SCIENTIFIC TEST & ANALYSIS TECHNIQUES
CENTER OF EXCELLENCE

Model Validation Levels: Methods and Implementation

February 2024

Corinne Stafford, Ctr
Kyle Provost, Ctr
Nicholas Jones, Ctr

To develop and apply independent, tailored Scientific Test & Analysis Techniques solutions to Test and Evaluation that deliver insight to inform better decisions.

About this Publication:
This work was conducted by the Scientific Test & Analysis Techniques Center of Excellence under contract FA8075-18-D-0002, Task FA8075-21-F-0074.

For more information:
Visit, www.AFIT.edu/STAT
Email, AFIT.ENS.STATCOE@us.af.mil
Call, 937-255-3636 x4736

Technical Reviewers:
Steve Oimoen
Tony Sgambellone

Scientific Test & Analysis Techniques Center of Excellence
2950 Hobson Way
Wright-Patterson Air Force Base, Ohio

Version: 1, FY24

Abstract

A Model Validation Level (MVL) is an objective, automatable metric scored from 0-9 that quantifies how much trust can be placed in the results of a model to represent the real world. MVLs aim to provide utility both for decision makers to quickly evaluate model risk, and for model developers to identify model improvement areas. Additionally, the MVL framework supports digital engineering via automation to continuously perform validation and by quantifying trust in preexisting models for new use cases. This paper walks through the entire MVL process, from determining MVL applicability to interpreting and acting on the results. This paper provides background on the key concepts of validation, discusses required data collection, and details how the MVL is calculated, with appendices containing full mathematical documentation. An MVL R tool and user guide are also available which automate the framework described here.

Keywords: model validation levels, validation, modeling and simulation, digital engineering, fidelity, test and evaluation

Table of Contents

Introduction

Modeling and simulation (M&S) has become pervasive throughout the Department of Defense (DOD), especially as the workforce further shifts toward digital engineering. These models are relied upon to aid in decision making, particularly in cases where live data is unsafe or prohibitively expensive to obtain. For models to be considered trustworthy for decision making, they must be validated for a given intended use case. DoDI 5000.61 defines validation to be, "the process of determining the degree to which a model or simulation and its associated data are an accurate representation of the real world from the perspective of the intended uses of the model" (2018). In practice, however, validation is often binary, meaning a model is either valid or not; static, meaning the model once valid is considered valid forever; and subjective, where validation is only a qualitative comparison between the model results and subject matter expert (SME) expectations. Many of these validation shortcomings result from the lack of a rigorous, easy-to-implement, broadly applicable approach to validation. Several validation frameworks have been developed to try to combat these issues; however, they are often either easy to use but lacking rigor or rigorous but difficult to apply (Ahner et al., 2023). Additionally, to support the shift toward digital engineering, a validation framework should be automatable, such that it can be integrated into digital infrastructure and support validation across a digital ecosystem.

The Scientific Test and Analysis Techniques Center of Excellence (STAT COE) developed Model Validation Levels (MVLs) to provide an objective, automatable metric that quantifies the degree of trust that can be placed in the results of a model to represent the real world. The MVL framework can be quickly applied to a broad range of predictive models, enabling continuous validation of models as they evolve and more data becomes available. In addition, the MVL framework enables legacy models to be evaluated for new intended use cases, helping to reduce duplicated efforts and stove-piping. Previous publications on MVLs have developed the conceptual and mathematical foundations for MVLs (Ahner et al., 2023; Provost et al., 2022; Weeks et al., 2022; Stafford et al., 2024b). This paper aims to provide comprehensive documentation and guidance for MVLs and all underlying methods, as well as discussing practical considerations for incorporating MVLs into validation plans. This paper can be used in tandem with the MVL R tool (Provost et al., 2024; Jones et al., 2024), which automates the calculation process described here. First, the *Background* section describes the conceptual foundations for MVLs, including the three pillars of validation: fidelity, referent authority, and scope. The *MVL Framework Applicability* section discusses when MVLs can and should be used, as well as the requirements for the calculation: a well-defined intended use case, model data, and referent data. The following sections, *Defining and Quantifying the Scope of Intended Use* and *Data Collection*, decribe how these requirements can be collected. Next, the *MVL Calculation* section walks through the MVL calculation process step-by-step, with references to appendices for full mathematical details. Finally, the *Outputs, Interpretations, and Actions* section discusses MVL framework outputs, their interpretation, and next steps for model improvement. Ultimately, the MVL will provide utility to decision makers by yielding a quickly interpretable level of trust on a standardized scale and to model developers by identifying areas for model improvements.

Background

The MVL framework defines validation in terms of three key pillars which must be carefully considered when determining the degree to which a model can be considered valid: fidelity, referent authority, and scope. Due to the many terms and definitions contained in this paper, Appendix A compiles key definitions for reference.

Fidelity

The first of the validation pillars is fidelity. Fidelity is the level of consistency between a model and a referent, defined in the three dimensions of accuracy, repeatability, and resolution. The referent is defined to be a codified body of knowledge representing real system behavior. Conceptually, fidelity can be understood to be a quantitative comparison between model outputs and referent data, where the model is treated as a black box: in other words, fidelity does not depend on the level of complexity or detail intrinsic to the model, only on the closeness of model outputs to referent data.

Fidelity is defined in terms of accuracy, repeatability, and resolution. Accuracy is the degree to which a parameter or variable, or a set of parameters or variables, within a model or simulation conforms exactly to reality or to some chosen standard or referent (Modeling and Simulation Enterprise, 2021). Repeatability refers to the similarity of the results obtained from the same model (or referent) over multiple observations under the same input conditions. Lastly, resolution is the degree of granularity with which a parameter or variable can be determined (Pace, 2015). Repeatability and resolution can be considered equivalent to the terms aleatory and epistemic uncertainty, respectively, which are commonly used in the field of uncertainty quantification. These dimensions are depicted in Figure 1. While accuracy is a comparison of model and referent, repeatability and resolution are properties of either a model or referent alone. When determining fidelity, both similarity in mean behavior (accuracy) and similarity in variability must be considered, where variability comprises both repeatability and resolution. In most cases, real system behavior shows random variation even when conditions are held constant. Models must be able to predict this variability to provide a complete picture of how the real system will behave.

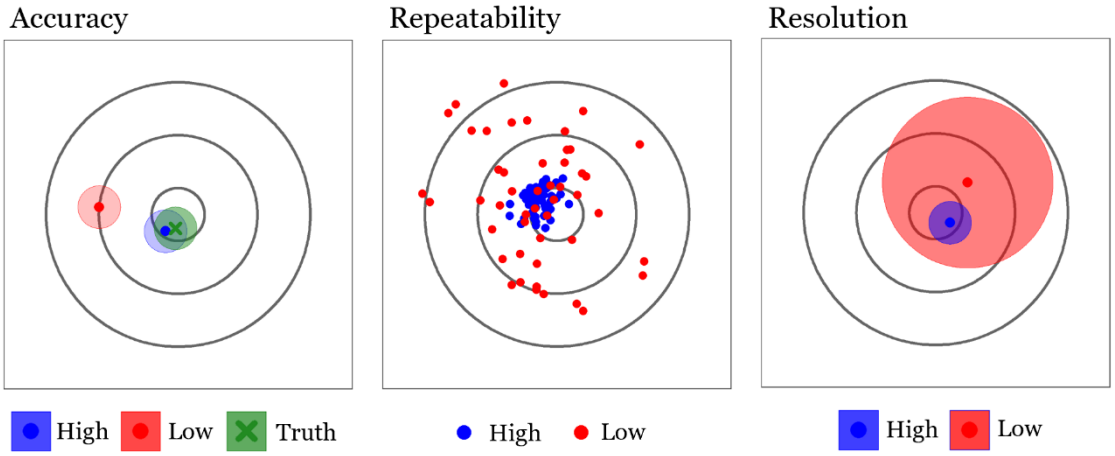

Figure 1
Dimensions of Fidelity

Referent Authority

Referent authority is the second pillar of validation and refers to the strength of credibility of a referent's claim to be a high-fidelity representation of reality.

For validation to serve as a means for building trust in models, it must be based on a comparison of the model to authoritative referents. Because a referent is a representation of reality, it has some authority, but not all referents are equally authoritative. For example, a set of recorded performance data for a system and the judgement of a subject matter expert (SME)

may both be drawn from observation of the same real-world event, but one is more objective than the other and is considered to be a closer representation of the real world. The amount of trust placed in a referent determines how much trust can be placed in a model that was validated against it.

For the MVL framework to objectively handle referent authority, referents must be assigned a quantifiable level of trust. To assign trust, the MVL framework leverages Technology Readiness Levels (TRLs), summarized in Table 1.

Table 1
Technology Readiness Levels

TRL	System Description
1	Idea, preliminary design, and/or documented requirements
2	Preliminary design using accepted physical principles & heuristics
3	Critical components or technologies demonstrated in lab environment
4	Basic integration of system components demonstrated in lab environment
5	Lab-scale integrated system demonstrated in relevant/simulated environment
6	Full-scale prototype system demonstrated in relevant/simulated environment
7	Full-scale prototype system demonstrated in operational test environment
8	Production-ready system demonstrated in operational test environment
9	Full-scale system deployed in real environment

TRLs are a standard tool for assessing the maturity of new technologies in Government acquisition and development programs (Government Accountability Office, 2020). Typically, TRLs are applied to technologies to indicate their technical maturity (e.g., a lab-scale prototype wing design demonstrated in a wind tunnel might have TRL 5). On the other hand, a referent is typically a body of data, not a technology (e.g., the *data* collected from testing the prototype wing in a wind tunnel), and so should not be thought of as having a TRL itself. However, an authority level for a referent can still be inferred from the TRL that referent would support assigning to the system that produced it (e.g., the wind tunnel prototype test was sufficient to decide if the wing design met TRL 5, so the *data* collected that supported that conclusion of the wing's TRL is a referent with an authority level of 5). This line of inference supports the assignment of authority levels to referents based on the maturity (or nearness to real-world operational expectations) of the systems from which they were derived. Ultimately, models will be used to make decisions about real-world system operations; thus, operational data, while potentially more noisy than more controlled referents, is the most authoritative referent for model validation because it is what the warfighter will experience. A set of representative referents that might be associated with each TRL and are therefore said to have a set of corresponding authority levels, is given in Table 2.

In practice, the MVL framework may be applied to many different types of simulations. However, regardless of the object represented by the simulation, the same authority scale can be applied by defining what is considered the 'system'. This concept is further discussed in Appendix B, which aims to answer common questions on how the authority level should be determined. The

application of these authority levels to determine model validity is discussed further in the MVL Calculation section.

Table 2
Referent Authorities from Relevant TRLs

Authority Level	Relevant Referent
1	SME Judgement
2	First Principles/Physics Predictions
3	Component Lab Test Data
4	Integrated Component Lab Test Data
5	Lab-Scale System Test Data
6	HWIL & SWIL Data
7	Prototype Field Test Data
8	Live System Test Data
9	Operational Real-World Data

Scope

The final pillar of validation is scope, which includes the set of inputs, outputs, assumptions, and limitations representing the mission-relevant system parameters, environmental conditions, constraints, and requirements, and their allowable values. The MVL framework considers three different scopes: the model scope, the referent scope, and the scope of intended use, as represented by Figure 2.

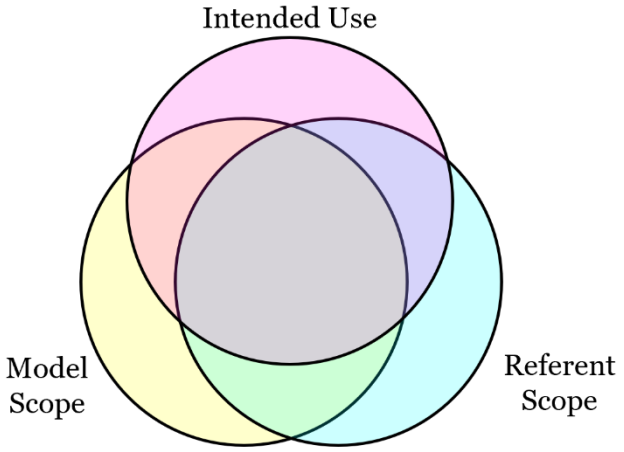

Figure 2
Intersecting Scopes of Model, Referent, and Intended Use

The model scope is defined first by where the model can be run, and second by where model

output has been obtained for validation. The referent scope is the scope where data has been obtained from any of the referents listed in Table 2. To validate a model, the model and referent scopes must be overlapping to assess fidelity of shared output(s). Fidelity can only be assessed when input conditions are shared. The intended use of a model is generally a description of the problem addressed by a model or simulation and its associated data, including the system or process being represented (US Department of Defense, 2012). The scope of intended use further specifies the set of dimensions, ranges, and assumptions of the model inputs and outputs needed to represent and assess system behavior. Definition of this intended use scope is further discussed later in this paper. The validity of a model is assessed over the scope of intended use, which ideally is covered by both model and referent scopes. If the model scope does not overlap with the scope of intended use, the model is either not suited to the intended use or there is a need to collect more outputs. If the referent scope does not overlap with the scope of intended use, more data needs to be collected to enable the model to be validated. Multiple referents may be used together to validate different regions of the intended use scope.

MVL

The MVL is mathematically derived from model and referent data with a defined scope of intended use. This process uses a combination of metrics based on methodical assessments of fidelity, referent authority, and scope. The result is a continuous score between zero and nine, rating the level of trust that can be placed in the outputs of a model for the specific intended use. The remainder of the paper will discuss the steps required to obtain the MVL as well as interpretation of the result. Additionally, a tool is available in R which automates the calculation of the MVL given model data, referent data, and a defined scope of intended use (Provost et al., 2024; Jones et al., 2024).

MVL Framework Applicability

M&S includes a diverse range of models across many disciplines with various objectives, levels of detail, and modeling mechanisms. The MVL framework aims to be broadly applicable to any model which produces a quantitative or measurable prediction about the behavior of a real system or object. This prediction can then be validated against referent data quantifying that same behavior. The MVL framework is not necessarily applicable to models without a predictive objective. For example, MBSE models supporting the conceptual design phase do not yet have algorithms implemented that predict system behavior. MBSE can help guide validation needs, such as defining the scopes for the model and the intended use, and later be validated itself as it evolves into a digital twin with detailed descriptions of system behavior. Another example of a non-predictive model could be mechanical or electrical computer-aided design (CAD) models that are purely descriptive and do not make predictions about behavior. Some CAD models may go further than pure description, such as a model of a full mechanism or electrical system with simulated behavior, and in these cases MVLs could be applied.

Predictive models exist at various levels of detail, from theater or mission level simulations down to subcomponent engineering models. For any of these models or simulations, they must have identifiable outputs or measures which can be validated against referent data. The types of referents which will be used to validate models will vary drastically depending on the object being validated. For example, miss distance of a projectile could be validated against live test data, while a red threat model may be validated against physics predictions based on system intelligence. Appendix B further discusses determining referent authority for various referents, accounting for the type of object being modeled, whether it be a system of systems, system, component, etc.

Many modeling mechanisms may be used within a model; however, because the MVL framework considers the model to be a black box and compares the outputs, a key classifier becomes whether the model is stochastic or deterministic. Stochastic models produce random outputs mimicking the variable behavior of reality, while deterministic models provide only point predictions of behavior. The MVL framework is best applied to stochastic models since fidelity is assessed in terms of both accuracy and match in variability. For deterministic models or outputs, the MVL framework can still be applied. Ideally, resolution and propagated uncertainties should be quantified for deterministic models to quantify variability. However, if uncertainty is not quantified, an accuracy MVL, MVL_a, may be most appropriate, where the MVL user accepts risk that the model cannot predict variability of real behavior.

