Towards a Credibility Assessment of Models and Simulations

Steve R. Blattnig,[*] Lawrence L. Green,[†] James M. Luckring, [‡] Joseph H. Morrison,[§] Ram K. Tripathi,[**] and Thomas A. Zang[††]
NASA Langley Research Center, Hampton, Virginia, 23681

A scale is presented to evaluate the rigor of modeling and simulation (M&S) practices for the purpose of supporting a credibility assessment of the M&S results. The scale distinguishes required and achieved levels of rigor for a set of M&S elements that contribute to credibility including both technical and process measures. The work has its origins in an interest within NASA to include a "Credibility Assessment Scale" in development of a NASA standard for models and simulations.

Nomenclature

AIAA = American Institute of Aeronautics and Astronautics
ASME = American Society of Mechanical Engineers
M&S = Models and Simulations
NASA = National Aeronautics and Space Administration

I. Introduction

THE literature on modeling and simulation contains many proposals for methods to assess various aspects of the quality of models and simulations. Some representative early contributions were those of Wang and Strong[1], Mehta[2], Tseng and Fogg[3], Logan and Nitta[4], and Balci[5]. The present paper reports a proposal that focuses on the aspect of the credibility of the results from the models and simulations (M&S). The proposal presented here grew out of the authors' response to a direct challenge from the NASA Chief Engineer in 2006 to include a so-called "Credibility Assessment Scale" in the NASA Standard for Models and Simulations[6] (then still in development), and it targets the goal of that standard, namely, "to ensure that the credibility of the results from models and simulations (M&S) is properly conveyed to those making critical decisions," where "critical decisions" refers to decisions that may affect human safety and mission success. The authors' original thinking on such a scale appeared in Appendix A2 of the interim version of the NASA standard[7]. The proposal described in the present paper represents the authors' more mature thoughts on the subject—it has the same structure as the "A2 Scale" of the interim standard, but differs in some details.

Development of a scale to assess models and simulations is very challenging for many reasons. First, a single scale suitable for all M&S may be infeasible since there are many types of models and simulations. Second, different organization cultures (or sub-cultures within a single agency/organization) may require different scales to be effective. Finally, scales can have many different purposes. There is certainly no single correct scale for all purposes, and there is probably no unique, correct scale for a single type of M&S, for a single purpose, or for a single organizational culture. This paper focuses on developing a scale with the primary objective of assisting in the evaluation of credibility of M&S results for high consequence decisions within a hierarchical organizational structure. The scale will help a decision maker, who may or may not have subject matter or M&S expertise, make decisions based on M&S results from others.

[*] Physicist, Durability, Damage Tolerance and Reliability Branch, Mail Stop 188E.
[†] Mathematician, Space Mission Analysis Branch, Mail Stop 462, Senior Member AIAA.
[‡] Senior Research Engineer, Configuration Aerodynamics Branch, Mail Stop 499, Associate Fellow AIAA.
[§] Assistant Head, Computational AeroSciences Branch, Mail Stop 128, Senior Member AIAA.
[**] Senior Research Scientist, Durability, Damage Tolerance and Reliability Branch, Mail Stop 188E, Associate Fellow AIAA.
[††] Chief Technologist, Systems Analysis & Concepts Directorate, Mail Stop 449, Associate Fellow AIAA.

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[8], Balci[5], Harmon and Youngblood[9], Oberkampf, Pilch, and Trucano[10], and Green *et al*[11]. Most of these were called maturity matrices rather than scales. (The authors consider the distinction between a scale and a maturity matrix to be primarily a semantic one; the term scale is used hereafter.) Subsequent to the dissemination of the original version of the present scale in the August 2006 draft of the M&S Standard, two alternative scales have been presented by some of those involved with the development of the M&S Standard (see Hale and Thomas[12] and Mehta[13]). Moreover, the final version of the M&S Standard contains a very different scale, which is the product of a team representing 9 NASA centers. Since, as noted earlier, there is no single correct scale, the present paper merely focuses on explaining the details and the rationale for the proposed scale.

II. Determination of Credibility

A. Goals and Benefits

The goal of the scale described in this paper is to enhance the decision makers' ability to determine the credibility of specific M&S results for the purpose of specific decisions. (The authors intentionally forgo attempting a technical definition of credibility, as we use this word in the usual sense of the English language.) The focus in this paper is on assessing the credibility of the very specific M&S results that are used for a very specific decision, not on the credibility of a broad set of results over the domain of intended use for the M&S. As noted earlier, the decisions of interest to this activity are those that affect human safety and mission success, *aka* critical decisions.

The present scale is meant to help a decision-maker (an individual or a group) evaluate the credibility of specific M&S results using common language for evaluating and communicating those aspects of M&S results that most influence credibility. This language is independent of the M&S type, of the specific application, and of the specific project requirements levied on the M&S. Such a language can form the basis of organization-wide, generic norms; having embedded project requirements would produce a scale that did little more than provide a checklist for verifying the relevant project requirements. On the other hand, the project should identify the scale ratings (using organization-wide norms) that are needed for the specific application.

Since credibility is subjective, different decision makers may well assign different degrees of credibility to the same M&S results; no one can be told by someone else how much confidence to place in something. The assessment of M&S credibility can be viewed as a two-part process. First, the M&S practitioner makes and conveys an assessment of the particular M&S results. Then, a decision maker infers the credibility of the M&S results presented to them in their particular decision scenario.

The authors' purpose here is to provide common language for the M&S practitioner to communicate the most important, objective contributors to a decision maker's assessment of the credibility of M&S results. Just to reinforce this fundamental point—the process described here does not purport to provide a measure of credibility. Its function is to enable clear communication between M&S practitioners and decision makers of the information needed for the decision maker to determine the credibility that he or she attaches to the results. The determination of whether a result is good enough to serve as a basis for a decision ultimately rests upon the person making that decision. A scale cannot replace good judgment. It is only a tool to communicate an important part of the information needed to make the decision.

In addition to the goal articulated above, the credibility assessment process has two ancillary benefits. The first is assisting the M&S project managers in measuring progress in the M&S capability. The second is assisting project managers in resource allocations to best improve the credibility of the M&S results. In the former case one can track the scale ratings on representative applications of the M&S. In the latter case, the scale provides a framework for identifying the specific areas where improvements in the M&S are most needed.

