Are Training Models and Simulations Credible?

A Straightforward Method for Answering that Question
Issues, Examples, and Lessons Learned

David H. Hall
SURVICE Engineering Company
Carlsbad, CA
Dave.Hall@survice.com

David J. Turner
SURVICE Engineering Company
Lexington Park, MD
Dave.Turner@survice.com

ABSTRACT

The Naval Air Warfare Center Aircraft Division (NAWCAD) Verification, Validation, and Accreditation (VV&A) Office has developed and is executing an innovative cost-effective, risk based VV&A process for models and simulations (M&S) used to support the Department of Defense (DoD). It is a systematic and straightforward way to determine whether a given M&S has the credibility to meet its intended uses regardless of the availability and quality of validation data. It is cost-effective because it focuses V&V activities on reducing the risk of applying the M&S to the specific intended use: only conduct V&V activities that are required to support the intended use. This process has proved to be suitable for application to any M&S, including Live, Virtual and Constructive Simulations, in support of system acquisition, testing, and training. This paper will present issues, examples, and lessons learned from demonstrating M&S credibility for a variety of programs, focusing on validation issues related to comparing M&S results with test data and with subject matter expertise.

ABOUT THE AUTHORS

David H. Hall has had a 55+ year career conducting modeling, simulation, and analysis for all the military services, but principally for the U.S. Navy. He is the Chief Analyst of the SURVICE Engineering Company for VV&A, providing VV&A support to numerous Navy programs. He retired after 34 years of Federal Civil Service, where he was the Chief Analyst of the Naval Air Warfare Center Survivability Division, Branch Head of the Survivability Analysis Branches, Director of the Joint Accreditation Support Activity (supporting VV&A of M&S for the Joint Services) and Chairman of the Survivability Methodology Subgroup for the Joint Technical Coordinating Group on Aircraft Survivability (JTCG/AS). He was Joint Program Manager of the OSD-funded Susceptibility Model Assessment with Range Test (SMART) program which developed the foundations of the VV&A process described in this paper. Dave holds B.S. and M.A. degrees in mathematics from California State University at Long Beach.

David J. Turner has had a 30+ year career primarily conducting M&S VV&A for numerous DoD programs and also working acquisition logistics and operations research for the U.S. Navy early in his career. He is the Patuxent River Area Operation Manager for the SURVICE Engineering Company, supporting a wide variety of Navy and Joint-Service programs at the Naval Air Warfare Center Aircraft Division as a long-term member of the VV&A Office within the Digital Analytics Infrastructure and Technology Advancement Group's Warfare Effectiveness Department. Dave holds a B.S. degree in Aerospace Engineering from Pennsylvania State University.

NAVAIR Public Release 2025-0383. Distribution Statement A – "Approved for public release; distribution is unlimited."

Are Training Models and Simulations Credible?

A Straightforward Method for Answering that Question
Issues, Examples, and Lessons Learned

David H. Hall
SURVICE Engineering Company
Carlsbad, CA
Dave.Hall@survice.com

David J. Turner
SURVICE Engineering Company
Lexington Park, MD
Dave.Turner@survice.com

INTRODUCTION

The Naval Air Warfare Center Aircraft Division (NAWCAD) Verification, Validation, and Accreditation (VV&A) Office has developed and is executing an innovative cost-effective, risk based VV&A process for models and simulations (M&S) used to support the Department of Defense (DoD). It is a systematic and straightforward way to determine whether a given M&S has the credibility to meet its intended uses regardless of the availability and quality of validation data. This process has proved to be suitable for application to any M&S, including Live, Virtual and Constructive Simulations, in support of system acquisition, testing, and training. This paper describes the process, and issues, examples, and lessons learned from demonstrating M&S credibility for a variety of programs, focusing on validation issues related to comparing M&S results with test data and with subject matter expertise.

CREDIBILITY AND VV&A

Before we begin a discussion of how to assess and improve the **credibility** of models and simulations, we will need to define it. A credible M&S is characterized by several key attributes. Most people think of simulation credibility in terms of three types of accuracy: software accuracy, data accuracy and output accuracy:

- **Software accuracy** refers to the error-freeness of software design and coding.
- **Data accuracy** refers to the accuracy and appropriateness of input and embedded data.
- **Output accuracy** refers to the degree to which simulation outputs match the real world.

These three aspects of M&S accuracy are easily recognized as the objects of verification and validation activities, and they give us an understanding of simulation accuracy. The DoD follows a **Verification, Validation, and Accreditation (VV&A)** process, and we use the standard DoD definitions for M&S VV&A (DODI 5000.61, 2024):

- **M&S Verification**: The process of determining that a model or simulation implementation and its associated data accurately represent the developer's conceptual description and specifications. (Does the M&S do what you intended, and is it relatively error free? Did you build the M&S right?)