To be able to compute the MVL for a model, the user must have:

1. Intended use: a well-defined scope of intended use describing the outputs and inputs for which the model must be validated.
2. Model data: output(s) collected from a predictive model, including the inputs or conditions under which those output(s) were obtained.
3. Referent data: data from one or more referents with matched inputs to model data and assigned referent authority levels for each referent.

Given these prerequisites, other qualifications determine which variation of an MVL may be calculated. The process to determine the applicable MVL is depicted in Figure 3.

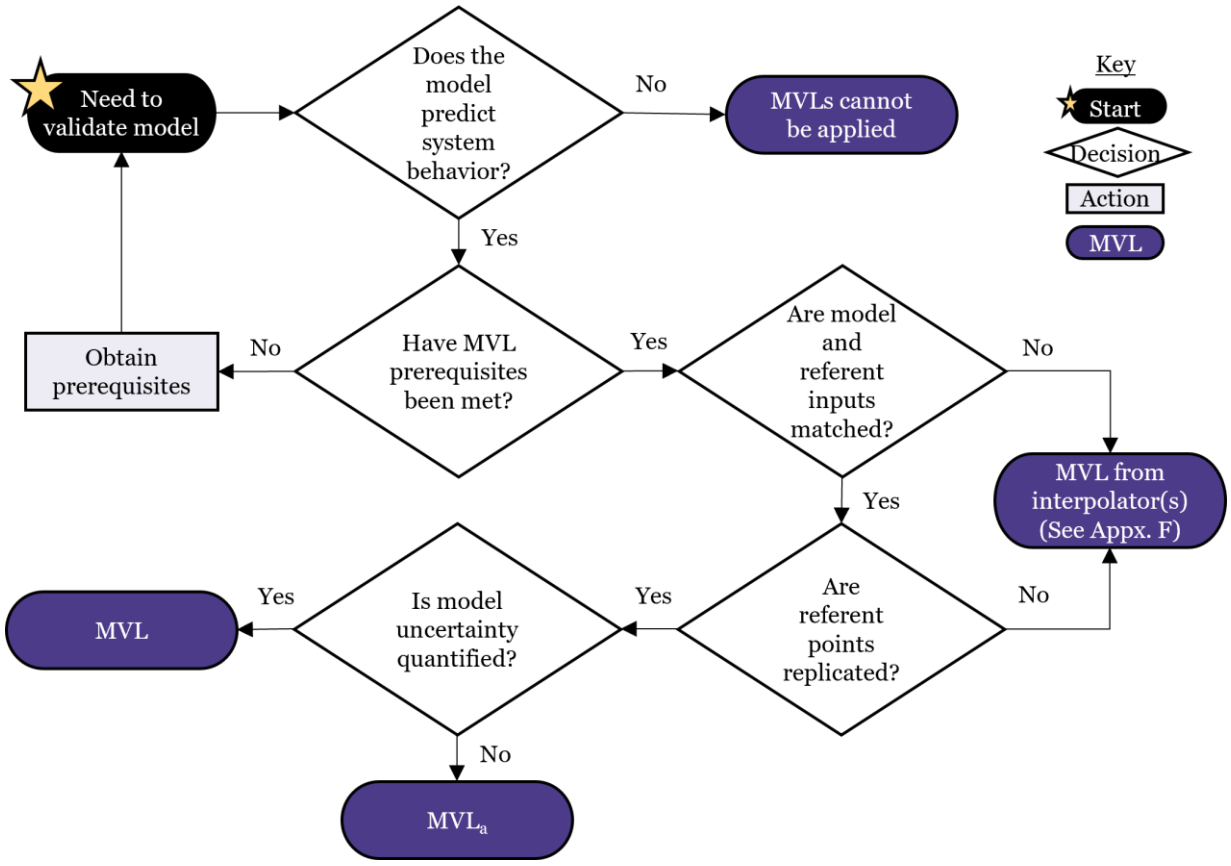

Figure 3
Flowchart to Determine Applicable MVL

The variations of MVLs include an MVL, MVL_a, or an MVL that uses a referent interpolator. An MVL_a is used as an alternative to the MVL when model uncertainty has not been quantified. Since model uncertainty is a key component for assessing a model's predictive ability, an MVL_a should be interpreted with this limitation in mind. Calculating an MVL_a only is best for deterministic models or when ability to predict variability is not desired. An MVL that uses a referent interpolator is necessary when model and referent inputs are not matched, and the referent must be interpolated to compare against the model. Additionally, when referent observations are not replicated, an interpolator (or statistical model) of the referent is needed to quantify referent variability, which is required to assess fidelity. In these cases, the MVL user must create the interpolator and accept any assumptions made in its construction. These variations allow model trust to be quantified when not all requirements are met for the MVL.

The objective of an MVL is to provide a quick, interpretable measure of model validity while additionally identifying areas for increasing model trust; however, the MVL framework is far from the only methodology for performing model validation. MVLs can and should be augmented with other statistical methods that can be tailored to the specific validation scenario. The Institute for Defense Analysis (IDA) Handbook on Statistical Design and Analysis Techniques for Modeling & Simulation Validation discusses many such validation methods (Wojton et al., 2019). Calculating an MVL is well suited to be an initial validation step that is easily performed for a wide range of scenarios, and it can enable a fast pace in a digital engineering environment. MVLs can provide initial insight to point to additional analysis and model improvement steps.

Defining and Quantifying the Scope of Intended Use

The scope of intended use defines the inputs and outputs over which the model needs to be validated and must be well-defined for an MVL to be calculated.

Responses

The first step in defining the scope of intended use is defining the outputs or responses which the model will be used to predict. Often due to the complex nature of DOD systems, decomposing system functions into smaller elements is key to understanding the performance of each critical component. This decomposition process can help with identifying responses, the measurable outputs of either a model or test event. A model will often have multiple responses that support one or more requirements and may vary from mission level outcomes to component behavior. The intended use specifies which responses are needed, first to help guide the level of detail required for model building and second to assess the validity of those responses. The MVL is calculated for each response, so each different response may have a different level of validity.

A key consideration in choosing a response is whether it is continuous or categorical. Continuous data types are preferred due to containing more information than categorical data types (e.g., Hit/Miss, Pass/Fail), meaning continuous data results in lower resource costs and improves analysis (Ortiz, 2018). Effort should be made to convert categorical responses to continuous whenever possible. The MVL can be computed for both continuous and categorical responses; however, it will be a more informative and reliable indicator of validity when continuous data is used.

Factors

In addition to responses, the inputs or factors that influence those responses must be documented as part of the scope of intended use, including the factor values or ranges over which the model will be validated. Factors could include system configurations, physical and ambient conditions, etc. They are known to have an effect on a response, where the effect is ideally both operationally meaningful and known to be statistically significant though results from a designed experiment. Different responses may have different factors that affect them and for which they need to be validated. Model factors should mirror the physical factors varying in testing and operations. There may be additional parameters that can be varied within the model that affect model outputs but have no physical equivalent (e.g. time-step, mesh size); therefore, these parameters should not be included in the MVL scope definition but should be recorded for documentation purposes and evaluated with sensitivity analysis. Like responses, factors should also be made continuous when possible due to the increased amount of information and ability to interpolate between factor values.

Factors in the intended use should be classified as varied or held constant. For factors which vary, a continuous range of values or categorical list of levels is part of the scope of intended use definition. Including a constant factor in the scope of intended use indicates that the model will only be used at that constant level and that the model should only be validated against referent data collected at that same factor level. In contrast, factors not included in the intended use may take on any value and still be incorporated into the validation.

Other Considerations

Additionally, constraints on factors should be documented as part of the scope of intended use. Constraints can be used to limit factor combinations where the model will be used (e.g. “the model will not be used for high speeds at low altitude”) as well as to exclude disallowed combinations (e.g. “aircraft model X does not have X radar installed”). These constraints can be defined through Boolean logic or mathematically defined relationships. These constraints combined with the varied factor ranges and constant conditions define the scope of intended use where the model will be validated.

The STAT COE Test Planning Guide (Adams et al., 2022) and IDA Handbook on Statistical Design and Analysis Techniques for Modeling & Simulation Validation (Wojton et al., 2019) provide additional insight for identifying responses (outputs) and factors (inputs) during test planning. The User Guide for the MVL R tool (Jones et al., 2024) provides templates for defining and structuring the scope of intended use for input into the MVL R tool.

Thus far, discussion has focused on defining a single scope of intended use for a model; however, multiple scopes of intended use may need to be defined to properly capture the complexity of model validation. For example, when a model has several responses that need validation, different factors could affect different responses, resulting in multiple scope domains for each unique factor space. Different responses may have different referents which pertain to their validation and apply to different scope domains. Scope domains of interest can be defined based on where data is available (e.g., a region where high authority data is available and a larger region with all available data). Scope division can also occur due to discontinuities or physical boundaries (e.g., phase changes, transonic/supersonic/hypersonic), so the response is expected to be smoothly varying or continuous across the scope. Multiple scope domains could also be defined based on user needs to address different model use cases (e.g., defining one scope of the critical mission space and another for all operations). The MVL is evaluated for each scope domain to assess model validity for each intended use case.

The intended use case(s) for a model may include more detail than what is needed for the scope definition, for example, the risk the program is willing to accept when using the model. However, this is outside of what factors into the MVL. Once the MVL is obtained, the user can evaluate the acceptable level of risk against the resulting MVL.

Data Collection

Validation requires both the model and referent data to be collected. This data collection is integrated into an overall test and verification, validation, and accreditation (VV&A) strategy. Additionally, tests should be designed according to rigorous Design of Experiment (DOE) principles.

Model Data Collection

Model data should be collected in accordance with the objective of model validation, using tailored designed experiments. Models can either be developed specifically for the intended use case, or alternatively, existing models from legacy systems can be assessed for a new system's use case. The IDA Handbook on Statistical Design and Analysis Techniques for Modeling & Simulation Validation (Wotjon et al., 2019) recommends different test designs for comparing the model to live data based on the amount of randomness present in the model, shown in Table 3. The handbook additionally provides suggested design types for exploring the model space, which can be used to verify model results are as expected, characterize the entire factor space, or generate predictions much more quickly than through live testing.

Table 3

Simulation Design Recommendations for Comparing to Live Data

Level of Randomness	Recommended Method
None (Deterministic)	Hybrid Design
Low (E.g., Physics-based with calibration factors)	Classical
High (E.g., Effects-based, Human-in-the-loop)	Classical with Replications

Note. Adapted from "Handbook on Statistical Design and Analysis Techniques for Modeling & Simulation Validation," by Wojton et al., 2019, Copyright 2019 by Institute for Defense Analyses.

Model data should be collected to match the conditions where live data is collected. Ideally, model output is generated to predict system behavior prior to live testing, so that predictive capabilities can be assessed (Miller, 2022). This order ensures the model is not influenced by the collected live data.

Additional Resources:

- Handbook on Statistical Design and Analysis Techniques for Modeling & Simulation Validation (Wotjon et al., 2019)
- Computer Experiments: Space-Filling Design and Gaussian Process Modeling (Natoli & Burke, 2018)
- Model Selection and Use of Empiricism in Digital Engineering (Jones et al., 2021)

Referent Data Collection

For any physical system acquisition program, live tests will be planned throughout contractor, developmental, and operational testing. These tests produce referent data that can serve to validate a model. As the program progresses through developmental and operational test, higher authority level referents become available. For example, during Technology Maturation and Risk Reduction, lab test data may be the most authoritative data available, while a system in operation will have operational real-world data for validation. Since high authority data is typically more difficult to obtain, many referents may be used in tandem to validate a model across its entire scope to cover regions where high authority data cannot be obtained. If only one set of referent data is available, the MVL may still be computed with that referent. Referents for red threats or of battlespaces may always have limited authority levels due to inability to obtain highly authoritative data. Thus, models of these systems and phenomena are expected to have lower MVLs than more exhaustively tested systems.

When no test data is available, model results can be compared against SME judgement, other M&S, or data from a legacy system. SME judgement data should be collected independently from model data and quantified in terms of estimated mean behavior and standard deviation at given inputs. One method for estimating standard deviation is to estimate the range within which 95% of the observations will fall. Then using a normal distribution assumption, the standard deviation is the range width divided by four (95% of normally distributed data falls within $\pm 2\sigma$ of the mean).

Models cannot be validated against referent data that was used to train, fit, or update the model. When referent data will influence the model, the referent data should be partitioned into separate training and validation (holdout) data sets to enable unbiased validation of the model. Ideally, model output predictions should be generated prior to obtaining live data so that live data does not bias the model and true predictive ability can be evaluated.

Referent data is most effective when testing is planned with the model validation effort in mind. The IDA Handbook on Statistical Design and Analysis Techniques for Modeling & Simulation Validation (Wotjon et al., 2019) states that direct matching of points between live testing and simulation provides the best validation strategy, and when points are not matched with a designed experiment, analysis techniques are limited and less powerful. Thus, referent data should be collected using designed experiments to the greatest extent possible. The model's scope of intended use should guide the factor conditions and ranges that need to be varied in physical testing. Test designs should also include replication so that the variability of real-world behaviors can be understood. The MVL framework requires referent replication and matched inputs between model and referent data to make a direct comparison and compute the MVL. However, it is possible to calculate an MVL using a referent interpolator when these conditions are not met (see Appendix E). All collected data should be inspected for erroneous data due to known collection errors.

The MVL framework can enable continuous validation across the lifecycle, meaning model validity can be reevaluated as the model evolves and more referents are obtained. In the continuum of test and evaluation (T&E), models can and should be revised and updated as more data is obtained on system behavior as part of an iterative cycle. VV&A strategies should identify when the model will be used to inform decisions and detail the validation occurring prior to those decision-making events, including what referents will be available. MVLs can be used to track model validity over time, supporting the shift toward digital engineering and T&E as a continuum.

Additional resources:

- Test Planning Guide (Adams et al., 2022)
- Handbook on Statistical Design and Analysis Techniques for Modeling & Simulation Validation (Wotjon et al., 2019)

MVL Calculation

This section will walk through the mathematical process for deriving the MVL from model data, referent data, and the scope of intended use. Since the MVL calculation is automated using the MVL R tool, this section focuses on guiding the reader through the process beginning to end while building intuition about the meaning of the MVL and lower-level metrics. Additional appendices augment this section to provide full mathematical details.

Prior to calculation, the user must provide codified model data, referent data, and the scope of intended use. Additionally, the user must assign the appropriate authority level to each referent. Based on the scope of intended use, an MVL will be calculated for each defined response, quantifying the level of trust that can be placed in each response from the model.

Figure 4 pictures an example scenario for calculating an MVL. In this scenario, the scope of intended use includes two factors of interest, Factor A and Factor B, and the relevant ranges for those factors, creating the square scope region shown. Model and referent data are available at each of nine validation points, which are points within the scope of intended use where a response was collected under the same input conditions from both the model and at least one referent. Both live system test data and SME judgement are used as referents to validate the model, with SME judgement used to augment live system test data where it is more challenging to obtain. The following sections will describe MVL calculation steps using Figure 4 for visualization. Note that while Figure 4 depicts a simple two-dimensional example, the MVL framework is extensible to much more complex cases.