B. Credibility Assessment Components

The objective contributors to M&S credibility are divided into two different parts, a comprehensive results statement and a rigor assessment. The comprehensive results statement consists of a best estimate (the "answer") and an uncertainty statement. The rigor assessment is based on a scale that describes the rigor of the processes used to produce the results. The uncertainty statement can itself be divided into two parts: an estimate of the uncertainty and a description of how well that uncertainty is known. Clearly, all else being equal, the smaller the uncertainty and the better that uncertainty is known, the more credible the M&S results are. But of these two aspects of uncertainty, only how well the uncertainty is known can be measured on a scale that is independent of the project requirements; in some cases the project may not even define a requirement on the size of the uncertainty. Since the amount of uncertainty that is acceptable is highly dependent upon the question being asked, the former aspect is clearly

dependent upon the context of the decision that the specific M&S results support; it is best left as a standalone piece of information provided to the decision maker that informs his (subjective) credibility assessment.

The other major piece of information provided to the decision maker is the assessment of the M&S results against a rigor scale. The fundamental premise of our approach is that as a general rule, the more rigorous were 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. Bear in mind that in the decision-making applications of interest in this paper, experimental data for the real system in its real environment are not available. (If they were, there would be no need for using the M&S results in the decision.) Statistically, there inevitably will be cases in which, *a posteriori*, low credibility results turn out to be closer to experimental data than high credibility results. For a well-chosen scale, these cases will be the exception rather than the rule.

Neither credibility nor rigor is an intrinsic property of M&S results. For example the result statement "4.0 ± 0.1 with 95% confidence" is neither credible nor non-credible by itself. Rigor is a property of the processes used to generate the results. Credibility is influenced by the rigor of the processes and the size of the uncertainty (in the context of the intended use).

There is a considerable literature on estimating uncertainty in the results from M&S. The focus in this paper is on a proposed rigor scale.

C. Desirable Features of a Scale to Assess Rigor

The proposed rigor scale was designed with certain desirable features in mind:

1. The scale should be objective.
2. The scale should be readily understandable by both M&S practitioners and decision makers.
3. The scale should be practical to implement.

The need for these features should be self-evident. Particular stress is placed on the first feature; different individuals should look at the same M&S processes and come up with similar rankings on the scale. The intention for this work was to produce a scale that is as simple as possible and yet still communicates the essential information needed to determine how much trust to place in an M&S result. In order to do this the scale cannot stand on its own. The scale is merely a summary. References and more detailed information need to be available in case the decision maker requires them.

III. Architecture of the M&S Rigor Scale

The generic architecture chosen for the scale is illustrated in Figure 1. The distinct aspects of the M&S rigor are grouped into N categories, and the M levels are the ordered, integer rating options for the degree of rigor obtained in each category. Each category consists of a small number of characteristics of the rigor used to produce the M&S results; an ideal set of categories would be an orthogonal set. A characteristic of the M&S is an aspect that can be assessed using a single type of criterion. Having many characteristics in a category complicates the scoring in the (frequent) case in which different scores are obtained on different characteristics. This makes the assignment of a single integer rating for category problematic. For this reason, having a small number of characteristics in each category is more practical to implement, more understandable and more objective. However, this desire for a small number of characteristics needs to be balanced by the need for completeness. Each category is scored independent of all the other categories.

The primary consideration for the choice of the number of categories is the ample evidence in the psychology literature buttressing Miller's classic observation[14] that humans can only grasp 5–9 distinct judgments at a time. Obviously, there are far more than 9 potential categories of M&S to include in the scale. We select only those categories that contain genuine show-stopping characteristics, ones for which a high score on all other categories is insufficient to establish the credibility of the results.

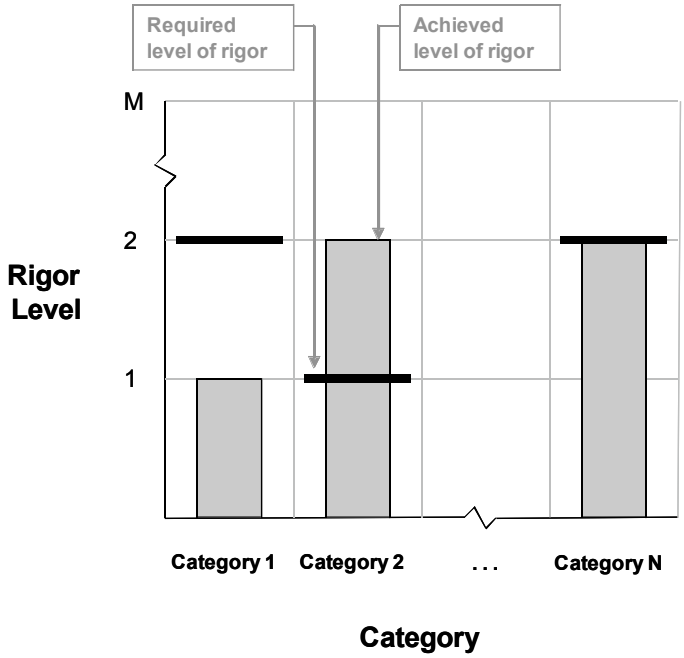

Figure 1. Architecture of the Scale

The primary considerations for the choice of the number of levels are as follows. If there are too many levels, then the distinctions between levels become fuzzy (and therefore the scale will be difficult to implement). If there are too few levels, then some important distinctions will be lost (decreasing the value of the scale). An odd number of levels creates the temptation to take the 'middle of the road' approach leading to an essentially 'no decision'. As discussed in the following subsections, the proposed scale has 7 categories and uses 4 levels.

For each level in each category a clear and concise level definition is provided to facilitate objectivity in ranking. These level definitions describe the conditions that must be met in order to qualify for a rating of at least that level. The rating given for a particular category corresponds to the highest level for which all the conditions in the level definitions are met. Since there is no "partial credit" for meeting the requirements of a given level, only integer rankings are used to describe the rigor in a category. In general, the criteria for the lower levels are based on subjective, informal evidence, whereas the criteria for the higher levels demand more objective and formal evidence. Once again, the ranking levels evaluate the M&S process (how the work was done) as opposed to the M&S outcome itself.

An explicit distinction is made for each category between the actual level (or rating) of rigor that was achieved in practice and the rigor that the project prescribed as required (or needed). In Figure 1, the gray bars report the achieved levels, whereas the thick black lines report the required levels. Recall that the distinction between required and achieved levels is made because the scale has been motivated by the desire to measure the M&S rigor consistently, even across projects with different requirements. This distinction can also provide guidance for M&S resource allocation to meet project requirements.

