OFFICE OF THE SECRETARY OF DEFENSE
1700 DEFENSE PENTAGON
WASHINGTON, DC 20301-1700

OPERATIONAL TEST
AND EVALUATION

MAR 14 2016

MEMORANDUM FOR COMMANDING GENERAL, ARMY TEST AND EVALUATION COMMAND
COMMANDER, OPERATIONAL TEST AND EVALUATION FORCE
COMMANDER, AIR FORCE OPERATIONAL TEST AND EVALUATION CENTER
DIRECTOR, MARINE CORPS OPERATIONAL TEST AND EVALUATION ACTIVITY
COMMANDER, JOINT INTEROPERABILITY TEST COMMAND

SUBJECT: Guidance on the Validation of Models and Simulation used in Operational Test and Live Fire Assessments

In some instances, modeling and simulation (M&S) has been and will be an important element contributing to my evaluations. For example, the testing of new systems, such as those designed to operate in an anti-access/area denial environment, as well as the testing of systems of systems, will involve the use of M&S to examine scenarios that cannot be created using live testing. Whenever M&S is used for operational test and evaluation, I need to have the same understanding of and confidence in the data obtained from M&S as I do any other data collected during an operational or live fire test. Thus, I expect to see validation, and accreditation approaches described in sufficient detail in Test and Evaluation Master Plans (TEMPs) and Test Plans, and the validation approach should employ statistically rigorous design and analysis techniques wherever possible.

Background

For the purposes of this memorandum, M&S includes any emulation of a system, entity, or environment that is essential to my evaluation of operational effectiveness, suitability, survivability, and lethality. This may include, but is not limited to, physics-based computer models, effects-based computer models, hardware-, software-, or operator-in-the-loop simulations, system integration labs, threat environment models, live virtual constructive environments (e.g. cyber ranges), or any combination of the above.

Validation is defined as "the process of determining the degree to which a model or simulation and its associated data are an accurate representation of the real world from the perspective of the intended uses of the model."[1] All M&S, when used to support operational tests and evaluations, should not be accredited until a rigorous comparison of live data to the model's predictions is done (if possible), and those predictions are found to have replicated live results with sufficient accuracy for the intended evaluation in the intended domain (region of the

[1] DoD 5000.61

operational envelope). Extrapolating M&S results to outside the domain in which they were validated---that is, into another domain in which the key physical effects, timelines, and threat considerations are significantly different from the domain in which the M&S validation has been conducted---is extremely dangerous and is likely to lead to invalid results. In cases where validation and accreditation are attempted unsuccessfully or inadequate data are available, either (1) the existing model should not be used and consideration should be given to whether the associated shortfalls in the model can be corrected to permit validation and accreditation; (2) effort should be made to collect the necessary data; or (3) the validation report and any results based on the M&S should be caveated with a clear explanation of which areas are not sufficiently validated and consideration should be given to increasing the scope of live testing.

I have noticed that the quality of, and methods for, validation vary greatly across the Services and across systems. A variety of methods are acceptable since the validation effort should always be tailored to the specific application, or "intended use" of the M&S. Regardless of the methodology, I expect the validation of M&S to include the same rigorous statistical and analytical principles that have become standard practice when designing live tests. In other words, the principles and techniques that comprise Design of Experiments methodologies, including formal statistical tests, should be employed as part of the process of determining what live data are needed for model validation, and in the process of determining how well the models/simulations reflect reality. If there are extraordinary circumstances prohibiting these principles from being used, the reasons why they were not used must be clearly articulated and the alternative approaches used to justify validation and accreditation cogently explained.

I recognize that validation is a complex process and there are many important elements, including documentation review, face validation, subject matter expert (SME) evaluation, and comparison to other models. While these processes typically are not statistical in nature, they provide useful information and can continue to be used. However, by themselves, they do not adequately address model validation. I have encountered several cases where a model is accredited after a visual comparison of live and simulation data is determined to be 'close enough,' often by some undefined measure of goodness. I expect the community to employ more rigorous methods going forward, applying formal statistical tests to these comparisons where possible. I also expect that, in addition to validating sub-components of the model, testers should also focus on the validation of the full system or environment being emulated. While validating every subcomponent that comprises the full system is ideal, I am most interested in ensuring that the integration of all the components and the environment together adequately represents the real-world system relevant to the intended use of the model.

TEMP and Test Plan Requirements

In light of the above, I expect TEMPs and Test Plans to describe M&S validation and accreditation approaches in sufficient detail.[2] In addition to describing the M&S capability and its intended use in detail, I expect a discussion of the following elements:

---

2 TEMPs and Test Plans may also reference other relevant documents such as Validation Plans and Accreditation Plans. If the operational or live-fire test intends to use existing data or results from validation for a previous test, the test plan should reference those data and summarize how they are relevant for the intended use of the M&S in the current test.

- The response variables and/or mission-level metrics of interest. What operationally relevant output from the M&S will be assessed for the purpose of evaluating effectiveness, suitability, survivability, or lethality? What metric(s) will be used to match live data with simulation data in order to determine whether the M&S should be validated?

- The range of conditions over which the M&S will be validated. Similar to the factors and levels in a designed experiment, it is essential to provide an outline of the operational space of interest for the M&S. Note, it is possible for an M&S to be considered valid under some conditions and not others, and whether I am concerned about the areas where it is not valid is tied to the intended use.

- The plan for collecting the necessary live and simulation data for M&S validation. The plan should articulate a method for strategically varying the factors that affect system performance with respect to the response variables of interest. It also should describe whether live data will cover the entire operational envelope to be explored with M&S or only a portion of the envelope. If only a portion of the envelope is covered, the plan should clearly describe which portion.

- An analysis of statistical risk. Where possible, statistical risk should be assessed using measures of merit such as power and confidence in order to scope the amount of data necessary for validation. Calculated uncertainties must be acceptable for accreditation for the intended use.

- The validation methodology. For each metric that is used to compare the live data with simulated data, describe the methodology that will be used to demonstrate validity. While simple qualitative or visual comparisons of plotted M&S outputs and live data may be part of the process, they are not sufficient. Where possible, the validation methodology should include a rigorous comparison using formal statistical tests that quantify risk, allow for sensitivity analyses, and objectively measure the magnitude of the differences between M&S and live data.

The validation process is complex and there is no one-size-fits-all solution. DOT&E will work with other members of the test communities to develop and refine best practices for rigorously validating M&S under a variety of situations. Since the methods for statistical comparison of M&S outputs and live data are not well-documented in the Defense community, I am developing additional forthcoming materials and case studies to aid the community in understanding and applying possible techniques. These materials will be posted on the DOT&E webpage. As always, I welcome feedback as we continue to improve the application of rigorous and scientific methodologies to operational testing, including M&S.

J. M. Gilmore

J. Michael Gilmore
Director