- **M&S Validation**: 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. (How well do M&S results match real world data, in the context of your needs? Did you build the right M&S?)

- **M&S Accreditation**: The official certification that a model or simulation and its associated data are acceptable for use for a specific purpose. (Does the Accreditation Authority have adequate evidence to be confident that an M&S is fit for the intended use? Did the ultimate user decide to accept it?)

It should be noted that verification and validation are ***processes*** and not end-states. Think of V&V as gradually shining light on the credibility of the M&S, rather than an "on/off" switch where the M&S is either "validated" or not. Accreditation is a *decision* that there is enough light shed on the subject to determine if you can proceed according to plan, or if instead you need to revise your plan based on what you see.

But does M&S accuracy by itself comprise a comprehensive assessment of M&S credibility?

Our experience suggests that M&S accuracy (software, data, and results) is a necessary but not a sufficient condition to characterize M&S credibility. The software may be demonstrably free of errors, the data may be demonstrably accurate and correct, and M&S outputs may correlate well with real world data/expectations. But does the M&S represent all the real-world phenomena that affect the problem you are trying to solve? Does it simulate these functions at the appropriate level of detail? Is the M&S (software, documentation, and data) managed well? Is it accompanied by enough support infrastructure to allow you to use it confidently and correctly, given the user's level of experience with the simulation?

After discussing what "M&S Credibility" means with various agencies, scientists, engineers and mathematicians who use the results of M&S to support their decision making, we have narrowed down to three basic criteria that influence the credibility of a simulation and whether it is acceptable for its intended uses:

1. **Capability**: A credible M&S has the functionality and fidelity to effectively represent and simulate the complexity and dynamics of the modeled system or phenomenon in accordance with the intended use.
2. **Accuracy**: A credible M&S ensures a high level of accuracy in its representation of the real-world system.
3. **Usability**: Usability is not simply "ease of use" but rather addresses whether sufficient resources are available to ensure that the M&S is not misused, including current documentation, user support, and effective M&S management.

These three criteria - capability, accuracy, and usability - are crucial for ensuring that the M&S are credible. So, our definition of ***credibility*** is:

> **"*A credible M&S has sufficient capability, accuracy (software, data, and output), and usability to support the intended use.*"**

OVERVIEW OF THE NAVAIR M&S CREDIBILITY PROCESS

Suppose that you are responsible for demonstrating the credibility of M&S used in a (hypothetical) simulator that will support operator training for a new surface-to-air missile system. You have been given a budget for the effort, but contractors are telling you that you'll need twice that amount to do the job right. How can you balance your budget and still have enough information to justify accrediting the M&S that you need to use? (Note that while this example is hypothetical, the figures below are based on real-world test and benchmarking results, recast as a trainer example).

The NAWCAD VV&A Office has developed a risk-based approach to demonstrating M&S credibility (Elele, Hall, Turner, 2022) that focuses activities on reducing the risk that the M&S we have selected will lead us astray (Figure 1). In this example, we want to describe and reduce the risk of negative training using the simulator. The answer to the question, "How much work do I need do to make sure we're not training operators incorrectly?" can be answered by determining how much risk is involved if the M&S gives you wrong answers. If someone might die if you make the wrong decision, then you need a lot of credibility evidence to support using the M&S tool. If the problem is likely to be self-correcting as training progresses, then you might not need as much supporting evidence. It all depends on the risk involved.

The following steps define the process to determine whether the M&S can meet its intended uses:

- Develop a Specific Intended Use Statement (SIUS).
- Determine M&S requirements (capability, accuracy, and usability) from the SIUS.
- Develop Acceptability Criteria with Metrics and Measures from those requirements.
- Assess the M&S against those metrics and measures by assessing the risk to the intended use, based on the credibility information available.
- Conduct a gap analysis to determine what additional credibility information is needed and develop V&V plans to obtain that information.
- Execute the plans to generate the necessary credibility information and develop V&V reports and any other reports necessary to document the results (CM plan, data dictionary, etc.).
- Conduct a final risk assessment based on the new information.
- Make a decision for M&S use based on that final risk assessment and document in an Accreditation Report.

