Consistent Credibility Criteria
Why have them, what are they, and how do you measure them?

Bobby Hartway
Alexia Joiner
Danny Thomas
AEgis Technologies Group
631 Discovery Drive
Huntsville, AL 35806
256-922-0802
bhartway@aegistg.com
ajoiner@aegistg.com
d.thomas@aegistg.com

Keywords:
"Credibility Assessment Scale", "NASA",

ABSTRACT: The National Aeronautics and Space Administration (NASA) has endeavored to develop a consistent method of assuring the credibility of simulation results. Requirements like "accurate enough for a particular use" present problems. How do you measure accuracy? What has to be accurate? What factors are measured? How do you quantify "enough for the particular use?" Which functions are measured? When during the simulated time are they measured? How do you state uncertainty requirements? These are but a few of the specifics that make writing defendable requirements for simulations many orders of magnitude more difficult that writing requirements for other software. NASA-STD-7009[1] presents a Credibility Assessment Scale (CAS) containing eight criteria grouped into three categories that attempts to quantify factors influencing simulation credibility. This paper identifies these criteria, provides straightforward definitions and introduces typical methods of assessment.

NASA Recognizes the Need for Consistent Measures of M&S Credibility

The genesis of NASA's recent emphasis on assuring the validity of simulations and the credibility of simulation studies is the Columbia Accident Investigation Report (CAIB)[2]. It called for NASA to "develop, validate, and maintain physics-based computer models to evaluate Thermal Protection System damage from debris impacts. These tools should provide realistic and timely estimates of any impact damage from possible debris from any source that may ultimately impact the Orbiter. Establish impact damage thresholds that trigger responsive corrective action, such as on-orbit inspection and repair, when indicated."

The Renewed Commitment to Excellence, or Diaz Report[3], broadened the scope beyond the Space Transportation System. Diaz action item four called for NASA to "develop a standard for the development, documentation, and operation of models and simulations (M&S):

