statistical Methods for Research Workers," Oliver and Boyd, 1925.

To compare a sample with size m and observed cumulative distribution function $F(x)$ and a sample with size n and observed cumulative distribution function $G(x)$, the test statistic is the maximum absolute difference between the two cumulative distribution functions, or:

$$D_{m,n} = \max_x |F(x) - G(x)|$$

The statistic is then compared to the critical value D_α (where α is the level of significance) to determine the probability of observing a D statistic at least as large as the one calculated if the null hypothesis is true. For large sample sizes, the critical value is $D_\alpha = c(\alpha)\sqrt{\frac{m+n}{m*n}}$, where the coefficient is given by the table below:

α	$c(\alpha)$
0.10	1.22
0.05	1.36
0.025	1.48
0.01	1.63
0.005	1.73
0.001	1.95

The null hypothesis H_0 is that both samples are drawn from populations with the same distribution. If $D_{m,n} \geq D_\alpha$, we reject the null hypothesis and conclude that the samples are drawn from populations with different distributions.

The KS test can be adapted to account for data collected under different conditions by scaling the data first. For each distinct condition:

$$scaled\ data = \frac{each\ individual\ data\ point - mean(all\ data\ in\ that\ condition)}{standard\ deviation(all\ data\ in\ that\ condition)}$$

Characterizations & Limitations of Method

One of the biggest strengths of the KS test is that it does not depend on the underlying distribution being tested. It is simple computationally and provides a good starting point to direct further analysis, as it considers the overall shape of a distribution rather than focusing specifically on central tendency or dispersion. It can also be modified easily to account for factors.

The main limitation of the KS test in the context of validating models is that multiple live data points are required in order to form a distribution.

Interpretation of Results in Context of Example

During live operational testing, 20 trials of the mission were conducted and the mean detection range was found to be 22 km, with a standard deviation of 1.8 km. The B 5000

X lab simulated the same mission scenario under the same conditions 100 times and found the mean detection range to be 20 km with a standard deviation of 1.5 km. The CDFs for each data set are shown below.

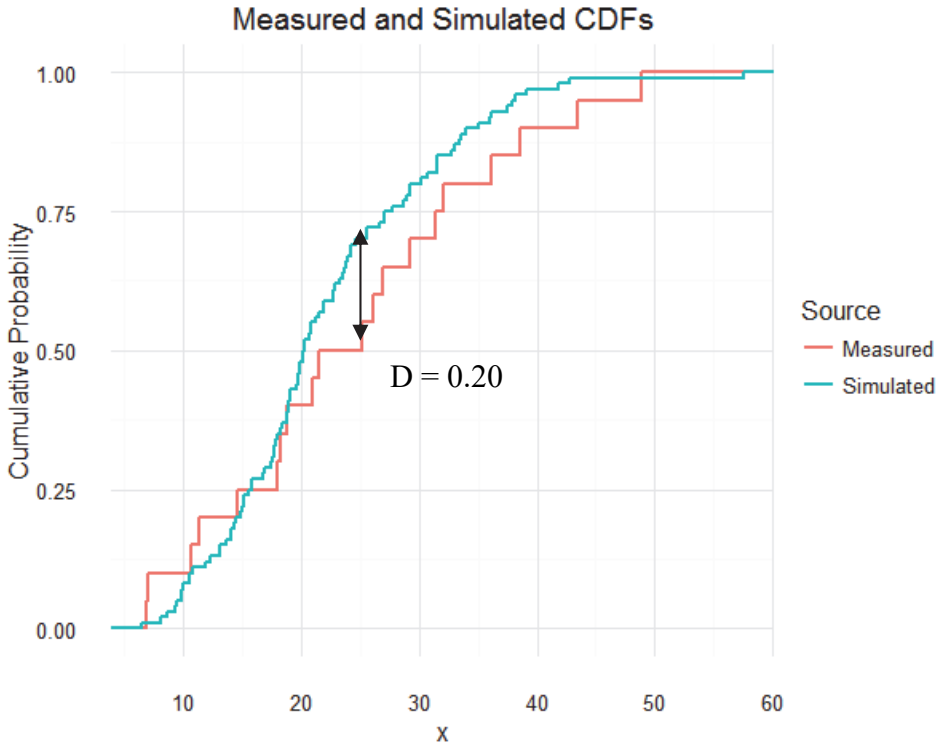

Figure A-3. Cumulative Probability Functions for Measured and Simulated Data

On visual inspection, the detection range data do not appear to be normally distributed. We can use the KS test to determine whether these samples come from populations with the same distribution. The D statistic is found to be 0.20, and the critical value D_α at the .05 significance level is calculated as:

$$D_\alpha = 1.36\sqrt{\frac{100+20}{100*20}} \approx 0.33$$

Because the D statistic is less than the critical value, we cannot reject the null hypothesis that both samples are drawn from populations with the same distribution. The simulation is considered to be valid.

