

# NASA Standard for Models and Simulations: Credibility Assessment Scale

Maria Babula<sup>1</sup>, William J. Bertch<sup>2</sup>, Lawrence L. Green<sup>3</sup>, Joseph P. Hale<sup>4</sup>, Gary E. Mosier<sup>5</sup>,  
Martin J. Steele<sup>6</sup>, Jody Woods<sup>7</sup>

## I. Introduction

As one of its many responses to the 2003 Space Shuttle Columbia accident<sup>i</sup>, NASA decided to develop a formal standard for models and simulations (M&S)<sup>ii</sup>. Work commenced in May 2005. An interim version<sup>iii</sup> was issued in late 2006. This interim version underwent considerable revision following an extensive Agency-wide review in 2007 along with some additional revisions as a result of the review by the NASA Engineering Management Board (EMB) in the first half of 2008. Issuance of the revised, permanent version, hereafter referred to as the M&S Standard or just the Standard, occurred in July 2008.

Bertch, Zang and Steele<sup>iv</sup> provided a summary review of the development process of this standard up through the start of the review by the EMB. A thorough recount of the entire development process, major issues, key decisions, and all review processes are available in Ref. v. This is the second of a pair of papers providing a summary of the final version of the Standard. Its focus is the Credibility Assessment Scale, a key feature of the Standard, including a new example of its application to a real-world M&S problem for the James Webb Space Telescope. The companion paper<sup>vi</sup> summarizes the overall philosophy of the Standard and an overview of the requirements. Verbatim quotes from the Standard are integrated into the text of this paper, and are indicated by quotation marks.

## II. Role of the Credibility Assessment Scale

Action #4 from the January 30, 2004 Diaz Team Report<sup>ii</sup> called for NASA to: “Develop a standard for the development, documentation, and operation of models and simulations.” None of the six specific objectives of this action call for the development of a Credibility Assessment Scale (CAS). Rather, this objective was levied by the NASA Chief Engineer (then Christopher Scolese), informally in a March 2006 meeting and formally in a September 2006 memo that stated that “the M&S standard will ... 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 affect human safety or mission success) using results from M&S.”

## III. Role of the CAS in the M&S Standard

The requirements for the use of the Credibility Assessment Scale fall in Sections 4.7 and 4.8 of the M&S Standard. The details of the CAS are given in its Appendix B. “The operational concept of the credibility assessment scale is that the presentation of any results from M&S to a decision maker include (1) the best estimate of the results, (2) a statement on the uncertainty in the results, (3) the evaluation of the results on the credibility assessment scale, and (4) any explicit caveats that accompany the results. (An example of such a caveat would be use of the model in violation of its assumptions.) The decision maker then makes his/her own assessment of credibility based upon all four pieces of information in the context of the decision at hand. Just to emphasize this fundamental point, the credibility assessment scale does not purport to measure credibility; rather, it assesses the M&S results, and the rigor of the processes used to produce them, against key factors that affect the credibility judgment. The

---

<sup>1</sup> NASA Glenn Research Center.

<sup>2</sup> NASA Jet Propulsion Laboratory.

<sup>3</sup> NASA Langley Research Center.

<sup>4</sup> NASA Marshall Space Flight Center

<sup>5</sup> NASA Goddard Space Flight Center.

<sup>6</sup> NASA Kennedy Space Center.

<sup>7</sup> NASA Stennis Space Center.

fundamental premise of this approach is that as a general rule, the more rigorous the key processes used for generating the M&S results, the greater the credibility of the M&S results, all else (including the estimated uncertainty) being equal.”

The particular reporting requirements related to the CAS are:

**Req. 4.7.1** – Shall assess the credibility of M&S results for each of the eight factors in the CAS described in Appendices B.2 and B.3.

**Req. 4.7.2** – Shall justify and document the credibility assessment for each of the eight factors referenced in Req. 4.7.1.

**Req. 4.7.3** – Shall perform the roll-up to an overall score according to the process described in Appendix B.4.

**Req. 4.8.3** – Reports to decision makers shall include the level of credibility for the M&S results and the subfactor weights, using the process specified in section 4.7.”

Several M&S scales had been proposed or were under development prior to the inception of this effort in July 2006, e.g., Balci, Adams, Myers and Nance<sup>vii</sup>, Harmon and Youngblood<sup>viii</sup>, Oberkampf, Pilch, and Trucano<sup>ix</sup>, and Green *et al*<sup>x</sup>. Several of those involved with the development of the CAS published papers on the two scales that appeared in the interim M&S Standard (Luckring *et al*<sup>xi</sup> and Hale and Thomas<sup>xii</sup>) and on a quite different alternative that was considered (Mehta<sup>xiii</sup>).

#### IV. Summary of the CAS

This section provides a high-level summary of the key features of the CAS.

##### Overview

“This CAS consists of eight factors grouped into three categories, as illustrated in Figure 1. The eight factors are Verification, Validation, Input Pedigree, Results Uncertainty, Results Robustness, Use History, M&S Management, and People Qualifications. The three categories are M&S Development (Verification, Validation); M&S Operations (Input Pedigree, Results Uncertainty, Results Robustness); and Supporting Evidence (Use History, M&S Management, People Qualifications). A five-level assessment of credibility is defined for each factor.”