IV. Elements of the M&S Rigor Scale

This section outlines the detailed elements of the M&S rigor scale along with reporting considerations of the rigor assessment to decision makers. The rigor scale categories and levels have been influenced by many sources, particularly Oberkampf, Pilch, and Trucano[10]. The categories chosen for the present scale are: (1) code verification, (2) solution verification, (3) validation, (4) prediction uncertainty, (5) level of technical review, (6) process control, and (7) operator and analyst qualifications. The first four categories represent technical distinctions of an M&S activity whereas the last three categories address more procedural distinctions.

A four-level measure of rigor is provided for each category. In general, the levels address an attribute of the particular metric along with a sense of coverage for conditions affecting critical decisions. An initial determination of conditions affecting critical decisions is best established before initiating the M&S. The distinctions specific to each category adhere to the following guidelines:

Level 1 – M&S results achieved but with no or ad hoc treatment of rigor category metrics
Level 2 –M&S rigor based on expert opinion and qualitative assessments; minimal coverage for conditions affecting critical decisions
Level 3 –Critical subset of Level 4 rigor achieved; coverage of more than half of the conditions affecting critical decisions
Level 4 –M&S rigor based on formal quantified assessments; extensive coverage for conditions affecting critical decisions

The nature of the levels is to (1) move from subjective to objective measures, and (2) move from minimal to extensive coverage of the conditions that affect critical decisions.

A. Description of categories and levels.

1. Code Verification

The American Institute of Aeronautics and Astronautics (AIAA)[15] and the American Society of Mechanical Engineers (ASME)[16] adopted the following definition of verification for M&S:

> Verification: The process of determining that a model implementation accurately represents the developer's conceptual description of the model and the solution to the model.

Roache[17] recognized that there are two parts to verification: (1) Code Verification, which focuses on activities that demonstrate that the code accurately predicts known exact solutions, is described in this section and (2) Solution Verification, which involves error estimation for a specific M&S application, is addressed in the next section.

The rigor of the Code Verification category is measured by the degree to which the M&S code and numerical algorithms have been tested for errors. Code verification consists of software quality assurance activities such as version control, configuration management, documentation, regression testing, and numerical algorithm testing. Numerical algorithm testing focuses on the correctness of specific implementations of the numerical algorithms using test cases such as exact analytical solutions. (Note: since it is not generally possible to prove a code correct, code verification is a continuing activity to develop confidence. Multiple application-relevant test problems should be developed with coverage for the domain of interest.).

The level descriptions for code verification are:
Level 1 – M&S results achieved with no or ad hoc code verification.
Level 2 –Software versions archived and results repeatable; numerical algorithm test suite with sparse coverage of required features and capabilities;
Level 3 –Software and test cases maintained in configuration control system; numerical algorithm test suite with moderate coverage of required features and capabilities;
Level 4 –Well defined and documented Software Quality Assurance (SQA) processes; numerical algorithm test suite with significant coverage of required features and capabilities.

2. Solution Verification

The rigor of the Solution Verification category is measured by estimates of numerical error associated with the M&S results. Examples of numerical errors for M&S results include discretization error, iterative error, round-off error, and statistical sampling errors. The acceptable level of numerical error should be specified before the M&S is performed.

The level descriptions for solution verification are:
Level 1 – M&S results achieved with no or ad hoc solution verification.
Level 2 –Expert opinion based on numerical errors estimated for similar problems. These error estimates are provided for some results that affect a critical decision.

Level 3 –Numerical errors estimated on actual application. These error estimates are provided for more than half of the results that affect a critical decision.
Level 4 –Rigorous numerical error bounds quantified for actual application. These error estimates are provided for all results that affect a critical decision.

3. Validation

The Validation category measures the degree to which outcomes of simulations agree with outcomes of physical experiments or observations. Validation is accomplished only at a finite number of experimental measurement points. More rigorous validation is achieved through improved characterization of both experimental and simulation uncertainties, and through improved coverage of the domain of interest. Characterizing experimental uncertainty allows the determination of accuracy of the model at that point. By further including a characterization of the model uncertainty in the validation comparison the model uncertainty is also validated or shown to be inaccurate.

The level descriptions for validation are:
Level 1 – M&S results achieved with no or ad hoc validation.
Level 2 –Quantified validation has been performed with estimates of experimental error; few aspects of the system have been validated.
Level 3 –Quantified validation has been performed with estimates of M&S uncertainty and experimental error; most aspects of the system have been validated.
Level 4 –Quantified validation has been performed with thorough determination of M&S uncertainty and experimental error; the system has been comprehensively validated.

4. Prediction Uncertainty

Prediction refers to an M&S result at any point other than a validation point, and we avoid the common but vague notion of a validation domain. Rather we speak of a *validation envelope*, by which we mean a boundary that contains all the validation points. One example of a geometrical definition of the validation envelope is the convex hull of validation points shown in Figure 2.

The Prediction Uncertainty category measures the rigor used to account for uncertainties in M&S predictions. A higher prediction uncertainty level corresponds to a better process of accounting for the prediction's uncertainty based on solution verification, validation activities, and inference from validation points to the prediction point. As noted in the ASME Guide, "Confidence in the model's predictions decreases as the conditions of application deviate from those used in the validation process." A prediction point may lie inside the validation envelope or outside the validation envelope. Some would say that the former is an interpolation point, and the latter is an extrapolation point. But where the point lies is less important than the size of the associated uncertainty (reported separately from the scale) and how well its prediction uncertainty is characterized (the focus of this category in the scale).

The level descriptions for prediction uncertainty are:
Level 1 – M&S results achieved with no or ad hoc estimate of prediction uncertainty.
Level 2 –Prediction uncertainties inferred from validation problems based on expert opinion and deterministic estimates; prediction uncertainties for some results that affect a critical decision have been quantified.
Level 3 –Prediction uncertainties inferred from validation problems using non-deterministic approach; prediction uncertainties for more than half of the results that affect a critical decision have been quantified.
Level 4 –Thorough determination of prediction uncertainty using non-deterministic approach; prediction uncertainties for most of the results that affect a critical decision have been quantified.

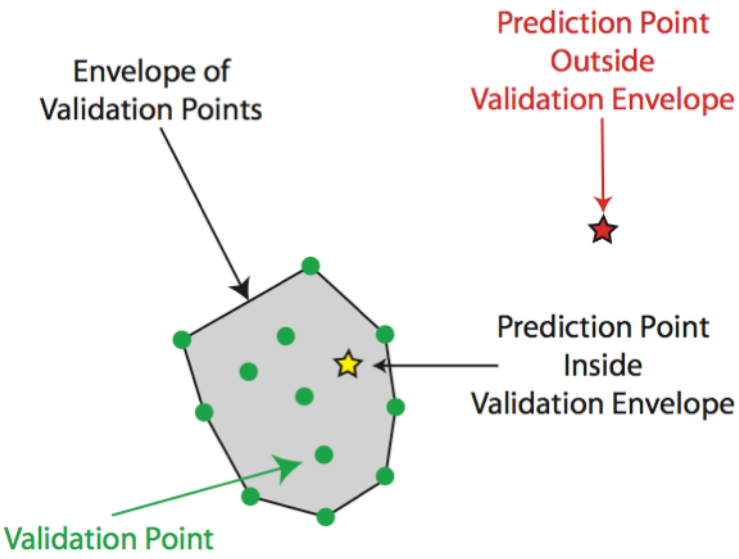

Figure 2. Distinction between validation points and prediction points.