R Code

```
ks.test(simulated, measured)

##
##                Two-sample        Kolmogorov-Smirnov        test
##
```

```
##        data:                simulated        and        measured
##       D       =       0.2,       p-value       =         0.4804
## alternative hypothesis: two-sided
```

b. Fisher's Combined Probability Test

Example

Consider an operational test of a generic air-to-air missile. Since each live missile shot is costly, missile simulations will be used to supplement the live data. Missile simulations typically employ a six-degree-of-freedom (6-DOF) model, which includes tactical code and simulated environments. In a typical model run, the 6-DOF model will simulate the flyout of the missile from launch to target intercept, producing a missile miss distance. Other simulations will then produce a warhead fuze time and calculate the effects of blast and fragments on the target to produce a Probability of Kill.

For each live missile firing, the model is run typically 50 to 100 times, producing one miss distance for each run. In order to validate the simulation, testers would like to determine whether the distribution of the simulated miss distances is consistent with live miss distances across all conditions tested.

Description of Method

Fisher's Combined Probability Test is a method for combining the results from several independent tests, each testing common hypotheses of interest. Essentially, the test is an analysis of analyses. For each independent test, a significance level, i.e., a p-value is calculated. These individual p-values, called p-tails below to differentiate from the final overall p-value, are combined into a joint test to address whether there is a collective bias. While not all of the p-tails individually are small, when taken collectively their trend might strongly support a statistical conclusion about the null hypothesis.

There are several possible methods of calculating these individual p-tails. In the context of our missile miss distance example, if retaining the two-dimensionality of the space is important, one could calculate mathematical contours of equal point density. The left side of Figure A-4 provides a notional example, with model runs in yellow, the live fire miss distance in red, and multiple colored contours, each containing the same number of points between contours. The p-tail for each individual shot is determined as the proportion of model runs that are further away from the target than the live missile firing, in this case 23/93, or 0.2473 (see right side of Figure A-4).

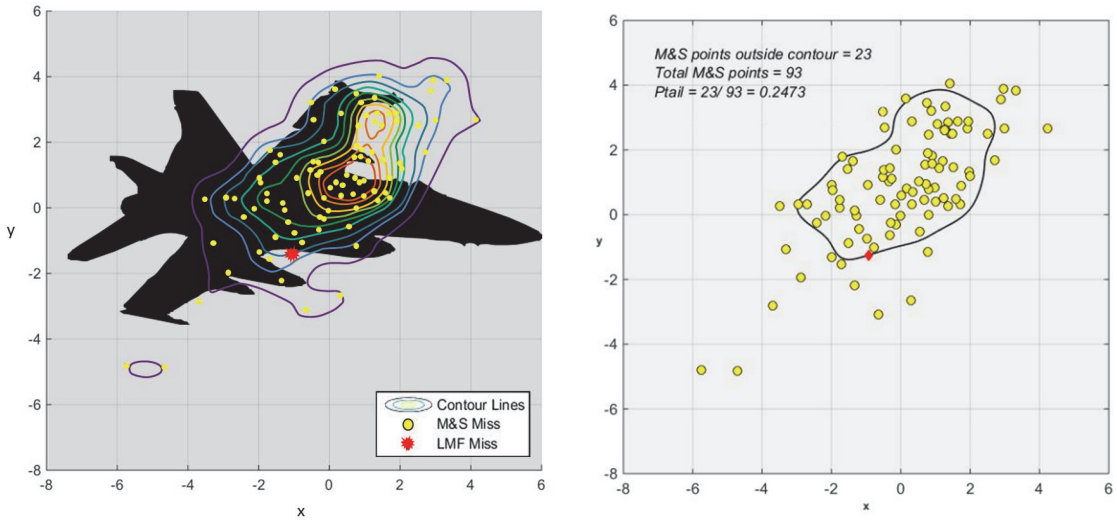

Figure A-4. Simulated Miss Distances Plotted on 2-dimensional Plane With Live Fire Result

If the two-dimensional aspect of the miss distance data is not important (e.g., all that matters is radial miss distance from the point of origin), then p-tails can be calculated via quantile from the one-dimensional radial miss distance distribution. Using the same notional data from Figure A-4, the histogram of distances from the origin is shown in Figure A-5, with the live missile firing miss distance indicated by a vertical red line. In this context, the p-tail is the proportion of radial miss distances greater than the observed live miss distance, which equals .6559.

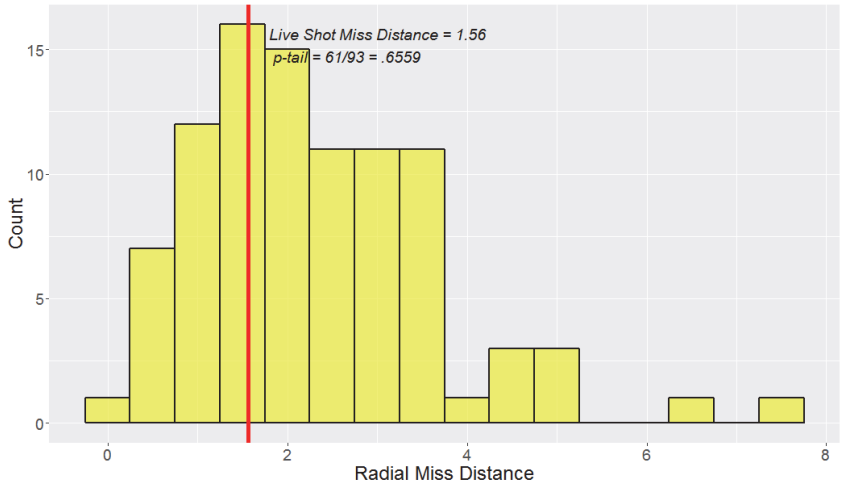

Figure A-5. Simulated Miss Distances Plotted in One Dimension With Live Fire Result