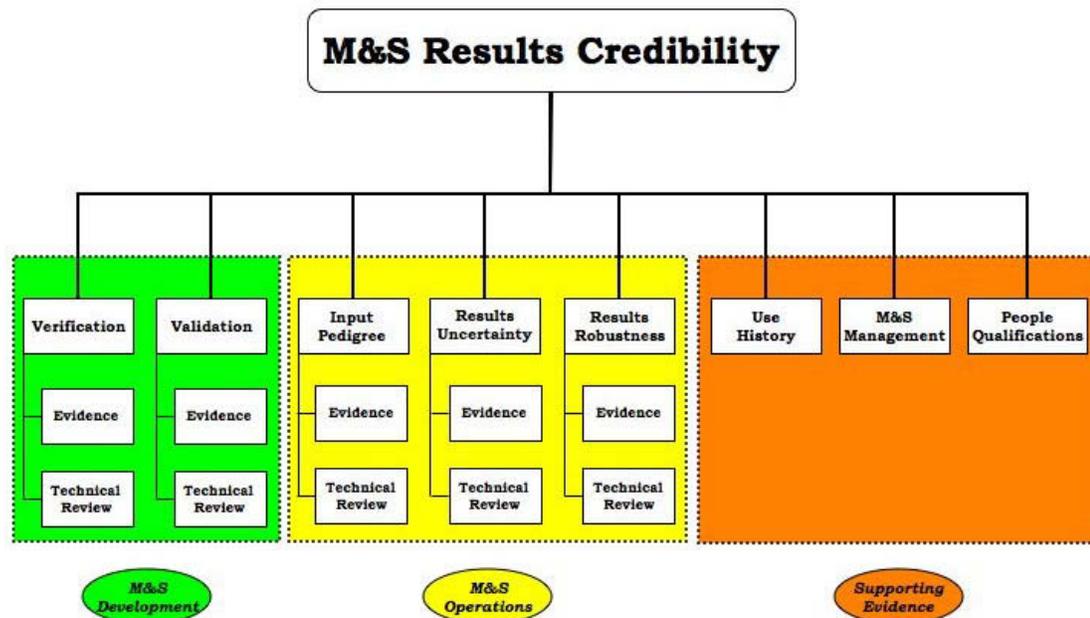


Figure 1. Credibility Assessment Scale

“These 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:

**a. 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?

**b. 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?

**c. 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 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.”

Table 1 gives a high-level summary of the evaluation criteria. The Appendix provides tables with the next level of detail. Neither Table 1 nor the information in Table 2–Table 5 in the Appendix is sufficient for interpreting the CAS. The full explanation covers 14 pages in the M&S Standard. However, some of the key explanations for interpreting the terms are as follows:

The phrase *insufficient evidence* is uniformly used for all factors to characterize level 0. It means either that no evidence exists for that factor, or that the evidence that does exist does not meet even the level 1 criteria for that factor.

The word *favorable* as used in the level definitions for several subfactors or factors (Verification Evidence, Validation Evidence, Input Pedigree Evidence and Use History) means that whatever relevant acceptance criteria have been deemed sufficient by the program/project in collaboration with the Technical Authority ... have been satisfied.

The phrase *real-world system* refers to the real system operating in its real environment.

A *problem of interest* refers to systems that are so close to the real-world system in its real environment that they capture most of the essential complexity of the real system and its environment (relevant to the current M&S application), and yet fall short of the real system in its real environment. This could be the real system in a similar environment, or a similar system in the real environment.

The phrase *unit problem* refers to problems that capture one or more physical phenomena relevant to the current M&S application. (Some disciplines use the phrase “building block” for what is referred to here as a unit problem.)”

**Table 1. Key Aspects of Credibility Assessment Levels**

(Factors with a Technical Review subfactor are underlined)

<b>Level</b>	<b>Verification</b>	<b>Validation</b>	<b>Input</b>	<b>Pedigree</b>	<b>Results</b> <u>Uncertainty</u>	<b>Results</b> <u>Robustness</u>	<b>Use History</b>	<b>M&amp;S Management</b>	<b>People Qualifications</b>
<b>4</b>	Numerical errors small for all important features.	Results agree with real-world data.	Input data agree with real-world data.	Non-deterministic & numerical analysis.	Sensitivity known for most parameters; key sensitivities identified.	De facto standard.	Continual process improvement.	Extensive experience in and use of recommended practices for this particular M&S.	
<b>3</b>	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 predictions were later validated by mission data.	Predictable process.	Advanced degree or extensive M&S experience, and recommended practice knowledge.	
<b>2</b>	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.	
<b>1</b>	Conceptual and mathematical models verified.	Conceptual and mathematical models agree with simple referents.	Input data traceable to informal documentation.	Qualitative estimates.	Qualitative estimates.	Passes simple tests.	Managed process.	Engineering or science degree.	
<b>0</b>	Insufficient evidence.	Insufficient evidence.	Insufficient evidence.	Insufficient evidence.	Insufficient evidence.	Insufficient evidence.	Insufficient evidence.	Insufficient evidence.	
<b>M&amp;S Development</b>			<b>M&amp;S Operations</b>			<b>Supporting Evidence</b>			

**Roll-up Processes**

“The primary focus of the CAS is on the scores for the eight factors; and the secondary focus is on the overall score, which is the minimum of the scores for the eight factors. The five factors in the M&S Development and M&S Operations categories are weighted averages of the associated Evidence and Technical Review subfactors. Nevertheless, the emphasis is on the scores at the factor tier; the Technical Review subfactor just serves to tune the evidence subfactor by the results of internal and external assessments.”

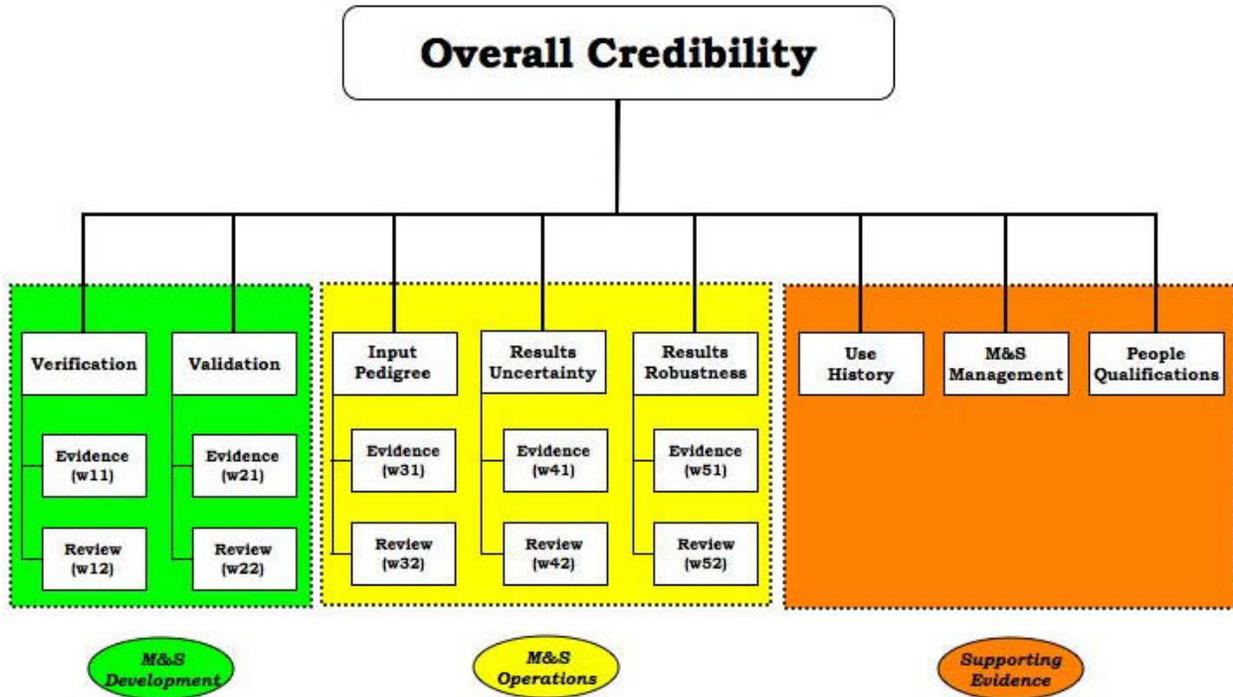


Figure 2. Subfactor Weights

Figure 2 illustrates the ten weights that are needed for the roll-up from the subfactor to the factor tier. The constraints on these weights are as follows:

- Each weight lies in the closed interval  $[0,1]$ .
- The sum of each subfactor pair, e.g.,  $w_{11}$  and  $w_{12}$ , is 1.
- 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 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.