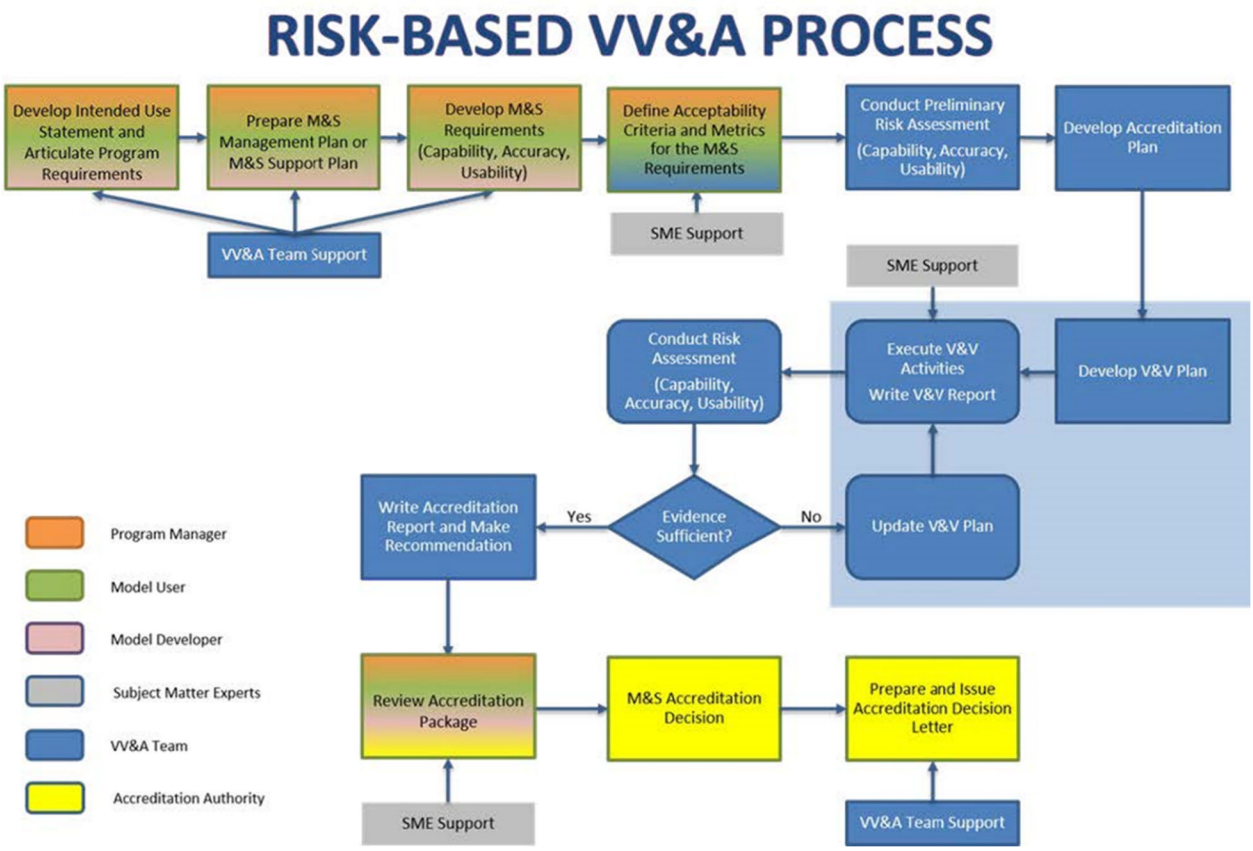

Figure 1 Risk-Based VV&A Process

This process is applicable to Live, Virtual and Constructive M&S as well as test facilities (Smith, Elele, Hall, Pedriani, 2011). It also will help you to determine what capability enhancements are required for your intended uses, and which are merely "nice to have" new features that someone else can pay for. The main point of this risk assessment process is to identify what tasks must be done to bring the risk down to an acceptable level.

The purpose of the intended use statement is to lay out your requirements for use of the M&S concisely and completely. A Specific Intended Use Statement (SIUS) must be developed in enough detail so that your requirements can be articulated; for example, do not simply say "I need a missile flyout model." You will need to list what you will be doing with that flyout model in sufficient detail to sort out whether the M&S meets your needs. Instead, say "I need the M&S to generate missile trajectories, under these conditions, with this level of detail, etc." "Carefully define the *specific* issues and the measures of performance that will be used for evaluation. Models are not universally valid, but are designed for specific purposes… A great model for the wrong problem will never be used…" (Law, 2015). For example, we develop **Table 1** to describe the SIUS in detail,

Intended Use	Intended Use Subcategory (Questions being Addressed)	Application	M&S Outputs Used
Support training for missile system operators	Can the simulator support the following learning objectives? 1) 2)	Train personnel to operate the system effectively under the following conditions: 1) 2)	• Target detection point(s) for a given target trajectory • Target tracking ranges • Valid and invalid launch condition indicators • Launch envelopes • Launch success metrics (hit/miss, miss distance…) • Etc.