If the p-tails for several flight tests are spread evenly between 0 and 1, the model is deemed a good predictor of miss distance. Figure A-6 below shows three of the possible cases. Each red diamond represents a single p-value calculated from a single flight test. In Case 1, p-tails tend low. This indicates that the model is producing miss distances that are smaller than reality. This is bad because the model is likely to give Probability of Kill

values that exaggerate performance. In Case 2, p-tails are spread evenly. This is the hoped-for situation since it indicates the overall model is neither exaggerating nor underestimating system performance. In Case 3, p-tails tend high. This indicates that the model is producing miss distances that are greater than reality. This is bad because the model is likely to give Probability of Kill values that underestimate performance.

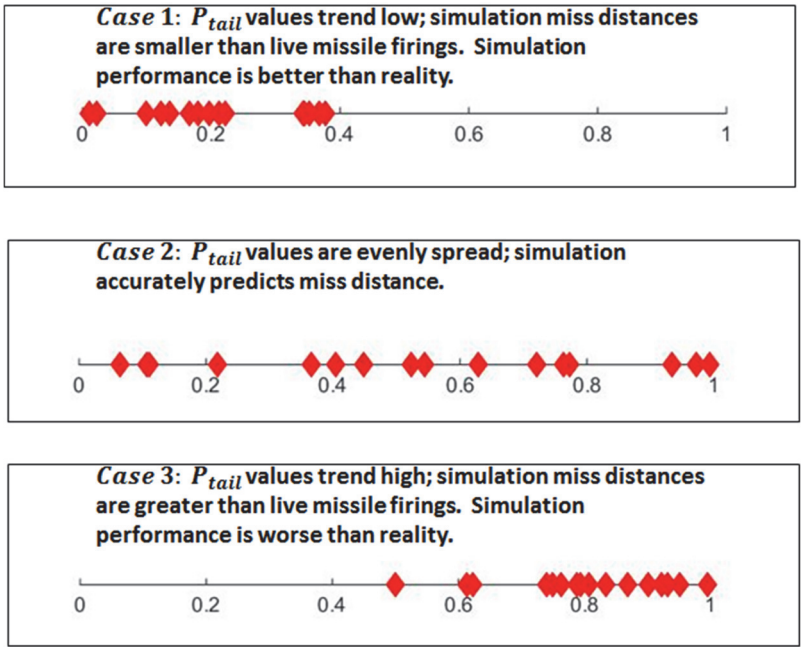

Figure A-6. Possible Distributions of p_{tails}

Once individual p-tails are calculated separately for each unique condition (or set of conditions) under test, they can be formally combined to produce an overall statistical conclusion regarding how well the simulation miss distances match the live fire results.

The statistic for the Fisher's Combined Probability Test is computed as:

$$X = -2\sum_{i=1}^{n} \ln(p_i)$$

where p_i is the p-tail of the ith hypothesis test. The test statistic follows a chi-square distribution with $2n$ degrees of freedom, from which a p-value for the global hypothesis can be obtained and used to determine whether the null hypothesis, H_o, can be rejected. Note the combined test represents the same type of test (one-sided vs. two-sided) of the p-tails from the individual tests. That is, if the p-tails before combination are one-tailed (as presented above), then the combined p-value represents these one-sided hypothesis tests. If the p-tails before combination are two-tailed, the combined p-value represents a two-sided hypothesis test.

Alternatively, one could simply use a KS test (see previous section) to test when the set of individual p-tails follows a uniform distribution. The choice between using Fisher's

test and the KS test comes down to what kind of differences are important. Consider the case where there is a clustered concentration of p-tails all in close proximity to 0.5 (as in Figure A-7). These data are clearly not uniform, so the KS test will reject the global null hypothesis. But is this the conclusion we would actually want to make? The live data are all right in the center of the simulation cloud so there is no evidence of the simulation being off in center, only perhaps in spread. Fisher's test is completely insensitive to such circumstances and will fail to reject the global null hypothesis in this case. So whether or not it is important for the simulation and live data to match closely in terms of spread for a specific intended use will dictate which statistical test on the p-tails is most appropriate.