Taking the minimum of the eight factor scores performs the roll-up of the eight factor scores into the overall score, following the philosophy that “a chain is only as strong as its weakest link.” Preliminary drafts of the M&S Standard used a weighted average of the factor scores. The decision to use the minimum score instead was made during the final review by the NASA Engineering Management Board.

Per Req. 4.7.1, Req. 4.7.3, and Req. 4.8.3, reporting of M&S results will be accompanied by reports of the eight factor scores and the single, overall score. Possible reporting formats for the factor scores are bar charts and radar plots. Also, a recommendation in the M&S Standard is that the achieved scores on the CAS be compared with the desired, or threshold, assessment levels. This facilitates a “gap analysis” to quickly identify factors where additional resources and effort, or new and improved methods, should be applied to improve the overall score and hence the credibility of the M&S results.

At this point, additional insight into the use of the CAS is best left to example.

## Example – M&S for the James Webb Space Telescope

### *Background*

The James Webb Space Telescope (JWST) is a large, infrared-optimized space telescope, scheduled for launch in 2013<sup>xiv</sup>. JWST will find the first galaxies that formed in the early Universe, connecting the Big Bang to our own Milky Way Galaxy. JWST will peer through dusty clouds to see stars forming planetary systems, connecting the Milky Way to our own Solar System. The instruments will be designed to work primarily in the infrared range of the electromagnetic spectrum, with some capability in the visible range. JWST will have a large mirror, 6.5 meters (21.3 feet) in diameter and a sunshield the size of a tennis court. Both the mirror and sunshade won't fit onto the rocket fully open, so both will fold up and open only once JWST is in outer space. JWST will reside in an orbit about 1.5 million km (1 million miles) from the Earth.

The JWST project is managed at NASA's Goddard Space Flight Center (GSFC) in Greenbelt, MD. The prime contractor for the observatory is Northrop Grumman Space Technology (NGST) in Redondo Beach, CA. The launch vehicle is an Ariane 5 ECA rocket, to be launched from Arianespace's ELA-3 launch complex at European Spaceport located near Kourou, French Guiana. The Space Telescope Science Institute in Baltimore, MD will manage Science and mission operations.

M&S has played an important role in the design and development of JWST, and will continue to play a critical role throughout integration and test (I&T) and pre-launch verification of the observatory. Many mission-critical requirements cannot be truly verified by test on the ground primarily due to the effects of gravity on the structure, the extreme environments that must be replicated (particularly the cryogenic thermal environment in which the telescope optics and instruments operate), and the inherent limitations of available test facilities and the sheer size of JWST with respect to those facilities. As prime contractor, NGST performs M&S to support design and verification, while GSFC performs independent M&S to cross-check the results. The overall M&S activity is broken up into numerous "threads" aligned with related sets of requirements and/or common M&S domains (e.g. thermal analysis, structural analysis, optical analysis). Many elements of the I&T program are specifically designed to support M&S Validation.

### *Example Problem Overview*

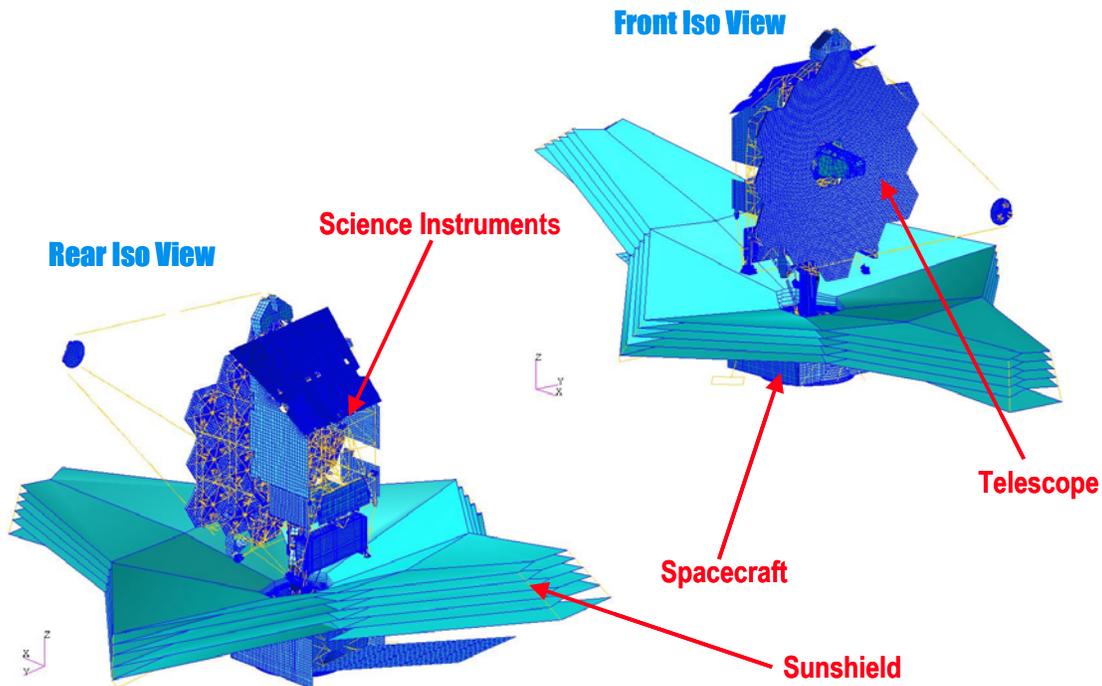
This assessment of the M&S for the "Deployed Dynamics" thread was conducted in 2008 using the JWST baseline Verification, Model Validation, and I&T plans. At the time of the assessment, JWST had just successfully completed its Mission Preliminary Design Review (PDR). The Deployed Dynamics analysis, which examines the impact on science imaging performance due to structural vibrations, had been repeatedly conducted as the observatory design matured<sup>xv</sup>. The CAS could have been used to score the results based on the post-PDR maturity of the M&S. Given that very little hardware testing had occurred by PDR, relatively low scores would have been achieved in the Validation and Input Pedigree factors. Instead, the scores and rationales given here represent a look ahead to the final pre-launch assessment of the M&S, assuming that all tests and reviews supporting the M&S take place as planned. This, then, serves to illustrate how a project M&S team might use the Standard and the CAS for planning purposes.