Table 1 Partial SIUS for Missile System Trainer

Whether an M&S is acceptable for an intended use is determined by how well it meets the requirements of that intended use. How well it meets those requirements must be determined by an assessment method, with criteria identified as to how the user will decide if it passes or fails. The acceptability criteria and metrics can be subjective or objective, depending on the assessment method. Some basic questions that underscore those steps are:

- **M&S Requirements**: what features and characteristics do M&S need to support your intended use?
- **Acceptability Criteria**: what quantitative or qualitative properties must M&S have to meet your requirements for the intended uses?
- **Metrics/Measures**: how will you tell if the acceptability criteria have been met?

The best way to define M&S requirements, criteria and metrics is in terms of capability, accuracy and usability features needed for the intended uses. For our missile system trainer, for example, we would develop a matrix such as partially shown in **Table 2**.

M&S Requirements	Acceptability Criteria	Metrics/Measures
General Capability		
The M&S shall support training in operation of the subject missile system	• M&S requirements documents and design are adequate for its intended use. • Input data required for the M&S Tool execution are available and the degree of their validity can be established. • The output parameters are appropriate for the intended use.	• Documentation of requirements and design are available and complete • Input data requirements are documented, understood • SME review determines that: design is adequate, input data are credible, and output parameters are of sufficient accuracy and fidelity to support decision-making needs.
Specific Capability/Accuracy		

M&S Requirements	Acceptability Criteria	Metrics/Measures
The M&S shall have the functional, fidelity and accuracy characteristics required to support the learning objectives, including: 1) Target acquisition system model (detailed requirements added) 2) Target tracking system model (detailed requirements added) 3) Missile flyout model (detailed requirements added) 4) Etc.	The M&S has the following capability and accuracy acceptability criteria: 1) The target acquisition model is well documented; has documented verification results; fidelity requirements have been met; comparisons between model results and available real world detection representations meet intended use requirements, etc. … 2) The target tracking model has… 3) The missile flyout model has… 4) Etc.	For all M&S requirements, SME review and approval of functionality and fidelity, assessment that the accuracy of M&S outputs meets the acceptability criteria: this will involve sensitivity analyses, benchmarking against another M&S with known credibility, face validation against SME expectations, and/or comparison with real-world test data.
General Accuracy		
Software Accuracy		
Software shall be tested adequately to demonstrate its proper operation against the requirements identified.	Appropriate software test results and verification activities have been successfully conducted; the software development environment has been well structured and documented. Appropriate Capability Maturity Model Integration (CMMI) process (or equivalent) products are developed.	SME review and acceptance of software test and verification results; SME review and acceptance of artifacts of the software development process.
Data Accuracy		
Input and embedded data shall be adequate and appropriate for the application and documented.	Input parameters are appropriate for the intended use. Data sources are documented, appropriate and authoritative. Data are sufficiently current for planned uses. Data transformations are verified, and input data are correctly formatted.	SME review and acceptance of input data and requirements, documentation, and data transformation verification.
Output Accuracy		
The outputs of the M&S shall be of sufficient fidelity and accuracy to support potential user requirements.	The dynamic behaviors are appropriate for the intended uses. Results are of appropriate fidelity for the intended use. Results compare favorably to other M&S (Benchmarking), SME expectations (Face Validation), and/or available empirical data (e.g., test data, operational data, historical data, performance specification data) (Results Validation).	Review and acceptance by SME; output parameters important to the intended use will be determined by SME review.
Processes and documentation are in place to ensure proper maintenance of the M&S.	Configuration Management (CM) processes are sufficient and adequately documented and being followed.	SME and User review and acceptance of: CM plans and artifacts
Usability		

M&S Requirements	Acceptability Criteria	Metrics/Measures
Processes and documentation shall be in place to ensure proper operation and appropriate interpretation and use of outputs.	Users are appropriately skilled and have the necessary training. User manuals and training are adequate to enable the user to properly execute the simulation. Analyst manuals and training are adequate to enable the user or analyst to properly understand and exploit outputs.	SME and User review and acceptance of user training, experience, and credentials; user and analyst manuals; and training materials.

Table 2 M&S Requirements, Acceptability Criteria and Metrics/Measures

We use a formal risk assessment process to evaluate the residual risk based on the available credibility information (including V&V results). We have developed a set of ten characteristics of M&S under the three categories of M&S credibility (Capability, Accuracy and Usability), and we evaluate the risk the M&S poses for the intended uses by evaluating the available credibility evidence in each of those ten areas. **Table 3** shows the ten characteristics and the criteria for a fully acceptable rating in each.