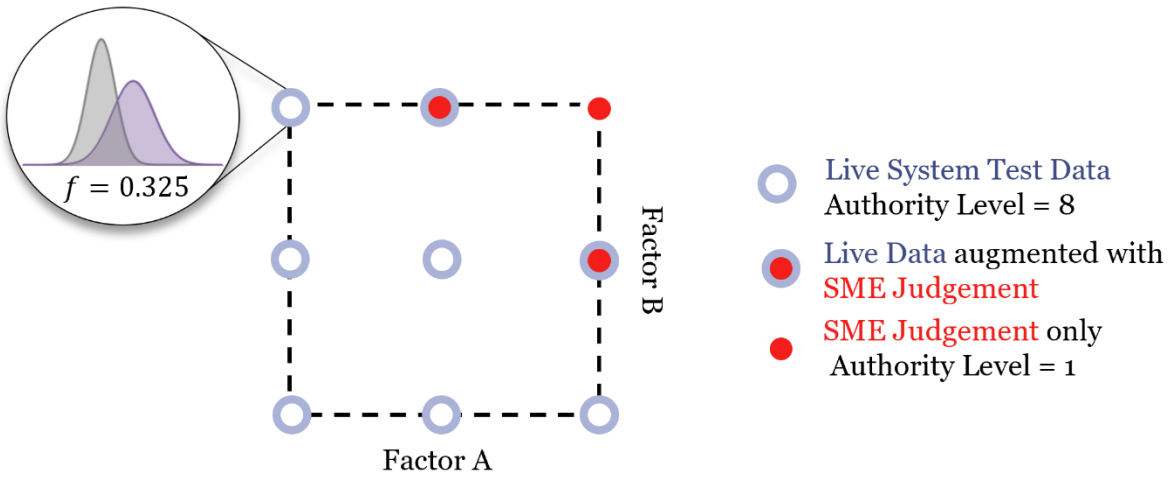

Figure 4
Example MVL Scenario

Finding validation points
Based on the provided observations and input conditions, the MVL algorithm finds validation

points where both model and referent data are available at the same combination of input conditions. For continuous responses, each validation point requires at least two model observations and at least two referent observations, from the same or different referents, to allow the MVL to be calculated. This amount of data allows both model and referent variability to be estimated. For validation points with only one model observation and at least two referent observations, the model variability cannot be estimated (meaning these points cannot support MVL calculation), but these validation points can still support calculation of an MVL_a. For cases where there are limited validation points due to limited replication or lack of points at the same input combinations, the MVL may be calculated using a referent interpolator (see Appendix E). If the response is known to follow a binomial, exponential, or Poisson distribution, one data point from both model and referent is sufficient for the MVL calculation (see Appendix C). Together, the validation points represent the intersection of the three scopes in Figure 2, the model scope, referent scope, and scope of intended use.

At each validation point

The validation points provide the locations where the fidelity can be determined by comparing the model and referent responses under the same inputs and quantifying their consistency. When validating a model, each validation point should be treated separately to first understand if a model is valid at that single point. This is pictured in Figure 4 with the magnification of one of the validation points. Then, analysis can be combined across many validation points to understand model validity across an entire scope domain. This section walks through the analysis that takes place at each validation point, including pooling referents with Bayesian power priors, computing fidelity, and determining model authority.

Pooling Referents with Bayesian Power Priors

If referent data is available from multiple referents, analysis first requires referent data to be pooled together to provide an overall understanding of system behavior based on the different referents available. For example, in Figure 4, two of the validation points have referent data from both live system test and SME judgement; these data sources must be pooled so that the model can be compared against a single body of data.

The MVL framework employs a Bayesian Power Prior method to pool referent data together in a manner consistent with the amount of authority each referent holds (Stafford et al., 2024b). Bayesian Power Priors allow information from different sources to be assigned a weight and pooled together to form a single distribution representing system behavior. The goal of this pooling is to determine a pooled referent mean, standard deviation, and resolution. These measures correspond to accuracy, repeatability, and resolution, and when quantified, they allow fidelity to be determined.

The weights assigned to different referents in the pooling are derived by applying the geometric scale in Equation 1 to the nine-level referent authority scale in Table 2. Table 4 shows the resulting weighted scale (Provost et al., 2022). Additionally, the weights assigned to each referent can be interpreted as modulating the value placed on a data point from that referent. For example, using the weighting scale, approximately 20 data points collected from a component lab test carry the same amount of influence as one observational real-world data point.

$$w_l = e^{-\frac{1}{2}(9-l)}, \quad l = 1, \ldots, 9 \tag{1}$$

Table 4

Weights of Referent Authority Levels and Effective Number of Equivalent Points

Authority Level	Relevant Referent	Weight	Equivalent Number of Data Points
1	SME Judgement	0.0183	54.60
2	First Principles/Physics Predictions	0.0302	33.12
3	Component Lab Test Data	0.0498	20.09
4	Integrated Component Lab Test Data	0.0821	12.18
5	Lab-Scale System Test Data	0.1353	7.39
6	HWIL & SWIL Data	0.2231	4.48
7	Prototype Field Test Data	0.3679	2.72
8	Live System Test Data	0.6065	1.65
9	Operational Real-World Data	1.000	1.00

Conceptually, this weighting scale means that whenever multiple referents are available at a validation point, the higher-level referents more heavily influence the resulting pooled distribution. When there is disagreement between referents, the pooled distribution expresses this disagreement through a larger pooled standard deviation than if referents agreed; in other words, because referents do not agree, there is not as much repeatability in system behavior.

To conduct referent pooling, the user must designate the appropriate distribution for each response. For example, continuous data (e.g., pressure) may be assumed to follow a normal distribution, while binary data (e.g., hit/miss) may be modelled with a binomial distribution. The type of data collected should guide the choice of distribution. As of this document's publication, the MVL R tool supports normal (continuous data), binomial (binary data), exponential (continuous data, failure time data), and Poisson (count data) distributions.

Appendix C contains the complete mathematical details for performing referent pooling with Bayesian Power Priors and for deriving the pooled referent mean, standard deviation, and resolution.

Computing Fidelity

Once referent data had been pooled, fidelity can be determined by comparing model outputs to pooled referent behavior. For example, Figure 4 shows a fidelity score that might be calculated at one validation point. Since model and referent behavior vary across the scope, fidelity is calculated independently at each validation point. If only one referent is used and referent pooling is not needed, the model is compared directly to that referent. In Figure 4, seven of the nine validation points have only one referent and do not require pooling.

Fidelity is based on both similarity in mean behavior (accuracy) and similarity in variability, where variability comprises both repeatability and resolution. The MVL framework uses a fidelity metric which is scored between zero and one, where zero indicates no fidelity and one is perfect fidelity between the model and referent (Weeks et al., 2022). The fidelity metric is given in

Equation 2, where $\bar{x}_{\mathrm{m}}$ is the model response mean, $\bar{x}_{\mathrm{p}}$ is the pooled referent mean, s_{m}^{*} is the model variability, and s_{p}^{*} is the pooled referent variability. Variability, s^{*}, is defined is Equation 3, where s is the sample standard deviation and δ is the resolution. Appendix C and Appendix D discuss how each of these inputs are obtained for the referent(s) and model, respectively.

$$f = e^{-\frac{1}{2}\left(\frac{\bar{x}_{\mathrm{m}}-\bar{x}_{\mathrm{p}}}{s_{\mathrm{p}}^{*}}\right)^{2}} e^{-\frac{(s_{\mathrm{m}}^{*}-s_{\mathrm{p}}^{*})^{2}}{s_{\mathrm{m}}^{*}s_{\mathrm{p}}^{*}}} \tag{2}$$

$$s^{*} = \sqrt{s^{2} + \frac{\delta^{2}}{12}} \tag{3}$$

The fidelity metric is the product of two lower-level metrics, the accuracy metric f_{a} and the variability metric f_{v}, defined in Equations 4 and 5, respectively.

$$f_{\mathrm{a}} = e^{-\frac{1}{2}\left(\frac{\bar{x}_{\mathrm{m}}-\bar{x}_{\mathrm{p}}}{s_{\mathrm{p}}^{*}}\right)^{2}} \tag{4}$$

$$f_{\mathrm{v}} = e^{-\frac{(s_{\mathrm{m}}^{*}-s_{\mathrm{p}}^{*})^{2}}{s_{\mathrm{m}}^{*}s_{\mathrm{p}}^{*}}} \tag{5}$$

The accuracy metric scores fidelity between zero and one based on the difference between the model and referent means, normalized by the amount of variability in the referent. In other words, the referent variability creates a scale to tell what difference in means is meaningful. Examples of accuracy metric scores as the model and referent means diverge are shown in Figure 5. Note that a separation of model and referent means that is equal to the pooled referent variability s_{p}^{*} (such that $f_{\mathrm{a}} = \exp(-1/2) = 0.6065$) results in a fidelity score that is the same as the base in the geometric sequence in Equation 1 (which defines the authority weights in Table 4). This means that a fidelity score of 0.6065 results in 1-level decrease in model authority from the referent to the model, regardless of the level of the referent.

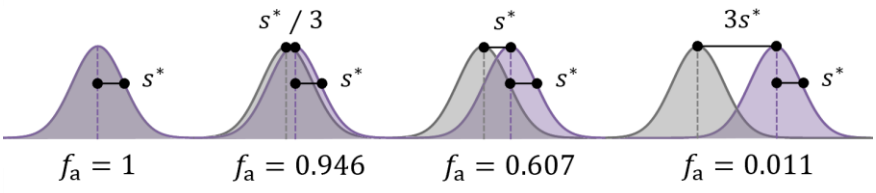

Figure 5
Visualizing the Accuracy Metric

Similarly, the variability metric scores fidelity between zero and one based on the difference between the model and referent variabilities. Figure 6 illustrates the variability metric for different cases.

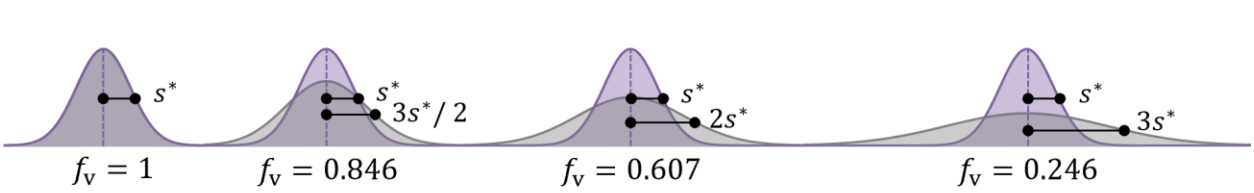

Figure 6
Visualizing the Variability Metric

Fidelity must be high in terms of both the accuracy and variability metrics to achieve a high fidelity score overall. If one or both metrics are low, the fidelity will also be low. Alternatively, an MVL_a could be calculated, which only takes the accuracy metric into account.

Figures 5 and 6 use a normal distribution to illustrate the model and referent distributions; however, the fidelity metric is broadly applicable to many types of distributions. Appendix C discusses how referent statistics are calculated for normal, exponential, binomial, and Poisson distributions. Appendix D discusses how model statistics are determined for those distribution types. These statistics can then be inputted into the fidelity metric.

Determining Model Authority

When referents have been pooled and the fidelity has been calculated, the final step at each validation point is to determine the amount of authority that is passed from the referent to the model. If a referent has high authority but poor fidelity with the model, the model cannot be considered very authoritative. Similarly, if the model has excellent fidelity with the referent, but the referent is not very authoritative, then the model also cannot be considered very authoritative. The model authority at a single validation point can be expressed as the product of the fidelity and the referent authority weight from Table 4. When multiple referents are pooled together, the referent authority at that validation point is the same as the authority of the highest-level referent incorporated into the pooling. Equation 6 gives the model authority weight w_m at a given validation point, where $\max(w_r)$ is the maximum referent authority at that point and f is the fidelity.

$$w_m = \max(w_r) \cdot f \tag{6}$$

Conceptually, the model authority weight quantifies the weight of authority that has been passed from the referent(s) to the model at a single validation point, and it is an intermediate score in the MVL calculation. The model authority weight is on the zero-to-one weighted scale, but it can be converted back to the zero-to-nine scale of Table 4, so the user can understand the level of authority which has passed to the model. This conversion from an authority weight to an authority level uses the inverse of Equation 1 and is given Equation 7. For example, in Figure 4, the magnified validation point has a level 8 referent with an authority weight of 0.6065; the fidelity between the model and referent at that point is 0.325, so the weight passed to the model is $w_m = 0.6065 * 0.325 = 0.197$. When inserted into Equation 7, the authority level of the model at that validation point is 5.75. The model authority level will never be higher than the level of the referent used to validate it.

$$l_m = 9 + 2\ln(w_m) \tag{7}$$

Note that because the geometric scale in Equation 1 uses a base of $\exp(-1/2) = 0.6065$, a fidelity score of 0.6065 results in a one-level drop in authority from the referent to the model.

This fidelity score is the same as the accuracy fidelity score obtained from the model and referent means being one s_{p}^{*} apart.

Across All Validation Points

Once the model authority weight has been determined at each validation point, the model can be assessed for validity across the entire scope of intended use. For example, in Figure 4, the model authority weight is calculated for each of the nine validation points to understand how much authority can be transferred to the model at those points; however, the validation points are not necessarily representative of the scope of intended use. Scope coverage must be assessed and combined with model authority to determine a model's MVL.

Quantifying Scope Coverage

For a model to be considered valid for a given intended use, it must be shown that the validation points produce enough evidence to demonstrate validity for the entire intended use scope, not just at validation points where data was collected. When a scope domain is well-covered with validation points, the validity determined at validation points also covers the space between points.

The MVL framework quantifies coverage of the scope of intended use using a metric scored between zero and one, where zero indicates no coverage and one indicates complete coverage of the scope domain. The details of how the metric is computed depends on the data types of the factors that define the intended use scope. When all factors are continuous, the coverage metric is broken down into two lower-level metrics, as in Equation 8, where C is the coverage, C_V is the volume coverage metric, and C_D is the density coverage metric (Provost et al., 2022). Both C_V and C_D are also metrics scored between zero and one, and both must be high in order to obtain a high scope coverage score.

$$C = C_V C_D \tag{8}$$

Conceptually, these lower-level metrics are understood as pictured in Figure 7, where volume coverage is high when it is possible to interpolate anywhere in the space from validation points, and density coverage is high when validation points densely populate the scope. For example, in the upper left case in Figure 7 (which matches the scenario in Figure 4), $C_V = 1$ and $C_D = 0.978$, resulting in $C = 0.978$. Appendix D provides complete mathematical formulations for each of these metrics.