The central component of this M&S is a modal (eigenvectors and eigenvalues) representation of the structural dynamics of the JWST observatory in its deployed, on-orbit configuration. The modal representation is subsequently integrated into two complementary, end-to-end M&S used to predict the jitter (or pointing stability) performance of JWST during science operations. The first is a time-domain simulation used to evaluate low-frequency jitter within the bandwidth of the active pointing control loop. The second is a frequency-domain simulation used to evaluate the uncompensated jitter at frequencies above the control bandwidth. This scoring example for the use of the Credibility Assessment Scale considers only the modal dynamics model, not these two end-to-end simulations.

The structure is modeled using linear Finite Element Methods (FEM) with MSC NASTRAN, a commercial Finite Element Analysis (FEA) code. A technical review of the conceptual models for the end-to-end simulations occurred early in the JWST development phase. The consensus of the reviewers was that the use of the linear normal modes solution provided by MSC NASTRAN was valid, based on the following observations:

- Disturbances from on board sources (reaction wheels and cryo-cooler compressors) are isolated such that the resulting displacements are very small (microns or less) compared to the size of the structure (tens of meters)
- Large pre-loads are used at jointed interfaces to prevent non-linear stick/slip behavior

Shown in Figure 3 are front and rear isometric views of the NASTRAN FEM, identifying key structural components of the JWST Observatory. The main elements are the Optical Telescope Element (OTE), Integrated Science Instrument Module (ISIM), Spacecraft Element (SCE), and Sun Shield (SS). This model is considerably smaller and less complex than other JWST structural models used for stress and thermal-elastic distortion analyses. Even so, it contains roughly 200,000 grid points and elements, with roughly 1,000,000 degrees-of-freedom. On the best single-processor workstations, extraction of the first 1000 natural frequencies, out to roughly 100 Hz, takes nearly 6 CPU hours.



**Figure 3. FEM of JWST used for dynamics analysis**

The primary sources of onboard disturbances are the reaction wheels (RW), located in the SCE. Each of the six wheels is mounted on a hexapod isolator assembled using passive viscoelastic struts. In the load path between the SCE and the optical payload – the combined OTE and ISIM structures – is an assembly of four large isolator struts arranged in a cruciform geometry, again built using passive viscoelastic materials. The OTE is a large, lightweight, flexible structure by comparison to previous designs such as the Hubble Space Telescope. Accurate modeling of the structural dynamics of the OTE and the isolators is the key to credible predictions of the jitter performance. The size and unconventional design, coupled with its operation at cryogenic temperatures, presents a challenge to the validation of the M&S.

#### *Application of CAS to JWST Example*

We now proceed to use the CAS to assess the credibility of M&S results for this example problem. This example will illustrate clearly that application of the CAS to real-world problems is not clear-cut, that justification for the scores must deal with “shades of gray” common to most real-world situations.

### *Technical Review Subfactors*

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. By a “peer review” we mean an assessment that is conducted by one or more persons of equal technical standing to person(s) responsible for the work being reviewed. An “informal peer review” is one that is not conducted pursuant to a process established by the reviewed or reviewing organization, whereas a “formal peer review” is one that is sanctioned by the program/project and conducted in accordance with rules explicitly established by the reviewed or reviewing organization. Peer reviews are classified as “internal” or “external” depending upon whether or not the panel members are drawn primarily from within the lead Center for the project. Table 5 of the Standard provides the level definitions for the Technical Review subfactor.

For this example we have assigned a weight of 0.3 to all five Technical Review subfactors. This is the maximum weight permitted for this factor by the standard. This large weight is chosen on the basis that the JWST project established, and maintains, a team based at NASA GSFC for purposes of continuously reviewing the M&S activities of the prime contractor. This parallel M&S effort provides more than technical review; it performs independent M&S in order to validate and reproduce Northrop Grumman’s results.

The M&S results are presented at all major design reviews and ultimately at the Flight Readiness Review. This is an external peer review, with representatives from other NASA programs, other agencies, other aerospace contractors, universities, and retired NASA experts as part of the panel. This review will not address the Verification factor. Reviews for the other factors are assumed favorable for purposes of this example, and hence all except Verification meet the CAS Level 3 criteria.

The M&S activity established by the JWST project at NASA GSFC does perform a thorough (again, assumed favorable for purposes of this example) evaluation of all Evidence subfactors. However, this is classified as an internal peer review as it is performed by the lead center for the project. On this basis, the Level 2 criteria are met for Technical Review for all 5 factors, including Verification. Were this parallel M&S activity performed at a different NASA center, or external to the agency, the Level 4 criteria would be met for all 5 factors.

In summary, the Technical Review subfactor scores are: Verification (2), Validation (3), Input Pedigree (3), Results Uncertainty (3), Results Robustness (3).

### *Verification Subfactor*

We now proceed to the evaluation of the factors themselves, starting with the Verification factor in the M&S Development category. This factor addresses the question “Were the models implemented correctly, and what was the numerical error/uncertainty?” This has both an Evidence subfactor and a Technical Review subfactor. Table 2 of the Standard provides the level definitions for the Verification Evidence subfactor. Note that for this subfactor one must “climb the ladder”, i.e., one must meet the criteria for Level 1 before qualifying for Level 2, and so on, up the levels. Section B.3.1.1 of the Standard provides more explanation for these criteria.

MSC-NASTRAN is a commercial code, and as such the evidence for verification at CAS Level 1 must be taken from the vendor’s documentation. This documentation provides evidence that results from the computational model accurately match results obtained from closed-form solutions to the mathematical models for simple, ideal structures such as beams and plates. For CAS Level 2, while it is assumed that evidence of unit and regression testing is available from the vendor, the JWST M&S team did not request such data and therefore cannot provide evidence. In fact, it is conceivable that MSC and other commercial vendors would not provide such evidence out of proprietary considerations. This is an issue with use of commercial codes that needs to be revisited in future versions of the M&S Standard. CAS Level 2 is therefore not achieved.