For the M&S intended to support training operators of a new surface-to-air missile system, of particular interest is whether the missile M&S is accurate enough to avoid negative training. In this case, negative training could result from the operator mistakenly believing based on his or her training that the system is more capable than it actually is, or conversely that it is not as capable as it actually is. In the first case, the operator might engage targets in real life outside of the true missile envelope; in the second case, the operator might not engage critical targets inside the true missile envelope. In either case, that is a risk area in using the missile M&S as part of training, causing a loss of warfighting capability.

Characteristic	Criterion
Capability	
Intended Use and Acceptability Criteria	The General and Specific Intended Use(s) of the M&S is/are clearly stated; the Acceptability Criteria and their Metrics are clearly articulated (Acceptability Criteria and measures articulate how the Requirements for The Intended Use(s) will be met)
Conceptual Model Validation	The conceptual model (framework, algorithms, data sources, and assumptions) is documented and correctly and adequately describes the needs and requirements of the intended use.
Model Fidelity (Function and Entity Level Decompositions)	The model's Functions, Entities, Interfaces, Data (framework, algorithms, data values, and assumptions) and environmental representation levels are documented and appropriate for the intended use.
Accuracy	
Design and Implementation Verification and Validation	The algorithms and/or mathematical formulations are correct and valid. The premises for the application of the algorithms and/or mathematical formulations are correct with no assumptions violated. Logical software implementation is correct and relatively error-free.
Input and Embedded Data	The simulation input and embedded data are credible, and subject to review and revision.
System Verification	The M&S architecture has been formally tested and/or reviewed and has been demonstrated to accurately represent the system simulated for which the specific intended use statement (s), requirements and acceptability criteria were articulated.
Output Validation	The M&S responses have been compared with known or expected behavior from the subject it represents and has been demonstrated to be sufficiently accurate for the specific intended use(s).
Usability	

Documentation	The M&S is well documented as to its capabilities, design and implementation, limitations and assumptions, and the documentation is readily available, up-to-date, and complete.
User Community	The M&S is designed and developed for the level of competency of the intended users. The users have access to documents such as user's manual, training manuals, and/or reference guides. User support is available from the M&S developer or proponent.

Table 3 Risk Level Categories and Criteria

We use the results of the preliminary risk assessment (which consists of an initial "gap assessment" of the available credibility information on the M&S) to focus any efforts to improve capability on the gaps identified and the risks to the intended uses that are associated with those gaps. This process leads to identifying requirements for additional information to be collected or generated to reduce those risks that were identified. These information requirements are then compared with any additional available information, and a list of credibility 'shortfalls' is compiled. Each element of this list is then evaluated for its impact on application risk. Unmet requirements for M&S credibility that have acceptable (i.e., low risk) workarounds are removed from the list. Unmet requirements for M&S credibility that have no acceptable workarounds generate a requirement for more detailed information in the appropriate category. This may include adding new verification and/or validation results, additional software documentation, the creation (and implementation) of a C/M Plan, establishing new user support functions, or enhancements to the functionality of the M&S to meet the requirements of the intended use. Activities required to generate this information are included in an Accreditation Plan and a V&V Plan.

We iterate the VV&A process as necessary by updating the risk assessment as tasks are completed in the V&V and Accreditation Plans; a "Final Risk Assessment" is developed and documented using the ten M&S Characteristics. If the final M&S characteristics risk assessment determines that the risk is low, then full accreditation is recommended as part of an accreditation report. If the risk is still moderate or high, then the accreditation recommendation might be accredited with limitations, accredited with modifications, or not accredited. The accreditation authority then will decide whether to use the M&S with the limitations or modifications and accept the risk. Conversely, the accreditation authority could decide to address the residual risk items with resources to conduct additional V&V or development work.

ISSUES WITH M&S VALIDATION

Of all the M&S credibility elements investigated in the tables above (capability, software accuracy, data accuracy, output accuracy, and usability), everyone really wants M&S **Results Validation**, meaning comparisons between M&S output and real-life data. But,

- It is not what everyone can afford,
- Usually, you can only get test data for a very limited set of conditions,
- Sometimes the test data have problems,
- Sometimes the test does not go as planned, or
- Sometimes you cannot get access to the data you need, either because the program conducting the test will not release their data, or it is at a higher level of classification than you have access to, or because you did not plan well enough to collect the right data.

Nevertheless, comparison to test data is what everyone thinks they want.

Even though there may be scant data to compare with the output of the M&S as a whole, there may be ample test data available to compare with the output of objects/modules within the overall M&S. Suppose you compare your M&S output with some test data, and it does not match very well: is the M&S totally useless? How do you determine what is wrong? To answer those questions under real-world cost and schedule constraints, you will need to divide the model into parts that can be analyzed, understood, and compared directly to test data, while at the same time keeping an eye on the overall M&S behavior in situations where the entire system can be tested.