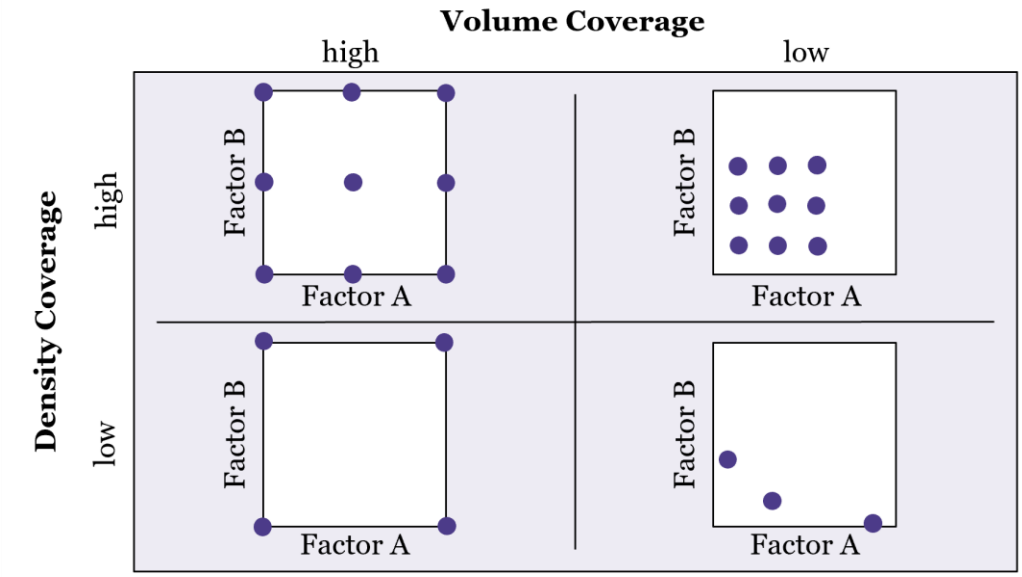

Figure 7
Volume Coverage Versus Density Coverage for Continuous Factors

When factors are categorical instead of continuous, coverage is assessed in terms of the fraction of combinations that are covered. When factors are a mix of continuous and categorical factors, a hybrid approach is used which assesses volume and density coverage within different categorical combinations. These approaches are described in detail in Appendix D.

Putting It All Together: Calculating the MVL

The MVL is mathematically derived from the methods and metrics quantifying each of the three model validation pillars: fidelity, referent authority, and scope. Specifically, the average authority transferred to the model at validation points combines with the coverage score to determine the MVL. Similar to Equation 7, which converts model authority weight at a single point to a zero-to-nine level, the weighted average of model authority and coverage are converted into a zero-to-nine score using the inverse of Equation 1.

Equation 9 defines the MVL, where C is the coverage, p is the number of validation points, $\max(w_{ri})$ is the maximum referent authority weight available at validation point i, and f_i is the fidelity between the model and referent(s) at validation point i.. The MVL is a continuous score between zero and nine, with higher scores indicating higher model validity and that more trust can be placed in the results of that model. The MVL can be up to as high, but no higher than, the level of highest authority data used to validate the model.

$$\mathrm{MVL} = \max\left[9 + 2\ln\left(\frac{C}{p}\sum_{i=1}^{p}\max(w_{ri})\cdot f_i\right), 0\right] \tag{9}$$

Equation 9 uses the maximum function so that the MVL can be no lower than zero, which would otherwise occur for poor scores across all pillars.

Outputs, Interpretations, and Actions

This section discusses the interpretation of several different MVL framework outputs, including the MVL itself as well as lower-level metrics and improvement metrics which provide additional information. Additionally, depending on the outputs generated, different risk-management and model improvement actions are discussed.

Interpreting the MVL

The MVL is a mathematically derived metric scored continuously between zero and nine that rates the validity of a model output for a given scope of intended use. The score is interpreted on an absolute scale, where the MVL indicates the level of trust that can be placed in model results by mapping to quantified levels of trust that are placed in different data sources. This interpretation is shown in Table 5. For example, a MVL of 5 indicates that the model results are as trustworthy lab-scale system test data.

Table 5

Interpretation of MVL in Terms of Trust Placed in Different Data Sources

MVL of:	Is as Trustworthy as:
1	SME Judgement
2	First Principles/Physics Predictions
3	Component Lab Test Data
4	Integrated Component Lab Test Data
5	Lab-Scale System Test Data
6	HWIL & SWIL Data
7	Prototype Field Test Data
8	Live System Test Data
9	Operational Real-World Data

While Table 5 shows a discrete one-through-nine scale, the MVL is scaled continuously between zero and nine and will likely fall somewhere between those discrete levels. The MVL can be no higher than the highest authority level of data used to validate the model. This property results from the concept that the model is only as trustworthy as the referent data used to validate it, and that trust is passed from the referent to become trust in the model through validation. Additionally, when interpreting the MVL, recall that the referent authority and MVL scales are interpreted with respect to what is considered the ‘system’ for the model’s intended use.

Models are often relied upon to make decisions about systems that cannot be tested or observed under some or all operational conditions due to safety or restrictive cost (e.g., space systems). In these cases, because the referents used for validation will be of lower authority, the model may never obtain a high MVL. While the model may in fact be highly representative of operations, this cannot be known until operational data is obtained. On the other hand, for a system where operational data is easy to obtain, the MVL could be as high as level nine. Comparing these two cases, one MVL is lower than the other; however, both models may be

suitable for their respective intended uses. MVLs are a tool for understanding the trust that can be placed in the results of a model on an absolute scale, but the acceptable level of risk varies between programs, thus changing how the MVL affects decision making. A low MVL does not necessarily mean that the model is not acceptable for the use case.

Since the MVL is a mathematically derived metric, it is not guaranteed to increase by progressing in the timeline of model development, though this is usually the goal. Instead, the MVL is based only on the model data, referent data, and scope of intended use available at the time of evaluation and may increase or decrease over time depending on the data. Since higher-level referents are typically obtained over time, if no severe fidelity changes occur, the MVL can be expected to increase over time. For instance, early in the system lifecycle, a low-MVL model may only be sufficiently trustworthy for selecting test points, while later, the now-high-MVL model may be used to evaluate the system against requirements.

Accuracy and Variability MVLs

Besides the MVL, many lower-level scores can also be reported to increase understanding of model validity and point to areas of improvement. Accuracy and Variability MVLs evaluate the model's validity using only one component of the fidelity metric. Recall, the fidelity metric is composed of accuracy and variability components which assess similarity of mean behavior and similarity of variabilities, respectively, between the model and the referent. The accuracy MVL can also be the only MVL reported in cases where the model uncertainty is not quantified, and the user accepts that the model's ability to predict behavior variability is not validated.

The accuracy MVL, $\mathrm{MVL_a}$, is defined in Equation 10, where C is the coverage, p is the number of validation points, $\max(w_{\mathrm{r}i})$ is the maximum referent authority weight available at validation point i, and $f_{\mathrm{a}i}$ is the accuracy fidelity between the model and referent(s) at validation point i.

$$\mathrm{MVL_a} = \max\left[9 + 2\ln\left(\frac{C}{p}\sum_{i=1}^{p}\max(w_{\mathrm{r}i})\cdot f_{\mathrm{a}i}\right), 0\right] \tag{10}$$

The variability MVL, $\mathrm{MVL_v}$, is defined in Equation 11, where $f_{\mathrm{v}i}$ is the variability fidelity between the model and referent(s) at validation point i.

$$\mathrm{MVL_v} = \max\left[9 + 2\ln\left(\frac{C}{p}\sum_{i=1}^{p}\max(w_{\mathrm{r}i})\cdot f_{\mathrm{v}i}\right), 0\right] \tag{11}$$

Each of these metrics (if/when they can be calculated) will be higher than the MVL since they quantify trust using only one aspect of fidelity, where the MVL takes both aspects into account, providing a more complete picture of validity.

Lower-level Metrics

While the MVL provides a high-level understanding of model trust, lower-level metrics are essential for gaining a deep understanding of model risk as well as for identifying opportunities to reduce that risk. Since the MVL framework is built on the three pillars of fidelity, referent authority, and scope, these pillars provide the areas for understanding a model's risk at a deeper level. Table 6 shows an example of a summary table that may be produced when calculating an MVL based on the scenario in Figure 4, including several lower-level metrics. This type of table can be produced for each of the responses of interest, and it is automatically

generated as part of the MVL R tool output when computing an MVL.

Table 6
Example MVL Summary for a Response

MVL	**6.690**
MVL_a	7.246
MVL_v	6.865
No. of validation points	9
Average fidelity	0.732
Average f_a	0.903
Average f_v	0.816
Average authority level	7.777
Coverage	0.978
C_V	1.000
C_D	0.978

In addition to the MVL, MVL_a, and MVL_v discussed above, this table first shows the number of validation points, which is useful to understand how many data points are present where the model and referent(s) can be directly compared, since it may differ from the total number of data points. Each validation point requires at least one model point and two referent points, from the same or different referents, to be counted. This amount of data allows f_a and therefore MVL_a to be calculated. For f_v, MVL_v, and the MVL to be calculated, at least two model points are required at a validation point, so that model variability can be calculated and compared to referent variability.

It is important to recall that fidelity of a model is not calculated directly for an entire model; instead, it is calculated at each validation point, and is understood at the model-level using the average fidelity, which summarizes fidelity across all validation points. The average f_a and f_v scores provide more insight into where poor fidelity may stem from. If f_a is low, the model has accuracy issues, if f_v is low, the model does not capture the real-world variability. These average fidelities are summarized in Table 6.

Because the MVL framework allows multiple referents to be used together to validate across the scope of intended use, the referent authority can be summarized using an average authority level of all referents used to validate, as in Table 6. This average uses the authority weights in Table 4 and is defined in Equation 11, where p is the number of validation points, and $\max(w_{ri})$ is the maximum referent authority weight available at validation point i.

$$\text{Average authority level } = 9 + 2\ln\left(\frac{1}{p}\sum_{i=1}^{p}\max(w_{ri})\right) \quad (12)$$

Finally, the coverage score(s) are essential for understanding the degree to which the data that has been collected can serve to validate the entire scope of intended use. Depending on the data types of the factors (e.g. continuous or categorical), different metrics can be reported to deepen understanding of coverage. Table 6 shows an example of metrics reported in the case that all factors are continuous. In this case, high C_{V} indicates no extrapolation is required to validate the entire scope of intended use, while high C_{D} indicates validation point coverage is dense. Appendix E provides more details on coverage outputs produced in other data-type cases.

Improvement Metrics

In addition to MVL summary and breakdown in Table 6, the MVL R tool also produces a table of improvement metrics that can help the user scope what actions would be most effective for increasing the MVL. An example of improvement metrics that could be generated based on the scenario in Figure 4 are shown in Table 7, including their interpretations.

Table 7

Improvement Metrics for an MVL

MVL = 6.690

Metric	Improved MVL	Change	Interpretation
Fidelity	7.732	1.042	If fidelity was 1 everywhere, the MVL would be 7.732
f_{a}	6.865	0.175	If f_{a} was 1 everywhere, the MVL would be 6.865
f_{v}	7.246	0.556	If f_{v} was 1 everywhere, the MVL would be 7.246
Authority	8.332	1.641	If data was level 9 everywhere, the MVL would be 8.332
Coverage	6.735	0.044	If coverage was 1, the MVL would be 6.735
C_{V}	6.690	0.000	If volume coverage was 1, the MVL would be 6.690
C_{D}	6.735	0.044	If density coverage was 1, the MVL would be 6.735

Essentially, the improvement metric table shows how the MVL might have been higher if a perfect score would have been obtained for different lower-level metrics. The change column shows the amount of improvement from the original score and can be used to help prioritize MVL improvement options based on which could have the most impact. In Table 7 improving authority is shown to have the biggest impact, followed by fidelity, where variability has the biggest impact. When interpreting these scores, recall that fidelity, authority, and coverage all depend on the model and referent data used to generate the MVL and in most cases do not change independently of each other. For example, if coverage were to be improved by gathering more referent data in a previously uncovered region, that new data would also influence the fidelity and authority (e.g., if the new data does not match model well or is of lower authority than previous data). Therefore, the “Improved MVL” may not be obtainable to the

degree in Table 7, because changes affect more than just a single lower-level metric and because improvement efforts often do not result in a “perfect” score (e.g., 1 for fidelity).

Actions for Risk Management and Model Improvement

The MVL and each of the accompanying outputs enable informed decision making on model use. Importantly, the MVL does not determine whether the model should be used; rather, it quantifies the level of risk, and allows the decision maker to decide if that level of risk is acceptable for the needs of the program.

The first possible outcome from an MVL assessment is that the model is acceptable for the intended use case and no further action is necessary. Recall, the acceptable level of risk varies between programs and between the models used by any given program, changing how the MVL affects decision making. A low MVL does not necessarily mean that the model is not acceptable for the use case.

However, if the MVL is deemed unacceptable for the intended use, risk reduction or model improvement actions are necessary. The MVL framework provides several outputs which guide the user in selecting the appropriate actions. Different actions are recommended based upon which MVL pillar needs improvement: fidelity, referent authority, and/or coverage. The summary table (Table 6) and improvement metric table (Table 7) guide which pillars carry the most risk and which will be most impactful on the MVL when improved. The recommended actions are summarized in Table 8. Any combination of these actions is also possible, as multiple pillars can be improved together.

Table 8

Appropriate Risk Reduction Actions by Risk Area

Risk Area	Action(s)
Low Fidelity	• Reduce scope to high fidelity regions • Improve model in low fidelity regions • Choose a different model (if an option) • Collect more replicates
Low Referent Authority	• Collect higher authority data • Reduce scope based on data availability
Low Coverage	• Reduce scope of intended use • Collect data in uncovered scope regions

One common action across all risk areas is reducing the scope of intended use. If fidelity is high for some input combinations within the scope and low for others, one option for reducing model use risk is to only use the model in the regions where fidelity is high. Fidelity outputs from the MVL R tool can be used to determine boundaries for these regions (such as through clustering methods). The MVL can then be recalculated for the newly defined scope of intended use. Similarly, if referent authority varies across the scope, the intended use can be reduced to only areas where high authority data is available. Finally, coverage can be improved by reducing the scope to be more covered by validation points, with less extrapolation.

If reducing the scope is not desired, alternative improvement actions are possible. To improve fidelity without reducing the scope, the model itself must be improved, whether by altering

modelling mechanisms, solvers, or using another model entirely. Another option is to use multiple models to cover the full scope, where each model may be valid for different regions. Fidelity can also potentially be improved by collecting more replicates to get better estimates of the mean and standard deviation. To improve referent authority, higher authority referents must be collected at existing or new validation points. When obtaining these referents is not feasible, improvement must occur in either fidelity or coverage, or the decision-maker may need to accept more risk due to the limited data. Recall the MVL may only be as high as the highest authority data available. Lastly, when coverage is poor, more data can be collected in uncovered regions, of any authority level, to reduce extrapolation and/or interpolation risk. Lower-level referents may be easier to collect in uncovered regions than higher-level referents, and can drastically increase coverage, although with moderate reduction to authority.

Discussion

The MVL framework provides a way to quickly communicate model trust to decision makers, who must evaluate the MVL against the level of trust needed to support decision making. However, there are risks MVLs do not account for such as tool availability, documentation, post processing required, model integration risks, etc., which must also be weighed in decision making.

When implemented throughout system and model development, the MVL can be used to track the level of trust as it changes with the system model. The Model-Validate-Design-Test-Validate paradigm summarizes how models can be built and validated (Collins, 2023). MVLs can be incorporated first in ‘Model’ to assess trust in historical models for a new intended use case. Then, in the first ‘Validate,’ MVLs can validate early model outputs on physics predictions, early technology demonstrations, and historical data. In the final ‘Validate’ step, MVLs can validate model outputs against test data obtained in the ‘Test’ step. As the state of digital engineering continues to evolve, the MVL framework can automate these processes through integration into digital infrastructures containing shared data and models.