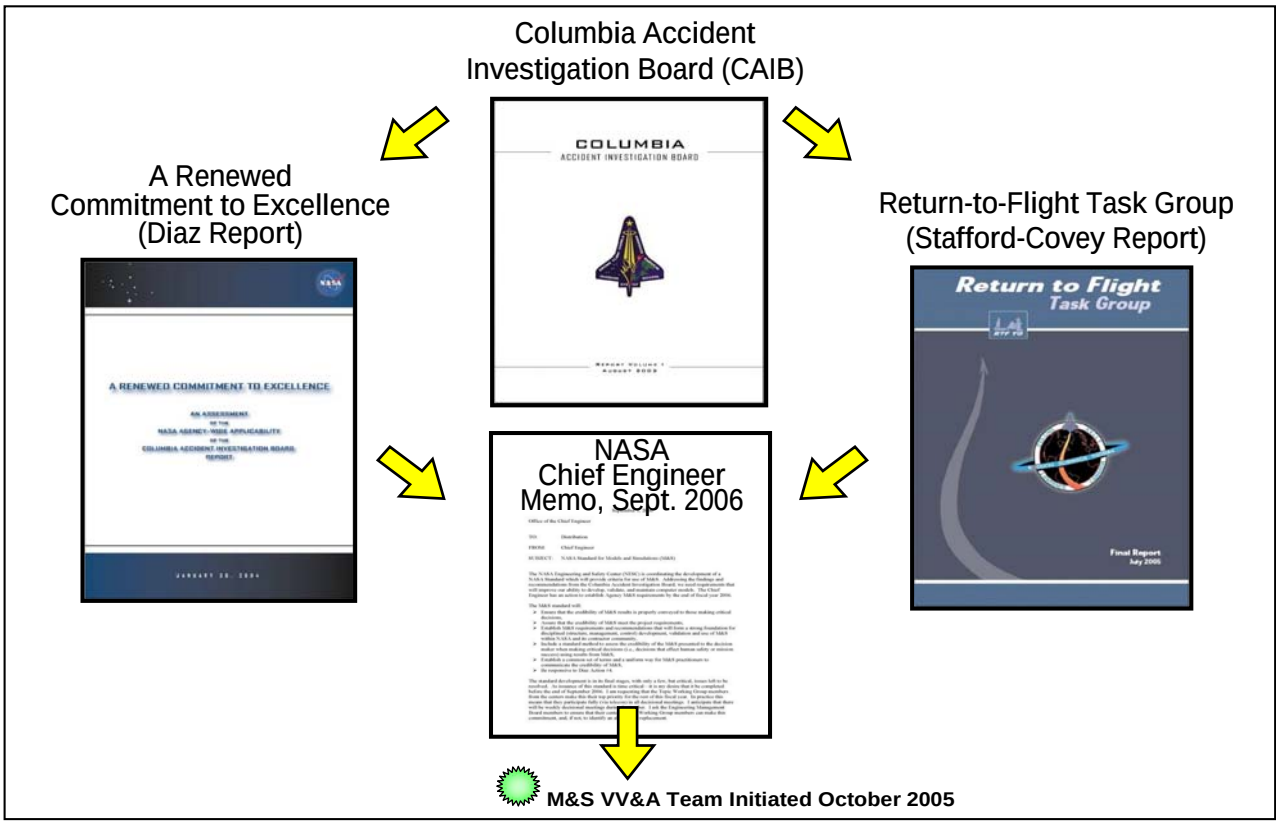
---

[1] This paper draws most of its information from NASA-STD-7009. Unless otherwise noted all references are to that standard. [1] http://standards.nasa.gov/released/NASA/NASA_STD_7009_APPROVED_2008_07_11.pdf
[2] http://caib.nasa.gov/
[3] http://www.nasa.gov/pdf/55691main_Diaz_020204.pdf

- Documentation, configuration management, and quality assurance
- Verification and validation, operational data and trending
- Tool management, maintenance, and obsolescence
- Training requirements
- Best practices for user interfaces
- User feedback when results appear unrealistic"

The Stafford-Covey Report [4] enjoined "formal development, verification and validation, and outside review plans". It added that "assumptions should be written down and consistently applied." Finally it required that "sensitivity analysis and careful analysis of uncertainty were to be performed." These commitments were to be agency wide.

The Chief Engineer Memo's required that the credibility of M&S results is properly conveyed to those making critical decisions, and that analysts should assure that the credibility of M&S meets the project requirements. NASA was to establish M&S requirements and recommendations that will form a strong foundation for disciplined (structure, management, control) development, validation and use of M&S within NASA and its contractor community, include a standard method to assess the credibility of the M&S presented to the decision maker when making critical decisions (i.e., decisions that effect human safety or mission success) using results from M&S, and establish a common set of terms and a uniform way for M&S practitioners to communicate the credibility of M&S. Figure 1 depicts these important drivers.



[4] http://www.nasa.gov/home/hqnews/2005/aug/HQ_m05138_StaffordCovey.html

Figure 1: Drivers for the NASA Modeling and Simulation Standard

The NASA Modeling and Simulation Standard

NASA's response was to develop a modeling and simulation standard that would "provide uniform engineering and technical requirements for processes, procedures, practices, and methods that have been endorsed as standard for M&S developed and used in NASA programs and projects, including requirements for selection, application, and design criteria of an item." The standard, designated NASA-STD-7009, was developed "as an indirect result of the Space Shuttle Columbia Accident". During the post-accident investigations, some findings indicated that the incorrect application of a model, simulation, and/or analysis method can lead to incorrect decision making. The resulting standard is distinct from existing software standards in that it focuses on the reporting of M&S results and the assessed credibility of those results."[5]

NASA-STD-7009 states "The main goal of the standard is to ensure that the decision maker is made aware of the key information regarding M&S results that is needed to infer their credibility. The information needed was broken down into three parts: the uncertainty of the results, the assessment on the scale, and any caveats that go along with the results." These "ensure that the decision maker has the information that is needed to determine the trustworthiness of the results." [6]