As discussed earlier, CAS Level 2 is achieved for Verification Technical Review.

### *Validation Subfactor*

The other factor in the M&S Development category is Validation. This factor asks “Did the M&S results compare favorably to the referent data, and how close is the referent to the real-world system?” It, too, has both an Evidence subfactor and a Technical Review subfactor. Table 2 of the Standard provides the level definitions for the Validation Evidence subfactor. Note that for this subfactor one must again “climb the ladder”. Section B.3.1.2 of the Standard provides more explanation for these criteria.

For many M&S types, the CAS Level 1 criterion “M&S conceptual and mathematical models compare favorably with ‘general problem’ and ‘textbook’ referents” may be interpreted as “M&S results compare favorably with ‘general problem’ and ‘textbook’ referents.” The CAS Level 1 evidence is again provided by the vendor documentation.

JWST dynamics model validation relies on a conventional approach to structural integration, test, and verification<sup>xvi</sup>. This includes “coupon” tests to measure important material properties such as modulus of elasticity and damping. These tests are performed over the full range of predicted operational temperatures, and the tests are repeated using a number of different samples in order to build up statistics to support uncertainty analysis. Structural components and sub-systems are subject to one or more of the following validation tests:

- frequency verification (ensures first mode exceeds its requirement)
- stiffness test (measures deflection resulting from a known static load)
- modal survey (multiple accelerometers are used to measure frequencies and mode shapes)
- transfer function test (measures displacements, velocities, accelerations, or interface forces resulting from a known dynamic load)

The material coupon tests and structural tests performed on discrete components such as the struts used to provide vibration isolation are deemed to be “unit problems”. CAS Level 2 criteria for validation are therefore met.

Structural tests performed at higher levels of integration, for example the modal survey and transfer function test of the OTE, are classified as “problems of interest”. However, here is the first “shade of gray” situation. The size of JWST precludes structural tests such as modal surveys to be performed at the highest (observatory) level of integration. Rather, these tests are performed on each of the major elements (OTE, SCE, ISIM, SS). Furthermore, the OTE and ISIM modal surveys are performed at room temperature even though the structures operate at cryogenic temperatures on orbit. Changes in stiffness and damping must be factored in using the results of the material testing, and one special cryogenic modal survey was performed on a small, flight-like structure assembly as a cross-check. It is certainly arguable that CAS Level 3 criteria for validation are therefore met.

The assessment was performed prior to launch of the JWST observatory, precluding a CAS Level 4 assessment on the basis that measurements from the real-world system in its real-world environment are not the referent.

As discussed earlier, CAS Level 3 is achieved for Validation Technical Review.

### *Input Pedigree Subfactor*

We now move to the M&S Operations category. The first factor there is Input Pedigree, again with an Evidence and a Technical Review subfactor. We want to know “How confident are we of the current input data?” Table 3 of the Standard provides the level definitions for the Input Pedigree Evidence subfactor. For this subfactor, it is not necessary to “climb the ladder.” One does not need to meet criteria at any lower level; one only needs to meet the criteria at the given level.

Clearly, as is the case with the Validation Evidence subfactor, CAS Level 4 can not be achieved prior to launch and on-orbit operations of JWST. But all significant input parameters are measured and/or test-correlated via “problems of interest”, and uncertainties in the data are established. Hence, the criteria for CAS Level 3 are met.

As discussed earlier, CAS Level 3 is achieved for Input Pedigree Technical Review.

### *Results Uncertainty Subfactor*

Next, we move to Results Uncertainty, again with an Evidence and a Technical Review subfactor, asking “What is the uncertainty in the current M&S results?” or perhaps better put, “How thoroughly are the uncertainties in the current M&S results known?” Table 3 of the Standard provides the level definitions for the Results Uncertainty Evidence subfactor. For this subfactor, it is not necessary to “climb the ladder.” One does not need to meet criteria at any lower level; one only needs to meet the criteria at the given level. Section B.3.2.2 of the Standard provides more explanation for these criteria.

The Results Uncertainty Evidence subfactor scores as CAS Level 2, as the size of the FEM model is problematic for non-deterministic (e.g. Monte Carlo) analysis. Uncertainty in the M&S results is largely evaluated using parametric sweeps on individual variables, over the expected ranges of those variables, for a relatively small number of variables. In a few instances, uncertainty is evaluated by changing the form or fidelity of the model.

As discussed earlier, CAS Level 3 was achieved for Results Uncertainty Technical Review.

### *Results Robustness Subfactor*

The final factor in this category is Results Robustness. This is the last of the 5 factors with both an Evidence and a Technical Review subfactor. Its question is: “How thoroughly are the sensitivities of the current M&S results known?” Table 3 of the Standard provides the level definitions for the Results Robustness Evidence subfactor. For this subfactor, it is not necessary to “climb the ladder.” One does not need to meet criteria at any lower level; one only needs to meet the criteria at the given level. Section B.3.2.3 of the Standard provides more explanation for these criteria. Consult the definitions for the distinction between uncertainty and sensitivity.

The Results Robustness Evidence subfactor scores as CAS Level 2, as the sensitivity of M&S results is determined only for a small number of parameters hand-picked by the analysts.

As discussed earlier, CAS Level 3 was achieved for Results Robustness Technical Review.

### *Use History Subfactor*

We now move to the factors in the Supporting Evidence category. These are factors that do not deal directly with the M&S results, but nonetheless have a significant bearing on their credibility. None of these have a Technical Review subfactor.

The issue for the Use History scoring is “Have the current M&S been used successfully before?” Table 4 of the Standard provides the level definitions for the Use History factor. Note that for this factor one must partially “climb the ladder”, i.e., one must meet the criteria for Level 2 before qualifying for Level 3, and one must meet the criteria for Levels 2 and 3 before qualifying for Level 4. However, one need not meet the criteria for Level 1 before qualifying for higher levels. Section B.3.3.1 of the Standard provides more explanation for these criteria.

The arguments for Levels 1-3 are easy to make. The M&S results for this particular application, implemented in MSC-NASTRAN using user-supplied material properties, geometric properties, and loads, compare favorably with data obtained from JWST technology development testbeds and from tests of the flight hardware (Level 1). MSC-NASTRAN is extensively used for critical decisions by the aerospace industry for most space vehicle and satellite FEA applications (Level 2). M&S results using MSC-NASTRAN have accurately predicted real-world performance for numerous space missions (Level 3).