A problem with only comparing end-to-end M&S outputs to test results lies in its failure to address contributing factors or functions/objects. **Figure 2** illustrates what can happen when initial target and launch data are used to drive the simulation of a missile flyout to a target. The measured target and missile ground tracks from the test are shown by short and long solid lines (respectively) and the modeled one by a dashed line. So, looking at the figure, is the flyout simulation representing the data?

Even though the tested and the modeled missiles flew a very different trajectory, they both guided to the target and missed. The model-predicted miss distance was half the miss distance calculated from the test data. This type of comparison only raises questions about the underlying causes of the obvious trajectory differences. Given that our model-level comparison is not convincing, we need to diagnose the underlying sources of error. We focused our attention on contributing objects/modules within the M&S that can be compared with other test data, as well as investigating possible issues with the test data.

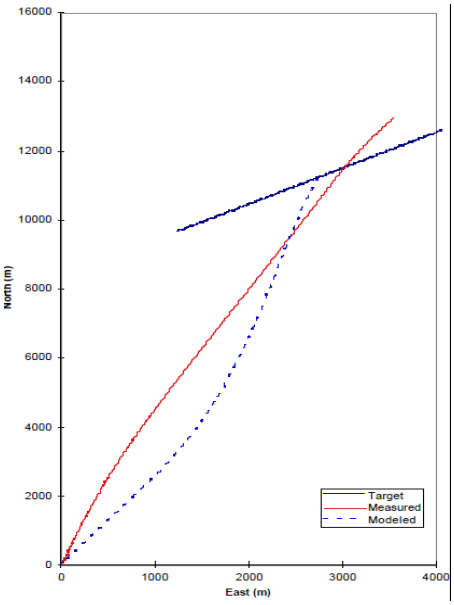

Figure 2 Typical End-to-End Missile Model-to-Test Comparison

In this example, autopilot errors contribute to lateral acceleration errors as shown in **Figure 3**, which also illustrates a typical test measurement problem between 2 and 5 seconds. During the initial phase of flight, on-board accelerometers recorded an unbelievable, noisy response (perhaps due to vibration) that settled out after 5 seconds. The M&S prediction is more believable than the test data in this region.

To see what the effect of the difference shown in **Figure 3** has on the overall missile flight prediction, we injected the guidance commands from the actual measured data into the M&S: **Figure 4** illustrates the resulting M&S-predicted altitude profile. The M&S flight profile exhibited slightly higher pitch through ten seconds of the flight but produced a lower altitude trajectory than was measured, which points to differences in the thrust and/or drag attributes between the modeled missile and the actual missile tested, which could be explored by sensitivity analysis. Whether this result is "close enough" depends on the requirements of your intended use and on how much round-to-round variations there are in the missiles being simulated. For our training simulator, it may be that the missile M&S only needs to ensure that the simulated flights are correctly inside or outside the launch envelope, and the M&S missile flight time is close to the actual missile flight time to intercept.