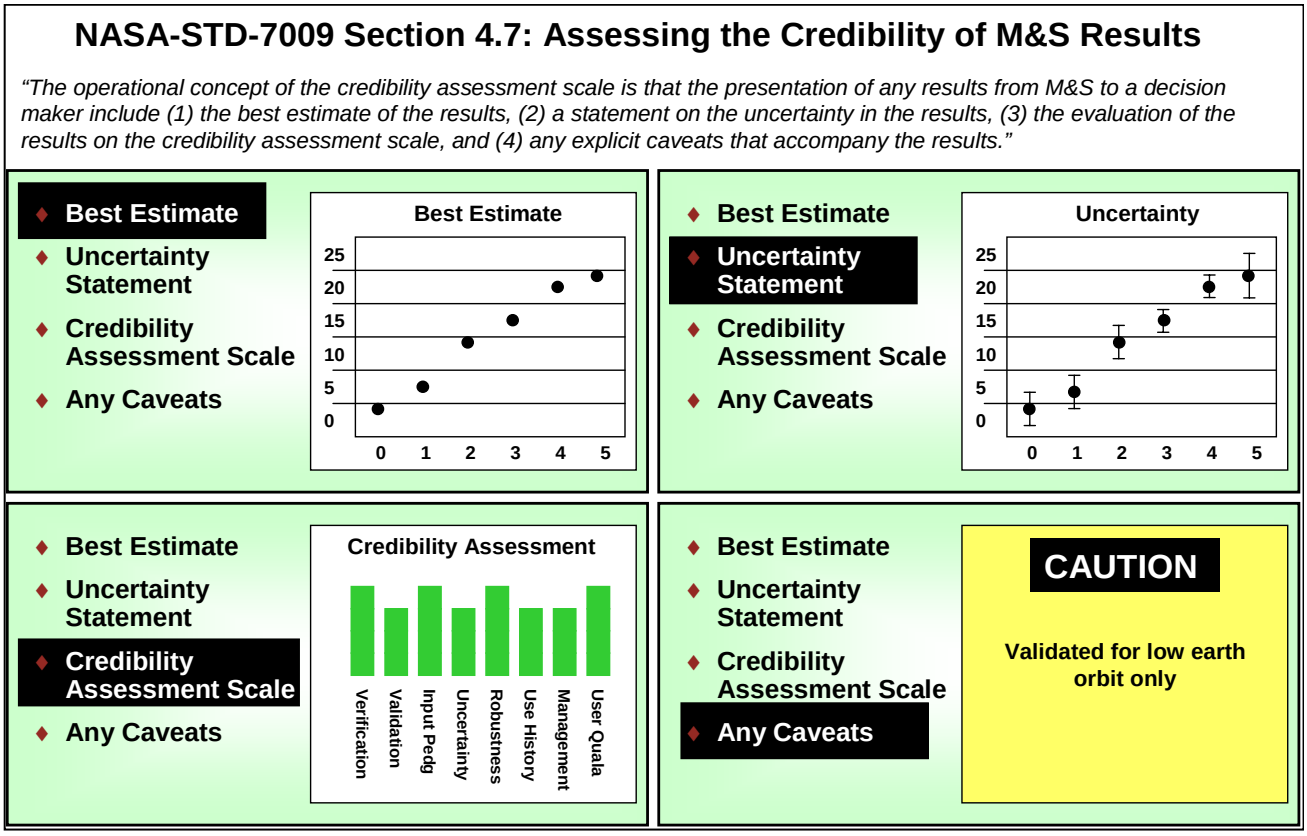


Figure 2: Four Items to be Reported to Management

[5] Martin J. Steele, http://portal.acm.org/citation.cfm?id=1357912
[6] Tom Zang, http://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20090007603_2009006269.pdf

NASA-STD-7009 also requires that the credibility assessment of the M&S results be assessed using the CAS shown in Figure 3. This CAS consists of eight factors grouped into three categories. The assessment process involves evaluating the M&S results on each of eight factors, and then rolling up these eight factor results into a single number that represents the summary credibility assessment. The M&S Development category captures those aspects of the M&S that pertain to the general assessment of the credibility of the M&S for their broad intended use; the M&S Operations addresses the aspects relevant to the current application of the M&S to generate the particular M&S results under assessment; and the Supporting Evidence category addresses three cross-cutting factors.