Regarding Level 4, there are numerous commercial FEA codes, and it is difficult to claim that any one is “the de-facto standard”. However, at the very least MSC NASTRAN is first-among-equals owing to its heritage in space applications and by virtue of NASA’s historical role in its development. Accordingly, the JWST project claims that Level 4 is achieved.

### *M&S Management Subfactor*

Next we consider M&S Management. The question is “How well managed were the M&S processes?” Table 4 of the Standard provides the level definitions for the M&S Management factor. Note that for this factor one must “climb the ladder”, i.e., one must meet the criteria for Level 1 before qualifying for Level 2, and so on, up the levels. Section B.3.3.2 of the Standard provides more explanation for these criteria.

JWST Project Management (NASA), Mission Systems Engineering (NASA), and Observatory Systems Engineering (NGST) defined clear roles and responsibilities for the M&S team, satisfying Level 1 for this factor. The M&S is developed, operated, and configuration controlled according to formal procedures established by the JWST project, including the “JWST Math Models Guidelines Document” and the “JWST System Modeling and Analysis and JWST Models Validation, Verification and Calibration Plan”, satisfying the Level 2 criteria. The M&S team periodically demonstrates repeatability of the M&S results via independent M&S performed by NASA GSFC, satisfying Level 3 criteria.

### *People Qualifications Subfactor*

We now come to the last of the 8 factors—People Qualifications. The question is “How qualified were the personnel?” Table 4 of the Standard provides the level definitions for the People Qualifications factor. For this factor, it is not necessary to “climb the ladder.” One does not need to meet criteria at any lower level; one only needs to meet the criteria at the given level. Section B.3.3.3 of the Standard provides more explanation for these criteria.

The argument for Level 4 is easy to make for this factor. All team and subteam leads have advanced engineering or science degrees. The team includes members with extensive (20+ years in most cases, with a minimum of at least 10 years) experience in structural dynamics modeling using MSC-NASTRAN for numerous prior space flight applications. Written best practices for using MSC-NASTRAN exist and are followed, including the MSC-NASTRAN User’s Manual and JWST-specific documentations (i.e. the “JWST Math Models Guidelines Document” and the “JWST System Modeling and Analysis and JWST Models Validation, Verification and Calibration Plan”).

### *Summary of Subfactor and Factor Scores, Overall Score and Reporting Formats*

Figure 4 summarizes all of the subfactor weights, subfactor scores, and the resulting factors scores for this example problem. The Standard specifies that the overall score for the M&S be determined by taking the minimum of the factor scores. For this example, the minimum score (1.3) was obtained for the Verification factor, therefore the overall score for this M&S is 1.3. If one looks at the complete set of eight scores, the JWST Deployed Dynamics M&S scores very well. The average score is 2.9, actually a very good score for a case in which neither the validation referent nor the input data are traceable to real world operational (on-orbit) measurements. The low Overall Score, per the Standard, results from an outlier. The decision maker(s) who rely on results from this M&S would clearly be aware of all of the scores and all of the evidence presented.

The standard suggests several options for graphical presentation of the scores, Bar Chart format and Radar Plot format, as illustrated in Figure 5. The standard suggests the use of color coded graphics (red for major deficiencies, yellow for minor deficiencies, green for meeting the threshold, blue for exceeding the threshold) to present the “gap analysis” based on required threshold scores for each of the eight factors. JWST has not established required levels for the factors, as, at this time, the Standard is not mandatory for NASA projects.

Factor	Evidence Subfactor Score	Evidence Subfactor Weight	Technical Review Subfactor Score	Technical Review Subfactor Weight	Factor Score
Verification	1.0	0.7	2.0	0.3	1.3
Validation	3.0	0.7	3.0	0.3	3.0
Input Pedigree	3.0	0.7	3.0	0.3	3.0
Results Uncertainty	2.0	0.7	3.0	0.3	2.3
Results Robustness	2.0	0.7	3.0	0.3	2.3
Use History					4.0
M&S Management					3.0
People Qualifications					4.0

Figure 4. Details of CAS Subfactor Scores for JWST Example

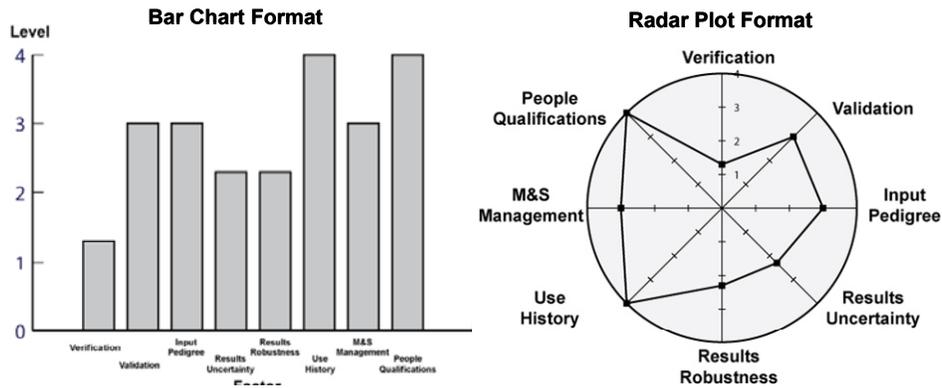


Figure 5. CAS Summary Scores for JWST Example – Bar Chart and Radar Plot Formats

## Conclusions and Future Work

The CAS measures the M&S results, and the rigor of the processes used to obtain them, against key factors that were judged to be best in form key decision makers and affect their credibility judgment. The factors used span a reasonably complete percentage of “credibility space”, and application of the CAS to “validation examples” such as the JWST M&S problem presented here has shown that it is not difficult, and in fact quite useful, to organize technical information and evidence into these categories.

Scoring each factor using the level definitions is reasonably straightforward for some factors, for example Use History and People Qualifications. In other cases, scoring requires considerable thought and judgment. For example, Input Pedigree, where in the JWST example there are thousands of parameters and it is impractical to measure each and every parameter including their uncertain ties. The resulting score was given based on judgment that measurements, including sensitivities, were made for all of the important parameters. Working within the framework of the CAS forces the analyst to prepare and organize evidence, and to make the arguments justifying the scores. These formalisms are beneficial to both analyst and decision maker alike.