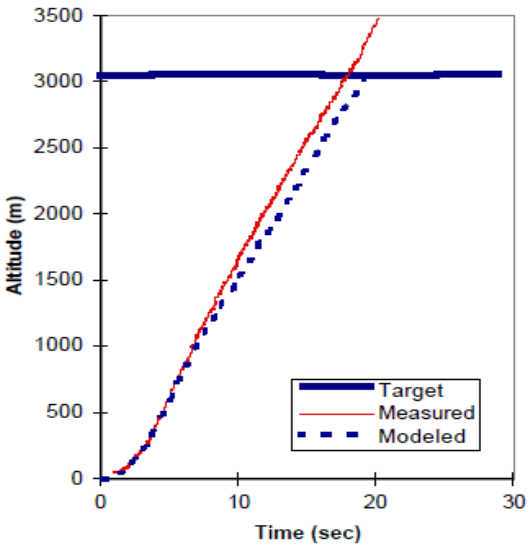

Figure 4 Results with Measured Guidance Commands Injected into the M&S

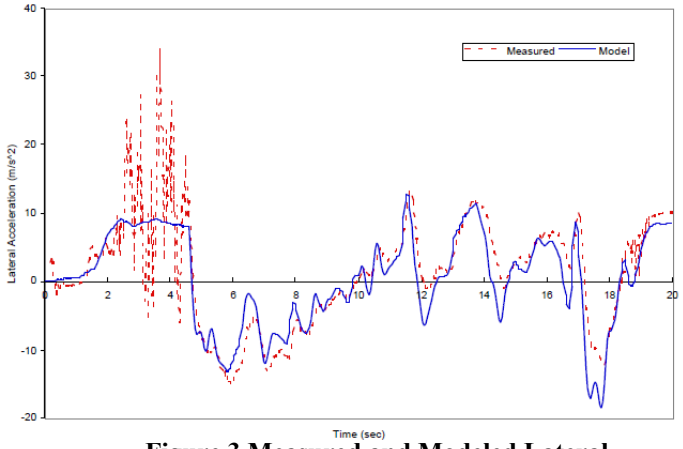

Figure 3 Measured and Modeled Lateral Accelerations

This example also illustrates the importance of having M&S proponents being involved in test data collection plans: if adequate data values are not collected during the test to support M&S validation, then it will be very difficult to compare M&S results with test data for all of the objects/modules that make up the M&S, or even for the M&S as a whole.

If we do not have much (or any) actual test data for comparison, or if the test data are hopelessly flawed, fortunately there are other comparisons to the "real world" we can use: comparisons to how reality is represented by another M&S with established credibility (**Benchmarking**), and comparisons with the real world as expected by subject matter experts (SME) based on their experience and on reviewing available (possibly limited) test data and extensive sensitivity analyses using the M&S of interest (**Face Validation**). SME are engineers, scientists or analysts with expertise in the M&S subject matter area,

Benchmarking is one way to represent a version of "reality" for comparison with M&S outputs as part of a validation effort. Benchmarking compares your M&S outputs with outputs of another M&S that is accepted as having demonstrated credibility for your intended use. The benchmark M&S can be an industry- or government-wide standard, an M&S that SMEs recognize as having sufficient accuracy for the intended use, etc. You might in fact prefer to use the benchmark M&S over yours, but the benchmark might be too costly, or take too long to run, be too complex to set up, may not cover all of the analysis domain required, or not be available to you. There are examples of these benchmarking issues in Computational Fluid Dynamics problems (Wroniszewski, Verschaeve, Pedersen, 2014) and Finite Element Modeling (Chen, Parker, Wan, Bravenec, 2013), to name two. In those cases, a simpler model is used, with simplifying assumptions, and the simpler model is benchmarked against a high-fidelity M&S that is based on first principles. Benchmarking can also include comparisons with established theoretical solutions, when those solutions are amenable to numerical solution even if for only a limited number of simplified cases.

Benchmarking can be very useful as a validation tool, but it must be done with careful planning and execution, it must be focused on the intended uses of the M&S, and it must be adequately documented.

For our missile system trainer, for example, we compared the results of our IR tracker atmospheric absorption M&S with an industry standard model, MODerate resolution atmospheric TRANsmission (MODTRAN) model versions 4 and 5. **Figure 5** shows an example of the comparison between the M&S for transmission vs. wavelength. While we obtained similar results, there were differences that needed explanation by subject matter experts, which led us to a Face Validation Review.

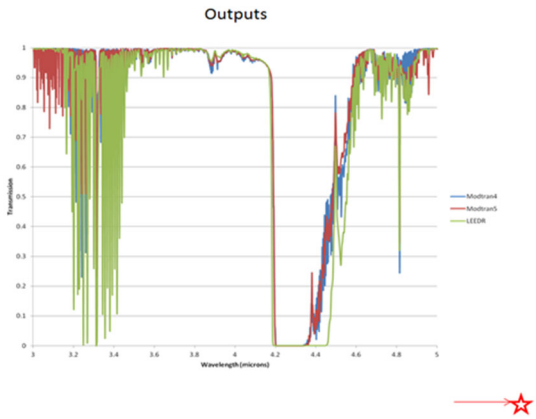

Figure 5 Atmospheric Transmission Benchmarking

Face Validation compares M&S results to SME expectations of those results. But if we ask SMEs to give us their opinion as to whether M&S results match their expectations, we need to provide them with some data for comparison, unless they have a ready source of data from other sources that they can use to bolster their view of what the "real world" should look like. For a successful face validation review, the SME will need some benchmarking data and/or whatever results validation data are available, plus a sufficient sensitivity analysis for their review to demonstrate how the simulation "behaves" under different assumptions and test conditions. The comparison should also capture SME opinion as to whether any M&S assumptions and limitations are acceptable from the perspectives of the intended uses.

For our missile trainer example, the SME focused on evaluating the requirements, acceptability criteria, metrics and measures (**Table 2**) to arrive at recommendations for use of the M&S, any improvements or corrections needed, limitations identified, etc. They based their review on the limited test data available, benchmarking against the older M&S, and their own expectations of how the missile system would behave under a variety of conditions (an extensive sensitivity analysis showing how the M&S represents the system under those conditions, and how the outputs vary with changes in the inputs). They recommended the M&S be used, but with several identified limitations on its ability to support training objectives.