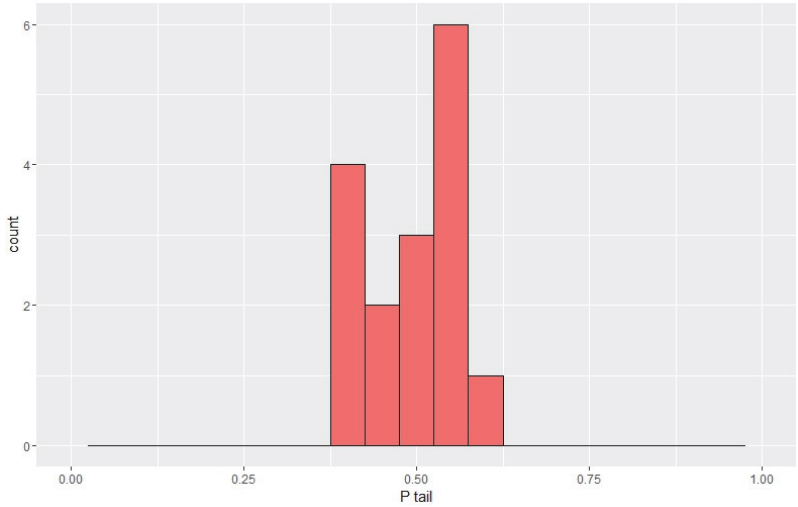

Figure A-7. Notional p-tail Values Clustered Around 0.5

Characterizations & Limitations of Method

Fisher's Combined Probability Test is a powerful technique for detecting differences in variance between the live data and simulation outcome when the variance of the live data is much larger than that of the simulation.

A limiting property of Fisher's Combined Probability Test methodology is that a single small p-tail can lead to a rejection of the null hypothesis even when the remaining n-1 p-tails are nominally large. Rather than simply dismiss the usefulness of Fisher's test in this situation, the "outlier" p-tail should drive one down a path of exploration and discussion as to how and why the "outlier" p-tail resulted and whether a single test data outcome should invalidate the model under scrutiny.

In addition, this test does not account for factor effects the way a statistical model would. If a structured designed experiment was performed in the live and simulated environments, this is not the best technique to apply.

Interpretation of Results in Context of Example

In this particular case, the simulation was assessed to not be systematically biased, thus the testers determined that one-dimensional miss distances were sufficient for their assessment. The notional data plotted in Figures A-4 and A-5 represent one operational condition; similar data was collected in 11 other conditions and produced the p-tails shown in Figure A-8.

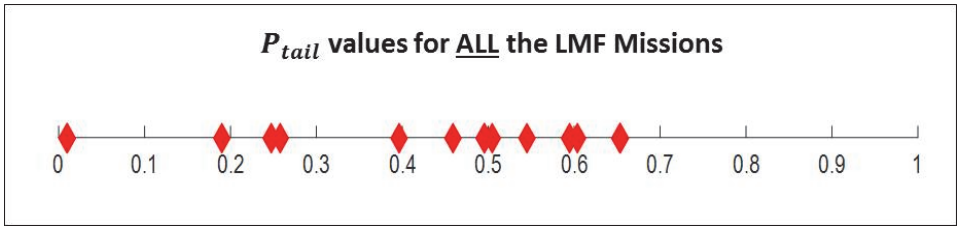

Figure A-8. Summary of P_{tail} results

The Fisher Test statistic is

$$X = -2\sum_{i=1}^{n} \ln(p_i) = 24.36$$

The test statistic follows a chi-square distribution with 2*n* degrees of freedom, from which a p-value for the global hypothesis can be obtained and used to determine whether the null hypothesis, H_o , can be rejected. In this example the p-value for the hypothesis test is 0.83, which signifies the observed data is likely consistent with the null hypothesis. Since the p-value for this example is greater than the 0.05 significance level (assuming a 95% confidence level) we fail to reject the null hypothesis and conclude the model output sufficiently represents the collection of live data.

The KS test for uniformity also fails to reject the null hypothesis and produces the same conclusion as Fisher's Test: that the model output sufficiently represents the live data.

While this statistical result is important and insightful, testers should always consider all sources of uncertainty, including knowledge uncertainty, before making an accreditation decision.

R Code

The Fisher's Combined Probability Test in R:

```
p.i <- c(0.49,0.69,0.79,0.81,0.99,0.02,0.99,0.69,0.92,0.91,0.79,0
.69,0.99,0.51,0.38,0.02)

Fisher.Statistic <- (-2*sum(log(p.i)))
```

```
degree.freedom <- 2*16

p.value.fisher <- pchisq(Fisher.Statistic,degree.freedom,lower.ta
il=FALSE)

p.value.ks <- ks.test(p.i, punif)
```

c. Combo Test

As the name implies, the "combo test" is a combination of other techniques. Specifically, the test involves performing the t-test (or log t-test in the case of skewed data), the KS test, and Fisher's Combined Probability Test, readjusting the type I error, and using the minimum p-value of those three tests. Readers are referred to the previous three sections for more information on these techniques. Overall, the combo test performs better than any one test alone.

3. Regression-based Techniques

Statistical regression methods are used to model and analyze several variables, or factors, at a time. If data are collected across a variety of conditions using a classical design approach, regression techniques can easily separate the effects of each individual factor on the response, as well as detect any interaction, or synergistic relationship, between factors.[56] In the modeling and simulation context, one way to leverage the benefits of such techniques is to combine the data from the live test with that of the model and fit a single regression model that can test for significant differences between live and model data, while controlling for other factors. Another regression-based technique to consider, especially if live data are limited or direct matching between model and live is not possible, is emulation and prediction.

a. Single Model Regression

Example

Consider an operational test of the capability of a ground-based radar to detect and track various types of threat targets. Testers use modeling and simulation to supplement live testing, which is limited due to safety and resource constraints. The testers choose to focus on detection range and aim to characterize the performance of the simulation across different threat types and orientations.