5. Level of Technical Review

The Level of Technical Review category measures the rigor of technical review of the models and/or results by experts, peers, and/or independent authorities. The review process plays a significant role in building objectivity in the rigor scale. This category also increases overall acceptability of M&S results for the particular application. Increasing the level requires more rigorous reviews with favorable outcomes.

Technical reviews will be conducted throughout the development lifecycle of an M&S, with each review focusing on a particular aspect of the work (e.g., requirements definition, M&S verification, M&S validation efforts, M&S uncertainty quantification, or the overall system being modeled). Different groups could conduct these reviews. Each review can be conducted at one of various levels of technical independence ranging from an informal self-review to a completely independent organization formally reviewing the work and, perhaps, duplicating key findings.

The Technical Review category includes two factors: 1) the level of technical independence of the review, and 2) the scope of subject matter of the review

The level descriptions for technical review are:
Level 1 – M&S results achieved with no or ad hoc technical review.
Level 2 – Informal subject matter expert review; some results that affect a critical decision are reviewed.
Level 3 – Formal internal peer review; more than half of the results that affect a critical decision are reviewed.
Level 4 – Formal external review board; all results that affect a critical decision are reviewed.

6. Process Control

Modeling and simulation requires a user to execute an M&S application to produce results. At a minimum, this process includes (1) an analysis of the problem to identify an appropriate model of the system, (2) preparation of data for the M&S application, (3) execution of the M&S application, and (4) interpretation of the results. Each step of this process is susceptible to human error. These human errors can include simple data entry errors, errors resulting from a misunderstanding of the program's requirements, and errors in understanding the applicability of the M&S. Configuration control is important to oversee the management of a potentially large number of computer files shared between a large number of users. These files are often generated as output from one segment of the M&S and passed along as input to another segment of the M&S.

There are two elements of human error that contribute to the rigor of the M&S that are included in the scale (1) the Process Control category addresses the M&S process and is described in this section and (2) the Operator and Analyst Qualification category addresses misunderstanding of the M&S analyst and is presented in the next section.

The Process Control category addresses the rigor of the level of formality and the extent of documentation of M&S processes used to generate the M&S result. Increasing the level of formality and the extent of documentation increases the level of repeatability and traceability of M&S results. Undocumented processes limit the ability to identify potential errors in the M&S and reproduce results. Complex M&S analysis often requires multiple users to analyze part of the overall problem. The final result is an amalgamation of the parts. Process control measures the rigor of the best practices that are defined for the M&S and the adherence to these processes. Better adherence to better processes increases the rigor of the M&S results.

The level descriptions for process control are:
Level 1 – M&S results achieved with no or minimal process control.
Level 2 – Informally documented, self monitored processes.
Level 3 – Formally documented, internally monitored processes.
Level 4 – Independently certified and audited processes.

7. Operator and Analyst Qualification

M&S requires a user to choose a model of the system, set up the analysis, execute the M&S application, and interpret the results. The M&S user must make a decision as to what features of the problem are important to model and which models are appropriate for the problem. Clearly the qualifications and experience of the M&S user influence these decisions. Program resource requirements such as schedule, budget, and available computer time influence the choice of parameters to study and the level of model applied. These requirements add to the difficulty in determining an appropriate M&S for all users but are even more difficult for the inexperienced user to manage.

The Operator and Analyst Qualification category addresses the rigor of the training and experience level of users and analysts conducting the M&S activities.

The level descriptions for operator and analyst qualifications are:
Level 1 – Minimal training and experience directly related to the M&S activity.
Level 2 – Moderate training or experience directly related to the M&S activity.
Level 3 – Extensive training and experience directly related to the M&S activity.
Level 4 –Independent certification for the specific M&S activity; Extensive training and experience directly related to the M&S activity.

A summary of the rigor level definitions for the seven categories is presented in Table 1.

Category	Code Verification	Solution Verification	Validation	Prediction Uncertainty	Technical Review	Process Control	Operator & Analyst Qualifications
Level 4	Well defined and documented Software Quality Assurance (SQA) processes. Numerical algorithm test suite with significant coverage of required features and capabilities.	Rigorous numerical error bounds quantified for actual application. These error estimates are provided for all results that affect a critical decision.	Quantified validation has been performed with thorough determination of M&S uncertainty and experimental error; The system has been comprehensively validated.	Thorough determination of prediction uncertainty using non-deterministic approach; Prediction uncertainties for most of the results that affect a critical decision have been quantified.	Formal external review board; All results that affect a critical decision are reviewed.	Independently certified and audited processes.	Independent certification for the specific M&S activity; Extensive training and experience directly related to the M&S activity.
Level 3	Software and test cases maintained in configuration control system. Numerical algorithm test suite with moderate coverage of required features and capabilities.	Numerical errors estimated on actual application. These error estimates are provided for more than half of the results that affect a critical decision.	Quantified validation has been performed with estimates of M&S uncertainty and experimental error; Most aspects of the system have been validated.	Prediction uncertainties inferred from validation problems using non-deterministic approach; Prediction uncertainties for more than half of the results that affect a critical decision have been quantified.	Formal internal peer review; More than half of the results that affect a critical decision are reviewed.	Formally documented, internally monitored processes.	Extensive training and experience directly related to the M&S activity.
Level 2	Software versions archived and results repeatable. Numerical algorithm test suite with sparse coverage of required features and capabilities.	Expert opinion based on numerical errors estimated for similar problems. These error estimates are provided for some results that affect a critical decision.	Quantified validation has been performed with estimates of experimental error; Few aspects of the system have been validated.	Prediction uncertainties inferred from validation problems based on expert opinion and deterministic estimates; Prediction uncertainties for some results that affect a critical decision have been quantified.	Informal subject matter expert review; Some results that affect a critical decision are reviewed.	Informally documented, self monitored processes.	Moderate training or experience directly related to the M&S activity.
Level 1	M&S results achieved with no or ad hoc code verification.	M&S results achieved with no or ad hoc solution verification.	M&S results achieved with no or ad hoc validation.	M&S results achieved with no or ad hoc estimate of prediction uncertainty.	M&S results achieved with no or ad hoc technical review.	M&S results achieved with no or minimal process control.	Minimal training and experience directly related to the M&S activity.