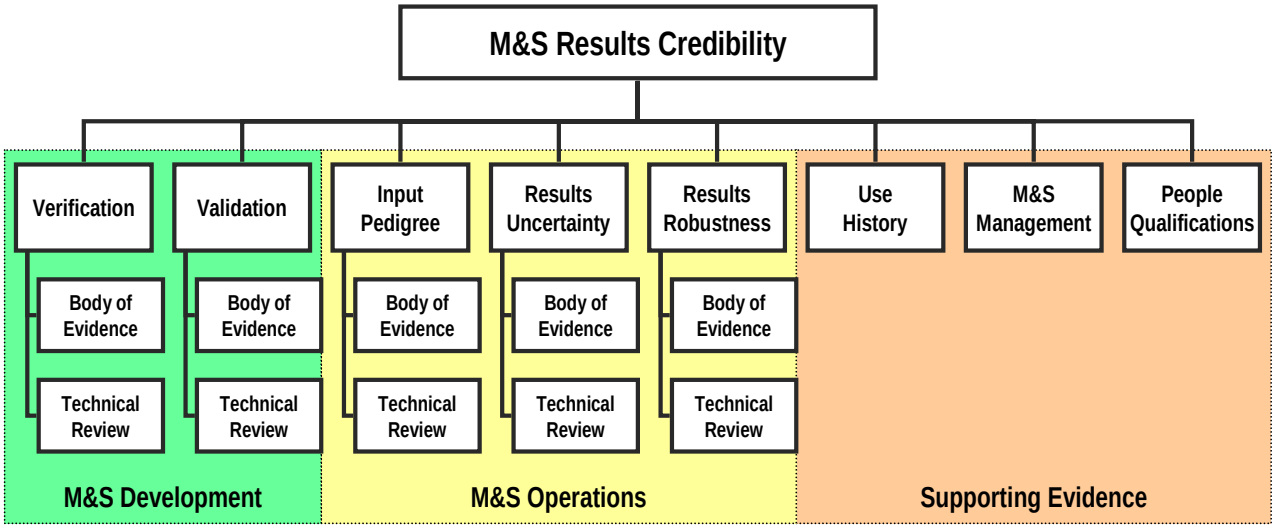


Figure 3: The Credibility Assessment Scale from NASA –STD-7009

The credibility assessment of the M&S results should be determined using the CAS. The eight factors were selected from a long list of factors that contribute to the credibility of M&S results because (a) individually they were judged to be the key factors in this list; (b) collectively they are nearly orthogonal, i.e., independent, factors; and (c) they can be assessed objectively. In short, the key aspects assessed by these eight factors are as follows:

M&S Development

1. Verification: Were the models implemented correctly, and what was the numerical error/uncertainty?
2. Validation: How well did the M&S results and the referent data compare?

M&S Operations

1. Input Pedigree: How confident are we of the current input data?
2. Results Uncertainty: What is the uncertainty in the current M&S results?
3. Results Robustness: How thoroughly are the sensitivities of the current M&S results known?

Supporting Evidence

1. Use History: Have the current M&S been used successfully before?
2. M&S Management: How well managed were the M&S processes?
3. People Qualifications: How qualified were the personnel?

The M&S Development category consists of Verification and Validation. Verification is the process of determining that a computational model accurately represents the underlying mathematical model and its solution from the perspective of the intended uses of the M&S. At its most elementary levels this involves assurance that the conceptual and mathematical models are correct. Validation is the process of determining the degree to which a model or a simulation is an accurate representation of the real world from the perspective of the intended uses of the model or the simulation.

The focus of the three factors in the M&S Operations category is an assessment of those M&S results that support the particular critical decision in question. The M&S Operations category consists of Input Pedigree, Results Uncertainty and Results Robustness. Input Pedigree involves the evaluation of all data that is used as input for the current M&S results. It includes not only data that is unique to the model, but also data that is produced by other simulations. Results Uncertainty is the quantification of the uncertainty in the current M&S results. Two important aspects of the uncertainty are (a) the size of the uncertainty, e.g., the size of the uncertainty interval; and (b) the confidence in or quality of the estimate of the uncertainty, e.g., a statistical confidence statement or the thoroughness used in the estimate. Results Robustness is the determination of how thoroughly the sensitivities of the current M&S results (to the variables and parameters of the M&S) are known. The purpose of considering robustness is to garner an understanding of the sensitivity of the real-world system to potential changes in the variables and parameters of the system.