[56] Douglas C. Montgomery, "Design and analysis of experiments," John Wiley & Sons, 1990.

Description of Method

Regression modeling is a method of validating simulation results by building a statistical model to formally compare simulation results to live test results, while controlling for all other factors of interest. It is appropriate in cases where a DOE approach has been used for both live and simulated data collection, and the factor spaces in both the live and simulation environment are the same, or nearly the same.

In cases with exact one-to-one matching between live and simulation data (matched pairs), regression analysis can be performed on the *differences* between live and simulation outcome, as a function of all other variables. Any differences significantly different from 0 signify a place where there is a difference between the model and the live data.

Alternatively, and even in cases where some data does not have a one-to-one match, regression analysis can be performed on the raw outcomes as a function of source (simulation vs. live) and all other variables, including at least second-order interactions between source and the other variables. This method is mathematically equivalent to the first if all possible interactions with source are included in the model and all data has a matched pair.

Use caution when performing regression analyses, as pooling live and simulated data may lead to misleading results unless certain criteria are met. First, the model must be able to explain major sources of variation, so do not use a regression approach unless all main effects and at least all two-way interactions with source can be mathematically modeled. Second, be mindful of sample size contributions from live vs. simulation data The number of data points from each source do not have to be identical, but a severe imbalance (e.g., 100 simulation points per single live data point) will cause confounding between variables of interest (see the example in Chapter 3 for details).

In the radar example introduced earlier, an appropriate statistical model can be expressed as:

$$\begin{aligned} Detection\ Range \\ &= \beta_0 + \beta_1 Source + \beta_2 Threat + \beta_3 Orientation \\ &+ \beta_4 (Source * Threat) + \beta_5 (Source * Orientation) \\ &+ \beta_6 (Orientation * Threat) + \in \end{aligned}$$

The statistical model for detection range is a function of three variables, *Source, Threat*, and *Orientation*. The *Source* variable indicates whether the data point comes from live testing or simulation, the *Threat* variable indicates the type of threat or target presented during testing, and the *Orientation* variable indicates the location of the threat in comparison to the face of the radar. The model also includes interaction terms between all three variables.

If the *Source* effect is not statistically significant, the overall sets of simulated data and live data are statistically indistinguishable and the simulation runs are, on average, providing consistent data. If an interaction term involving *Source* is significant, the differences between the live data and simulated data depend on the threat and/or orientation. This indicates that the simulation may be providing good results in certain conditions but not others.

Different types of regression may be used depending on the nature of the observed data. Linear regression is appropriate for symmetric data, lognormal regression for skewed data, and logistic regression for binary data.

Characterizations & Limitations of Method

Given sufficient statistical power, this technique is a good way to detect differences in means when a designed experiment was executed. It also supports various types of uncertainty quantification, including confidence intervals and prediction intervals. However, limited live test points hinder the ability to differentiate between bias and variance problems in the model.

This technique should also only be used when the factor spaces in the live and simulation environments are the same, or nearly the same. In addition, correlation problems are encountered if the sample sizes between the live and simulation data are very uneven. In this case, the recommended procedure is to sample with replacement (e.g., bootstrap) from the simulation dataset such that the live and simulation sample sizes are the same size, fit the appropriate model to obtain p-values, and repeat this procedure thousands of times to obtain distributions of p-values, coefficient estimates, and model predictions. This approach eliminates the potential confounding between main effects and interaction terms that arises with unbalanced models.

Interpretation of Results in Context of Example

Returning to the motivating example, testers execute one live test per threat type, and 30 runs of the simulation. Since these sample sizes are quite unbalanced, a bootstrapped regression is performed. Using the regression equation above and 1000 bootstrap iterations, analysts obtain the following median p-values for each parameter:

Parameter	*Median p-value*
$\boldsymbol{\beta_1}$ *(Source)*	0.503
$\boldsymbol{\beta_2}$ *(Threat)*	0.251
$\boldsymbol{\beta_3}$ *(Orientation)*	0.122
$\boldsymbol{\beta_4}$ *(Source * Threat)*	0.048
$\boldsymbol{\beta_5}$ *(Source * Orientation)*	0.493
$\boldsymbol{\beta_6}$ *(Orientation * Threat)*	0.415

Note that the *Source* main effect is not statistically significant, but the *Source*Threat* interaction effect is significant at the .05 level. This indicates that, on average, the simulation may represent reality fairly well, but there is a particular threat or threats where the simulation matches significantly less well. To visualize such effects, we can plot the bootstrapped distributions of p-values or beta estimates. Or perhaps more interpretably, we can look at the difference in estimated detection range between live and simulation, based on our regression model. Figure A-9 depicts such differences across the factor space, with 80% bootstrapped confidence intervals. Here we can clearly see that the simulation underestimates actual detection range by nearly 8 units on average for Threat C. The confidence intervals for threats A and B contain zero, which indicates no significant difference between live and simulation for those particular threats. Note that the displayed confidence intervals represent a quantification of statistical uncertainty in the results.