Table 1. Rigor level definitions

B. Communicating Assessment of Rigor to Decision Makers

The Columbia Accident Investigation Board[18] pointed out the importance of clear and concise communication of results to critical decision makers. Figure 3 provides an approach for reporting a summary of the rigor to guide an M&S credibility assessment. This summary figure is to be provided with the M&S results. The standard reporting format enables decision makers to quickly evaluate the results and ask key questions to clarify issues. The individual rigor categories are displayed on the abscissa, while the level of rigor is shown on the ordinate. The required level of rigor is noted on the chart with a line; the achieved level of rigor is plotted as color-coded bar. The color-coding uses green to identify elements that meet or exceed the required level, yellow to indicate elements that achieve one level lower than the required level, and red to identify elements that fail to achieve the required level by two or more levels. The key ideas presented in this format are:

- Separate assessment of each rigor component allowing for individual show-stoppers to be readily identified.
- Separate documentation of required and achieved levels of rigor.
- Simple reporting graphic.

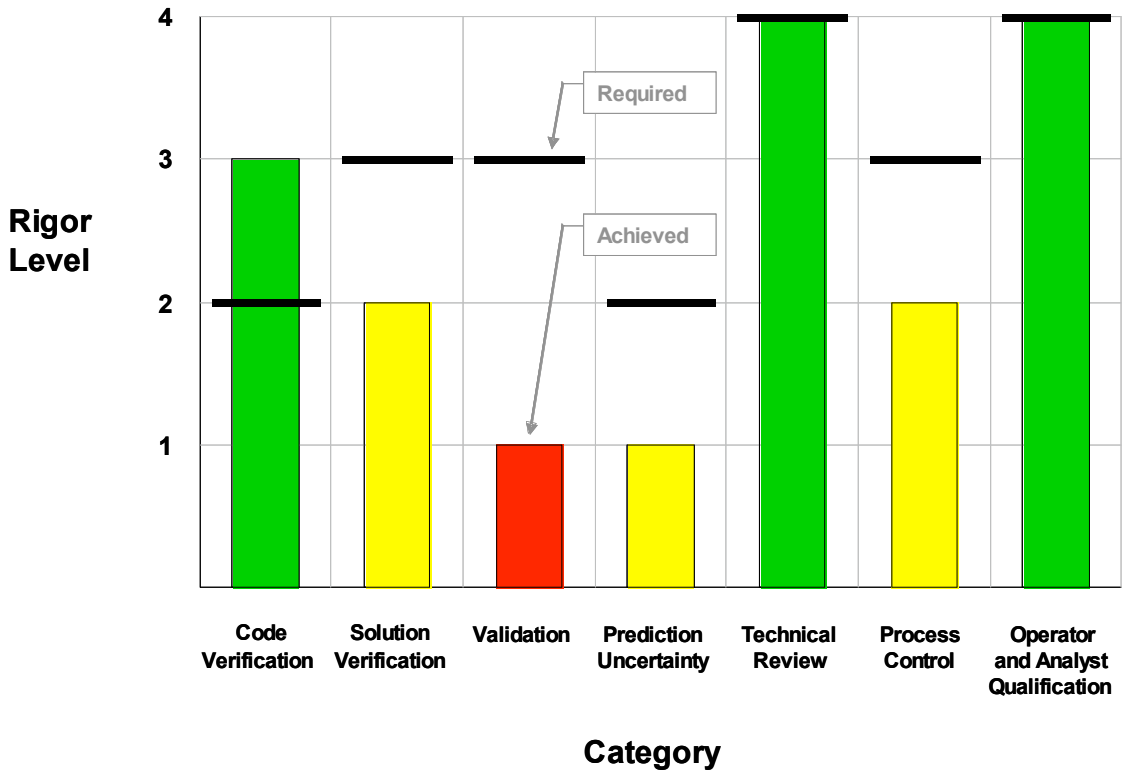

Figure 3. Summary rigor reporting

V. Examples/Application

The M&S Rigor Scale was applied to two M&S studies. The first example, a statistical uncertainty analysis of historical hurricane data, establishes both required and achieved levels of rigor. The second example, a physics-based radiation transport simulation, establishes only the required levels of rigor. These two quite different examples demonstrate that this scale is applicable to a wide range of M&S.

A. Uncertainty Analysis of Historical Hurricane Data Example

A statistical ANalysis Of VAriance (ANOVA) uncertainty analysis study was conducted for a set of historical hurricane data dating back to 1851, which was obtained from the U. S. Department of Commerce National Oceanic and Atmospheric Administration (NOAA)[19]. The study was both an application of the existing ANOVA uncertainty analysis technique provided in the Design-Expert software, version 6, from Stat-Ease, Inc., as well as a research study to document the process of conducting a large uncertainty analysis and the applicability/benefits of such a study. The study took place at a low-level of effort over the course of about one year, while the NASA Standard for Models and Simulations was being developed. Motivation to conduct this uncertainty analysis study came from three sources: 1) the desire to exercise and assess the requirements enumerated within the NASA Interim Standard for Models and Simulations[7] , 2) the desire to use the ANOVA technique with a large, publicly available collection of data exhibiting great variability, which makes the forecasting of future states, from current and previous states, difficult, and 3) the availability of substantial, high-fidelity validation data.

Although only one software product was used, numerous sub-studies were conducted and reported. Each sub-study involved the identification of: 1) a data set (a subset of all the data available), 2) a choice of independent variables (factors) and 3) one or more dependent variables (responses). The factors could be either continuous (numerical) or discrete (categorical). Typically several factors were identified as potentially affecting each of the responses. Different factor sets could be applied to the same data set and responses, to investigate which factors had the most influence on the response(s). The software provided a regression fit of the responses as a polynomial function of the factors used; this regression fit, the "predictive model", could be used to calculate values of the responses anywhere within the domain of the factors, whether or not data was provided at the point of interest. In addition, the software provides numerous diagnostics of the data set and ANOVA application.