Our experience has shown that SME reviews are essential to establishing the credibility of M&S for any intended use. The key to making an SME review successful is to require the SME to focus on whether the M&S satisfies the requirements and acceptability criteria established by the intended use. Without that focus, regardless of the review technique, the review can devolve into SME discussing how they would have modeled the problem differently, or how their M&S would have been preferable, etc.

Also, it is very important to carefully select the participants. The Review Team should represent a cross-section of experts in the M&S application area, the system(s) being modeled, and the technical underpinnings of the M&S (science, engineering, economic theory, etc.). Participants can be included for political reasons, but the SMEs need to cover adequately the important technical bases. Participants should have adequate time to prepare and fully participate in the entire review. We have found that if the SME review panel has fewer than 8 or 10 active participants it helps to keep lengthy and (usually immaterial) discussions to a minimum.

LESSONS LEARNED

M&S credibility efforts must be focused on demonstrating whether M&S meet the requirements, acceptability criteria, metrics and measures derived from the specific intended use: you must define in detail how you will decide whether an M&S meets your needs before you proceed with any testing or other methods of demonstrating whether the M&S is suitable for your intended use.

If you intend to use test data as part of an M&S validation effort, M&S analysts need to be involved in planning for test data collection; otherwise, there may be insufficient data collected for the analyst to replicate the test conditions and evaluate the various objects/modules within the M&S that will contribute to differences in outcomes between the M&S and the test. End-to-end M&S validation with test data is difficult to achieve without evaluating the various functions within the model against the same test conditions.

Benchmarking can be a useful tool in situations where test data cannot be collected, as long as the benchmark simulation is carefully selected, and the benchmark comparisons are evaluated by SME.

In the end, almost all validation efforts should be adjudicated by SME review (Face Validation): someone who understands the system being simulated and the environment in which it operates will have to pass judgement on any comparisons with validation data to make an informed decision about whether the M&S is suitable for the intended use. Even if you have a wealth of statistically significant comparisons with test data, you still will need SMEs to review the results with respect to the requirements of your intended use. The key to making an SME review successful is to focus the discussion on whether the M&S satisfies the M&S requirements and acceptability criteria of the intended use, using the agreed upon metrics.

SUMMARY

The NAVAIR M&S VV&A process has evolved over the last 30+ years through successful application to a wide variety of system acquisition programs, evaluating the credibility of Live, Virtual and Constructive M&S and test facilities. The process applies equally well to M&S used in training as well as earlier phases of system acquisition. Validation can be accomplished through three techniques: benchmarking, face validation, and results validation, but all three require review by SME to evaluate the results as compared to the requirements, acceptability criteria and metrics/measures derived from the specific intended use of the M&S.

REFERENCES

1. Chen, Y., Parker, S. E., Wan, W., Bravenec, R., *Benchmarking gyrokinetic simulations in a toroidal flux-tub*e, University of Colorado at Boulder, Fourth-State Research, published on-line 23 Sept 2013, Physics Of Plasmas 20, 092511 (2013) http://dx.doi.org/10.1063/1.4821982
2. Department of Defense Instruction, *DODI 5000.61 DoD Modeling and Simulation (M&S) Verification, Validation, and Accreditation (VV&A),* September 17, 2024, Glossary G.2: Definitions

3. Elele, James N. PhD, Hall, David H., Turner, David J., *Risk-Based VV&A: A Systematic Approach*, The ITEA Journal of Test and Evaluation, March 2022, Volume 42, No.1, PP. 24-33, International Test and Evaluation Association
4. Law, Averill M., *Simulation Modeling and Analysis*, 5th edition, 2015, pg. 249, McGraw Hill, ISBN 978-0-07-340132-4
5. Smith, Jeremy, Elele, James, Hall, David, Pedriani, Charles, Society of Photo-Optical Instrumentation Engineers (SPIE), *Applying risk-based M&S VV&A techniques to test and laboratory facilities*, SPIE Proceedings Volume 8060, Modeling and Simulation for Defense Systems and Applications VI; 80600K, 20 May 2011
6. Wroniszewski, Paweł A., Verschaeve, Joris C.G., Pedersen, Geir K., *Benchmarking of Navier–Stokes codes for free surface simulations by means of a solitary wave*, Department of Mathematics, University of Oslo, Norwegian Geotechnical Institute (NGI), Norway Coastal Engineering, 2014, 0378-3839/© 2014 Elsevier B.V. http://dx.doi.org/10.1016/j.coastaleng.2014.04.012