While MVLs are excellent for quickly quantifying model trust for a broad range of models, they can and should be augmented with other statistical methods that can be tailored to the specific validation scenario. These tailored methods can help overcome limitations of MVLs created by their generality. For example, hypothesis tests tailored to the data type can inform on the statistical significance of differences between model and referent behavior.

This paper describes the MVL methodology and is augmented by the MVL R tool and user guide which automate the calculation described here. For case studies showing the application of MVLs to different models, see Stafford et al, 2024a. Future work includes updating this paper and the R tool as needed to incorporate lessons learned from the M&S and T&E communities as MVLs are applied.

Conclusion

As M&S continues to be used and relied upon to inform system decisions, particularly in the shift toward digital engineering, the MVL framework provides an essential tool for quickly quantifying and communicating the amount of trust (and therefore risk) held by model results. MVLs provide a detailed understanding of model validity in terms of the three pillars of validation: fidelity, referent authority, and scope. The summary of lower-level MVL scores assessing each of these pillars provides vital information for identifying risk-reduction actions

when the model does not meet the trust level required for decision making. The automation of MVLs through the MVL R tool enables users to easily incorporate MVLs into validation plans, as well as into digital environments, to understand model validity and how it changes over time. MVLs can be implemented to assess trust in a wide variety of models, across all stages of system development and operation, providing utility to both decision makers and model developers to use M&S in the DOD more effectively.

References

Adams, W., Divis, E., Jones, N., Kershner, C., Lazarus, J., Marshall, M., McBride, A., Mott, T., Natoli, C., Oimoen, S., Provost, K., Ramert, A., Sgambellone, A., Sigler, G., Theimer, J., Truett, L., & Weeks, C. (2022). *Test Planning Guide*. https://www.afit.edu/images/pics/file/Final%200930_Test%20Planning%20Guide_2_2%20(1).pdf

Ahner, D. K., Jones, N., Adams, W., Key, M., Weeks, C., & Provost, K. (2023). *Model Validation Levels: A New Framework Enabling Model Assessment and Use.* [Manuscript submitted for publication].

Burke, S. (2017). *Model Building Process Part 1: Checking Model Assumptions V 1.1.* Best Practice, Scientific Test & Analysis Techniques Center of Excellence. https://www.afit.edu/images/pics/file/Model%20Building%20Process%20Part%201%20Checking%20Model%20Assumptions%20V2.pdf

Burke, S. (2018). *Model Building Process Part 2: Factor Assumptions*. Best Practice, Scientific Test & Analysis Techniques Center of Excellence. https://www.afit.edu/images/pics/file/Model%20Building%20Process%20Part%202%20Factor%20Assumptions.pdf

Burke, S. (2020). *The Model Building Process Part 3: Model Goodness Metrics*. Best Practice, Scientific Test & Analysis Techniques Center of Excellence. https://www.afit.edu/images/pics/file/Model%20Building%20Process%20Part%203%20Model%20Metrics%20Final.pdf

Collins, C. (2023, July 18–20). *The Test & Evaluation Continuum in Support of Multi-Domain Operations* [Conference presentation]. Multi-Domain Operations in an Extended Range Environment, Ventura, CA, United States.

Government Accountability Office. (2020). Technology Readiness Assessment Guide: Best Practices for Evaluating the Readiness of Technology for Use in Acquisition Programs and Projects. Government Accountability Office.

Helton, J. C., & Johnson, J. D. (2011). Quantification of margins and uncertainties: Alternative representations of epistemic uncertainty. *Reliability Engineering & System Safety*, *96*(9), 1034–1052. https://doi.org/10.1016/j.ress.2011.02.013

Ibrahim, J. G. & Chen, M. (2000). Power Prior Distributions for Regression Models. *Statistical Science*. 15(1), 46-60. https://doi.org/10.1214/ss/1009212673

Ibrahim, J. G., Chen, M., Gwon, Y., & Chen, F. (2015) The power prior: theory and applications. *Stat Med*. 34(28), 3724-3749. https://doi.org/10.1002/sim.6728

Jones, N., Adams, W., & Burke, S. (2021). Model Selection and Use of Empiricism in Digital Engineering. Best Practice, Scientific Test & Analysis Techniques Center of Excellence. https://www.afit.edu/images/pics/file/Model%20Selection%20and%20Use%20of%20Empiricism%20in%20Digital%20Engineering.pdf

Jones, N., Provost, K., & Stafford, C. (2024). Model Validation Level (MVL) R Tool User Guide.

User Guide, Scientific Test & Analysis Techniques Center of Excellence.

Lazarus, J., Weeks, C., Jones, N., Provost, K., Sigler, G. (2021). Uncertainty Quantification: An Overview. Best Practice, Scientific Test & Analysis Techniques Center of Excellence. https://www.afit.edu/STAT/statcoe_files/4_0328_LazarusUQBP_2_2.pdf

Miller, R. D. (2020). *Guidance-Based M&S V&V Checklist* [Memorandum]. Institute for Defense Analyses Operational Evaluation Division. https://osd.deps.mil/org/dote-extranet/Guidance/Modeling%20and%20Simulation%20Guidance/MandS%20VandV%20Guidance-Based%20Checklist.pdf

Modeling and Simulation Enterprise. (2021, August). Department of Defense Modeling and Simulation Glossary. Retrieved from Department of Defense Modeling and Simulation Enterprise: https://www.msco.mil/MSReferences/Glossary

Natoli, C., & Burke, S. (2018). Computer Experiments: Space Filling Design and Gaussian Process Modeling. Best Practice, Scientific Test & Analysis Techniques Center of Excellence. https://www.afit.edu/images/pics/file/Computer%20Experiments-%20Space%20Filling%20Designs%20and%20Gaussian%20Process%20Modeling.pdf

Ortiz, F. (2018). *Categorical Data in a Designed Experiment Part 1: Avoiding Categorical Data.* Best Practice, Scientific Test & Analysis Techniques Center of Excellence. https://www.afit.edu/images/pics/file/Categorical%20Data%20in%20a%20Designed%20Experiment%20Part%201%20Avoiding%20Categorial%20Data1.pdf

Pace, D. K. (2015). Fidelity, Resolution, Accuracy, and Uncertainty. Modeling and Simulation in the Systems Engineering Life Cycle, Simulation Foundations, Methods and Applications, 369. Retrieved from https://doi.org10.1007/978-1-4471-5634-5_10

Provost, K., Stafford, C., & Jones, N. (2024). MVL R Tool. Tool. Scientific Test & Analysis Techniques Center of Excellence.

Provost, K., Weeks, C., Jones, N., & Sieck, V. (2022). *Elements of a Mathematical Framework for Model Validation Levels*. Best Practice, Scientific Test & Analysis Techniques Center of Excellence. https://www.afit.edu/images/pics/file/0930_ProvostMVLBP_2_2.pdf

US Department of Defense. (2012). MIL-STD 3022 Department of Defense Standard Practice Documentation of Verification, Validation, and Accreditation (VV&A) for Models and Simulations.

US Department of Defense. (2018). Department of Defense Instruction 5000.61.

Weeks, C., Jones, N., & Key, M. (2022). *Constructing a Metric for Fidelity in Model Validation*. Best Practice, Scientific Test & Analysis Techniques Center of Excellence. https://www.afit.edu/docs/0831_AFIT2022ENS08106_WeeksFidelityBP_2_2.pdf

Stafford, C., Provost, K., & Jones, N. (2024a). Case Studies on Model Validation Levels. Case Study, Scientific Test & Analysis Techniques Center of Excellence.

Stafford, C., Provost, K., & Jones, N. (2024b). *Model Validation Levels: An Automatable Framework for Model Validation.* [Manuscript submitted for publication].

Wojton, H., Avery, K. M., Freeman, L. J., Parry, S. H., Whittier, G. S, Johnson, T. H., & Flack, A. C. (2019). *Handbook on Statistical Design & Analysis Techniques for Modeling & Simulation Validation.* Institute for Defense Analyses. https://www.ida.org/research-and-publications/publications/all/h/ha/handbook-on-statistical-design-and-analysis

Appendix A
Key Definitions

To ensure a common understanding of the subject, the following definitions are used throughout this paper:

accuracy: the degree to which a parameter or variable, or a set of parameters or variables, within a model or simulation conforms exactly to reality or to some chosen standard or referent (Modeling and Simulation Enterprise, 2021).

aleatory uncertainty: uncertainty arising from an inherent randomness in the properties or behavior of the system under study (Helton, 2011).

convex hull: the smallest possible convex space that contains a set of data points.

epistemic uncertainty: uncertainty derived from a lack of knowledge about the appropriate value to use for a quantity that is assumed to have a fixed value in the context of a particular analysis (Helton, 2011).

fidelity: the level of consistency between a model and a referent, defined in the three dimensions of accuracy, repeatability, and resolution.

model: a physical, mathematical, or otherwise logical representation of a system, entity, phenomenon, or process (US Department of Defense, 2018).

model validation level (MVL): an objective, automatable metric that quantifies how much trust can be placed in the results of a model to represent the real world.

modeling and simulation (M&S): the use of models, including emulators, prototypes, simulators, and stimulators, either statically or over time, to develop data as a basis for making managerial or technical decisions (Modeling and Simulation Enterprise, 2021).

referent: a codified body of knowledge representing real system behavior.

referent authority: the strength of credibility of a referent's claim to be a high-fidelity representation of reality.

repeatability: the similarity of the results obtained from the same model (or referent) over multiple observations under the same input conditions.

resolution: the degree of granularity with which a parameter or variable can be determined (Pace, 2015).

scope: the set of inputs, outputs, assumptions, and limitations representing the mission-relevant system parameters, environmental conditions, constraints, and requirements, and their allowable values.

scope of intended use: the set of dimensions, ranges, and assumptions of the model inputs and outputs needed to represent a system's relevant mission parameters, environmental conditions, constraints, and requirements, combined with the additional constraints imposed by the target modeling environment.

simulation: a method for implementing a model over time (US Department of Defense, 2018).

validation: the process that determines the degree to which a model has fidelity relative to an appropriate referent(s) for a specific intended use.

validation point: a factor combination where both model and referent data are available to validate a response.

validity: the fidelity of a model over a pre-specified scope relative to an appropriate referent(s).

Appendix B

Determining Referent Authority using the Referent Authority Scale

This appendix aims to provide guidance on choosing the appropriate referent level for a given referent. Table B1 contains the referent authority scale for ease of reference. The application of these authority levels is covered in the MVL Calculation section, which discusses the mapping of referent levels to a continuous absolute scale and the method of combining data from multiple referents to build a consolidated expectation of performance while respecting the authority levels of the source referents.

Table B1

Referent Authorities from Relevant TRLs

Authority Level	Relevant Referent
1	SME Judgement
2	First Principles/Physics Predictions
3	Component Lab Test Data
4	Integrated Component Lab Test Data
5	Lab-Scale System Test Data
6	HWIL & SWIL Data
7	Prototype Field Test Data
8	Live System Test Data
9	Operational Real-World Data

In practice, the MVL framework may be applied to many different types of simulations. However, regardless of the object represented by the simulation, the same authority scale can be applied. Whether the simulation represents a system of systems, a component of a larger system, or an environment containing many interacting systems (such as a mission engineering simulation or wargame covering a full battlespace), the primary object being modeled should be considered the “system” for the purposes of applying the referent scale. In the case of a model of a sub-system or component of a larger system, the referents used should pertain to that component. In this case, the referent levels (and MVL produced) for that model would view that component as a system. Since data for the behavior of a component does not describe how the component interacts with other components in a larger target system (or any emergent behaviors that come from those interactions), when it is used as a referent for the full system it has a lower authority than when used as a referent for only the component.

In some cases, the nature of the model or the system being modeled may raise questions about how to interpret some of the levels. Some of the more common cases are addressed here.

First, not all referent levels may be relevant or obtainable for every type of system. For example, a model of a subcomponent may not have referents for components of that subcomponent, if it cannot be further decomposed. Alternatively, for a red system threat, lab-scale data certainly exists, but it is unlikely the user has access to that data. Thus, the user should evaluate what

referents may be available for their given system.

In the case of SME judgement, the MVL framework requires that SME judgement be quantified in terms of a predicted mean value and standard deviation for defined input values. This data should be collected independently from model results. The scale makes no distinction between SMEs based on experience or education level. This is because it is difficult to quantify the impact of experience or education on the SMEs' ability to correctly predict the expected mean and spread of values that might be observed in the response of a system at various conditions. In many cases, this may only be quantifiable by referring to data for the events on which the SMEs' knowledge is based, but if such data is available, it should be preferred as a more authoritative referent.

The treatment of predictions from physics equations and/or other first principles may also present confusion, as many such models are widely used and trusted. Broad use of physics models built on known first principles is often considered an objective goal for DE. Indeed, many high-quality physics models have been built and used extensively in many fields with excellent results. However, physics models have limitations. The first principles represented by any physics model are typically subject to a set of bounding assumptions, and no single set of first principles completely describes all sources of variability that would act on a real system. Despite their esteem, physics and other first-principles models receive authority level 2 because they fail to account for these other sources of variability. If higher authority referents are available to validate that a physics model is robust to certain outside sources of variation, then the model's MVL may be determined and used in place of the default level.

Another potential point of confusion is the use of legacy system data as a referent for a new system model. The legacy data can act as a surrogate referent for similar systems, or for upgrades or replacements to the original system. A legacy system is not the system of interest and does not represent the new system in every respect. The similarity between the systems, and therefore the applicability of the legacy system data to the new system model, is best captured by the similarity in scope between the new system and the legacy system. The legacy data is rated on the referent authority scale according to how it was gathered for the legacy system, but the MVL may ultimately be discounted due to the dissimilarity in scope, even if the legacy data is high authority. Additionally, a user might desire to reuse a legacy model as a representation of a new system, in which case the legacy model should be validated with an MVL against legacy system referents to determine its authority level; however, even if the MVL of the legacy model is high, due to the dissimilarity in scope the new system model MVL could be much lower than the MVL of the legacy model used to validate it.

Appendix C
Referent Inputs to Fidelity Calculation

This appendix contains the complete mathematical for deriving the pooled referent mean, standard deviation, and resolution (Stafford et al., 2024), which are used to calculate fidelity (Equation 2). Bayesian power priors allow referents of various authorities to be pooled together to form a single distribution representing system behavior. Additionally, this appendix discusses how to calculate fidelity inputs when only one referent is used for validation.

Bayesian Power Prior Pooling of Referents

Bayesian power priors provide a framework for incorporating previous information into analysis, where previous data is weighted relative to "current", more authoritative data (Ibrahim & Chen, 2000). In the context of the MVL framework, all referents are pooled together, with weights assigned relative to highest authority referent being pooled. Note the highest authority referent may or may not be the most current data, but it sets the standard by which other referents are weighted. Table C1 shows the relative scale for referent weighting during pooling, which is derived from the absolute scale in Table 4 but defines weights relative to the authority level of the highest available referent at a given validation point. These weights are determined by Equation C1, where w_r is the weight of referent r, l_r is the 1-9 authority level of referent r, and R is the total number of referents at a given validation point.