Each sub-study consisted of a series of uncertainty analyses for the given data set, in which more knowledge about the contributing factors was incrementally and systematically brought to bear on the analysis. The first uncertainty pass through a given data set assumed the response(s) were simply a function of a random variable; the predictive model was simply the mean value of the data set. Subsequent uncertainty passes through a given data set considered the effects of each factor one at a time, then in pairs of two factors at a time, then three at a time (if applicable) until all the relevant factors were considered together; in each of these passes a linear, quadratic, or cubic predictive model in each of the factors was fit to the data. The standard deviation of the data about the predictive model was tracked throughout each sub-study to identify the amount of uncertainty that was accounted for within a given data set by the set of factors considered within that sub-study. Ideally, the predictive model incrementally improves with the consideration of more factors, while the standard deviation decreases.

The factors available within the data set were the longitude, latitude, and wind intensity of the storm given at 6-hour intervals over the entire life of each storm, as well as several metrics that could be constructed from the date and time stamp of each reading. The responses considered were the location and intensity of the storm 24, 48, 72, 96, and 120 hours beyond a given state, as recorded in the data base and assuming no error in the recorded states. That is, future states of a given storm were predicted from previous states of the same storm (or perhaps some set of storms).

The process was verified with a number of simple examples in which specified amounts of uncertainty about a given polynomial form were considered. Validation of the methods was provided both by comparison of the results with those from NOAA, as well as internally within each sub-study. For example, the diagnostics for a given data set consider the effects of removing a given data point from the predictive model and comparing those results to the model with the same information included. Also, the uncertainty for each prediction was readily available by comparing the computed results with the correct information at a later point in the same data set.

The required level of rigor was determined to be unity across all the categories. The achieved levels of rigor were assessed and reported by the researcher as:

Code Verification = 1; the effort lacked the use of a formal configuration management system and would have benefited greatly from such simply to help organize the many sub-study input data sets and results. The effort also

lacked the use of formal software quality assurance practices within this study; it is not known if the software developer used such practices.

Solution Verification = 3; numerical errors estimated based on present analysis of actual application were repeatedly produced and examined.

Validation = 1; no error was assumed in the reported hurricane track and intensity data and such could have easily been included. Because of this simplifying assumption, the practitioner felt that this limited the study to Level 1, although the remainder of the effort was believed to achieve Level 4.

Prediction Uncertainty = 2; the reporting did not provide uncertainty distributions for most of the results, simply single values comparisons, although the data was available for reporting probability distribution functions (PDFs) or cumulative distribution functions (CDFs).

Technical Review = 3; the effort was subjected to NASA's formal publication review process and was also informally reviewed by two experts from other institutions. The achieved level of rigor in this category may be open to further debate; the formality of the review process was less than that required for a journal publication or a NASA Technical Publication (TP).

Process Control = 3; the process was documented in detail within a conference paper and consistently followed.

Operator and Analyst Qualification = 3; the practitioner had extensive training and experience related to the M&S activity.

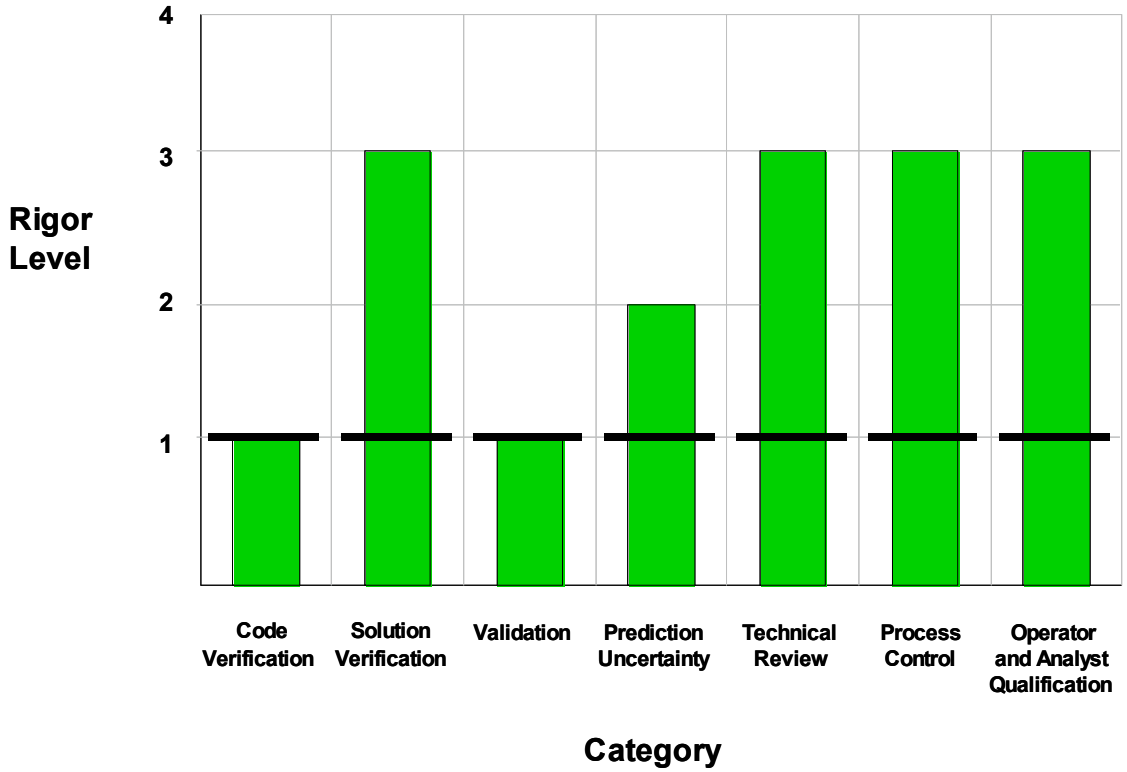

Figure 4. Summary rigor scale for hurricane data analysis.

In several ways, the effort exceeded the Level 1 requirements established for this research. The effort lacked much of the formalization required for a higher level of rigor with notable deficiencies in the use of configuration management and software quality assurance. Also, no uncertainty was assumed for the measured data, although this assumption was reported and used merely for convenience and clarity, rather than because such was not available. Finally, full PDFs or CDFs were not presented for all the results, but could have been. The percentage additional effort imposed by complying with the Level 3 rigor criteria was estimated to be between 1% and 100%, with a most

likely value of 25% increase over the baseline project level of effort. Some of the Level 3 criteria not achieved (e.g. configuration management) would have added minimal overhead, but would have added significant value to the project. Other Level 3 criteria not achieved (e.g. formal software quality assurance and external formal technical review) would have at least doubled the respective efforts with little added value. Describing the conceptual and mathematical models, the uncertainty processes, and the validation details beyond what was done would have added substantial overhead.