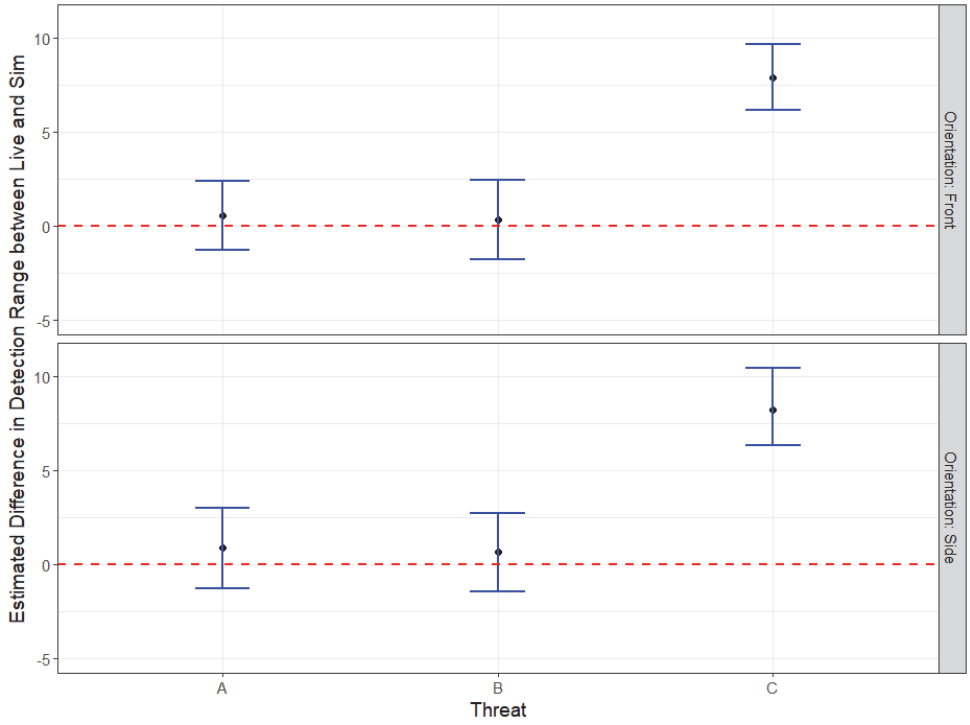

Figure A-9. Estimated differences in detection range between live and simulation across threat and orientation, with 80% confidence intervals.

R Code

```
dat_list <- vector(mode = "list", length = nboot)
# create list of nboot data sets, matching live and sim sample si
ze
for (i in 1:length(dat_list)){
  dat_list[[i]] <-
```

```
dat %>% split(list(.$Threat, .$Source, .$Orientation)) %>%
    map(~ .x[sample(1:nrow(.x), 1, TRUE), ]) %>%
    do.call(rbind, .)
}
# run regression on each of nboot data sets
model <- dat_list %>%
  map(~lm(Range ~ (Threat + Source + Orientation)^2, data = .x))

# Save model terms and p-values
terms <- model %>%
  map(summary) %>% map(tidy) %>% do.call(rbind, .) %>%
  filter(term != "(Intercept)")

# Save predictions and differences
preds <- model %>%
  map(augment) %>%
  do.call(rbind,.) %>%
  mutate(case = str_c(Threat, Orientation, sep=" "),
         boot = rep(1:nboot, each=nrow(conditions)*2)) %>%
  group_by(case, boot) %>%
  select(.fitted, Source) %>%
  spread(Source, .fitted) %>%
  ungroup() %>%
  separate(case, c("Threat", "Orientation"), sep = " ") %>%
  rename(fit.live = Live, fit.sim = Sim) %>%
  mutate(fit.diff = fit.live-fit.sim) %>%
  select(-boot)
```

b. Emulation and Prediction

Example

Consider a hardware-in-the-loop simulation capability for lightweight and heavyweight torpedoes. Testers use the simulation to predict values of a continuous

variable, such as the time required for a torpedo to reach a target. The operational environment is complex, with many factors affecting the response. In order to validate the simulation, the testers would like to compare live data collected during a designed experiment to the outputs from the simulation under the same conditions.

Description of Method

Emulation and prediction is a method for validating simulation results capable of testing for factor effects, detecting differences in variance, and identifying bias. Outputs from the simulation are used to build an empirical emulator, or statistical model, characterizing the response as a function of each factor. As live data points become available, they are compared to the prediction interval generated from the emulator under the same conditions. If a live point falls within the prediction interval, we have evidence that the simulation is performing well under those conditions. Conversely, if a live point falls outside of the prediction interval, we should investigate why the emulator is failing under certain conditions and test for any systematic patterns.

Emulation and prediction enables an iterative process of improving the simulation and informing live testing. Once the live data points have been used in validation, they can be used to update the simulation and further "train" the model, and future testing can focus on conditions under which the simulation performed poorly.