Table C1

Referent Weights According to Difference in Authority Level from Highest Authority Referent

Authority Level Difference	Weight	Equivalent Number of Data Points
0	1	1
1	0.6065	1.65
2	0.3679	2.72
3	0.2231	4.48
4	0.1353	7.39
5	0.0821	12.18
6	0.0498	20.09
7	0.0302	33.12
8	0.0183	54.6

$$w_r = e^{-\frac{1}{2}[\max(l_1,\cdots,l_R)-l_r]} \tag{C1}$$

The general equation for deriving the pooled distribution (posterior distribution) of parameter(s) θ is given in Equation C2. The parameter(s) θ described by the pooled distribution depend on the type of data being pooled; for instance, binomially distributed data has one parameter of interest, p, the probability of success, while normally distributed data has two parameters of interest, μ and σ, the mean and standard deviation, respectively. In Equation C2, $P(\theta|D_1,\ldots,D_R)$

is the posterior distribution of θ given observed data sets $D_1, \ldots, D_R$, $L(\theta|D_r)$ is the likelihood of observing the data set D_R for different parameter values, and $\pi_0(\theta)$ is the prior distribution. The form of the likelihood function depends on the data distribution (e.g. binomial, normal), and the prior is chosen to be non-informative. The weights w_r in Equation C2 serve to weight the influence of each referent's likelihood on the posterior distribution.

$$P(\theta|D_1, \ldots, D_R) \propto \left(\prod_{r=1}^{R} L(\theta|D_r)^{w_r}\right) \cdot \pi_0(\theta) \quad \text{(C2)}$$

Figure C1 shows an example of pooling exponentially distributed data (commonly used for modeling reliability), where the parameter of interest is the mean time between failures (MTBF) for a system. This example demonstrates how higher-level referents have a greater impact on the pooled distribution and on the predicted MTBF. Additionally, Figure C1 demonstrates how the number of data points factors into the impact a referent has on the pooled distribution: a lower-level referent has more impact when it contains more data. Specifically, in Figure C1, the level 8 referent outweighs the level 9 referent in its impact on the MTBF since it has more data points. On the other hand, in the level 1 referent case, the increased data cannot overcome the drastic difference in authority.

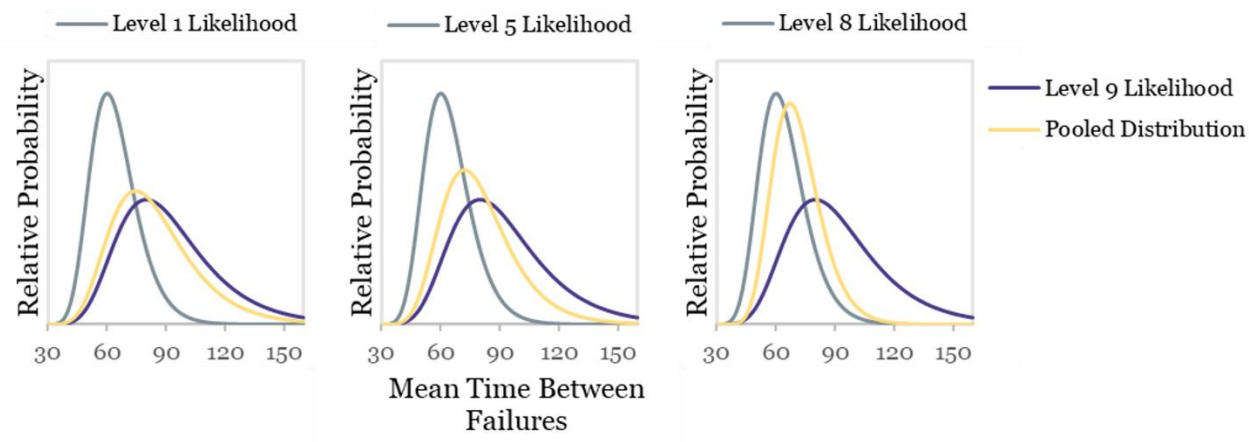

Figure C1
Example of Pooling a Level 9 Referent (15 data points) with a Lower-Level Referent (30 data points)

Due to the wide variety of data in DOD testing, the MVL R tool is tailored to support four different data distribution types: normal, exponential, binomial, and Poisson. The choice of distribution defines the likelihood component in Bayesian pooling. The MVL user can specify the distribution type for a response, and referent data can then be pooled together under that distribution assumption. The normal distribution is the most common distribution used to model continuous data and can be applied in many situations. The exponential distribution is a continuous distribution commonly used in reliability to model failure times. The distribution has one parameter, λ, the rate parameter, which for reliability applications can be interpreted as the inverse of the MTBF. The binomial distribution is commonly used to model binary data (e.g., hit/miss data), and gives the probability of observing x successes in n trials, given the probability of success p. Note that the terms 'success' and 'failure' are used generally here and can be understood as the two possible binary outcomes for any given situation. The Poisson

distribution can be used to model count data, specifically the number of events occurring in a fixed interval of time. The rate parameter, λ, indicates the average number of events in the given time interval.

Given a specific data distribution for the likelihood, the prior is chosen to be a conjugate prior, which allows the pooled distribution to be obtained in closed form. It is also chosen to be non-informative, so the pooled distribution is driven by referent data rather than the chosen prior. The priors for the four distributions covered here are given in Table C2. Jeffrey's prior is a common type of non-informative prior that is used for the exponential, binomial, and Poisson distributions, while the normal distribution prior is defined such that the pooled estimates of mean and standard deviation are unbiased. Due to the properties of conjugate priors, the pooled distribution can be derived in closed form as a function of the collected referent data.

Once the pooled distribution has been obtained, it can be used to make a point estimate, or "best estimate", for the parameter(s). While θ denotes the parameter distributed by the pooled distribution, $\hat{\theta}$ will be used to denote the point estimate for that parameter. The parameter point estimates derived for each of the four distributions discussed here are given in Table C2. In each of these cases the expected value of the pooled distribution (the distribution mean) is used to make the point estimate. These parameter point estimates then allow for calculation of the inputs to the fidelity metric, the pooled mean, $\bar{x}_{\mathrm{p}}$, and the pooled variance, ${s_{\mathrm{p}}}^2$. For the normal distribution, the parameters of mean and variance directly correspond to the inputs into the fidelity calculation. For other data distributions, mean and variance are properties of the distribution, and while they may not directly correspond to the distribution parameter(s), they can be directly calculated from the parameters(s). For example, the variance of the exponential distribution is $1/\lambda^2$, where λ is the rate parameter of the exponential distribution. Table C2 gives the expressions for obtaining $\bar{x}_{\mathrm{p}}$ and ${s_{\mathrm{p}}}^2$, so they can be used to calculate fidelity.

The expressions for the normal distribution point estimates in Table C2 are formulated in terms of individual data points or observations; however, they can also be formulated in terms of summary statistics (mean and standard deviation of each referent). This alternate formulation is given in Equations C3 and C4, and it is especially useful for referents which are only expressed in terms of summary statistics, such as a SME estimate of mean and standard deviation. When a SME estimate is used, $n_r = 2$ should be used in Equations C3 and C4 to represent the number of pieces of information provided (mean and standard deviation). If the SME estimates for mean and standard deviation were derived from more pieces of information (e.g., three quantile estimates) another value for n_r may be appropriate.

$$\bar{x}_{\mathrm{p}} = \frac{\sum_{r=1}^{R} w_r n_r \bar{x}_r}{\sum_{r=1}^{R} w_r n_r} \tag{C3}$$

$${s_p}^2 = \frac{1}{\sum_{r=1}^{R} w_r n_r - 1}\left[\sum_{r=1}^{R} w_r {s_r}^2 (n_r - 1) + \sum_{r=1}^{R} w_r n_r {\bar{x}_r}^2 - \frac{\left(\sum_{r=1}^{R} w_r n_r \bar{x}_r\right)^2}{\sum_{r=1}^{R} w_r n_r}\right] \tag{C4}$$

Table C2

Parameter Point Estimates and Fidelity Inputs for Pooling Different Distribution Types

Distribution Type	Parameter(s)	Prior	Point Estimate(s)[1]	Pooled Mean	Pooled Variance
Normal	Mean: μ Variance: σ^2	$\pi_0(\mu,\sigma^2) \propto \frac{1}{\sigma^4}$	$\hat{\mu} = \frac{\sum_{r=1}^{R}\sum_{i=1}^{n_r} x_{r,i} * w_r}{\sum_{r=1}^{R} w_r n_r}$ $\hat{\sigma}^2 = \frac{\sum_{r=1}^{R}\sum_{i=1}^{n_r}\left(x_{r,i}-\hat{\mu}\right)^2 * w_r}{\sum_{r=1}^{R} w_r n_r - 1}$	$\bar{x}_p = \hat{\mu}$	${s_p}^2 = \hat{\sigma}^2$
Exponential[2]	Rate parameter: λ	$\pi_0(\lambda) \propto \frac{1}{\lambda}$	$\hat{\lambda} = \frac{\sum_{r=1}^{R} w_r n_r}{\sum_{r=1}^{R} w_r n_r \bar{x}_r}$	$\bar{x}_p = \frac{1}{\hat{\lambda}}$	${s_p}^2 = \frac{1}{\hat{\lambda}^2}$
Binomial[3]	Success probability: p	$\pi_0(p) \propto \frac{1}{\sqrt{p(1-p)}}$	$\hat{p} = \frac{\sum_{r=1}^{R} w_r k_r + \frac{1}{2}}{\sum_{r=1}^{R} w_r n_r + 1}$	$\bar{x}_p = \hat{p}$	${s_p}^2 = \hat{p}(1-\hat{p})$
Poisson	Rate parameter: λ	$\pi_0(\lambda) \propto \sqrt{\frac{1}{\lambda}}$	$\hat{\lambda} = \frac{\sum_{r=1}^{R} w_r n_r \bar{k}_r + \frac{1}{2}}{\sum_{r=1}^{R} w_r n_r}$	$\bar{x}_p = \hat{\lambda}$	${s_p}^2 = \hat{\lambda}$

[1]Where R is the number of referents, w_r is the weight of referent r, and n_r is the number of observations. Normal: $x_{r,i}$ is the observed response for the i^{th} observation in the r^{th} referent. Exponential: $\bar{x}_r$ is the mean time to event. Binomial: k_r is the number of successes in referent r. Poisson: $\bar{k}_r$ is the mean number of events observed in n_r time intervals for referent r.

[2]The exponential point estimate expression can also be used for censored data, where n_r is the number of non-censored events and $\bar{x}_r$ is the total test time (censored and non-censored events) divided by n_r, the number of non-censored events.

[3]The pooled mean and variance expressions assume one trial so the mean is a proportion bounded between zero and one.

Determining Pooled Resolution

While the previous section describes how to obtain the pooled mean $\bar{x}_{\mathrm{p}}$ and the pooled variance ${s_p}^2$ for the fidelity metric, this section discusses further how to obtain the pooled resolution.

Recall resolution is defined as the degree of granularity with which a parameter or variable can be determined (Pace, 2015), and it can be considered equivalent to the epistemic uncertainty, which is defined as the uncertainty derived from a lack of knowledge about the appropriate value to use for a quantity that is assumed to have a fixed value in the context of a particular analysis (Helton, 2011). Lazarus et.al provide an overview of uncertainty quantification, where epistemic uncertainty is a key concept, in the T&E context (2022). In the MVL framework, resolution should be specified for both the model and each referent for a given response. For example, a model with numerical approximations may have a known quantifiable error associated with its results, which contributes to the epistemic uncertainty on the response. This uncertainty is reducible, such as by using a higher order approximation or a finer numerical mesh. On the other hand, aleatory uncertainty, which represents the inherent noise in the system, is not reducible. In addition to numerical approximations, epistemic uncertainty of model inputs contributes to output uncertainty; therefore, uncertainties should be propagated through the model to quantify the epistemic uncertainty of a model output. For referent data collected from a physical test, epistemic uncertainty may be present due to known measurement error, significant figures, etc.; however, uncertainties are often mixed, meaning the epistemic uncertainty is confounded with the inherent noise. In this case, measured noise may be a good estimate for the total uncertainty.

For models and referents with a known resolution, the MVL R tool currently supports entering resolution for arbitrary continuous or normally distributed data only. The resolution, in this case, is understood as the width of an interval around observed data points. In other words, each data point $x_{r,i}$ in a data set can be understood to have upper and lower bounds as in Equation C5, where δ is the resolution.

$$(\text{lower bound}, \text{upper bound}) = \left(x_{r,i} - \frac{\delta}{2}, x_{r,i} + \frac{\delta}{2}\right) \tag{C5}$$

This understanding of each observation allows the epistemic uncertainty of individual observations to be propagated through the Bayesian pooling, to determine a resolution for the pooled data. This process uses the following steps:

1. Pool referent data “as-is” to calculate the pooled mean, $\bar{x}_{\mathrm{p}}$, and the pooled variance, ${s_p}^2$.
2. Pool the referent data using all the lower bounds to obtain a lower bound on the pooled mean, $\bar{x}_{\mathrm{p,l}}$.
3. Pool the referent data using all the upper bounds to obtain an upper bound on the pooled mean, $\bar{x}_{\mathrm{p,u}}$.
4. Calculate the pooled resolution, $\delta_{\mathrm{p}} = \bar{x}_{\mathrm{p,u}} - \bar{x}_{\mathrm{p,l}}$.

Equation C6 shows how this pooled resolution can then be combined with the pooled variance to determine the pooled variability, ${s_p}^*$, which is input into the fidelity equation (Equation 2).

$$s_{\mathrm{p}}^* = \sqrt{{s_{\mathrm{p}}}^2 + {\delta_{\mathrm{p}}}^2/12} \tag{C6}$$

Single Referent Inputs to Fidelity Calculation for Different Data Types

When only one referent is used for validation, referent statistics are calculated from only that referent to be inserted into the fidelity metric. While calculating the mean and sample standard deviation uses well known equations, when the distribution of data is known, mean and standard deviation can be derived from the maximum likelihood estimators. The equations for $\bar{x}_{\mathrm{p}}$ and ${s_{\mathrm{p}}}^2$ used in the MVL framework for just one referent are summarized in Table C3.

Note that for the binomial and Poisson cases, the MVL framework uses Bayesian estimation with a Jeffrey's prior, similar to when pooling multiple referents, such that non-zero variance estimates are still possible when, for example, only successes are observed, or no events are observed during a time period.

Table C3

Single Referemt Fidelity Inputs for Different Distribution Types

Distribution Type	Model Mean	Model Variance
Arbitrary Continuous or Normal[1]	$\bar{x}_{\mathrm{p}} = \frac{\sum_{i=1}^{n_{\mathrm{r}}} x_{\mathrm{r},i}}{n_{\mathrm{r}}}$	${s_{\mathrm{r}}}^2 = \frac{\sum_{i=1}^{n_{\mathrm{r}}} \left(x_{\mathrm{r},i} - \bar{x}_{\mathrm{r}}\right)^2}{n_{\mathrm{r}} - 1}$
Exponential[2]	$\bar{x}_{\mathrm{r}} = \frac{\sum_{i=1}^{n_{\mathrm{r}}} x_{\mathrm{r},i}}{n_{\mathrm{r}}}$	${s_{\mathrm{r}}}^2 = \frac{\sum_{i=1}^{n_{\mathrm{r}}} x_{\mathrm{r},i}}{n_{\mathrm{r}}}$
Binomial[3]	$\bar{x}_{\mathrm{r}} = \frac{k_{\mathrm{r}} + \frac{1}{2}}{n_{\mathrm{r}} + 1}$	${s_{\mathrm{r}}}^2 = \frac{\left(k_{\mathrm{r}} + \frac{1}{2}\right)\left(n_{\mathrm{r}} - k_{\mathrm{r}} - \frac{1}{2}\right)}{(n_{\mathrm{r}} + 1)^2}$
Poisson[4]	$\bar{x}_{\mathrm{r}} = \frac{\sum_{i=1}^{n_{\mathrm{r}}} k_{\mathrm{r,i}} + \frac{1}{2}}{n_{\mathrm{r}}}$	${s_{\mathrm{r}}}^2 = \frac{\sum_{i=1}^{n_{\mathrm{r}}} k_{\mathrm{r,i}} + \frac{1}{2}}{n_{\mathrm{r}}}$

[1]Where n_{r} is the number of observations and $x_{\mathrm{r},i}$ is the observed response for the i^{th} observation.
[2]Where n_{r} is the number of non-censored observations and $x_{\mathrm{r},i}$ is the observed response for the i^{th} observation.
[3]Where n_{r} is the number of observations and k_{r} is the number of successes.
[4]Where n_{r} is the number of time intervals and $k_{\mathrm{r},i}$ is the number of events observed in the i^{th} time interval.