B. A Radiation Transport Example

1. Background

Radiation in space is much more intense than it is on the surface of the earth which has a thick atmosphere and a magnetic field to protect it. Currently, the only technologically feasible way to provide protection from this radiation is to place mass between sensitive sites and the external radiation environment. It is expensive and difficult to launch large amounts of extra mass for radiation shielding into space, so radiation constraints are being included in the design process to optimize the design and minimize parasitic shielding. This is done through choice and placement of materials, taking advantage of existing structures and equipment. The present example involves setting scale level requirements for a radiation design analyses for a high consequence decision such as whether to include a mass allocation for parasitic shielding during a preliminary design review (PDR) for a lunar or Mars mission.

2. Model Architecture

A radiation analysis consists of five main components as shown in Figure 5: external radiation environment models, radiation transport models, atomic and nuclear physics models, vehicle and human body models, and radiation effects models. First, an external radiation environment is calculated. Next, the levels of radiation at different depths in shielding materials are calculated using the transport model with input from the atomic and nuclear physics models. These results are then used in conjunction with the body and vehicle models and response functions to estimate levels of exposure at points in the vehicle for the given environment. These exposures are then compared to design limits taking into account the ALARA principle which requires that radiation exposures be kept As Low As Reasonably Achievable.

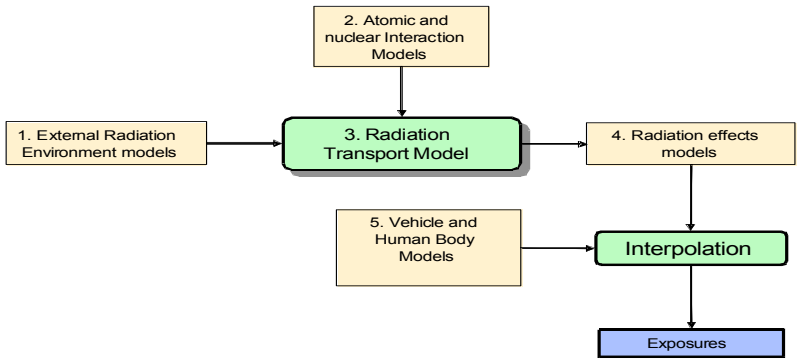

Figure 5. Components of Radiation Analysis.

The following required levels of rigor were established:

Category 1 Code Verification - In order to ensure a high level of confidence in results, a high level of code verification is required to remove potential bugs from the software. However, the large cost associated with having formal documented SQA processes would add little benefit to an activity this small. Also, numerical algorithm testing is important but many of the algorithms are standard and applied to situations without significant difficulties (e.g. integration over 1d smooth functions) therefore the benefit gained from creating test suits for all of these is outweighed by the cost. The requirement is "Level 3 - Software and test cases maintained in configuration control system; numerical algorithm test suite with moderate coverage of required features and capabilities."

Category 2 Solution Verification - As described in Anderson *et al*.[20], the rate of geometric convergence of a solution is highly dependent on the particular vehicle model. Therefore, it is necessary to perform such convergence tests on the actual calculation under consideration. Other solution verification techniques such as energy grid or spatial step size convergence generalize well over large classes of problems and therefore don't need to be performed for each individual problem. The requirement is therefore "Level 3 - Numerical errors estimated on actual application. These error estimates are provided for more than half of the results that affect a critical decision."

Category 3 Validation - Most of the space flight data have been obtained in Low Earth Orbit (LEO) but the main applications of interest are in free space or on lunar or Martian surfaces. Since the application has significantly different radiation environment than the validation, the model error estimates need to be included in the validation. This leads to a requirement of "Level 3 - Quantified validation has been performed with estimates of M&S uncertainty and experimental error; most aspects of the system have been validated."

Category 4 Prediction Uncertainty - Uncertainty estimations are an important element in determining risk from radiation. However, vehicle design requirements have already been set as a deterministic point calculation which places the required level to be "Level 2- Prediction uncertainties inferred from validation problems based on expert opinion and deterministic estimates; prediction uncertainties for some results that affect a critical decision have been quantified."

Category 5 Level of Technical Review - Review is essential to make certain that the results make sense and to catch potential errors made in the calculations. However, the vehicle analysis will be performed through a website such that most of the calculation will be automated including some error checking and recording of what was actually run. This means technical review will be relied upon less to catch mistakes. The requirement is "Level 2 - Informal subject matter expert review; some results that affect a critical decision are reviewed."

Category 6 Process Control - The purpose of process control is to ensure reliability and repeatability of results. Much of the process control is automatically handled by the website and related software taking the burden off the user. The requirement is "Level 2 - Informally documented, self monitored processes."

Category 7 Operator and Analyst Qualification - The vehicle analysis will be performed through a website such that most of the calculation will be automated including some error checking and recording of what was actually run. This will allow people with less experience to perform these calculations more reliably making the requirement "Level 2 - Moderate training or experience directly related to the M&S activity."

The summary rigor scale for this application is shown in Fig. 6. In this case, the results constitute a set of requirements to be met by subsequent computations.