The focus of the three factors in the Supporting Evidence Category is the assessment of three elements of the M&S process that may indirectly affect the credibility of the M&S results. The factors included are Use History, M&S Management, and People Qualifications. The Use History factor describes the extent of any prior use of the M&S in similar situations for critical decisions. The M&S Management factor assesses the level of formality applied by the program or project to the oversight of the M&S. The People Qualifications factor assesses the training and experience of the developers, operators, and analysts conducting the M&S activities. Five of the eight factors have a Technical Review subfactor, which assesses the level of peer review that has been successfully completed relevant to that factor.

Evaluating the Credibility Factors

Figure 4 gives a high level summary of the evaluation criteria – level 1 being the lowest and level 4 being the highest with a 0 reserved for situations were a scope could not be determined because of insufficient evidence, either no evidence exists for that factor, or the evidence that does exist does not meet even the level 1 criteria for that factor.

Level	Verification	Validation	Input Pedigree	Results Uncertainty	Results Robustness	Use History	M&S Management	People Qualifications
4	Numerical errors small for all important features	Results agree with real-world data	Input data agree with real-world data	Non-deterministic and numerical analysis	Sensitivity known for most parameters; key sensitivities identified	De facto standard	Continual process improvement	Extensive experience in the use of and recommended practices for this particular M&S
3	Formal numerical error estimation	Results agree with experimental data for problems of interest	Input data agree with experimental data for problems of interest	Non-deterministic analysis	Sensitivity known for many parameters	Previous predications were later validated by mission data	Predictable process	Advanced degree or extensive M&S experience, and recommended practice knowledge
2	Unit and regression testing of key features	Results agree with experimental data or other M&S on unit problems	Input data traceable to formal documentation	Deterministic analysis or expert opinion	Sensitivity known for a few parameters	Used before for critical decisions	Established process	Formal M&S training and experience, and recommended practice training
1	Conceptual and mathematical models verified	Conceptual and mathematical models agree with simple referents	Input data agree traceable to informal documentation	Qualitative estimates	Qualitative estimates	Passes simple tests	Managed process	Engineering or science degree
0	Insufficient evidence	Insufficient evidence	Insufficient evidence	Insufficient evidence	Insufficient evidence	Insufficient evidence	Insufficient evidence	Insufficient evidence
	M&S Development		M&S Operations			Supporting Evidence		

Figure 4: Values of the Credibility Assessment Scale

Each of the factors is evaluated based on the available evidence by a technical review process. The scores are determined by the minimum score for any of the factors. Figure 3 illustrates the 10 weights that are needed for the roll-up from the subfactor to the factor tier. The constraints on these weights are as follows:

a. Each weight lies in the closed interval [0,1].
b. The sum of each subfactor pair, e.g., w11 and w12, is 1.
c. The subfactor weight for Technical Review is further constrained to be no more than 0.3.

The achieved score at the lowest tier (factor or subfactor) is based on the objective assessment of the documented evidence against the level definition. In the M&S Development and M&S Operations categories the achieved factor score is the Evidence score times the Evidence weight plus the Review score times the Review weight. Constraint c, listed above, limits the amount by which Technical Review can increase or decrease the factor score with respect to the Evidence

subfactor score. In the most extreme case, with an Evidence score of 0 and Technical Review score of 4.0, the factor score is 1.2.

The choice of the weights within a category is necessarily subjective. The roll-up of the 8 factor scores into the overall score is performed by taking the minimum of the 8 factor scores. It can be argued that this scoring method is unnecessarily conservative. For example if a simulation study produced results that scored all 4 in every category except Supporting Evidence where it scored only a 1, then the credibility assessment reported would be only a 1. Two important observations must be cited. First, the single number was never intended to be reported alone. The credibility assessment scale was intended to be a communication tool between decision makers and the analysts who advise them. Scores for specific criteria can be indicators where priorities should be set. Secondly, when dealing with the critical decisions that the standard was developed to address, the lowest score may be the best indicator of risk – sort of "weakest link" thinking.