Another issue flagged by the application to the JWST example relates to the use of COTS (Commercial Off-The-Shelf) software for M&S, and in particular the inter-relationship between the Verification and Use History factors. The level definitions and the “climb the ladder” rule for Verification work quite well for, and in fact were entirely motivated by, the case in which the source code for the computational model is controlled. For COTS software, this is obviously not the case. In particular, we have quite a “catch 22” scenario in this example involving the use of a de-facto standard COTS tool that scores low for Verification, as a result of lack of visibility into proprietary software development processes, yet scores high for Use History. This issue needs to be addressed in future revisions to the standard.

Attempts to apply the CAS to the two end-to-end dynamics simulations proved difficult, resulting in the decision to score only the NASTRAN FEM that is common to both simulations. This problem with the CAS was in fact understood at the time the standard was released, and resulted in a specific recommendation, captured in the report<sup>v</sup> to the NASA Engineering and Safety Center (NESC), stating that

***“NASA should refine how submodels are treated in the CAS. The present version of the M&S Standard makes no distinction between individual models and integrated models consisting of multiple submodels. The roll-up of assessments of the individual submodels into the assessment of the integrated model is primarily an issue for the credibility assessment scale. The credibility assessment should eventually be refined to account for the additional issues associated with integration of submodels.”***

As the M&S Standard was only recently completed and approved, it has yet to be applied to any NASA programs or projects. As the Standard gains acceptance and becomes more widely used, feedback from the real-life M&S applications will be crucial in refining the Standard, and in particular the CAS. Accordingly, an additional recommendation made in the NESC report states that

***“Information regarding credibility assessment scale usage should be collected to determine effectiveness and provide data for further revision. In general, scales measuring the rigor, credibility, or similar aspects of M&S results have not received much use, and there is no consensus on such assessments. In particular, the credibility assessment scale in the M&S Standard has not been used. The immaturity of this particular field necessitates close monitoring of the impact of credibility assessment scale usage by NASA programs and the use of that information to update the credibility assessment scale. This is not a criticism of the present credibility assessment scale, but merely an acknowledgment of the state of such assessments; operational use is essential to advance the state-of-the-art.”***

## Acknowledgments

The authors wish to acknowledge the contributions of Steve R. Blattning<sup>2</sup>, Lawrence L. Green<sup>2</sup>, James M. Luckring<sup>2</sup>, Joseph H. Morrison<sup>2</sup>, Ram K. Tripathi<sup>2</sup>, and Thomas A. Zang<sup>2</sup> of the M&S Standard Development Team, and Ummeel B. Mehta<sup>8</sup> of the M&S Standard Topic Working Group, for discussions on the topic of scales in general. In addition, we are indebted to the work of William Oberkamp<sup>8</sup>, Martin Pilch<sup>9</sup>, David Peercy<sup>9</sup>, and Francois Hemez<sup>9</sup> on the Predictive Capability Maturity Model, and for their feedback on the CAS during a 2007 technical interchange meeting in Albuquerque, NM.

The authors wish to thank the NASA Office of the Chief Engineer and the NASA Engineering and Safety Center (NESC) for sponsorship of this work. In particular, we are grateful to the NASA Chief Engineer from 2005-2007, Christopher Scolese, for motivating and championing the development of a credibility assessment scale, to the NASA Chief Engineer from 2007 onwards, Mike Ryschkewitsch, for continuing to champion the standard and recommending the key approach of using a risk-informed approach to resolving the thorny scope question, and to Harold Bell and Dawn Schaible for facilitating the project.

Finally, the authors wish to acknowledge the work done by Allen Bronowicki and Parker Lin at Northrop Grumman Space Technology for leading the development of the JWST Deployed Dynamics model and the associated modeling guidelines, analysis plans, and model verification and validation plans.

Portions of the research described in this paper were carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.

---

<sup>8</sup> NASA Ames Research Center

<sup>8</sup> Sandia National Laboratories

## Appendix

The following tables provide further details on the CAS. Still further details are provided in the M&S Standard itself. What's given here is just summary information. A thorough reading of the M&S Standard is necessary for interpreting and applying the CAS.

**Table 2. Level Definitions for Evidence Subfactors in the M&S Development Category**

Level	Verification Evidence	Validation Evidence
<b>4</b>	Reliable error estimation methods are used to quantitatively assess numerical errors. These estimates show that the errors are small from test suites, which exercise all important algorithms, all important features and capabilities, and all important couplings (physics, modules, etc.) of the full computational model.	M&S results compare favorably for the real-world system at validation points by comparison of M&S results to an acceptable referent, which is measurements on the real-world system.
<b>3</b>	Some formal method is used to assess numerical errors associated with unit testing with significant coverage of the code.	M&S results compare favorably for problems of interest at validation points by comparison of M&S results to an acceptable referent, which is experimental measurements on problems of interest.
<b>2</b>	Favorable results from unit and regression testing of key features of the computational model.	M&S results compare favorably for unit problems at validation points by comparison of M&S results to an acceptable referent, which is either experimental measurements or higher-fidelity M&S results.
<b>1</b>	Favorable evidence of verification for conceptual and mathematical models.	M&S conceptual and mathematical models compare favorably with “general problem” and “textbook” referents.
<b>0</b>	Insufficient evidence.	Insufficient evidence.

**Table 3. Level Definitions for Evidence Subfactors in the M&S Operations Category**

<b>Level</b>	<b>Input Pedigree Evidence</b>	<b>Results Uncertainty Evidence</b>	<b>Results Robustness Evidence</b>
<b>4</b>	The input data compare favorably with measured data from the real-world system, or the input data came from M&S with a summary credibility rating above 3.5. Uncertainty associated with the input data is known.	Uncertainty estimates are quantitative and based upon nondeterministic and numerical analysis.	Sensitivity of the M&S results for the real-world system is quantitatively known for most of the variables and parameters, including all of the most sensitive variables and parameters.
<b>3</b>	The input data compare favorably with acceptable measured referent data from problems of interest, or the input data came from M&S with a summary credibility rating above 3.0. Uncertainty associated with the input data is known.	Uncertainty estimates are quantitative and based upon nondeterministic analysis.	Sensitivity of the M&S results for the real-world system is quantitatively known for many variables and parameters.
<b>2</b>	The input data is traceable to formal documentation, or the input data came from M&S with a summary credibility rating above 2.0.	Uncertainty estimates are quantitative and based upon deterministic analysis or expert opinion.	Sensitivity of the M&S results for the real-world system is quantitatively known for a few variables and parameters.
<b>1</b>	The input data is traceable to informal documentation, or the input data came from M&S with a summary credibility rating above 1.0.	Uncertainty estimates are qualitative.	Sensitivity of M&S results for the real-world system is estimated by analogy with the quantified sensitivity of similar problems of interest.
<b>0</b>	Insufficient evidence.	Insufficient evidence.	Insufficient evidence.