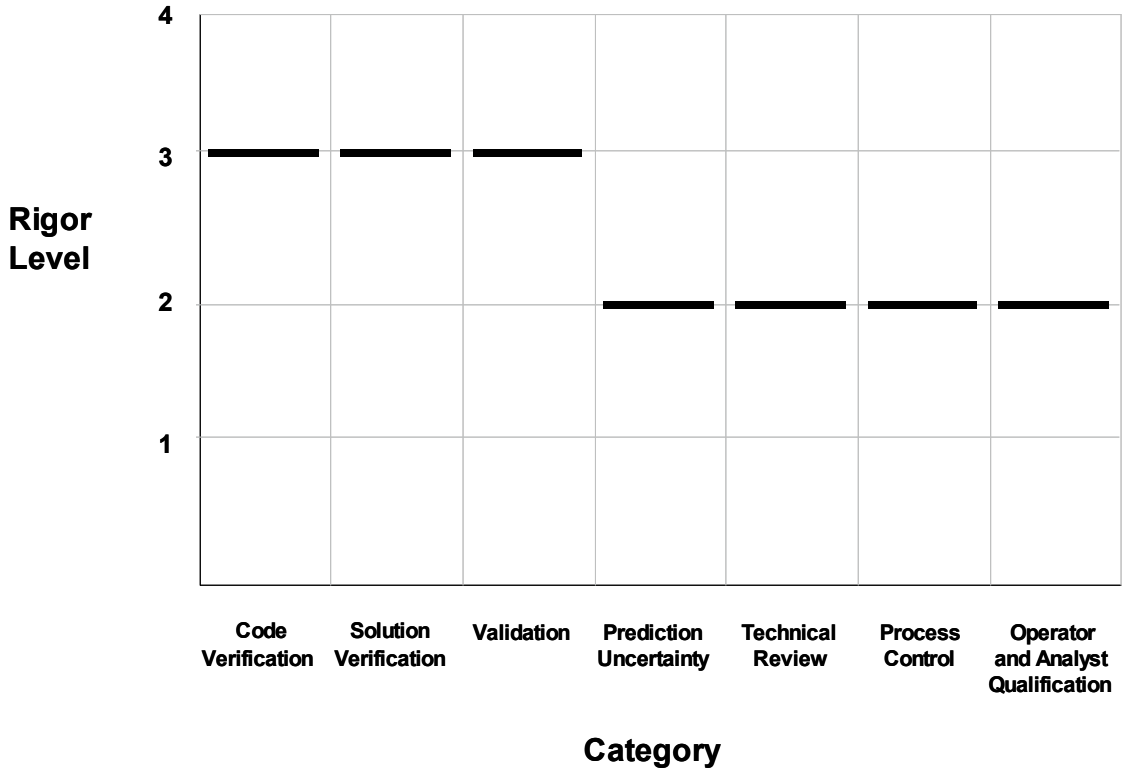

Figure 6. Summary rigor scale for radiation transport example.

VI. Concluding Remarks

This paper has demonstrated a scale to communicate the rigor of models and simulations to decision makers for critical decisions. Two different M&S applications were evaluated with the scale. Based on the results from the two examples, it appears that application of the scale to additional cases would be of interest. There are several issues of this or any scale that need additional study. Some facets that should be studied are:

1. Roll-up of multiple scale measures across a program - Critical decisions in programs or projects typically require input from many different models and simulations. The case studies in this paper apply the current scale to single M&S results. Programs and projects must define an approach to integrate this information from many different M&S applications to evaluate the decision. The authors do not favor a simple approach of averaging or taking the minimum of all of the individual results; it is important to keep the reported measures separate. However, this approach needs to be vetted in a real decision process.
2. Integration of M&S applications from simpler sub-models - This scale is intended to be applied to the M&S at the level that results are used for a critical decision. However, it would be reasonable to apply this scale successively to sub-models. The results of each sub-model would then be used to communicate the rigor of the complete M&S. The case studies presented in this paper did not approach this problem. The approach to roll up the sub-model rigor into an integrated rigor has not been addressed and is beyond the scope of this paper. A good discussion on this subject can be found in Oberkampf *et al.*[10]
3. Evaluating the effectiveness of the scale and the processes that are used to make decisions - This requires that the current scale, and possibly others, be applied to realistic problems for critical decisions.

Acknowledgments

The authors have benefited from many valuable discussions in the course of this work, and they wish to thank the NASA Office of the Chief Engineer and the NASA Engineering and Safety Center for sponsorship of this work, then NASA Chief Engineer, Christopher Scolese, for motivating and championing the development of a rigor assessment scale, Harold Bell and Dawn Schaible for facilitating the project, Martin Pilch and William Oberkampf for sharing a preliminary version of their scale, William Oberkampf, Timothy Trucano and David Peercy for their comments during the development of the scale described in this paper, Scott Harmon for steering us towards a concise description of the goals, assumptions, and architecture of the scale, and the Topic Working Group responsible for the review of NASA-STD-(I)-7009 for numerous discussions which have clarified our thinking on the scale presented in this paper.

References

[1]Wang, R.Y. and Strong, D.M., "Beyond Accuracy: What Data Quality Means to Data Consumers," *J. Mgmt. Info. Systems*, Vol. 12, No. 4, pp. 5–34, 1996.

[2]Mehta, U.B., "Guide to Credible Computer Simulations of Fluid Flows," *AIAA Journal*, vol. 12, no. 5, pp. 940–948, 1996.

[3]Tseng, S. and Fogg, B.J., "Computing and Computing Technology," *Comm. ACM*, vol. 42. no. 5, pp. 39-44, 1999.

[4]Logan, R.W. and Nitta, C.K., "Validation, Uncertainty, and Quantitative Reliability at Confidence (QRC)," AIAA Paper 2003-1337, 2003.

[5]Balci, O., "Quality Assessment, Verification, and Validation of Modeling and Simulation Applications," Proceedings of the 2004 Winter Simulation Conference, R.G. Ingalls, M.D. Rossetti, J.S. Smith, and B.A. Peters, eds., pp. 122–127, 2004.

[6]Standard for Models and Simulations, NASA-STD-7009, (currently under final agency review).

[7]Standard for Models and Simulations, NASA-STD-(I)-7009, Dec. 1, 2006 (available at ftp://ftp.sonic.net/pub/users/usacm/PTC60/NASA_STD_I_7009.pdf).

[8]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.

[9]Harmon, S.Y. and Youngblood, S.M., "Simulation Validation Quality and Validation Process Maturity," Simulation Interoperability Workshop, Paper No. 04S-125, March 2004.

[10]Oberkampf, W.L., Pilch, M., and Trucano, T.G., "Predictive Capability Maturity Model for Computational Science and Engineering," Sandia Report SAND2007-5948, October 2007.

[11]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.

[12]Thomas, D. and Hale, J., "A Characterization Taxonomy for Integrated Management of M&S," Paper 06F-SIW-091, September 2006.

[13]Mehta, U.B., "Simulation Credibility Level," 5th Joint Army-Navy-NASA-Air Force Modeling and Simulation Subcommittee Meeting, May 14-17, Denver, CO.

[14]Miller, G.A., "The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information," *Psychological Rev.,* Vol. 63, pp. 81–97, 1956.

[15]AIAA, *Guide for the Verification and Validation of Computational Fluid Dynamics Simulations*, American Institute of Aeronautics and Astronautics, AIAA G-077-1998, 1998, Reston, Va.

[16]ASME, *Guide for Verification and Validation in Computational Solid Mechanics*, American Society of Mechanical Engineers, ASME V&V 10-2006, 2006, New York, NY.

[17]Roache, P.J., "Verification of Codes and Calculations," *AIAA Journal*, Vol. 36, No. 5, pp. 696-702, May 1998.

[18] NASA, Columbia Accident Investigation Board Report, Volume 1, August, 2003. [http://caib.nasa.gov].

[19]Green, L.L., "Uncertainty Analysis of Historical Hurricane Data," AIAA Paper 2007-1101, Reno, Nevada, January 2007.

[20]Anderson, B.M., Blattnig, S.R., and Clowdsley, M.S., "Numerical Uncertainty Quantification of Radiation Analysis Tools," International Conference on Environmental Systems, July 2007, Chicago, IL, USA http://www.sae.org/technical/papers/2007-01-3110.