Appendix D
Model Inputs to Fidelity Calculation

This appendix describes how to obtain the model mean, $\bar{x}_{\mathrm{m}}$, and the model variance, ${s_{\mathrm{m}}}^2$ for inputting into the fidelity metric. While calculating the mean and sample standard deviation uses well known equations, when the distribution of data is known, mean and standard deviation can be derived from the maximum likelihood estimators. The equations for $\bar{x}_{\mathrm{m}}$ and ${s_{\mathrm{m}}}^2$ used in the MVL framework are summarized in Table D1. Note that some distribution types use different methods depending on whether the model is stochastic or deterministic. In the stochastic cases, the MVL framework uses Bayesian estimation with a Jeffrey's prior, similar to in Appendix C, such that non-zero variance estimates are still possible when, for example, only successes are observed, or no events are observed during a time period.

Table D1
Model Fidelity Inputs for Different Distribution Types

Distribution Type	Model Mean	Model Variance
Arbitrary Continuous or Normal[1]	$\bar{x}_{\mathrm{m}} = \frac{\sum_{i=1}^{n_{\mathrm{m}}} x_{\mathrm{m},i}}{n_{\mathrm{m}}}$	${s_{\mathrm{m}}}^2 = \frac{\sum_{i=1}^{n_{\mathrm{m}}} \left(x_{\mathrm{m},i} - \bar{x}_{\mathrm{m}}\right)^2}{n_{\mathrm{m}} - 1}$
Exponential[2]	$\bar{x}_{\mathrm{m}} = \frac{\sum_{i=1}^{n_{\mathrm{m}}} x_{\mathrm{m},i}}{n_{\mathrm{m}}}$	${s_{\mathrm{m}}}^2 = \frac{\sum_{i=1}^{n_{\mathrm{m}}} x_{\mathrm{m},i}}{n_{\mathrm{m}}}$
Binomial[3]	Deterministic: $\bar{x}_{\mathrm{m}} = \frac{k_{\mathrm{m}}}{n_{\mathrm{m}}}$ Stochastic: $\bar{x}_{\mathrm{m}} = \frac{k_{\mathrm{m}} + \frac{1}{2}}{n_{\mathrm{m}} + 1}$	Deterministic: ${s_{\mathrm{m}}}^2 = \frac{k_{\mathrm{m}}(n_{\mathrm{m}} - k_{\mathrm{m}})}{n_{\mathrm{m}}^2}$ Stochastic: ${s_{\mathrm{m}}}^2 = \frac{\left(k_{\mathrm{m}} + \frac{1}{2}\right)\left(n_{\mathrm{m}} - k_{\mathrm{m}} - \frac{1}{2}\right)}{(n_{\mathrm{m}} + 1)^2}$
Poisson[4]	Deterministic: $\bar{x}_{\mathrm{m}} = \frac{\sum_{i=1}^{n_{\mathrm{m}}} k_{\mathrm{m},i}}{n_{\mathrm{m}}}$ Stochastic: $\bar{x}_{\mathrm{m}} = \frac{\sum_{i=1}^{n_{\mathrm{m}}} k_{\mathrm{m,i}} + \frac{1}{2}}{n_{\mathrm{m}}}$	Deterministic: ${s_{\mathrm{m}}}^2 = \frac{\sum_{i=1}^{n_{\mathrm{m}}} k_{\mathrm{m},i}}{n_{\mathrm{m}}}$ Stochastic: ${s_{\mathrm{m}}}^2 = \frac{\sum_{i=1}^{n_{\mathrm{m}}} k_{\mathrm{m,i}} + \frac{1}{2}}{n_{\mathrm{m}}}$

[1]Where n_{m} is the number of observations and $x_{\mathrm{m},i}$ is the observed response for the i^{th} observation.
[2]Where n_{m} is the number of non-censored observations and $x_{\mathrm{m},i}$ is the observed response for the i^{th} observation.
[3]Where n_{m} is the number of observations and k_{m} is the number of successes.
[4]Where n_{m} is the number of time intervals and $k_{\mathrm{m},i}$ is the number of events observed in the i^{th} time interval.

Appendix E
Scope Coverage Methodology

Since many DOD systems and models have both continuous and categorical factors which affect system behavior, coverage must be computed differently depending on the types of factors present. This appendix walks through how coverage is quantified in three cases, separated by factor types present: (1) continuous factors only, (2) categorical factors only, and (3) both continuous and categorical factors. All three cases produce a coverage metric C between 0 and 1, where 0 indicates the given scope domain is not covered at all and 1 indicates full coverage of the scope domain.

Continuous Factors Only
In the first case, where only continuous factors are present in the scope domain, coverage can be decomposed into two dimensions of volume coverage and density coverage. One reason they are broken down this way is because continuous factors allow for interpolation, which is not the case with categorical factors. The first dimension, volume coverage, quantifies the volume of the domain where the response could be predicted through interpolation. Interpolation is critical to quantify since extrapolation carries much more risk. The second, density coverage, quantifies the density of validation points across the scope of interest, which ensures the whole domain can be validated, with small amounts of interpolation and extrapolation. The distinction between the two metrics is conceptually shown in Figure E1.

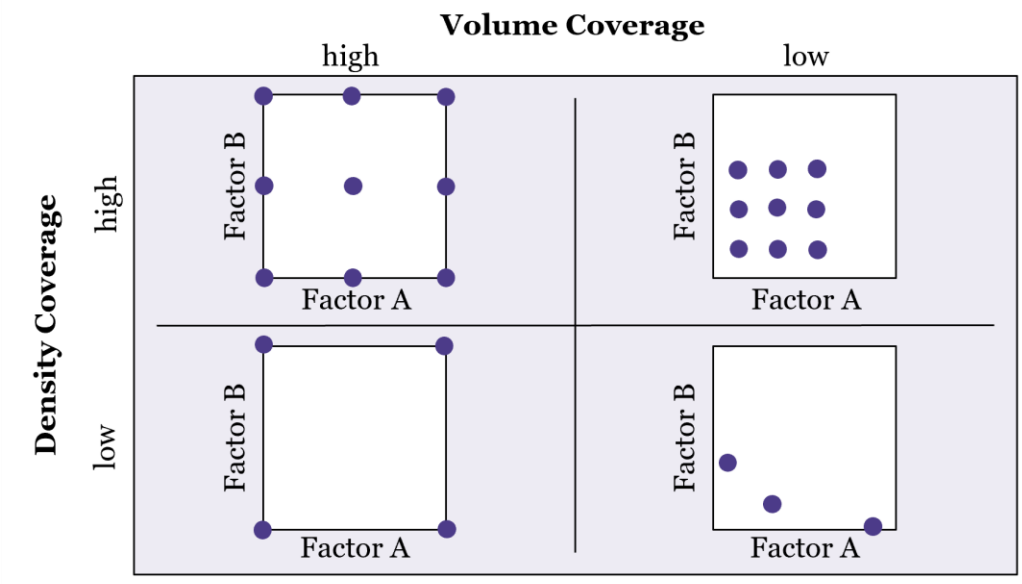

Figure E1
Volume Coverage versus Density Coverage for Continuous Factors

The coverage metric for continuous factors is mathematically comprised of a volume coverage component, C_{V}, and a density coverage component, C_{D}, as in Equation E1, where both components are metrics rating coverage between 0 and 1.

$$C = C_V C_D \tag{E1}$$

Volume Coverage

The volume coverage metric is derived from the interpolation volume between validation points. The interpolation volume is equivalent to the volume of the convex hull generated by the validation points, where the convex hull is the smallest possible convex space that contains a set of data points. In other words, the convex hull defines the region of interpolation between validation points. The volume metric is generally defined in Equation E2.

$$C_{\mathrm{V}} = \left(\frac{\text{Covered Volume}}{\text{Scope Volume}}\right)^{1/\text{dimensions}} \tag{E2}$$

Provost et al. (2022) describe the mathematical details and construction of this metric. Note that to obtain nonzero coverage with d continuous factors, at least $d + 1$ unique validation points must be gathered. For example, in two dimensions, 3 points are needed to make a shape with nonzero area.

Validation points can either be inside or outside of the scope. A validation point is inside if it falls within all the factor bounds and constraints defined in the intended use; otherwise, it is outside. Outside points can increase the interpolation volume, however only the interpolation region within the scope contributes to the coverage metric, as seem in Figure E2. The covered volume is therefore the volume of the overlapping region of the interpolation convex hull, which was generated from all validation points, and the scope convex hull, which is typically a hypercube defined by the factor ranges.

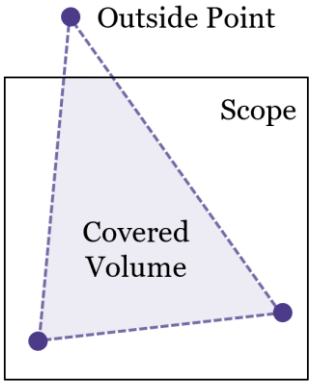

Figure E2

Covered Volume resulting from Inside and Outside Validation Points

Figure E2 shows a simple two-dimensional example, however the concept is extended to higher dimensions. When the number of continuous factors, d, is greater than 5, the metric described above is no longer computationally feasible, due to the increasing time complexity of computing the convex hull. Thus, the MVL framework uses an approximation of the above metric when $d > 5$, which calculates the convex hull in 5-dimensional projections of the validation points. The complete volume coverage metric is given in Equation E3, where d is the number of continuous factors, $p = (d \text{ choose } 5)$ is the number of 5-factor combinations, and $V_{\text{covered},i}$ and $V_{\text{scope},i}$ are the 5-dimensional covered volume and scope volume, respectively, in the i^{th} unique 5-factor projection. Each volume is calculated with a convex hull around a random sample of validation points, where the random sample size is the number of validation points, n, scaled

down to be approximately proportional to the reduction in dimensions: $n_{\text{sample}} = \text{ceiling}(n * 5/d)$. Note all validation points (inside and outside the scope) are projected down, where the volume covered is again determined by the overlapping region of the validation point convex hull and the scope convex hull.

$$C_V = \begin{cases} \frac{1}{p}\sum_{i=1}^{p}\left(\frac{V_{\text{covered},i}}{V_{\text{scope},i}}\right)^{1/5} & \text{for } d > 5 \\ \left(\frac{V_{\text{covered}}}{V_{\text{scope}}}\right)^{1/d} & \text{for } d \leq 5 \end{cases} \quad \text{(E3)}$$

Density Coverage

The second dimension of coverage for continuous factors is density coverage. Provost et al. (2021) also describe the construction of this metric. The scope domain and validation points must first be rescaled for each factor to be between 0 to 1, with 0 representing the minimum of the factor range and 1 representing the maximum of the factor range. This rescaling standardizes the density metric between different systems. The density metric uses nearest neighbor methods to determine coverage by generating a Latin Hyper Cube (LHC), which has high coverage by design, and mapping each LHC point to the nearest validation point within the scope. High distance between an LHC point and the nearest validation point indicates poor coverage, while low distance indicates good coverage. To provide an intuitive scaling, LHC points with distances greater than $M = \sqrt{d}/2$ receive a score of 0, and points with distances less than $L = \sqrt{d}/6$ receive a score of 1. Coverage is scored linearly with distance between these two bounds. This process is illustrated in Figure E3.

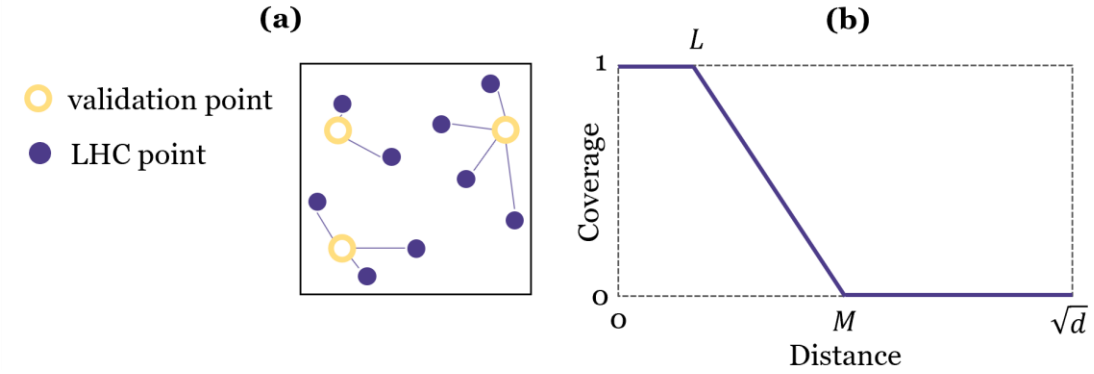

Figure E3

Illustration of finding nearest neighbors (a) and mapping distance to coverage (b)

The density coverage metric is the average of coverage scores for the q LHC points and is given in Equation E4, where r_i is the distance from the i^{th} LHC point to the nearest validation point inside the scope, and q is the total number of grid points.

$$C_{\text{D}} = \frac{1}{q}\sum_{i=1}^{q} c_i \quad \text{where} \quad c_i = \begin{cases} 1 & \text{for} \quad r_i \leq L \\ \frac{r_i - M}{L - M} & \text{for} \quad L < r_i \leq M \\ 0 & \text{for} \quad r_i > M \end{cases} \quad \text{(E4)}$$

When multiplied together, the volume coverage and density coverage metric fully quantify coverage as an overall score between 0 and 1.

Categorical Factors Only

When only categorical factors are present, the volume and density coverage metrics described for continuous factors are no longer applicable. Instead, coverage for categorical factors is based on the proportion of possible combinations which are covered by the validation points. This is demonstrated for a simple example in Figure E4, with two categorical factors, where a validation point is indicated with an 'X'.

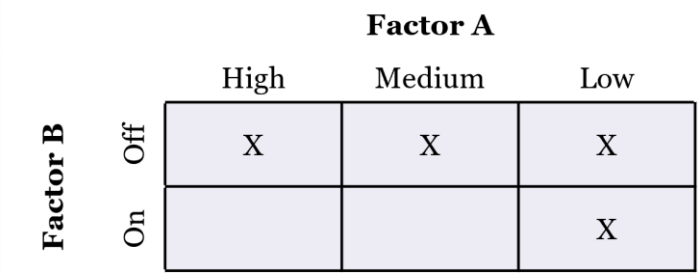

Figure E4
Example of Coverage with Two Categorical Factors

In Figure E4, four out of the six possible combinations are covered with a validation point. Thus, the coverage for this example would be $C = 4/6 = 0.667$.