**Table 4. Level Definitions for Factors in the Supporting Evidence Category**

<b>Level</b>	<b>Use History</b>	<b>M&amp;S Management</b>	<b>People Qualifications</b>
<b>4</b>	De facto standard.	Continuing Process Improvement: The M&S effort is using measurements on M&S processes to improve the repeatability of the M&S results.	Possesses an advanced engineering or science degree or extensive work experience in M&S, has extensive experience with the development and use of the M&S being reviewed, and has employed specific recommended practices relevant to current application.
<b>3</b>	Post-decision real-world events have been accurately represented in results (e.g., validated by mission data).	Predictable Process: The M&S effort is measuring repeatability of the M&S results generated by the M&S processes.	Possesses an advanced engineering or science degree or extensive work experience, has general M&S training, has specific experience with the M&S being reviewed, and has been trained on specific recommended practices relevant to the current application.
<b>2</b>	Used previously to perform analysis upon which critical decisions have been made.	Established Process: The M&S effort has established a documented process for M&S development and operations.	Possesses an engineering or science degree, has received formal training in formulation of M&S and generic training in recommended practices for M&S, and has developed M&S products.
<b>1</b>	Specific scenarios have been created to test application, or results compare favorably with outputs from other similar tools.	Managed Process: The M&S roles and responsibilities have been defined.	Possesses an engineering or science degree, has been introduced to the topic of M&S, and has been exposed to generic recommended practices in M&S.
<b>0</b>	Insufficient evidence.	Insufficient evidence.	Insufficient evidence.

**Table 5. Level Definitions for the Technical Review Subfactors**

<b>Level</b>	<b>Technical Review</b>
<b>4</b>	Favorable external peer review accompanied by independent <i>factor evaluation</i> .
<b>3</b>	Favorable external peer review.
<b>2</b>	Favorable formal internal peer review.
<b>1</b>	Favorable informal internal peer review.
<b>0</b>	Insufficient evidence.

## References

- <sup>i</sup> Columbia Accident Investigation Board Report, Volume 1, August 2003 (<http://caib.nasa.gov/>)
- <sup>ii</sup> A Renewed Commitment to Excellence. An Assessment of the NASA Agency-wide Applicability of the Columbia Accident Investigation Board Report, January 30, 2004 (P. A-5, R13) ([http://www.nasa.gov/pdf/55691main\\_Diaz\\_020204.pdf](http://www.nasa.gov/pdf/55691main_Diaz_020204.pdf))
- <sup>iii</sup> NASA-STD D-(I)-7009, Standard for Models and Simulations, December 1, 2006 ([ftp://ftp.sonic.net/pub/users/usacm/PTC60/NASA\\_STD\\_I\\_7009.pdf](ftp://ftp.sonic.net/pub/users/usacm/PTC60/NASA_STD_I_7009.pdf))
- <sup>iv</sup> Bertch, William, Zang, Thomas A., and Steele, Martin J.: “Development of NASA’s Models and Simulations Standard,” Paper 08S-SIW-037, 2008 SISO Spring Simulation Interoperability Workshop Providence, RI, April 13–18, 2008.
- <sup>v</sup> T.A. Zang et al, NASA Standard for Models and Simulations (M&S): Development Process and Rationale, NASA Engineering and Safety Center Technical Report RP-08-118, Nov. 20, 2008 (URL: <http://www.nasa.gov/offices/nesc/reports/index.html>)
- <sup>vi</sup> Blattnig, S.R. Luckring, J.M., Morrison, J.H. Tripathi, R.K. Zang, T.A. and Sylvester, A.J., “NASA Standard for Models and Simulations: Philosophy and Requirements Overview,” AIAA Paper 2009-1010, Jan. 2009
- <sup>vii</sup> Balci, O., Adams, R.J., Myers, D.S., and Nance, R.E., “A Collaborative Evaluation Environment for Credibility Assessment of Modeling and Simulation Applications,” Proceedings of the 2002 Winter Simulation Conference, E. Yücesan, C.-H. Chen, J.L. Snowdon, and J.M. Charnes, eds., pp. 214–220.
- <sup>viii</sup> Harmon, S.Y. and Youngblood, S. M., “Simulation Validation Quality and Validation Process Maturity,” Simulation Interoperability Workshop, Paper No. 04S-125, March 2004.
- <sup>ix</sup> Oberkampf, W.L., Pilch, M., and Trucano, T.G., “Predictive Capability Maturity Model for Computational Science and Engineering,” Sandia Report SAND2007-5948, October 2007.
- <sup>x</sup> Green, L.L., Blattnig, S.R., Luckring, J.M. and Tripathi, R., “An Uncertainty Structure Matrix for Models and Simulations,” AIAA Paper 2008-2154, 2008.
- <sup>xi</sup> Blattnig, S. R.; Green, L. L.; Luckring, J. M.; Morrison, J. H.; Tripathi, R. K.; and Zang, T. A., “Towards a Credibility Assessment of Models and Simulations,” AIAA 2008-2156, April, 2008.
- <sup>xii</sup> Thomas, D. and Hale, J., “A Characterization Taxonomy for Integrated Management of M&S,” Paper 06F-SIW-091, September 2006.
- <sup>xiii</sup> Mehta, U.B., “Simulation Credibility Level,” 5th Joint Army-Navy-NASA-Air Force Modeling and Simulation Subcommittee Meeting, May 14-17, Denver, CO.
- <sup>xiii</sup> Blattnig, S.R., Green, L.L., Luckring, J.M., Morrison, J.H., Tripathi, R.K., and Zang, T.A., “NASA Standard for Models and Simulations: Philosophy and Requirements Overview,” this conference
- <sup>xiv</sup> <http://www.jwst.nasa.gov/>
- <sup>xv</sup> Hyde, T.T., Ha, K.Q., Johnston, J.D., Howard, J.M. and Mosier, G.E., “Integrated Modeling Activities for the James Webb Space Telescope: Optical Jitter Dynamics Analysis”, SPIE 5487 – Astronomical Telescopes and Instrumentation, 2004
- <sup>xvi</sup> Mosier, G.E., Howard, J.M., Johnston, J.D., Parrish, K., Hyde, T.T., McGinnis, M.A., Bluth, M., Kim, K., and Ha, K.Q., “The Role of Integrated Modeling in the Design and Verification of the James Webb Space Telescope”, SPIE 5528 – Space Systems Optomechanics and Dynamics, 2004