Characterizations & Limitations of Method

This method is most effective when used in conjunction with a designed experiment, and when multiple simulation runs can be performed for each condition. Its strength is powerfully detecting when the variance of the live data is larger than that of the simulation data. The general approach is applicable to any amount of live data, but multiple live points are required to separate bias and variance issues.

Interpretation of Results in Context of Example

Returning to the example presented above, testers perform multiple runs of the simulation under two conditions, build a statistical model to characterize time to target as a function of each factor (condition 1 and condition 2), then generate prediction intervals for each condition. These intervals represent the uncertainty in future observations. The simulation results under each condition are shown in the figure below as blue points. The prediction intervals generated from the emulator are shown by the error bars. A small number of live tests were then performed under each condition, shown as red points.

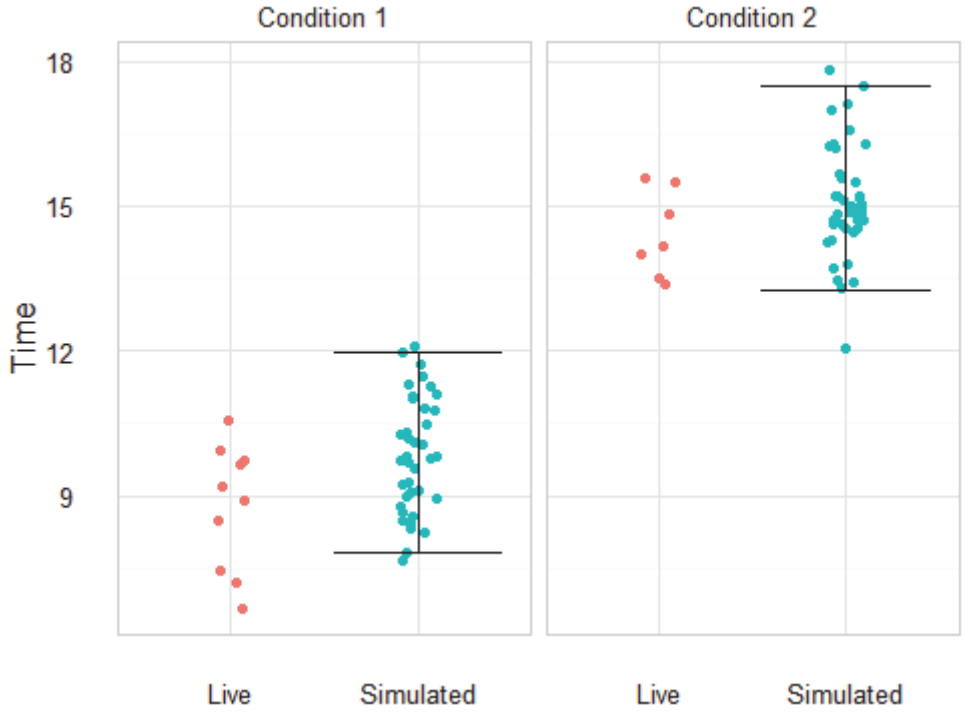

Figure A-10. Live Data Compared to Simulation Output and Corresponding Prediction Intervals

Since most of the live points under condition 2 fall within the prediction interval of the emulator, the testers have strong evidence that the simulation is performing well for this condition. Under condition 1, however, some points fall outside of the emulator prediction interval. The testers decide to perform additional analysis of the results under condition 1 to understand why the simulation performs poorly under this condition, and make changes to the simulation if necessary.

R Code

If *ld* is the live data, *dm* represents the design matrix for the condition of interest, and *mf* is the model fit to the simulation data, the emulation and prediction procedure is outlined below. Note that this routine has been simplified for length reasons; depending on the size and shape of the data, indexing and loops may be required.

```
pred.interval <- predict(dm, mf, interval="predict", level=.8)

if( (pred.interval[1] < ld) && (pred.interval[2] > ld) )
{print('data point within interval')} else {print('data point NOT
in interval')}
```

4. Binomial techniques

Most of the statistical techniques for validation discussed thus far are only appropriate for continuous responses. If the outcome of interest is binary (pass/fail) or categorical in nature, a different class of approaches is needed. Some examples of such techniques include Fisher's Exact Test, chi-squared tests, and permutation tests[57].

a. Fisher's Exact Test

Example

Consider an operational test of a financial system in which testers measure a binary success/failure response. Simulations of the operational test are also conducted under the same mission scenario and environment as the live test. There are no factors considered, and the sample size is small. The results of the live and simulated tests are shown in the 2x2 contingency table below. In order to validate their simulations, the testers aim to determine whether the data provide sufficient evidence to indicate that the success rate differs between the live and simulated tests.

	Success	*Failure*	
Sim	a = 8	b = 14	$a + b = 22$
Live	c = 1	d = 3	$c + d = 4$
	$a + c = 9$	$b + d = 17$	$n = 26$

Description of Method

Fisher's Exact test is a nonparametric test of independence most often used when data are binary or categorical, there are no factors, and the sample size is small. It assumes the individual observations are independent, and that the marginal totals are fixed.