In general, coverage for categorical factors is given in Equation E5, where k is the number of categorical factors, v is the number of unique validation points, and l_i is the number of levels for the i^{th} categorical factor. Note this equation could be modified to account for disallowed combinations which need not be covered.

$$C = \frac{v}{\prod_{i=1}^{k} l_i} \tag{E5}$$

Additional Categorical Coverage Diagnostics

In addition to the coverage metric in Equation E5, which assesses coverage of all possible combinations, metrics that address coverage of lower-level combinations can also be calculated. These metrics can help identify the factors or factor combinations that contribute to a low coverage score. Specifically, C_t gives the fraction of t-way factor combinations that are covered. For example, $C_4 = 0.50$ indicates that only 50 percent of the 4-way factor combinations are covered. The MVL R tool reports C_t metrics for $t = 1 \ldots k$, where C_k is equivalent to the coverage score in Equation E5.

Both Continuous and Categorical Factors

When both continuous and categorical factors are present, a combined approach is used to calculate the coverage. As seen in Figure E5, the scope domain as well as the validation points are first segmented into the possible combinations of categorical factors. In each possible combination, the validation points which meet those conditions still vary over the remaining continuous factors.

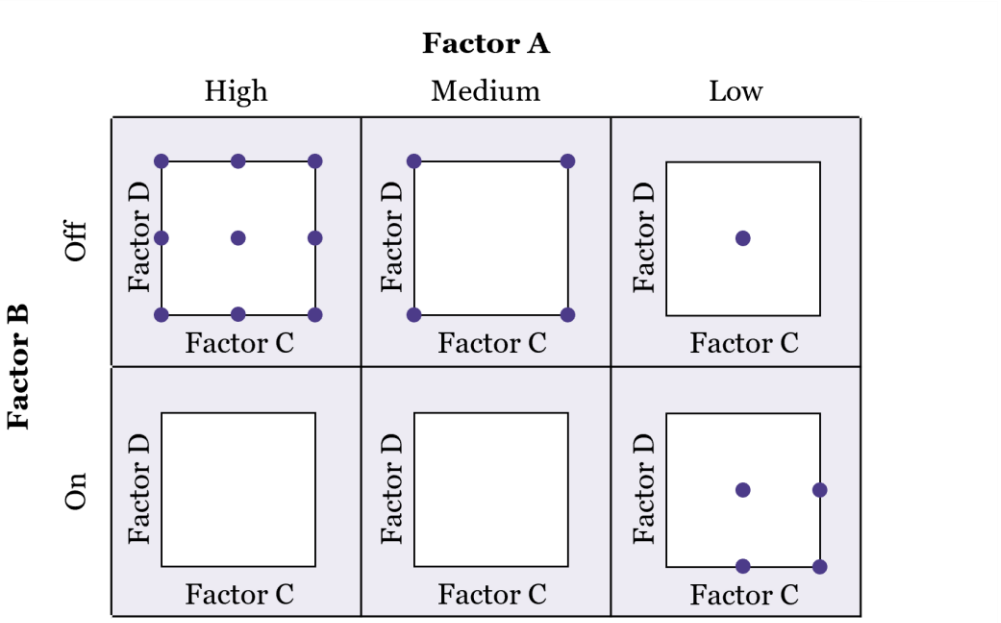

Figure E5
Example of Coverage with Categorical and Continuous Factors

The MVL framework uses the previously described continuous metrics ($C = C_{\mathrm{V}}C_{\mathrm{D}}$) to determine the coverage in each categorical combination; the coverage metrics in each combination are then averaged together to give an overall coverage metric. The equation for the case of both continuous and categorical factors is given in Equation E6, where $C_{\mathrm{V},i}$ and $C_{\mathrm{D},i}$ are the volume and density metrics for the validation points which match the categorical factors in the i^{th} categorical combination.

$$C = \frac{1}{h}\sum_{i=1}^{h} C_{\mathrm{V},i}C_{\mathrm{D},i} \quad \text{where } h = \prod_{i=1}^{k} l_i \tag{E6}$$

The methods for computing C_{V} and C_{D} in Equation E6 may still include the use of overlapping convex hulls and 5-dimensional projections when the conditions for those methods are met.

Additional Mixed Continuous and Categorical Coverage Diagnostics

In addition to the coverage metric in Equation E6, the MVL R tool reports the average volume coverage, average density coverage, and C_t scores for $t = 1 \ldots k$. The average volume and density coverage scores average $C_{\mathrm{V},i}$ and $C_{\mathrm{D},i}$ over all h combinations and can identify whether a poor coverage score is due to low volume coverage or low density. Additionally, the C_t scores can identify if coverage is poor due to failure to cover categorical combinations, and they identify at which level of factor interaction that might occur.

Appendix F
Calculating an MVL Using a Referent Interpolator

This appendix describes methods for calculating an MVL either when model and referent inputs are not matched or when referent points are not sufficiently replicated. In both cases, a referent interpolator, or statistical model of the referent, is needed in order to directly compare the model mean behavior and variability to the referent mean and variability. The interpolator serves to predict referent behavior where referent data was not collected, and it also must include an estimate of the amount of noise or variability in the response. An example of a referent interpolator for one factor is shown in Figure F1. As shown in the figure, a referent interpolator cannot be used to make predictions that extrapolate beyond the collected referent data. When multiple referents are used, multiple referent interpolators may be necessary to compare the model against multiple referents at single points in space.

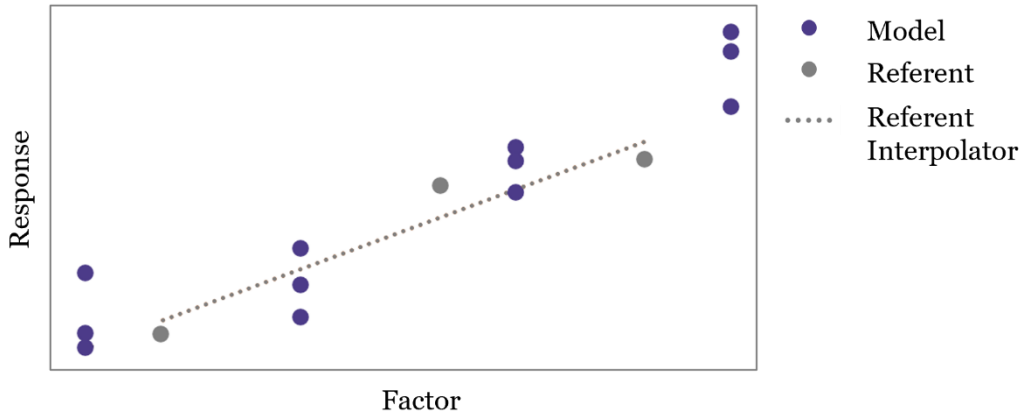

Figure F1
Example of a Referent Interpolator for One Factor

To use these methods, the user is responsible for constructing the interpolator and accepting any assumptions required to construct it. Assumptions often include the form of the model (e.g., linear, quadratic), independence, constant variance, or normality (Burke, 2017). Due to the additional assumptions required to calculate the MVL with an interpolator, it is preferred, when possible, to collect the data necessary such that the interpolator is not needed.

The remainder of this appendix discusses applicability of these methods, considerations for constructing an interpolator, and changes to the MVL calculation.

MVL from Interpolator Applicability
Using the flowchart in Figure 3, MVL users can determine if an interpolator is required to assess the MVL. Additionally, the flowchart in Figure 3 is continued in Figure F2 to determine if an MVL or MVL_a would be most appropriate.

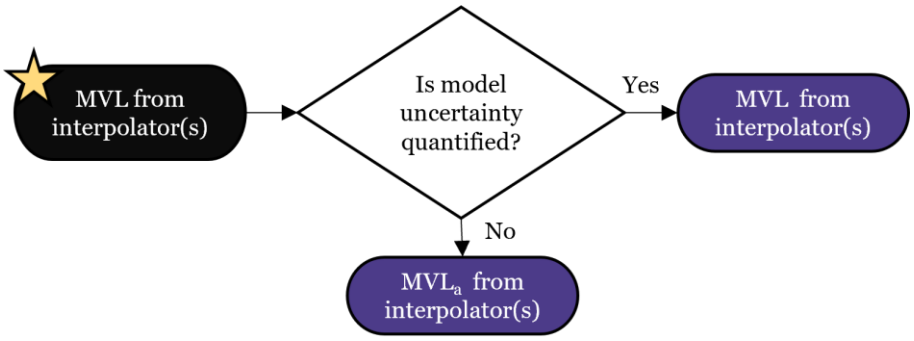

Figure F2
Flowchart to Determine Applicable MVL from Interpolator

To calculate the MVL using the MVL R tool, the user must provide:

1. Intended use: a well-defined scope of intended use describing the outputs and inputs for which the model must be validated.
2. Model data: output(s) collected from a predictive model, including the inputs or conditions under which those output(s) were obtained.
3. Referent data: data from one or more referents with assigned referent authority levels for each referent, including data that was used to construct the interpolator(s).
4. Interpolator Predictions: The mean predicted by the interpolator at model comparison points and the variance predicted by the interpolator.

Note that the MVL R tool does not require the interpolator code or equation, only the predictions produced by that interpolator. For example, in Figure F1, the user would provide the scope of intended use, the 12 model data points, the three referent data points, the referent interpolator prediction at the two comparison points in the interpolation region, and the variance predicted by the interpolator. The MVL R tool also checks for interpolation/extrapolation to comparison points, which becomes important for high numbers of factors where it is challenging to manually define the interpolation region.

When possible, additional data should be collected such that the interpolator is not needed. For example, in Figure F1, additional referent data could be collected to replicate the referent data already available, and additional model runs could be conducted at the same inputs as where there is referent data. This additional data would allow both model and referent means and variabilities to be compared without any interpolation or extra assumptions. With early VV&A planning, tests can be coordinated for collection of matched model and referent points. However, when validation is limited by data availability or budget, interpolator methods still allow an MVL to be calculated. In these cases, all assumptions made in constructing the interpolator should be thoroughly documented and reported with the MVL.

Constructing a Referent Interpolator

A referent interpolator can be constructed through a variety of methods, so long as the interpolator predicts both mean behavior and behavior variability. The recommend method is to use regression techniques to build a statistical model of the response based on factor levels. The statistical model can be used to predict mean behavior at untested factor combinations, and the mean square error (MSE) in a regression can be used to estimate the variance. Model building should be accompanied by rigorously evaluating regression assumptions.

Resources:

- Model Building Process Part 1: Checking Model Assumptions (Burke, 2017)
- Model Building Process Part 2: Factor Assumptions (Burke, 2018)
- Model Building Process Part 3: Model Goodness Metrics (Burke, 2020)

MVL Calculation with Interpolator

While the MVL calculation process remains largely the same when using a referent interpolator, there are a few changes required due to the different type of information given by the user. These differences fall into a few areas: determining validation points, Bayesian pooling, assessing referent authority, and calculating coverage. Additionally, these differences do not affect the reportable results summary and improvement metrics.

Determining Validation Points

In the standard MVL framework, validation points are defined as points where both model and referent data are available to validate a response, and they require at least one model observation and two referent observations, from the same or different referents. Interpolator methods should only be used when sufficient numbers of model and referent data points are not available together. Given this situation, validation points are redefined to be points where model data can be compared against the prediction of a referent interpolator. Intrinsic to this definition is that an interpolator cannot be used to extrapolate referent data beyond the bounds of where it was collected. For example, in Figure F1, there are two validation points, which are the points with model data inside the referent interpolation region. When a combination of categorical and continuous factors are present, a referent interpolator can still be used, but interpolation is usually limited because interpolation is not possible between different categorical levels or combinations. Within a given categorical combination, interpolation is possible for any continuous factors. The validation points determined in this way are the points where referent data is pooled, fidelity is computed, and authority is transferred to the model.

Bayesian Pooling

When using multiple referents to validate a model, using multiple referent interpolators, or using referent interpolator(s) combined with raw referent data, the Bayesian pooling (Appendix C) uses the summary statistics predicted by the interpolator, since raw data is not available in these cases. For normally distributed cases, the number of data points n_r at a predicted referent interpolator point is assumed to be at least two (one degree of freedom each for mean and standard deviation) or equal to the ratio of raw data points to interpolated points, if it is larger than two. For binomially distributed data, n_r is one or equal to the ratio of raw data points to interpolated points, whichever is larger. For exponential or Poisson data, the rate predicted at each interpolated point is scaled such that the total time observed in raw data is equally divided among interpolated points.

In the case that the interpolator is the only referent, the MVL framework uses the interpolator statistics directly for inserting into the fidelity metric, and no pooling takes place. For different distribution types, the equations for calculating statistics are the same as those for calculating model statistics in Appendix D, using the deterministic case when applicable.

Assessing Referent Authority

To construct a referent interpolator, the user must make assumptions about the behavior of a system between points where data was collected; therefore, the interpolator does not hold the same degree of authority as the raw referent data. One of the key assumptions that is required for the use of an interpolator is that the behavior for each factor varies continuously between

referent data points. The MVL framework accounts for this difference by reducing the authority based on the distance of interpolation from the raw referent data to each interpolator prediction point. Thus, each interpolator prediction point may have a different referent authority depending on the amount of interpolation required. This construct draws on the same methods used to define the density coverage score described in Appendix F and uses nearest neighbor methods to determine the distance between the interpolator prediction point and the nearest raw referent data point. If categorical factors are present, only referent points with the same combination of categorical factor levels are used when assessing amount of interpolation to the prediction point. This will result in a reduction score between zero and one, where zero means the authority is reduced to zero, while one means the interpolator retains all authority of the referent used to construct it. To derive the amount of authority reduction, the factor values must first be rescaled for each factor to be between zero and one, with zero representing the minimum of the factor range and one representing the maximum of the factor range. This rescaling standardizes the reduction amount between different systems. Additionally, following the same methods as the density metric, a distance greater than $M = \sqrt{d}/2$ receives a score of zero, and points with distances less than $L = \sqrt{d}/6$ receive a score of one, where d is the number of continuous factors. Coverage is scored linearly with distance between these two bounds. Equation F1 gives the expression for deriving the reduction factor, α_i, for prediction point i, where r_i is the distance from distance between interpolator prediction point i and the nearest referent data point.

$$\alpha_i = \begin{cases} 1 & \text{for} \quad r_i \leq L \\ \dfrac{r_i - M}{L - M} & \text{for} \quad L < r_i \leq M \\ 0 & \text{for} \quad r_i > M \end{cases} \tag{F1}$$

Once, the authority reduction score is determined, the authority inherited by the referent interpolator prediction can be calculated. The amount of authority depends on the initial authority level held by the referent. Note that each interpolator prediction point may have a different amount of authority depending on the amount of interpolation needed. The authority weight held by the ith interpolator prediction point, $w_{\mathrm{int}i}$, is given in Equation F2, where w_{r} is the authority weight of the referent (determined by Equation 1).

$$w_{\mathrm{int}i} = \alpha_i w_{\mathrm{r}} \tag{F2}$$

Calculating Coverage

The MVL framework uses validation points to assess coverage of the scope of intended use. This does not change when using interpolator methods; however, because validation points are defined differently when using interpolator methods, the coverage score is affected. For example, in Figure F1, recall there are two validation points where there is model data inside the referent interpolation region. These two validation points are used to assess coverage, so even though the model and referent alone each cover more of the domain, the points that are used to validate do not cover as much. This difference between validation point coverage and model or referent coverage could result in a coverage score lower than the user might expect.