To compare variables X and Y, with m and n observed states, respectively, form an m x n matrix of observed counts. Fisher's Exact test is most commonly performed on 2x2 tables. Assuming the margins of the matrix are fixed, the conditional probability of observing a matrix with cells a, b, c, d and marginal totals $(a + b), (c + d), (a + c)$ and $(b + d)$ equals:

$$P_{cutoff} = \frac{(a+b)! * (c+d)! * (a+c)! * (b+d)!}{n! * a! * b! * c! * d!}$$

[57] Alan Agresti, "Categorical Data Analysis," 2nd edition, John Wiley & Sons, 2002.

The P-value of the test is determined by calculating the conditional probabilities for all other possible matrices with the same marginal totals, then summing those whose conditional probabilities are $\leq P_{cutoff}$ calculated above (i.e., matrices that are more extreme than observed). The null hypothesis H_0 is that the relative proportions of one variable are independent of the second variable. If the P-value of the test is less than the significance level α, we can reject the null hypothesis.

Characterizations & Limitations of Method

Fisher's Exact Test is the recommended procedure for assessing differences in a binary response variable between live and simulation when there are no factors under consideration.

Interpretation of Results in Context of Example

In the context of the example provided above, the null hypothesis H_0 is that the proportion of successes and failures is independent of whether the data are live or simulated.

We calculate the conditional probability of observing the exact table as:

$$P_{cutoff} = \frac{22! * 4! * 9! * 17!}{26! * 8! * 14! * 1! * 3!} = 0.4094$$

The conditional probabilities of observing other possible matrices with the same marginal totals are calculated below.

9	13
0	4

$$\frac{22! * 4! * 9! * 17!}{26! * 9! * 13! * 0! * 4!} = 0.1592$$

7	15
2	2

$$\frac{22! * 4! * 9! * 17!}{26! * 7! * 15! * 2! * 2!} = 0.3275$$

6	16
3	1

$$\frac{22! * 4! * 9! * 17!}{26! * 6! * 16! * 3! * 1!} = 0.0955$$

5	17
4	0

$$\frac{22! * 4! * 9! * 17!}{26! * 5! * 17! * 4! * 0!} = 0.0084$$

The two-tailed P-value for the test is the sum of all probabilities ≤ 0.409. In this case, all probabilities are ≤ 0.409, so the P-value is $0.4094 + 0.1592 + 0.3275 + 0.0955 + 0.0084 = 1$. We cannot reject the null hypothesis. There is not enough evidence to indicate

that the proportion of successes and failures differs between the live and simulated trials, and the simulation is considered to be valid.

R Code

Fisher's Exact test is performed in R using `fisher.test(x)`, where `x` is a two-dimensional contingency table in matrix form.

```
example_data <- matrix(c(8, 1, 14, 3), ncol = 2)
row.names(example_data) <- c("sim", "live")
colnames(example_data) <- c("success", "failure")



example_data

##      success failure
## sim        8      14
## live       1       3



fisher.test(example_data)

##
##  Fisher's Exact Test for Count Data
##
## data: example_data
## p-value = 1
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
## 0.1116273 101.0207966
## sample estimates:
## odds ratio
## 1.681372
```

b. Logistic Regression

See Section 3 (Regression) above

Approved for public release; distribution is unlimited.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
xx-02-2019	OED Draft	

4. TITLE AND SUBTITLE

Handbook on Statistical Design & Analysis Techniques for Modeling & Simulation Validation

5a. CONTRACT NUMBER

HQ0034-14-D-0001

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Kelly M. Avery (OED); Laura J. Freeman (OED); Samuel H. Parry (OED); Gregory S. Whittier (OED); Thomas H. Johnson (OED); Andrew C. Flack (OED)

5d. PROJECT NUMBER

BD-9-2299

5e. TASK NUMBER

229990

5f. WORK UNIT NUMBER

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

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

8. PERFORMING ORGANIZATION REPORT NUMBER

D-10455-NS

H 2019-000044

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

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

10. SPONSOR/MONITOR'S ACRONYM(S)

DOT&E

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

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release. Distribution is unlimited.

13. SUPPLEMENTARY NOTES

Project Leader: Wojton, Heather

14. ABSTRACT

This handbook focuses on methods for data-driven validation to supplement the vast existing literature for Verification, Validation, and Accreditation (VV&A) and the emerging references on Uncertainty Quantification (UQ) . The goal of this handbook is to aid the test and evaluation community in developing test strategies that support model validation (both external validation and sensitivity analysis) and uncertainty quantification. Following an introductory chapter, the handbook discusses the VV&A process as it applies to operational testing, statistical analysis strategies for validation, and design of experiments techniques for test planning.

15. SUBJECT TERMS

Modeling and Simulation (M&S) Validation; Statistics; Design of Experiments (DOE); Uncertainty quantification

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Heather Wojton (OED)
Unclassified	Unclassified	Unclassified	Unlimited	101	19b. TELEPHONE NUMBER *(Include area code)* (703) 845-6811

Approved for public release; distribution is unlimited.